ADVANCES IN MATHEMATICAL AND COMPUTATIONAL ONCOLOGY

EDITED BY: Doron Levy, George Bebis, Russell C. Rockne, Ernesto Augusto Bueno Da Fonseca Lima, Katharina Jahn and Panayiotis V. Benos

PUBLISHED IN: Frontiers in Physiology, Frontiers in Genetics, Frontiers in Applied Mathematics and Statistics, Frontiers in Oncology and Frontiers in Immunology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-178-6
DOI 10.3389/978-2-88974-178-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ADVANCES IN MATHEMATICAL AND COMPUTATIONAL ONCOLOGY

Topic Editors:
Doron Levy, University of Maryland, College Park, United States
George Bebis, University of Nevada, Reno, United States
Russell C. Rockne, Beckman Research Institute, City of Hope, United States
Ernesto Augusto Bueno Da Fonseca Lima, University of Texas at Austin, United States
Katharina Jahn, ETH Zürich, Switzerland
Panayiotis V. Benos, University of Pittsburgh, United States

Citation: Levy, D., Bebis, G., Rockne, R. C., Da Fonseca Lima, E. A. B., Jahn, K., Benos, P. V., eds. (2022). Advances in Mathematical and Computational Oncology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-178-6

Table of Contents

frontiers | Frontiers in Physiology

EDITORIAL
published: 07 April 2022
doi: 10.3389/fphys.2022.889198

Editorial: Advances in Mathematical and Computational Oncology

George Bebis[1]*, *Doron Levy*[2], *Russell Rockne*[3], *Ernesto Augusto Bueno Da Fonseca Lima*[4] *and Panayiotis V. Benos*[5]

[1]Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, United States, [2]Department of Mathematics, University of Maryland, College Park, MD, United States, [3]Department of Computational and Quantitative Medicine, Beckman Research Institute, City of Hope, Pasadena, CA, United States, [4]Center for Computational Oncology, Oden Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX, United States, [5]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, United States

Keywords: mathematical oncology, computational oncology, machine learning, mathematical modeling, computational modeling

Editorial on the Research Topic

Advances in Mathematical and Computational Oncology

Cancer is not a single disease, it is a complex and heterogeneous disease which leads to the second cause of death worldwide. Although all cancers manifest themselves as an uncontrolled growth of abnormal cells, they are actually distinct neoplastic diseases that possess different genetic and epigenetic alterations, underlying molecular mechanisms, histopathologies and clinical outcomes. Understanding the origins and growth of cancer requires understanding the role of genetics in encoding proteins that form phenotypes and molecular alterations at multiple levels (e.g., gene, cell, and tissue).

Advanced mathematical and computational models could play a significant role in examining the most effective patient-specific therapies. Tumors, for example, undergo dynamic spatio-temporal changes, both during their progression and in response to therapies. Multiscale advanced mathematical and computational models could provide the tools to make therapeutic strategies adaptable enough and to address the emerging targets. Similarly, understanding the interrelationship amongst complex biological processes requires analyzing very large databases of cellular pathways. High-performance computing, big data analytics solutions, data-intensive computing, and medical image analysis techniques could be critical in addressing these challenges. Therefore, there is pressing need to design and develop mathematical and computational strategies to harness cancer data in an accurate and efficient fashion.

This special issue includes contributions to the state of the art and practice in mathematical and computational oncology addressing some of the challenges and difficulties in this field, as well as prototypes, systems, tools, and techniques.

Identifying potential biomarkers with prognostic value for various cancers has been a challenging research problem. Xu et al. (Screening and Identification of Potential Prognostic Biomarkers in Adrenocortical Carcinoma) investigate this problem in silico for the case of adrenocortical carcinoma (ACC) by integrating protein interaction networks with gene expression profiles. By looking for the most significantly differentially expressed genes in three microarray datasets from the Gene Expression Omnibus (GEO) database, they identified 150 genes that overlapped in the three datasets. These 150 significant genes were further analyzed using DAVID, KEGG and other methods resulting in 24 hub-genes, which were then used for downstream analysis and validation. Using these 24 hub-genes for Disease Free Survival (DFS) and Overall Survival (OS) in a cohort of 76 ACC cases from the TCGA revealed

OPEN ACCESS

Edited and reviewed by:
Raimond L. Winslow, Johns Hopkins University, United States

***Correspondence:**
George Bebis
bebis@unr.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *03 March 2022*
Accepted: *15 March 2022*
Published: *07 April 2022*

Citation:
Bebis G, Levy D, Rockne R, Lima EABDF and Benos PV (2022) Editorial: Advances in Mathematical and Computational Oncology. Front. Physiol. 13:889198. doi: 10.3389/fphys.2022.889198

that 5 of these hub-genes were significantly correlated with either DFS or OS. By performing univariate and multivariate Cox regression, as well as Kaplan-Meier survival analysis, the authors demonstrated the potential prognostic value of the mRNA overexpression of these 5 genes in ACC.

In an effort to better understand why rectal cancer patients, even with the same tumor stage, have different response to radiotherapy, Pham et al. (Image-Based Network Analysis of DNp73 Expression by Immunohistochemistry in Rectal Cancer Patients) investigate the predictive value of DNp73 in patients with rectal adenocarcinoma using image-based network analysis. Using Fuzzy Weighted Recurrence Networks (FWRN), they analyzed the immunohistochemistry images of DNp73 expression from a small cohort of rectal cancer patients who underwent radiotherapy before surgery. Their analysis showed that the primary tumors-to-biopsy ratios of two FWRN parameters, namely the clustering coefficient and the characteristic path length, can correlate DNp73 expression with increased survival time.

McKenna et al. (Leveraging Mathematical Modeling to Quantify Pharmacokinetic and Pharmacodynamic Pathways: Equivalent Dose Metric) provide a framework for leveraging mathematical modeling to quantify pharmacokinetic and pharmacodynamic (PK/PD) pathways. Standard treatment response assays compare cell survival at a single timepoint to applied drug concentration overlooking drug PK/PD properties in developing treatment response assays. Addressing this oversight, McKenna et al. utilize mathematical modeling to decouple and quantify PK/PD pathways. They propose a notion of an "equivalent dose metric", a metric that is derived from a mechanistic PK/PD model and provides a biophysically-based measure of drug effect. An equivalent dose is defined as the functional concentration of drug that is bound to the nucleus following therapy. This metric can be used to quantify drivers of treatment response and potentially guide dosing of combination therapies. Examples are provided through studying the response of cells to time-varying doxorubicin treatments, modulating doxorubicin pharmacology with small molecules that inhibit doxorubicin efflux from cells and DNA repair pathways. This approach can be leveraged to quantify the effects of various pharmaceutical and biologic perturbations on treatment response.

Evaluating the dynamics interactions of multiple angiogenic factors involved in tumor angiogenesis was attempted by Li and Finley (Exploring the Extracellular Regulation of the Tumor Angiogenic Interaction Network Using a Systems Biology Model). In this context, they propose a new computational framework to understand the extracellular distribution of angiogenic factors in tumor tissue and generate new insights into the regulation of the angiogenic factors' interaction network. The model describes the distribution of two potent pro-angiogenic factors and two important anti-angiogenic factors in tumor tissue. The model predicts that most of VEGF and FGF2 is bound to the cell surface and in signaling forms, while most of TSP1 and PF4 is in the interstitial space and in non-signaling forms that are trapped by HSPGs or inactive due to proteolysis. Moreover, it predicts that increasing the secretion of PF4 in tumor tissue can lead to two counterintuitive results: an increase in interstitial FGF2 and VEGF levels and greater formation of pro-angiogenic complexes, particularly in the VEGF signaling pathway. The study provides mechanistic insights into these counterintuitive results and highlights the role of heparan sulfate proteoglycans in regulating the interactions between angiogenic factors.

SNARE proteins facilitate membrane fusion and as such they have a multifaceted role in the life of cells and have been associated to multiple diseases, including cancer. Thus, identifying SNARE proteins from sequence composition only has become quite important. Previous methods rely on motif identification and some use position-specific scoring matrices, obtained from alignments, as (image equivalent) inputs to a 2D convolutional neural network (CNN). Le and Huynh (Identifying SNAREs by Incorporating Deep Learning Architecture and Amino Acid Embedding Representation) propose a new method for identifying new SNARE proteins. First, they used protein motif information to extract 26,789 non-redundant "SNARE superfamily" proteins from NCBI; and equal number of non-SNARE proteins. Second, they used the NLP algorithm "fastText" (used by FaceBook) to create amino acid embedding representations. Finally, they used these representations as input to a 1D CNN to predict whether a protein belongs to the SNARE family or not. Given that the peptide "words" do not contain "spaces" like in human languages, they tested n-grams of size n = 1,2,3,4,5. Unsurprisingly, they found that n = 4 or 5 achieved the best performance. They used n = 5 results to compare their method to the other CNN based on position-specific scoring matrices and they found to perform substantially better across metrics (sensitivity, specificity, accuracy, MCC).

Hochman et al. (Metastases Growth Patterns in vivo—A Unique Test Case of a Metastatic Colorectal Cancer Patient) explore colorectal cancer (CRC) lung metastases growth patterns. Available mathematical tumor growth models rely mainly on primary tumor data, and rarely relate to metastases growth. The study is based on a data set of a metastatic CRC patient, for whom 10 lung metastases were measured while untreated by seven serial computed tomography (CT) scans, during almost 3 years. Three mathematical growth models (Exponential, logistic, and Gompertzian) were fitted to the actual measurements. The study explores factors affecting growth pattern including size, location, and primary tumor resection. This study provides evidence that exponential growth of CRC lung metastases is a reliable approximation, and encourages focusing research on short-term effects of surgery on metastases growth rate.

Zhu et al. (Genetic Alterations and Transcriptional Expression of m6A RNA Methylation Regulators Drive a Malignant Phenotype and Have Clinical Prognostic Impact in Hepatocellula) study how genetic alterations and transcriptional expression of m6A RNA methylation regulators derive a malignant phenotype and have clinical prognostic impact in hepatocellular carcinoma (HCC). This study is conducted on data collected from 371 HCC patients from the Cancer Genome Atlas database. Techniques used are survival analysis and gene set enrichment analysis. Machine-learning tools were used on

selected regulators to develop a risk signature, m6Ascore. This score is based on four m6A regulators, predicting HCC prognosis well at three or five years. Zhu et al. further show that mutations and copy number variations of m6A regulators, conferring worse survival, are strongly associated with TP53 mutations in HCC.

The ARF/MDM2/p53 is one of the regulatory networks that is heavily solicited in cancer therapy, thus there are important challenges for controlling the output of this network (high p53, low mdm2 with desirable consequences on cell fate). In Suarez et al. (Pinning Control for the p53-Mdm2 Network Dynamics Regulated by p14ARF), a mathematical model of p53-Mdm2 dynamics is used to explore how "pinning" of p14ARF can enable control of gene regulatory network dynamics under biological contexts. The pinning is introduced using a control systems approach where the control dynamics is solved in conjunction with the systems dynamics. In this context, the control system takes the place of either an external perturbation such as DNA damage or an internal reset due to transcription (gene expression). They tested their methodology to confirm 1) the behaviors induced by p53 as DNA damage response to gamma radiation and apoptosis, and 2) the behavior consequence on mdm2 levels and the feedback regulation of p53 levels by mdm2 mediated degradation. Using this approach, the authors propose to computationally model and stir the dynamics of a gene regulatory network such as the p14ARF/MDM2/p53 network to a desired state particularly to reproduce either the coordinated fluctuation behavior of p53 transcription that is usually generated by irradiation and/or to recover the tumor suppressor effect of p53 (cell cycle arrest and apoptosis).

Clear cell renal cell carcinoma (ccRCC) is the most common kidney cancer. Its 5-years survival prognosis increases by 5-fold (from 10% to 50–69%) if it is diagnosed early, when the cancer is small. Therefore, identifying biomarkers of ccRCC is very important. In their study, Yang et al. (Identification of KIF18B as a Hub Candidate Gene in the Metastasis of Clear Cell Renal Cell Carcinoma by Weighted Gene Co-expression Network Analysis) perform a bioinformatic analysis on publicly available gene expression dataset to identify such biomarkers. Specifically, they used a GEO dataset with 265 samples, and after pre-processing and quality control (during which they excluded samples with no meta-data and outliers), they used the popular weighted gene co-expression network analysis (WGCNA) package to identify cancer stage-related gene modules and corresponding "hub" genes. Overall, they identified 10 such genes with high maximal clique centrality (MCC), and KIF18B was at the top of this list. They then validated this finding on the TCGA database where they did not only found KIF18B to be differentially expressed ccRCC patients and controls, but they also found that its high expression was significantly associated with worse survival.

Yu et al. (Predicting Relapse in Patients with Triple Negative Breast Cancer (TNBC) Using a Deep-Learning Approach) performed a preliminary study on predicting the risk of relapse for patients with Triple Negative Breast Cancer (TNBC) using machine learning (ML). By examining the spatial distribution of $CD8^+$ T cells and cancer cells in immunofluorescence (IF) images, they derived a prognostic score for predicting early relapse. Using a small dataset, the authors demonstrated that the relative infiltration of CD8 cells into cancer cell islands is associated with good prognosis. The approach could possibly be generalized to other types of cancers.

In every computational model, parameter estimation is an important component. This is also true in systems biology, where biological processes are represented by a set of ordinary differential equations (ODEs). Bianconi et al. (A New Bayesian Methodology for Nonlinear Model Calibration in Computational Systems Biology) present a new Bayesian method for parameter estimation, named Conditional Robust Calibration (CRC). The authors consider the parameter vector as a random variable in the parameter space. The way that CRC works is that it first simulates a fixed number of samples, given a parameter vector. Then it calculates the posterior of the parameters given the data and compares this to the posterior of the observed data. In the next iteration, the parameter vector changes based on the distance (error) of the two. They benchmarked CRC against three other state-of-the-art algorithms (ABC-SMC, profile likelihood, DRAM) on two different systems (Lotka-Voltera model, EpoR system, multiple myeloma model). In the Lotka-Voltera model, all algorithms performed well, with CRC being more accurate although it required one more iteration. In the EpoR system, the CRC found a different solution thatn the other two algorithms, but the authors report their solution was more reliable because of many missing values in this dataset. Finally, the multiple myeloma system is a high-dimensional ODE model. In that system the authors showed that their method outperformed the others.

Storey et al. (Modeling Oncolytic Viral Therapy, Immune Checkpoint Inhibition, and the Complex Dynamics of Innate and Adaptive Immunity in Glioblastoma Treatment) propose an ordinary differential equation model of treatment for a lethal brain tumor, glioblastoma, using an oncolytic Herpes Simplex Virus. Thea authors use a mechanistic approach to model the interactions between distinct populations of immune cells, incorporating both innate and adaptive immune responses to oncolytic viral therapy (OVT), and including a mechanism of adaptive immune suppression via the PD-1/PD-L1 checkpoint pathway. They focus on the tradeoff between viral clearance by innate immune cells and the innate immune cell-mediated recruitment of antiviral and antitumor adaptive immune cells. The model suggests that when a tumor is treated with OVT alone, the innate immune cells' ability to clear the virus quickly after administration has a much larger impact on the treatment outcome than the adaptive immune cells' antitumor activity. Even in a highly antigenic tumor with a strong innate immune response, the faster recruitment of antitumor adaptive immune cells is not sufficient to offset the rapid viral clearance. This motivates the subsequent incorporation of an immunotherapy that inhibits the PD-1/PD-L1 checkpoint pathway by blocking PD-1, which is combined with OVT within the model. The combination therapy is most effective for a highly antigenic tumor or for intermediate levels of innate immune localization. Extreme levels of innate immune cell activity either clear the virus too quickly or fail to activate a

sufficiently strong adaptive response, yielding ineffective combination therapy of GBM. This work shows that the innate and adaptive immune interactions significantly influence treatment response and that combining OVT with an immune checkpoint inhibitor expands the range of immune conditions that allow for tumor size reduction or clearance.

In the work by Conforte et al. (Modeling Basins of Attraction for Breast Cancer Using Hopfield Networks), bulk RNA-Seq data from 70 paired breast cancer and control samples were analyzed with Hopfield network modeling. Hopfield networks are a form of recurrent artificial neural network which does not require kinetic parameter rates or knowledge of underlying protein-protein interactions. The authors leverage these properties to analyze the high-dimensional RNA-Seq data to find a correlation with the distance between the cancer and control attractors with overall survival. They then use the Hopfield network to identify potential therapeutic gene targets and validate their approach with single-cell sequencing data collected from HER2+ breast cancer patients. This work showcases the power of a theory-driven approach to analysis of complex gene sequencing data and predicts novel therapeutic interventions.

In their paper, Chen et al. (Bioinformatics Analysis of Prognostic miRNA Signature and Potential Critical Genes in Colon Cancer) report the results of bioinformatic analysis they performed in two omics colon cancer datasets. Specifically, they analyzed data form TCGA and another publicly available GEO dataset and identified an 8-miRNA signature predictive of colon cancer prognosis and 14 (mRNA) genes that seem to play a critical role in carcinogenesis. Differential gene/miRNA expression analysis, initially, identified 472 miRNAs and 563 mRNAs that vary significantly between cancer and controls. Further Cox regression analysis found 12 of those miRNAs and 8 of them were used to build the predictive signature (AUROC = 0.729). These 8 miRNAs have a total of 112 target genes, which are also differentially expressed. Downstream pathway analysis for these 112 genes was performed using pathway enrichment and WGCNA. Finally, protein-protein interaction analysis identified 14 of these genes as critically important for colon cancer.

A central challenge of mathematical and computational oncology is the prediction of response to therapy, which often hinges on resistance to treatment as a primary mechanism of treatment failure. Perez-Velazquez and Rejniak (Drug-Induced Resistance in Micrometastases: Analysis of Spatio-Temporal Cell Lineages) use a hybrid agent-based computational model to investigate the role of the tumor microenvironment in the evolution of resistance in micrometastases. The authors examine the dynamics of drug distribution and oxygen gradients in the tissue to simulate response at single cell resolution. Their modeling suggests that resistant cell clones need not exist prior to treatment administration, rather, that they may emerge, and even be induced, by the treatment itself. The authors conclude that successful treatment strategies may mirror those from the field of microbial resistance to antibiotics, where the goal of treatment is to mitigate the emergence of resistance, rather than complete eradication of the cancer cells. Agent-based modeling helps gain insight into dynamics often not directly observable and is a valuable approach in cancer research.

Accurate and quick prediction of stable peptide binding to Class I HLAs is important for designing immunotherapies but also for evaluating the immunogenicity of different peptides. Abella et al. (Large-Scale Structure-Based Prediction of Stable Peptide Binding to Class I HLAs Using Random Forests) attempted to predict stable binding for class I HLAs from pHLA structures using machine learning. The structures are generated from a large set of pHLA sequences using the authors' previously developed method named the Anchored Peptide-MHC Ensemble Generator. These structures are then transformed into feature vectors for training using random forest classifier to distinguish binders from non-binders. The model achieves competitive performance using significantly less data when compared to other popular sequence-based, pan-allele models with the advantage of interpretability.

Tao et al. (Neural Network Deconvolution Method for Resolving Pathway-Level Progression of Tumor Clonal Expression Programs With Application to Breast Cancer Brain Metastases) develop computational methods to infer clonal heterogeneity and dynamics across progression stages via deconvolution and clonal phylogeny reconstruction of pathway-level expression signatures in order to reconstruct how these processes might influence average changes in genomic signatures over progression. In this work, the authors show, via application to a study of gene expression in a collection of matched breast primary tumor and metastatic samples, that the method can infer coarse-grained substructure and stromal infiltration across the metastatic transition. Their results suggest that genomic changes observed in metastasis, such as gain of the ErbB signaling pathway, are likely caused by early events in clonal evolution followed by expansion of minor clonal populations in metastasis, a finding that may have translational implications for early detection or prevention of metastasis.

Molecular disease subtypes characterized by relevant clinical differences, such as survival, are difficult to differentiate. Tran et al. (A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis) have introduced a new method based on Consensus Factor Analysis (CFA) for disease subtyping and risk assessment using multi-omics data. The new method is capable of exploiting complementary signals available in different types of data to improve the subtypes. Using a large dataset of 30 different cancers from TCGA, it outperformed existing approaches in discovering novel subtypes with significantly different survival profiles. In particular, the authors demonstrated that the new method was able to predict risk scores that are highly correlated with vital status and survival probability. The accuracy of risk prediction was shown to improve as more data types were integrated.

Although most mathematical models deal with a primary cancer, metastasis is the most fatal and clinically challenging phase of cancer progression. This is particularly the case in colorectal cancer (CRC), with the added challenge that metastatic lesions often may not be detectable. To address this challenge, Hochman et al. (Metastasis Initiation Precedes Detection of Primary Cancer—Analysis of Metastasis Growth

in vivo in a Colorectal Cancer Test Case) analyze growth rates derived from serial computed tomography (CT) scans collected from a single patient with several metastatic lesions arising from CRC. Making calculations of growth rates using three different mathematical models, the authors estimate that metastasis may have occurred 4–5 years before the primary tumor diagnosis, suggesting that metastasis may be an early event in CRC. This provocative hypothesis has potentially dramatic implications for clinical management of CRC by suggesting that the primary lesion may already have spread through the body by the time it is detected. Here, the use of mathematical modeling aided the investigators in rolling the clock back in time to calculate the likely order of events which are not otherwise clear.

One of the main contribution of the work developed in Manjunath et al. (ABC-GWAS: Functional Annotation of Estrogen Receptor-Positive Breast Cancer Genetic Variants) is the creation of the database "Analysis of Breast Cancer GWAS" (ABC-GWAS), available at http://education.knoweng.org/abc-gwas/. ABC-GWAS is an interactive database of functional annotation of estrogen receptor-positive breast cancer genome-wide association studies (GWAS) variants. This resource provides useful practical results and conceptual approaches to the functional genomics community in general and breast cancer researchers in particular. Over the past decade, hundreds of GWAS have implicated genetic variants in various diseases, including cancer. However, only a few of these variants have been functionally characterized to date, mainly because the majority of the variants reside in non-coding regions of the human genome with unknown function. A comprehensive functional annotation of the candidate variants is thus necessary to fill the gap between the correlative findings of GWAS and the development of therapeutic strategies. By integrating large-scale multi-omics datasets such as the Cancer Genome Atlas (TCGA) and the Encyclopedia of DNA Elements (ENCODE), the authors performed multivariate linear regression analysis of expression quantitative trait loci, sequence permutation test of transcription factor binding perturbation, and modeling of three-dimensional chromatin interactions to analyze the potential molecular functions of 2,813 single nucleotide variants in 93 genomic loci associated with estrogen receptor-positive breast cancer. To facilitate rapid progress in functional genomics of breast cancer, they have created ABC-GWAS. This resource includes expression quantitative trait loci, long-range chromatin interaction predictions, and transcription factor binding motif analyses to prioritize putative target genes, causal variants, and transcription factors. An embedded genome browser also facilitates convenient visualization of the GWAS loci in genomic and epigenomic context. ABC-GWAS provides an interactive visual summary of comprehensive functional characterization of estrogen receptor-positive breast cancer variants. The web resource will be useful to both computational and experimental biologists who wish to generate and test their hypotheses regarding the genetic susceptibility, etiology, and carcinogenesis of breast cancer. ABC-GWAS can also be used as a user-friendly educational resource for teaching functional genomics.

Over the last 50 years, glioblastoma has remained one of the most difficult cancers to treat. A frequent and primary cause of treatment failure in glioblastoma is drug delivery and transport across the blood-brain-barrier (BBB), which is designed to protect the brain from harmful substances. Using patient-derived xenograft models with tumor burden followed over time with bioluminescence imaging, Massey et al. (Quantifying Glioblastoma Drug Response Dynamics Incorporating Treatment Sensitivity and Blood Brain Barrier Penetrance From Experimental Data) estimate parameters for a predictive mathematical model to suggest that BBB permeability may be more important in determining response to treatment than relative sensitivity of the glioblastoma cells to treatment. This prediction underscores the challenges in treating glioblastoma and suggests that response may be improved with therapeutic approaches which do not rely on transport across the BBB. Mathematical models informed by experimental data and motivated by clinical challenges is the essence of mathematical and computational oncology.

The work developed by Dinh et al. (Application of the Moran Model in Estimating Selection Coefficient of Mutated CSF3R Clones in the Evolution of Severe Congenital Neutropenia to Myeloid Neoplasia) focuses on the transition from severe congenital neutropenia (SCN) to pre-leukemic myelodysplastic syndrome (MDS). Stochastic mathematical models have been conceived that attempt to explain the transition of SCN to MDS, in the most parsimonious way, using extensions of standard processes of population genetics and population dynamics, such as the branching and the Moran processes. The authors previously presented a hypothesis of the SCN to MDS transition, which involves directional selection and recurrent mutation, to explain the distribution of ages at onset of MDS or acute myeloid leukemia (AML). Based on experimental and clinical data and a model of human hematopoiesis, a range of probable values of the selection coefficient s and mutation rate μ have been determined. These estimates lead to predictions of the age at onset of MDS or AML, which are consistent with the clinical data. In this work, based on data extracted from published literature, we seek to provide an independent validation of these estimates. The goal of this work is twofold: 1) to determine the ballpark estimates of the selection coefficients and verify their consistency with those previously obtained and 2) to provide possible insight into the role of recurrent mutations of the G-CSF receptor in the SCN to MDS transition.

Subbalakshmi et al. (NFATc Acts as a Non-Canonical Phenotypic Stability Factor for a Hybrid Epithelial/Mesenchymal Phenotype), employ an integrated computational-experimental approach, and show that the transcription factor nuclear factor of activated T-cell (NFATc) can inhibit the process of complete epithelial–mesenchymal transition, thus stabilizing the hybrid E/M phenotype. Reversible transitions between epithelial and mesenchymal phenotypes—epithelial–mesenchymal transition (EMT) and its reverse mesenchymal–epithelial transition (MET)—form a key axis of phenotypic plasticity during metastasis and therapy

resistance. Unlike previously identified phenotypic stability factors (PSFs), NFATc does not increase the mean residence time of the cells in hybrid E/M phenotypes, as shown by stochastic simulations; rather it enables the co-existence of epithelial, mesenchymal and hybrid E/M phenotypes and transitions among them. Clinical data suggests the effect of NFATc on patient survival in a tissue-specific or context-dependent manner. The results of Subbalakshmi et al. suggest that NFATc behaves as a non-canonical PSF for a hybrid E/M phenotype.

Liu et al. (Identification and Validation of Two Lung Adenocarcinoma-Development Characteristic Gene Sets for Diagnosing Lung Adenocarcinoma and Predicting Prognosis) identify and validate two Lung adenocarcinoma (LUAD)-development characteristic gene sets that can be used for diagnostics and prognosis. Lung adenocarcinoma (LUAD) is one of the main types of lung cancer. LUAD has low early diagnosis rate, poor late prognosis, and high mortality. The study identified 84 genes that were associated with LUAD survival and named as LUAD-unfavorable gene set. 39 genes were associated with LUAD survival and named as LUAD-favorable gene set. The LUAD-unfavorable genes were significantly involved in p53 signaling pathway, Oocyte meiosis, and Cell cycle. The study was conducted on 512 LUADs from The Cancer Genome Atlas and validated and data sets from Gene Expression Omnibus. Functional enrichment analysis was used to explore the potential biological functions of LUAD-unfavorable genes.

In an effort to enable novel, quantitative measures of the tumor microenvironment, Mi et al. (Digital Pathology Analysis Quantifies Spatial Heterogeneity of CD3, CD4, CD8, CD20, and FoxP3 Immune Markers in Triple-Negative Breast Cancer) developed a computational platform and workflow for digital pathology analysis. Using triple-negative breast cancer (TNBC) as an example, the authors demonstrate how their analysis platform can automatically characterize important features from a digital pathology slide, including the invasive front, central tumor, and normal tissue. This automated analysis is critical in order to quantify spatial heterogeneity of prognostic markers in TNBC, including immune cell density and local tumor immuno-architecture. The authors show how quantitative analyses generated from their workflow can be associated with treatment outcomes to predict response and also to inform mechanistic computational models which can use these spatial maps as inputs for calibration or validation. Quantitative spatial analyses such as those provided by this tool are essential to extract the most information from precious rare clinical samples, and also critically important for predictive mathematical and computational models.

In Stiehl et al. (Computational Reconstruction of Clonal Hierarchies From Bulk Sequencing Data of Acute Myeloid Leukemia Samples), the authors develop a computational algorithm that allows identifying all clonal hierarchies that are compatible with bulk variant allele frequencies measured in a patient sample. The clonal hierarchies represent descendance relations between the different clones and reveal the order in which mutations have been acquired. The proposed computational approach is tested using single cell sequencing data that allow comparing the outcome of the algorithm with the true structure of the clonal hierarchy. The authors investigate which problems occur during reconstruction of clonal hierarchies from bulk sequencing data. The algorithm proposed by the authors provides a tool to better understand the ambiguity of such reconstructions and their sensitivity to measurement errors. Their results suggest that in many cases only a small number of possible hierarchies fits the bulk data. This implies that bulk sequencing data can be used to obtain insights in clonal evolution.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Physiology

ORIGINAL RESEARCH
published: 22 May 2019
doi: 10.3389/fphys.2019.00616

Leveraging Mathematical Modeling to Quantify Pharmacokinetic and Pharmacodynamic Pathways: Equivalent Dose Metric

Matthew T. McKenna[1,2], *Jared A. Weis*[3,4], *Vito Quaranta*[5] *and Thomas E. Yankeelov*[6,7,8,9,10]*

[1] *Vanderbilt University Institute of Imaging Science, Vanderbilt University, Nashville, TN, United States,* [2] *Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States,* [3] *Department of Biomedical Engineering, Wake Forest School of Medicine, Winston-Salem, NC, United States,* [4] *Comprehensive Cancer Center, Wake Forest Baptist Medical Center, Winston-Salem, NC, United States,* [5] *Department of Cancer Biology, Vanderbilt University School of Medicine, Vanderbilt University, Nashville, TN, United States,* [6] *Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, United States,* [7] *Department of Diagnostic Medicine, Dell Medical School, The University of Texas at Austin, Austin, TX, United States,* [8] *Department of Oncology, Dell Medical School, The University of Texas at Austin, Austin, TX, United States,* [9] *Oden Institute for Computational and Engineering Sciences, The University of Texas at Austin, Austin, TX, United States,* [10] *Livestrong Cancer Institutes, The University of Texas at Austin, Austin, TX, United States*

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Marcel Schilling,
German Cancer Research Center (DKFZ), Germany
Alexey Goltsov,
Abertay University, United Kingdom

***Correspondence:**
Thomas E. Yankeelov
thomas.yankeelov@utexas.edu

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *10 January 2019*
Accepted: *01 May 2019*
Published: *22 May 2019*

Citation:
McKenna MT, Weis JA, Quaranta V and Yankeelov TE (2019) Leveraging Mathematical Modeling to Quantify Pharmacokinetic and Pharmacodynamic Pathways: Equivalent Dose Metric. Front. Physiol. 10:616. doi: 10.3389/fphys.2019.00616

Treatment response assays are often summarized by sigmoidal functions comparing cell survival at a single timepoint to applied drug concentration. This approach has a limited biophysical basis, thereby reducing the biological insight gained from such analysis. In particular, drug pharmacokinetic and pharmacodynamic (PK/PD) properties are overlooked in developing treatment response assays, and the accompanying summary statistics conflate these processes. Here, we utilize mathematical modeling to decouple and quantify PK/PD pathways. We experimentally modulate specific pathways with small molecule inhibitors and filter the results with mechanistic mathematical models to obtain quantitative measures of those pathways. Specifically, we investigate the response of cells to time-varying doxorubicin treatments, modulating doxorubicin pharmacology with small molecules that inhibit doxorubicin efflux from cells and DNA repair pathways. We highlight the practical utility of this approach through proposal of the "equivalent dose metric." This metric, derived from a mechanistic PK/PD model, provides a biophysically-based measure of drug effect. We define equivalent dose as the functional concentration of drug that is bound to the nucleus following therapy. This metric can be used to quantify drivers of treatment response and potentially guide dosing of combination therapies. We leverage the equivalent dose metric to quantify the specific intracellular effects of these small molecule inhibitors using population-scale measurements, and to compare treatment response in cell lines differing in expression of drug efflux pumps. More generally, this approach can be leveraged to quantify the effects of various pharmaceutical and biologic perturbations on treatment response.

Keywords: mathematical modeling, breast cancer, pharmacokinetic modeling, pharmacodynamics, doxorubicin, treatment response

INTRODUCTION

The parameterization of *in vitro* treatment response data is central to biomarker and drug discovery and the quantitative study of cancer therapies. With recent exceptions (Hafner et al., 2016; Harris et al., 2016), investigation of treatment response *in vitro* has been limited to cell survival assays that assess cell viability at a single, specified timepoint following treatment with a temporally constant concentration of drug. A range of drug concentrations are evaluated in these assays, and the results are conventionally summarized by Hill function parameters, which quantify cell survival with respect to applied drug concentration (Fallahi-Sichani et al., 2013). While this approach has yielded significant insights into cancer biology, it is fundamentally limited by the coarseness of parameters used to summarize treatment response. In particular, these parameters do not explicitly characterize the dynamics of treatment and subsequent response. Further, response metrics are reported relative to the extracellular concentration of drug in the assay, overlooking drug exposure times and variable cell line pharmacologic properties. This not only impairs analysis of *in vitro* treatment response data, but also presents a challenge in translating these therapies *in vivo*.

There are a host of biochemical processes that modulate a tumor cell's response to therapy. For example, the accumulation of drug within cells can be altered by drug metabolism or modification of surface proteins that regulate drug flux through the membrane (Larsen and Skladanowski, 1998; Larsen et al., 2000). Indeed, the multi-drug resistance protein 1 (MDR1) is a well-studied mechanism of resistance to cytotoxic therapies (Clarke et al., 2005). This ATP-dependent pump actively effluxes drug from cells, decreases drug accumulation within cells, and confers resistance to anthracyclines, taxanes, and several other agents (Mechetner et al., 1998). Similarly, pharmacodynamic response to therapies can be altered through modulation of signaling pathways downstream of the therapeutic target. With respect to DNA-damaging agents, changes in DNA repair pathways, which are activated in response to treatment, can alter sensitivity to those agents (Fink et al., 1998; Bouwman and Jonkers, 2012). For example, DNA-dependent protein kinase (DNA-PK) plays a major role in the repair of double strand DNA breaks *via* non-homologous end joining (Smith and Jackson, 1999). Increased expression of DNA-PK has been shown to confer resistance to doxorubicin, an anthracycline commonly used clinically (Shen et al., 1998). Fundamentally, cell line-specific pharmacokinetic and pharmacodynamic properties, such as those described above, drive observed treatment responses. Using conventional methods, these processes are conflated by the parameters used to summarize *in vitro* dose response data (Prentice, 1976; Fallahi-Sichani et al., 2013). The resulting parameters are imprecise measures of drug efficacy, which limits the biological insights to be gained from the data.

More precise technologies are required to advance systems approaches to studying cellular response to therapy (Anderson and Quaranta, 2008). We posit that a mechanistic, mathematical modeling framework is essential to maximize the knowledge gained through treatment response studies (Yankeelov et al., 2013, 2015). In this paradigm, biologically-motived mathematical models are constructed to describe observed behaviors of the system under investigation. The model is then fit to experimental data, yielding a set of parameter values that provide mechanistic insight into observed data. There exist several models in the literature that explicitly incorporate drug pharmacokinetics (PK) and pharmacodynamics (PD) to describe treatment response. *In vitro*, transit compartment models have been used to describe the temporal relationship between drug application and effects (Lobo and Balthasar, 2002). More biologically-motivated PK/PD models have been employed to study specific pharmacokinetic and pharmacodynamic parameters (Lankelma et al., 2003, 2013). PK/PD models have also been developed to investigate treatment response *in vivo* (Simeoni et al., 2004; Sanga et al., 2006; Wang et al., 2015). Recently, we proposed and validated a coupled PK/PD model of doxorubicin treatment response *in vitro* (McKenna et al., 2017). The model incorporates measured doxorubicin pharmacokinetics and pharmacodynamics and predicts response to a specified treatment timecourse on a cell line-specific basis. The model behaves consistently across a wide spectrum of treatment protocols and cell lines, thereby demonstrating that the response dynamics of cancer cell lines to doxorubicin is predictable within this framework. Specifically, the PK model-estimated concentration of doxorubicin bound to the cell nucleus is predictive of cell line pharmacodynamic rates. We further noted a mismatch of drug uptake and response among the investigated cell lines, suggesting that each cell line has an intrinsic sensitivity to stress induced by doxorubicin. By explicitly modeling both drug uptake and subsequent effect, these processes can be independently quantified to study each component of treatment response (McKenna et al., 2017).

It is the goal of the present effort to demonstrate the utility of a mechanistic, mathematical modeling framework in quantifying treatment response and PK/PD pathways. We leverage mathematical models to filter experimental data to yield quantitative measures of specific cellular processes. Specifically, we experimentally perturb doxorubicin pharmacokinetics and pharmacodynamics with chemical inhibitors of each process. We modulate pharmacokinetics in an MDR1 over-expressing cell line and modulate pharmacodynamics *via* DNA-PK in a BRCA1-mutated cell line. These data are analyzed with the proposed PK/PD model to yield quantitative measures of these pathways. We further illustrate the utility of our approach by proposing the equivalent dose metric, which we derived from the PK model. The equivalent dose is analogous to that in radiation therapy, which is used to compare radiation fractionation schedules (Fowler, 1992). In the context of chemotherapy, we define equivalent dose as a functional measure of drug exposure. We specify that for a given equivalent dose, treatment response dynamics are similar. As this approach accounts for variable pharmacologic properties, we posit that it allows for more precise comparisons among cell lines relative to metrics based on extracellular drug concentration. We demonstrate this experimentally through comparison of treatment response in cell lines differing only in MDR1 expression. The modeling-based framework proposed in this work can be leveraged to more precisely quantify the effects of various pharmaceutical and biologic perturbations on treatment response.

MATERIALS AND METHODS

Mathematical Model of Doxorubicin Treatment Response

Doxorubicin is an anthracycline that remains standard-of-care therapy for several cancers (Tacar et al., 2013). Ultimately, doxorubicin induces a host of cellular stress responses which either inhibit further DNA synthesis allowing for cellular recovery, or initiate a cascade leading to cell death (Gewirtz, 1999). At high doxorubicin concentrations, extensive DNA damage often results in cell death *via* apoptosis. Low to moderate concentrations of doxorubicin induce cell senescence and cell death *via* mitotic catastrophe (Chang et al., 1999; Eom et al., 2005). Whereas, apoptosis is immediate (on the order of hours to days), mitotic catastrophe is a relatively protracted process (on the order of several days).

We previously developed and validated a parsimonious treatment response model to describe doxorubicin pharmacokinetics and pharmacodynamics (McKenna et al., 2017). Briefly, a three-compartment model was employed to describe the uptake and binding of doxorubicin in cancer cells. This process is modeled *via* mass conservation through Equations (1–3):

$$\frac{dC_E(t)}{dt} = k_{FE}\frac{v_I}{v_E}C_F(t) - k_{EF}C_E(t) \quad (1)$$

$$\frac{dC_F(t)}{dt} = k_{EF}\frac{v_E}{v_I}C_E(t) - k_{FE}C_F(t) - k_{FB}C_F(t) \quad (2)$$

$$\frac{dC_B(t)}{dt} = k_{FB}C_F(t) \quad (3)$$

where $C_E(t)$, $C_F(t)$, and $C_B(t)$ are the concentrations of doxorubicin in the extracellular, free, and bound compartments, respectively, at time t. The free compartment represents drug that has diffused into the cell, while the bound compartment represents drug that has bound to DNA. The k_{ij} parameters are rate constants that describe the movement of doxorubicin between the i^{th} and j^{th} compartments; for example, k_{FE} describes the rate of drug transfer from the free, intracellular compartment to the extracellular compartment. Similar definitions apply to k_{EF} and k_{FB}. The volumes of the extracellular and intracellular compartments are denoted by v_I and v_E, respectively (see **Table 1** for a full list of model parameter definitions). We note that in this model, several intracellular processes, including doxorubicin metabolism and dissociation from DNA, are not explicitly considered. In previous work (McKenna et al., 2017), we evaluated the performance of several candidate models in describing our experimental data (described in section Doxorubicin Uptake Imaging and Image Processing) with the Akaike information criterion. Of the proposed models, we found that Equations (1–3) best balanced accuracy and the number of model parameters.

A logistic growth model, Equation (4), modified by either of two empirical time-dependent response functions, Equations (5, 6), reflecting distinct mechanisms of cell death, was proposed to describe population level response to doxorubicin therapy as follows:

$$\frac{dN_{TC}(t)}{dt} = \left(k_p - k_d(t,D)\right)N_{TC}(t)\left(1 - \frac{N_{TC}(t)}{\theta(D)}\right) \quad (4)$$

$$k_d(t,D) = \begin{cases} 0 & t < 0 \\ k_{d,a}(D) & t \geq 0 \end{cases} \quad (5)$$

$$k_d(t,D) = \begin{cases} 0 & t < 0 \\ k_{d,b}(D)\, r(D)\, te^{1-r(D)t} & t \geq 0 \end{cases} \quad (6)$$

where k_p and k_d are the proliferation and dose-specific death rates, respectively, D is the delivered dose [defined to be the bound concentration of drug, C_B, calculated with Equations (1–3)], r is a dose-specific constant describing the rate at which treatment induces an effect, θ is the dose-specific carrying capacity describing the maximum number of cells that can be observed in the experimental system, and $N_{TC}(t)$ is the number of cells at time t. Logistic growth models have traditionally been used to describe growth of a variety of biological species whose total size is limited (Gerlee, 2013; Jarrett et al., 2018). This equation accurately describes our experimental system (described in section Doxorubicin Treatment Response Imaging), in which cell population is limited by the surface area of the experimental platform. Prior to treatment (i.e., $t < 0$), cells are modeled to have a constant proliferation rate, k_p. Following treatment at $t = 0$, Equation (5) assumes an immediate induction of a stable, post-treatment death rate ($k_{d,a}$). Equation (6) allows for a smooth induction of drug effect following treatment to a maximum death rate of $k_{d,b}$, while ultimately allowing for recovery of the cell population. The dynamics of this induction and decay is governed by r. A weighted averaging approach is used to incorporate both Equations (5, 6) in the treatment response model. This model was designed to describe cell death *via* apoptosis Equation (5) and mitotic catastrophe Equation (6) and fit experimental data well. Further details on the model can be found in McKenna et al. (2017).

TABLE 1 | Model parameter definitions.

Model Parameter	Units	Definition
k_{EF}	h^{-1}	Rate of drug influx into cell
k_{FE}	h^{-1}	Rate of drug efflux from cell
k_{FB}	h^{-1}	Mixed rate of drug binding and DNA repair
C_E	nM	Extracellular doxorubicin concentration
C_I	nM	Intracellular, extranuclear doxorubicin concentration
C_B	nM	Concentration of doxorubicin bound to the nucleus
D_{eq}	nM	Equivalent dose
N_{TC}	Count	Number of cells
k_p	h^{-1}	Proliferation rate of cells
θ	Count	Carrying capacity of experimental system
$k_{d,a}$	h^{-1}	Death rate assumed in Equation (5)
$k_{d,b}$	h^{-1}	Death rate assumed in Equation (6)
r	h^{-1}	Rate of induction and decay of death rate in Equation (6)

Complete listing of model parameter definitions.

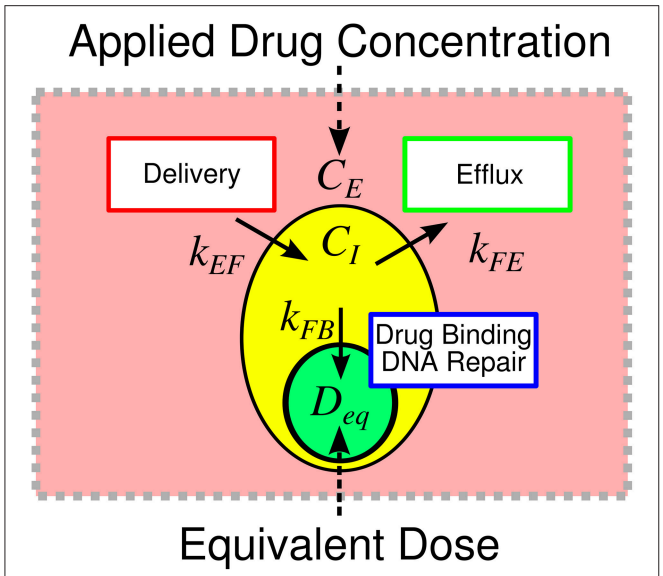

FIGURE 1 | Overview of equivalent dose metric. The response to therapy is determined by the applied drug concentration and cell-specific pharmacologic properties. Traditionally, therapeutic response is summarized relative to the applied extracellular concentration of drug. We propose the equivalent dose metric, D_{eq}, to summarize the contributions of various pharmacologic properties in shaping treatment response. We define equivalent dose as a measure of the functional drug concentration that enters the cell. The equivalent dose is calculated through a mechanistic biophysical model that considers several sources of variability in shaping treatment response. The metric consolidates variable drug uptake (quantified with k_{EF}), efflux (quantified with k_{FE}), and binding (quantified with k_{FB}) into a single descriptor of treatment. The equivalent dose summarizes pharmacologic properties to provide biological insight into treatment response and allows for more precise comparison of treatment response among cell lines.

Equivalent Dose

We define equivalent dose (D_{eq}) as a functional measure of therapy, a summary statistic connecting the amount of drug delivered with the biological effect of that drug. In the context of doxorubicin therapy, we define equivalent dose as the functional concentration of drug that is bound to the nucleus following therapy. To calculate D_{eq} for a specified treatment condition (i.e., extracellular drug concentration timecourse), Equations (1–3) are populated by cell-line- and treatment-specific k_{EF}, k_{FE}, and k_{FB} parameters derived from experimental data (described below). The model is then simulated using the experimentally-defined treatment condition. D_{eq} is the maximum concentration of bound drug (C_B) as predicted by the simulation. We hypothesize that the equivalent dose metric can account for variable cell line pharmacologic properties through its explicit consideration of k_{EF}, k_{FE}, and k_{FB} rates, and it can be leveraged to quantify the effect of agents that modulate those properties. Notably, in proposing the equivalent dose, we lump pharmacodynamic effects into the k_{FB} term. Specifically, k_{FB} is a mixed measure of doxorubicin binding and DNA repair and describes the functional net binding rate. The equivalent dose is illustrated in **Figure 1**.

Cell Lines

The MDA-MB-468 and SUM-149PT cell lines were obtained through American Type Culture Collection (ATCC, http://www.atcc.org) and maintained in culture according to ATCC recommendations. Cell lines were passaged no more than 30 times before being discarded. To facilitate automated image analysis for identifying and quantifying individual nuclei in time-lapsed microscopy experiments (described below), each cell line was modified to express a histone H2B conjugated to monomeric red fluorescent protein (H2BmRFP; Addgene Plasmid 18982) as previously described (Quaranta et al., 2009; Tyson et al., 2012).

To specifically modulate doxorubicin pharmacokinetics, the H2BmRFP-expressing MDA-MB-468 cell line (MDA-MB-468$_{H2B}$) was transduced to express a green fluorescent protein (GFP)-tagged MDR1 protein (ABCB1 gene, Origene Technologies, Rockville, MD). Following transduction, the cell line was cultured in 100 nM doxorubicin for 2 weeks to select a doxorubicin-resistant phenotype (MDA-MB-468$_{MDR1}$). These cells were serially imaged to ensure that all surviving cells stably expressed GFP.

The SUM-149PT cell line possesses a BRCA1 2288delT mutation (Elstrodt et al., 2006). BRCA1 is involved in maintaining genome stability through its role in repairing double strand DNA-breaks *via* homologous recombination (Gudmundsdottir and Ashworth, 2006). The BRCA1 mutation causes an increased reliance on alternate DNA damage repair pathways, such as non-homologous end joining (Farmer et al., 2005). The DNA damage repair pathway mediated by DNA-PK was targeted with a small molecule inhibitor to specifically modulate doxorubicin pharmacodynamics in the SUM-149PT cell line.

Chemicals

Doxorubicin was purchased from Sigma Aldrich (St. Louis, MO) and dissolved to a 1 mM stock concentration in sterile saline for subsequent experiments. Tariquidar (TQR) is a third-generation MDR1 inhibitor that non-competitively inhibits MDR1 function (Mistry et al., 2001). TQR is leveraged to modulate doxorubicin pharmacokinetics in the MDA-MB-468$_{MDR1}$ cell line. NU7441 is a DNA-PK inhibitor that has been investigated as a means to improve treatment response to DNA-damaging agents (Zhao et al., 2006; Ciszewski et al., 2014). NU7441 is used to modulate doxorubicin pharmacodynamics in the SUM-149PT cell line. TQR and NU7441 were both purchased from Selleckchem (Boston, MA). Each was dissolved to a 1 mM stock concentration in DMSO. We subsequently refer to these therapies (TQR and NU7441) as sensitizers. All solutions were stored in 250 μL aliquots at -80°C.

Doxorubicin Uptake Imaging and Image Processing

Time resolved fluorescent microscopy was employed to characterize the uptake of doxorubicin by each cell line (MDA-MB-468$_{H2B}$, MDA-MB-468$_{MDR1}$, and SUM-149PT) using a modification of the previously-published drug uptake assay (McKenna et al., 2017). The method leverages the intrinsic fluorescence of doxorubicin to quantify the movement of

doxorubicin from the extracellular space into cells. Briefly, each cell line was introduced into 96-well microtiter plates at ∼10,000 cells per well. Each well was imaged at 20–25 min intervals *via* fluorescent microscopy with a 20× objective in 2×2 image montages on a BD Pathway 855 Bioimager (BD Biosciences, San Jose, CA). Imaging began 1 h prior to and continued for approximately 24 h following application of 1 μM of doxorubicin. After 8 h, doxorubicin was removed *via* media replacement. This timeframe allowed for an extended observation of drug uptake without inducing morphological changes and cell death that would limit the effect of the measurement. To measure the effect of TQR and NU7441 on drug uptake kinetics in the MDA-MB-468$_{MDR1}$ and the SUM-149PT cell lines, respectively, each sensitizer was applied over a range of concentrations (250–2 nM for TQR and 2 μM−16 nM for NU7441 both *via* a 2-fold dilution series) 1 h prior to doxorubicin application. At least three replicates of each treatment condition were collected.

The collected images were subsequently post-processed to correct for uneven background illumination and to isolate the contribution of each fluorophore in the experiment. First, the illumination function for each image was estimated (Jones et al., 2006). The image is defined:

$$I = L\left(C + b\right)$$

where I is the image, L is the illumination function, C is signal from cells, and b is the background. The signal from cells was removed from each image through use of a median disc filter with a radius of 50, isolating b. To estimate L, the background-only images in each well were averaged over all timepoints. A smooth surface was fit to this averaged image, and the surface was normalized to a maximum value of 1. Each image in the time series was divided by this surface (L) to correct for uneven illumination. Following illumination correction, a threshold-based approach was used to segment each cell.

To account for the various fluorophores in the experiment (H2BRFP, MDR1GFP, and doxorubicin), a linear unmixing approach was employed to isolate the signal from each fluorophore to more precisely quantify doxorubicin accumulation (Zimmermann, 2005). The approach leverages spectral imaging data collected at multiple excitation and emission wavelengths to isolate the signal from each fluorophore. This method can also be used for background subtraction by modeling the background (here, the signal from cell culture media) as an additional fluorophore. For these experiments, we define four fluorophores of interest: MDR1GFP, doxorubicin, H2BRFP, and background. The observed images are modeled as a linear combination of the signals from each of these fluorophores:

$$\left[S_{H2B}\ S_{MDR}\ S_{Dox}\ S_{background}\right] T_{4\times n} = \left[I_1\ I_2\ \ldots\ I_n\right]$$

where S_{H2B} is the signal from the H2BRFP, S_{MDR} is the signal from the GFP-tagged MDR1, S_{Dox} is the signal from doxorubicin, and $S_{background}$ is the background signal from cell culture media. T is the transformation matrix that estimates the contribution from each fluorophore in creating each image I. In this work, five images ($n = 5$) were collected at each timepoint. The excitation, dichroic, and emission filters for each image are listed in **Table 2**.

TABLE 2 | Filter settings.

Image	Excitation (nm)	Dichroic (nm)	Emission (nm)
I_1	470/40	515, longpass	515, longpass
I_2	470/40	515, longpass	570, longpass
I_3	470/40	515, longpass	575/25
I_4	470/40	515, longpass	540/50
I_5	548/20	595, longpass	645/75

Fluorescence imaging filter sets used to collect pharmacokinetics data.

To construct T, images of each fluorophore were collected from control samples. Specifically, control images of GFP, H2BRFP-positive cells, doxorubicin, and background were collected. For each fluorophore, the image with the highest intensity is assumed to be the true image; i.e., the corresponding entry in T is set to 1. The relative intensity of the other four images with respect to the true image are then estimated. This normalized spectrum is deposited into the row of T corresponding to the current fluorophore. T is estimated at each timepoint to compensate for any temporal changes in fluorophore intensity.

With an estimate of T and a spectral image set for each well at each timepoint, the underlying signals (i.e., S_{H2B}, S_{MDR}, S_{Dox}, $S_{background}$) can be estimated using QR decomposition [implemented in MATLAB (Mathworks, Natick, MA)]. This can be done on a per-pixel basis as shown in **Supplementary Figure 1**. However, as we are only interested in the intracellular and extracellular doxorubicin signals, the average value from each image in the intracellular and extracellular ($I_{i,I}$, $I_{i,E}$) space was calculated using a cell segmentation (as detailed above). Each signal can then be recovered:

$$\begin{bmatrix} S_{H2B,I} & S_{MDR,I} & S_{Dox,I} & S_{background,I} \\ S_{H2B,E} & S_{MDR,E} & S_{Dox,E} & S_{background,E} \end{bmatrix} T_{4\times n} = \begin{bmatrix} I_{1,I} & \cdots & I_{5,I} \\ I_{1,E} & \cdots & I_{5,E} \end{bmatrix}$$

where $S_{Dox,I}$ and $S_{Dox,E}$ are the signals from doxorubicin in the intracellular and extracellular spaces, respectively. Similar definitions apply for the other signals S.

Finally, S_{Dox} is converted into doxorubicin concentration. We assume that doxorubicin signal is linearly proportional to its concentration, [Dox] (McKenna et al., 2017):

$$S_{Dox} = a\,[Dox] + b$$

To calibrate this model, images are collected on a series of wells containing a range of known doxorubicin concentrations. Estimates of a and b were obtained by fitting the doxorubicin signal equation to these control data. The image processing pipeline is illustrated in **Supplementary Figure 1**.

Doxorubicin Treatment Response Imaging

Using the previously-published dose-response assay, each cell line was treated with a range of doxorubicin concentrations

(5,000–10 nM *via* a 2-fold dilution series) for 24 h as monotherapy. Additionally, the sensitizing effects of TQR and NU7441 in the MDA-MB-468$_{MDR1}$ and the SUM-149PT cell lines, respectively, were investigated by applying those therapies over a range of concentrations 1 h prior to application of doxorubicin. TQR concentrations in a 2-fold dilution series from 250 to 2 nM were used for the MDA-MB-468$_{MDR1}$ cell line, and NU7441 concentrations in a 2-fold dilution series from 2 μM to 15 nM were used for the SUM-149PT cell line. These combination studies were each performed at three doxorubicin concentrations. All drug (doxorubicin and sensitizer) was removed from each well *via* media replacement at 24 h. These cells were imaged daily *via* fluorescent microscopy for at least 15 days following treatment. For these studies, fluorescence microscopy images were collected using a Synentec Cellavista High End platform (SynenTec Bio Services, Münster, Germany) with a 20× objective and tiling of 25 images. To generate images, the H2BmRFP fluorophores were excited with 529 nm light for 650 ms, and emissions were collected at 585 nm. Nuclei were segmented and counted in ImageJ (http://imagej.nih.gov/ij/) using a previously-described, threshold-based method (Frick et al., 2015) to quantify cell population. Six replicates of each treatment condition were collected. Media was refreshed every 3 days for the duration of each experiment to ensure sufficient growth conditions for surviving cells. Data were manually truncated when cell populations reached carrying capacity. At this point, signals from neighboring nuclei overlap, and the cell counting algorithm becomes unreliable.

Model Fits

The three-compartment model described in Equations (1–3) was fit to the uptake data under each treatment condition (doxorubicin monotherapy and doxorubicin combination with sensitizer) for each cell line using a non-linear least squares optimization implemented in MATLAB. Of note, each cell line is assumed to have a single set of compartment model parameters (k_{EF}, k_{FE}, and k_{FB}) for each sensitizer concentration; i.e., a parameter set for doxorubicin monotherapy and a set for each sensitizer concentration. The mean errors of the best-fit model across all timepoints and treatment conditions with respective standard deviations are reported. Similarly, the pharmacodynamic model described by Equations (4–6) was fit to the dose response data from all treatment conditions (i.e., doxorubicin monotherapy and doxorubicin combination with sensitizer) for each cell line. Each treatment condition in each cell line was fit independently, yielding cell line- and treatment condition-specific parameter values. This was also accomplished through a non-linear least squares optimization implemented in MATLAB, and we report the mean percent errors of the best-fit models across all timepoints. For additional details on the model fitting procedure see McKenna et al. (McKenna et al., 2017).

Measurement of Pharmacologic Properties With Equivalent Dose

We assume, by definition, that each unique treatment response timecourse corresponds to a specific equivalent dose. As the equivalent dose is perfectly known for doxorubicin monotherapy [i.e., the equivalent dose is simply C_B, which can be directly calculated with k_{FE}, k_{EF}, and k_{FB} values measured from drug uptake studies], the equivalent dose for co-treatment conditions can be estimated by comparing treatment response dynamics from co-treatment conditions to those from doxorubicin monotherapy treatments. With appropriate experimental design to isolate each equivalent dose parameter (i.e., k_{FE}, k_{EF}, and k_{FB}), this approach can quantify the effect of each sensitizing agent on their PK/PD pathway. Specifically, by assuming the effect of each sensitizing therapy is limited to a single equivalent dose parameter, the effect of TQR on k_{EF} and the effect of NU7441 on k_{FB} can be measured. As response under all treatment conditions (i.e., doxorubicin monotherapy and co-treatment with a sensitizer) can by summarized by the parameters in Equations (4-6) (i.e., $p = [k_{d,a}, k_{d,b}, r]$), we use model parameters to compare treatment response timecourses.

The response parameters (p) from doxorubicin monotherapy experiments are first interpolated with respect to equivalent dose *via* a local linear approach. This yields a continuous set of parameters (p_{est}) across all possible equivalent doses in the range from no treatment to maximal doxorubicin dose. The fit parameter values (p_{fit}) for each of the m co-treatment conditions are then matched to the interpolated parameters from doxorubicin-only treatment conditions (p_{est}) to estimate the equivalent dose (D_{est}) for each co-treatment condition. Specifically, D_{est} is the set of equivalent doses that correspond to the best matches between p_{fit} and p_{est} in the L_2 norm sense (i.e., $\min \left\| p_{est} - p_{fit} \right\|_2$). This process is illustrated in **Supplementary Figure 2**. The following constrained objective function, $G(k_x)$, can then be used to estimate k_x (the equivalent dose parameter under investigation; e.g., k_{EF} and k_{FB}) for each of the n sensitizer concentrations:

$$G(k_x) = \min_{k_x} \sum_{i=1}^{m} \left(D_{est,i} - D_i(k_x) \right)^2$$
$$such\ that\ k_{x,q+1} - k_{x,q} \geq 0\ \forall\ q = [1, ..., n] \qquad (7)$$

where the D_i is the equivalent dose calculated for each co-treatment condition as described below, and $D_{est,i}$ is the estimated equivalent dose for the i^{th} co-treatment condition. Specifically, in calculating $D_i$for the NU7441 experiments, we fix k_{EF} and k_{FE} values and optimize k_{FB} values corresponding to each sensitizer concentration in the co-treatment conditions. The constraints in the objective function ensure that k_{FB} increases monotonically with sensitizer concentration. Similarly, for the TQR experiments, we fix k_{EF} and k_{FB} and optimize k_{FE} for each sensitizer concentration. This objective function was minimized *via* a constrained optimization routine implemented in MATLAB. The non-parametric interpolation and optimization procedures were utilized as we did not assume any functional relationships between model parameters and equivalent dose. While this fitting procedure could have been made more robust by proposing such functional relationships, we implemented this non-parametric approach to allow for greater generalizability.

Comparison of Cell Lines With Equivalent Dose

As the MDA-MB-468$_{MDR1}$ line was engineered from the MDA-MB-468$_{H2B}$ line, we hypothesize that the response of these cell lines to doxorubicin therapy is not significantly different when compared *via* equivalent dose. Specifically, the mechanism of action of MDR1 is to increase drug efflux, which effectively reduces the equivalent dose in the MDA-MB-468$_{MDR1}$ line for a given treatment timecourse. Indeed, the proposed equivalent dose metric was developed to account for the differing pharmacokinetic properties between these cell lines to more precisely compare their respective responses to therapy. To test this hypothesis, survival of the parental MDA-MB-468$_{H2B}$ cell line is compared to that of the MDA-MB-468$_{MDR1}$ cell line. This comparison is made utilizing a conventional treatment response assay in which survival is assessed 72 h following treatment. Specifically, each cell line was treated with a range of doxorubicin concentrations (5,000–10 nM *via* a 2-fold dilution series) for 24 h as monotherapy, and survival was assessed *via* cell counting. Survival data for each cell line was fit with a pair of Hill functions. The first of these Hill functions assumed the dose to be the applied doxorubicin concentration. The second utilized the equivalent dose (D_{eq}) calculated with cell-line specific k_{EF}, k_{FE}, and k_{FB} values. We report the EC_{50} (drug concentration at half-maximal effect) for each cell line as measured *via* extracellular doxorubicin concentration and equivalent dose.

RESULTS

Treatment Response in MDA-MB-468$_{MDR1}$ Cell Line

The measured intracellular doxorubicin concentration timecourses for the MDA-MB-468$_{MDR1}$ cell line under doxorubicin monotherapy and combination therapy with TQR are shown in **Figure 2A**. Intracellular doxorubicin increases with TQR concentration. The average intracellular concentration at the end of each experiment, estimated with the last 10 timepoints, is significantly different among the treatment groups (one-way ANOVA, $p < 1e\text{-}5$). Equations (1–3) are fit to these data, and the best-fit models are overlaid on the timecourses. The mean error of the best-fit pharmacokinetic models was 45.6 (±47.4) nM across all timepoints and treatment conditions, and the corresponding model parameters are shown in **Figures 2B–D**. Increasing TQR concentrations decrease doxorubicin efflux in the MDA-MB-468$_{MDR1}$ cell line in a dose-dependent manner. For example, the efflux rate (k_{FE}) is decreased from 0.216 (±0.028) h^{-1} to 0.046 (±0.008) h^{-1} as TQR increases from 2 to 250 nM (the bounds here and below correspond to the 95% confidence interval of the parameter estimates). k_{EF} values varied with TQR concentration, all falling within $[1.63, 3.46] \times 10^{-6}$ h^{-1}.

Treatment response timecourses for the MDA-MB-468$_{MDR1}$ cell line under doxorubicin combination therapy with TQR are shown in **Figures 2E–G**. Equations (4–6) are fit to these data, and the best-fit models are overlaid on the observed cell counts. Model parameters are shown in **Figures 2H–J**. For a fixed concentration of doxorubicin, increasing concentrations of TQR incrementally sensitize cells to doxorubicin. For example, at a fixed dose of 156 nM doxorubicin, increasing the TQR concentration from 0 to 250 nM increased the death rate ($k_{d,a}$) from -0.16 (±0.23) $\times 10^{-2}$ h^{-1} to 2.21 (±0.1) $\times 10^{-2}$ h^{-1}. TQR monotherapy did not affect the growth of these cells as shown in **Supplementary Figure 3**. Treatment response timecourses of the MDA-MB-468$_{MDR1}$ line to doxorubicin monotherapy are shown in **Figure 3A**. These data are fit with Equations (4–6), and the best-fit models are overlaid on the observed cell counts. The mean percent error of the best-fit model across all timepoints and treatment conditions is 10.3%. Prior to treatment, the MDA-MB-468$_{MDR1}$ line demonstrated a proliferation rate (k_p) of 2.12 (±0.03) $\times$ 10^{-2} h^{-1}. Treatment response varied smoothly with doxorubicin concentration, and this response is quantified by the parameters in **Figures 3B–D**. Notably, high variance in parameter estimates is observed as values of r approach 0.05 h^{-1} and values of $k_{d,b}$ approach 0 h^{-1}. There exists intrinsic uncertainty at this limit as the rapid dynamics (r) coupled with small $k_{d,b}$ effects cannot be resolved by the current data. This uncertainty in r for small $k_{d,b}$ does not affect model predictions as demonstrated by a sensitivity analysis in previous work (McKenna et al., 2017).

By leveraging the proposed mechanistic model and equivalent dose statistic, k_{FE} values for each TQR concentration can be estimated using the measured treatment response data and the optimization routine outlined in section Measurement of Pharmacologic Properties With Equivalent Dose. To make these measurements, the equivalent dose for each doxorubicin monotherapy condition was first calculated with the PK model parameters measured in the doxorubicin uptake studies. Specifically, k_{FE}, $k_{EF,}$ and k_{FB} were measured to be 0.313 h^{-1}, 3.08 $\times$ 10^{-6} h^{-1} and 0.0212 h^{-1}, respectively. The equivalent dose statistic was then estimated for each co-treatment condition. To perform this estimation, treatment response parameters from co-treatment conditions were matched to those from doxorubicin monotherapy conditions (**Figures 4A–C**). As the equivalent doses for all monotherapy conditions are perfectly known (i.e., $C_B = D_{eq}$ for doxorubicin monotherapy), the equivalent dose for each co-treatment condition can be estimated with the matching process illustrated in **Supplementary Figure 2**. Briefly, parameter values are estimated across a range of equivalent doses utilizing parameters from the doxorubicin monotherapy experiments. The equivalent dose for each treatment condition can then be estimated by matching measured parameter values to those estimates. To demonstrate the efficacy of the parameter matching in comparing treatment response timecourses, a subset of responses from doxorubicin monotherapy and co-treatment conditions are color-coded to their estimated equivalent dose (**Figures 4D–F**). Note similar dynamics for similarly-colored data, indicating the efficacy of the parameter matching in comparing treatment response timecourses. With estimates of equivalent dose for all co-treatment conditions, the k_{FE} value for each TQR concentration was estimated with the optimization routine summarized by Equation (7). As we hypothesized that the effect of TQR is limited to k_{FE} (**Figure 4G**), k_{EF} and k_{FB} values were fixed to the values reported above in the optimization routine. The optimized k_{FE} values for all

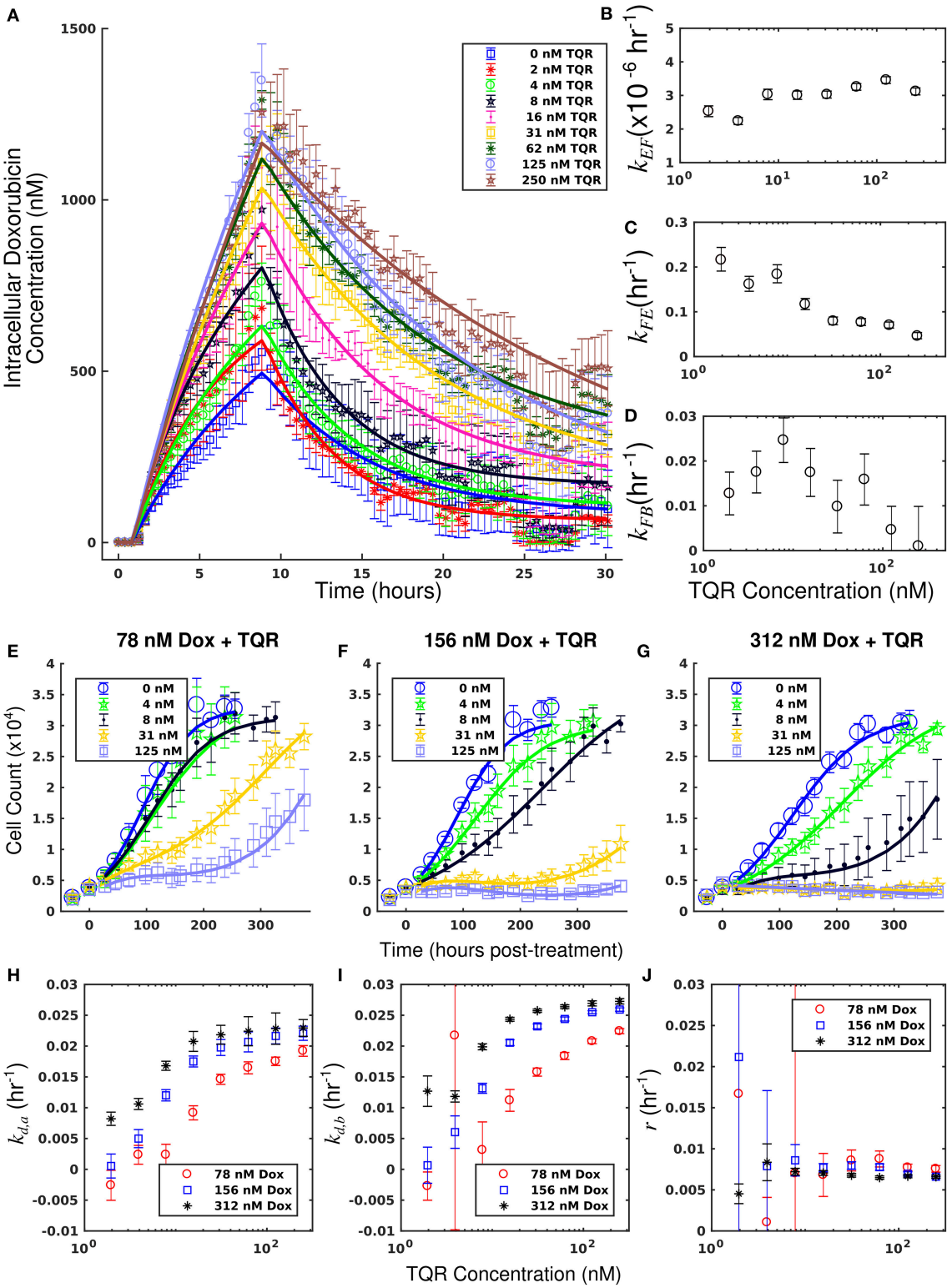

FIGURE 2 | Doxorubicin and TQR combination studies in the MDA-MB-468$_{MDR1}$ cell line. Timecourses of the mean intracellular concentration of doxorubicin with corresponding standard deviations are shown for each treatment condition in **(A)**. Doxorubicin accumulation increases along with TQR concentrations. Equations (1–3) were fit to the data, and the best-fit models are overlaid on the data (smooth lines) in a. Model parameter fits corresponding to the best-fit models are shown in **(B–D)**. Similar k_{EF} and k_{FB} vales are observed across all TQR concentrations. There is a trend of decreasing k_{FE} values with increasing TQR concentrations

(Continued)

FIGURE 2 | (C), consistent with MDR1 inhibition by TQR. Cell counts of MDA-MB-468$_{MDR1}$ following combination treatment with TQR and doxorubicin are show in panels **(E-G)**. In each plot, a fixed concentration of doxorubicin is applied with variable TQR concentrations. These counts are fit with Equations (4–6) as described in section Model Fits, and the best-fit model is overlaid on the cell counts [smooth lines in panels **(E–G)**]. Error bars represent the 95% CI from six experimental replicates for each treatment condition. Model parameters with corresponding 95% CI are shown in **(H–J)** as a function of TQR concentration. For each doxorubicin concentration, the death rate ($k_{d,a}$ and $k_{d,b}$) increased with TQR concentration **(H,I)**.

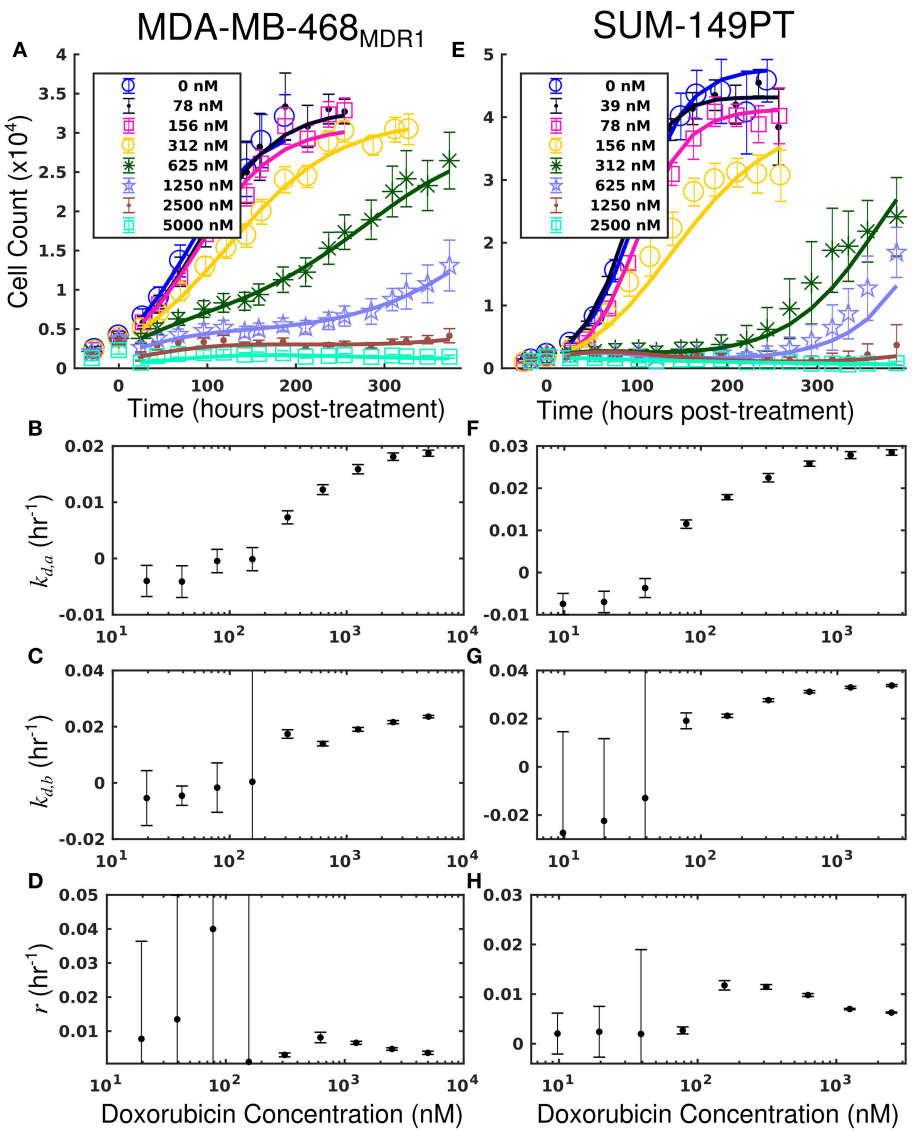

FIGURE 3 | Treatment response in MDA-MB-468$_{MDR1}$ (left column) and SUM-149PT (right column) cell lines under doxorubicin monotherapy. The top row [panels **(A,E)**] shows cell counts over time from treatment response studies for each cell line. For these studies, cells were treated with a fixed concentration of doxorubicin for 24 h. These counts are fit to Equations (4–6) as described in section Model Fits, and the best-fit model is overlaid on the cell counts [smooth lines in **(A,E)**]. Error bars represent the 95% CI from six experimental replicates for each treatment condition. Model parameters with corresponding 95% CI are shown in the bottom three rows as a function of doxorubicin concentration. Panels **(B–D)** show fits from the MDA-MB-468$_{MDR1}$ experiments, and panels **(F–H)** show fits from the SUM-149PT experiments. For each doxorubicin concentration for each cell line, the death rate ($k_{d,a}$ and $k_{d,b}$) increased with increasing doxorubicin concentrations.

TQR concentrations are shown in **Figure 4H**. Decreasing k_{FE} values were observed with increasing TQR concentrations, matching the measurements from the uptake studies in **Figure 2**.

The equivalent dose can summarize all treatment conditions in the MDA-MB-468$_{MDR1}$ cell line and is predictive of response. Further, this statistic can be leveraged to quantify the effect of TQR.

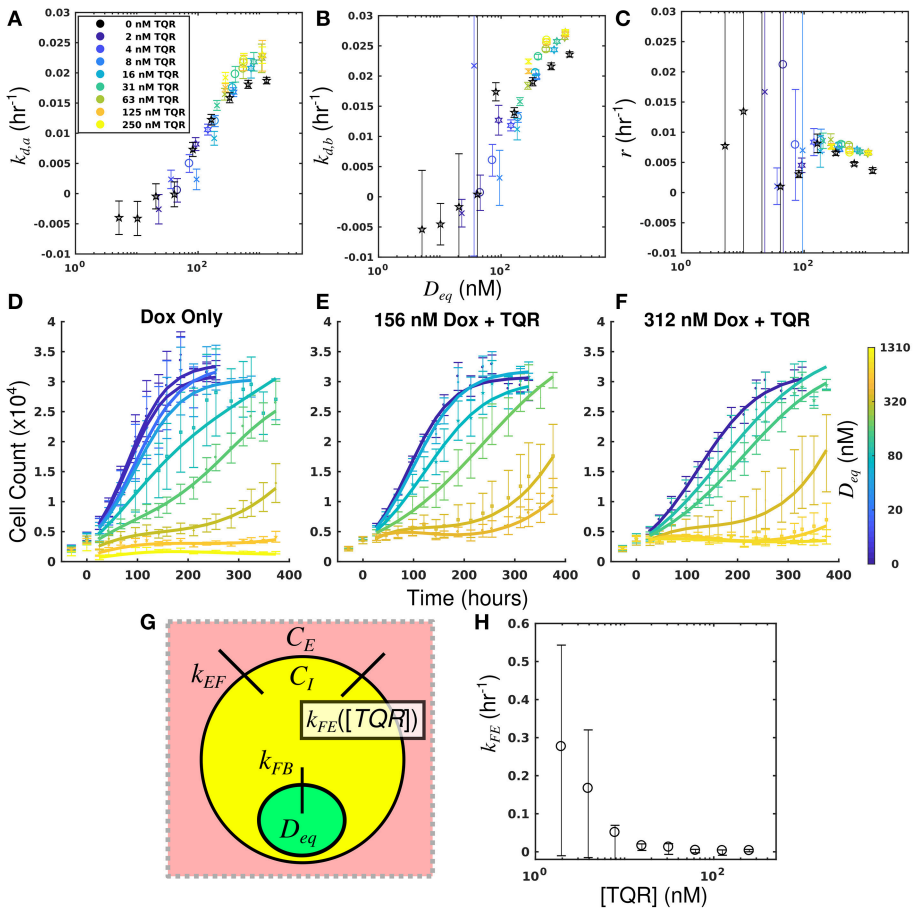

FIGURE 4 | Leveraging equivalent dose to estimate the effect of TQR in the MDA-MB-468$_{MDR1}$ cell line. The equivalent dose for each doxorubicin monotherapy condition was first calculated with the PK model parameters measured in the doxorubicin uptake studies. The equivalent dose statistic was then estimated for each co-treatment condition by matching treatment response parameters from co-treatment conditions to those from doxorubicin monotherapy conditions. Parameter values from all doxorubicin monotherapy and co-treatment conditions are plotted as a function of equivalent dose **(A-C)**. A subset of responses from doxorubicin monotherapy and co-treatment conditions are color-coded to their estimated equivalent dose **(D–F)**. Similar dynamics are observed with similarly-colored data, demonstrating the efficacy of the parameter matching in comparing treatment response timecourses. As TQR impairs the function of the MDR1 pump, we hypothesized the effect of TQR is limited to the k_{FE} parameter **(G)**. With estimates of equivalent dose for all treatment conditions, the k_{FE} value for each TQR concentration was estimated with the optimization routine summarized by Equation (7) **(H)**. These values, calculated with treatment response data, agree well with direct measurements of k_{FE} reported in **Figure 2**. We note the large confidence intervals are a result of the optimization approach, in which the value ($1/k_{FE}$) was optimized.

Treatment Response in SUM-149PT Cell Line

The measured intracellular doxorubicin concentration timecourses for the SUM-149PT cell line under doxorubicin monotherapy and combination therapy with NU7441 are shown in **Figure 5A**. NU7441 treatment did not affect intracellular doxorubicin accumulation following treatment. The average intracellular doxorubicin concentration at the end of each experiment, estimated with the last 10 timepoints, did not demonstrate significant differences at the $p = 0.05$ level (one-way ANOVA). Equations (1–3) are fit to the uptake data, and the best-fit model is overlaid on the timecourses. The corresponding model parameters are shown in **Figures 5B–D**. The mean error of the best-fit pharmacokinetic model was 77.9 ($\pm$71.4) nM across all treatment conditions and timepoints. Further, similar values of k_{FE}, k_{EF}, and k_{FB} are observed across all NU7441 concentrations (**Figures 5B–D**). Given its effect on DNA-PK, NU7441 is not expected to affect intracellular doxorubicin accumulation.

Treatment response timecourses for the SUM-149PT cell line under doxorubicin co-treatment with NU7441 are shown in **Figures 5E–G**. Equations (4–6) are fit to these data, and the best-fit models are overlaid on the observed cell counts. Model parameters are shown in **Figures 5H–J**. For a fixed concentration of doxorubicin, increasing concentrations of NU7441 incrementally sensitized cells to doxorubicin. For example, with a fixed dose of 156 nM doxorubicin, NU7441 concentrations increased the death rate ($k_{d,a}$) from 0.25 ($\pm$0.16) $\times$ 10^{-2} h^{-1} to 2.00 ($\pm$0.06) $\times$ 10^{-2} h^{-1}. NU7441 monotherapy did not affect the growth of these cells as shown in **Supplementary Figure 3**. Treatment response timecourses of the SUM-149PT line to doxorubicin monotherapy are shown in

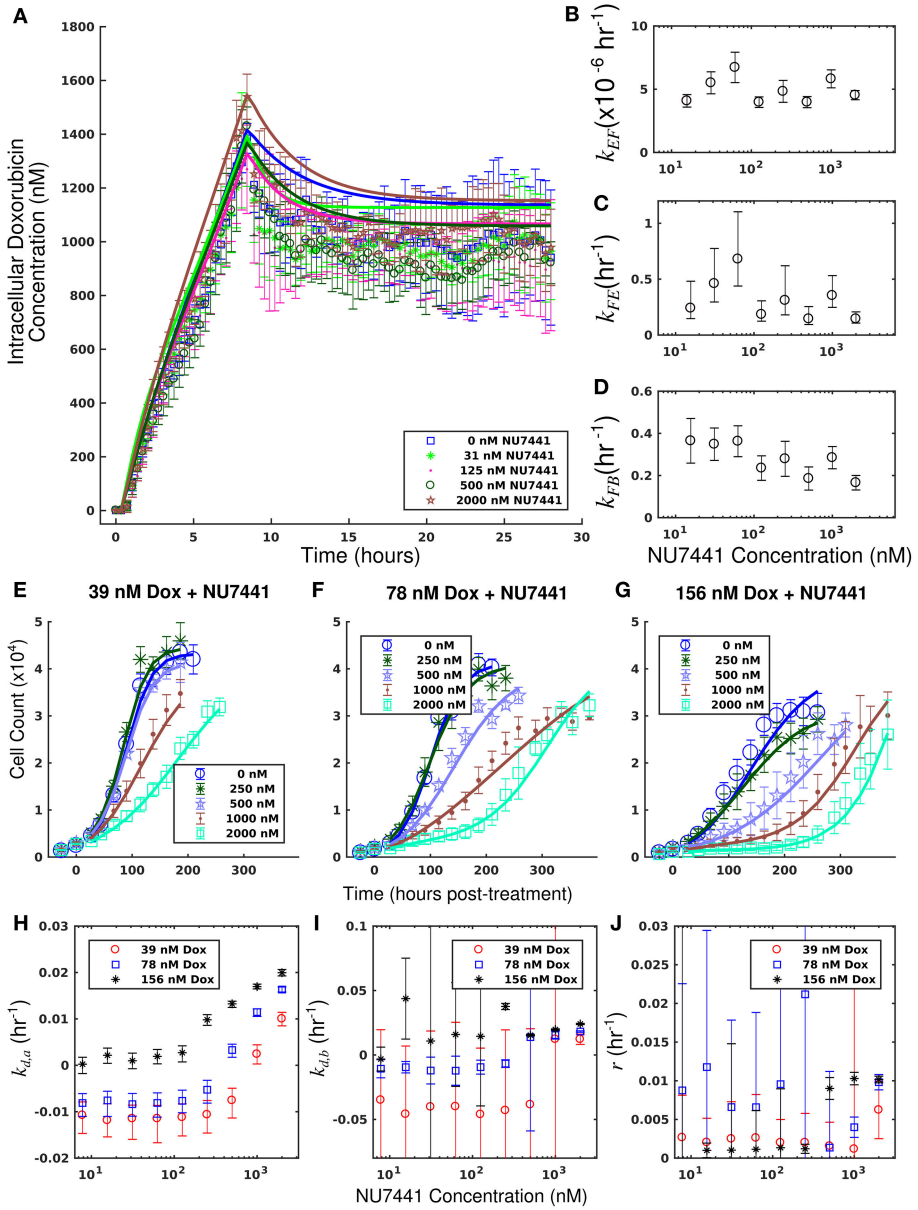

FIGURE 5 | Doxorubicin and NU7441 combination studies in the SUM-149PT cell line. Timecourses of the mean intracellular concentration of doxorubicin with corresponding standard deviations are shown for each treatment condition in a. No significant difference in doxorubicin accumulation was observed as a function of NU7441 concentration. Equations (1–3) were fit to the data, and the best-fit models are overlaid on the data (smooth lines) in **(A)**. Model parameter fits corresponding to the best-fit models are shown in **(B–D)**. For each model parameter, similar vales were observed across all NU7441 concentrations, consistent with the similar intracellular doxorubicin timecourses in **(A)**. Counts of SUM-149PT cells following combination treatment with NU7441 and doxorubicin are show in panels **(E–G)**. In each plot, a fixed concentration of doxorubicin is applied with variable NU7441 concentrations. These counts are fit with Equations (4–6) as described in section Model Fits, and the best-fit models are overlaid on the cell counts [smooth lines in panels **(E–G)**]. Error bars represent the 95% CI from six experimental replicates for each treatment condition. Model parameters with corresponding 95% CI are shown in panels **(H–J)** as a function of NU7441 concentration. For each doxorubicin concentration, the death rate ($k_{d,a}$) increased with NU7441 concentration **(H)**. The parameters shown in panels **(I,J)** are unable to be resolved with the current data as discussed in section Model Fits.

Figure 3E. These data are fit with Equations (4–6), and the best-fit models are overlaid on the observed cell counts. The mean percent error of the best-fit model across all treatment conditions is 11.9%. Prior to treatment, the SUM-149PT line demonstrated a proliferation rate (k_p) of 2.58 ($\pm$0.03) $\times$ 10^{-2} h^{-1}. Treatment response varied smoothly with doxorubicin concentration, and this response is quantified by the parameters in **Figures 3F–H**.

By leveraging the proposed mechanistic model and equivalent dose statistic, k_{FB} values for each NU7441 concentration can be estimated using the measured treatment response data

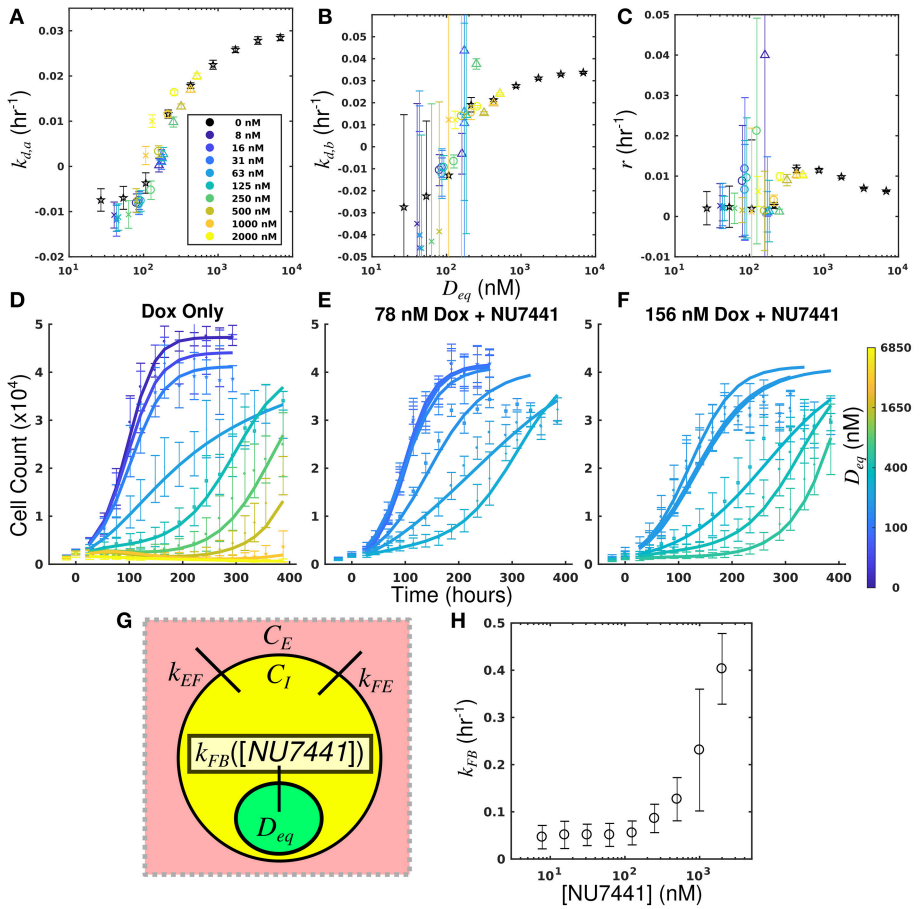

FIGURE 6 | Leveraging equivalent dose to estimate the effect of NU7441 in the SUM-149PT cell line. The equivalent dose for each doxorubicin monotherapy condition was first calculated with the PK model parameters measured in the doxorubicin uptake studies. The equivalent dose statistic was then estimated for each co-treatment condition by matching treatment response parameters from co-treatment conditions to those from doxorubicin monotherapy conditions. Parameter values from all doxorubicin monotherapy and co-treatment conditions are plotted as a function of equivalent dose **(A-C)**. A subset of responses from doxorubicin monotherapy and co-treatment conditions are color-coded to their estimated equivalent dose **(D–F)**. Similar dynamics are observed with for similarly-colored data, demonstrating the efficacy of the parameter matching in comparing treatment response. As NU7441 impairs the function DNA-PK, we hypothesized the effect of NU7441 is limited to the k_{FB} parameter **(G)**. With estimates of equivalent dose for all treatment conditions, the k_{FB} value for each NU7441 concentration was estimated with the optimization routine summarized by Equation (7). Increasing values of k_{FB} are observed with increasing NU7441 concentrations, indicating an increase in functional drug bound **(H)**. These values cannot be directly observed with the uptake study, demonstrating the utility of the equivalent dose in estimating parameters that cannot be directly measured with current techniques.

and the optimization routine summarized by Equation (7). To make these measurements, the equivalent dose for each doxorubicin monotherapy condition was first calculated with the PK model parameters measured in the doxorubicin uptake studies. Specifically, k_{EF}, k_{FE}, and k_{FB} were measured to be 4.00 × 10^{-6} h^{-1} and 0.165 h^{-1}, and 0.236 h^{-1}, respectively. These were calculated by fitting the SUM-149PT uptake studies assuming constant parameters for all NU7441 concentrations. The equivalent dose statistic was then estimated for each co-treatment condition. To perform this estimation, treatment response parameters from co-treatment conditions were matched to those from doxorubicin monotherapy conditions (**Figures 6A–C**). As the equivalent doses for all monotherapy conditions are perfectly known, the equivalent dose for each co-treatment condition can be estimated with this matching process. To demonstrate the efficacy of the parameter matching in comparing treatment response timecourses, a subset of responses from doxorubicin monotherapy and co-treatment conditions are color-coded to their estimated equivalent dose (**Figures 6D–F**). Note similar dynamics for similarly-colored data. With estimates of equivalent dose for all co-treatment conditions, the k_{FB} value for each NU7441 concentration was estimated with the optimization routine summarized by Equation (7). As we hypothesized that the effect of NU7441 is limited to k_{FB}(**Figure 6G**), k_{FE} and k_{EF} values were fixed to the values reported above in the optimization routine. The optimized k_{FB} values for all NU7441 concentrations are shown in **Figure 6H**. Increasing k_{FB} values were observed with increasing NU7441 concentrations, indicating the functional increase in drug with NU7441, mediated through its effect on DNA-PK. We note that the DNA

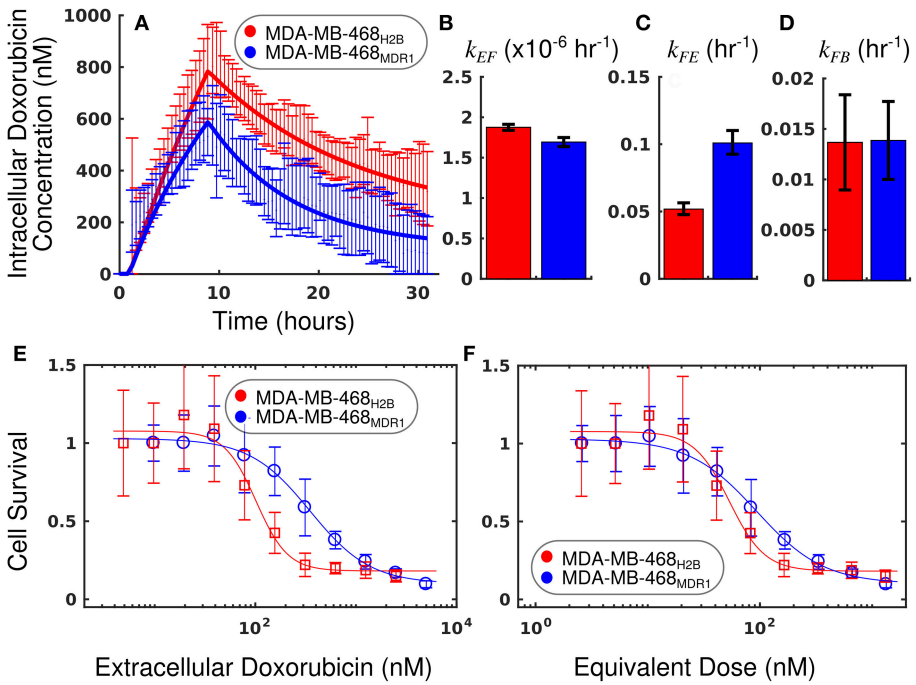

FIGURE 7 | Comparison of MDA-MB-468$_{H2B}$ and MDA-MB-468$_{MDR1}$ cell lines using equivalent dose. The intracellular doxorubicin concentration with 95% CI for each cell line is shown in **(A)**. The MDA-MB-468$_{H2B}$ line demonstrates increased intracellular accumulation of doxorubicin relative to the MDA-MB-468$_{MDR1}$ line. Equations (1–3) are fit to the doxorubicin uptake data, and the best-fit models are overlaid on the data in a (smooth line). The corresponding parameters with 95% CI are shown in **(B–D)**. The MDA-MB-468$_{H2B}$ data are shown in red, and the MDA-MB-468$_{MDR1}$ data are shown in blue. Notably, the efflux of drug from the MDA-MB-468$_{MDR1}$ (k_{FE}) line is significantly greater than the corresponding rate in the MDA-MB-468$_{H2B}$ line ($p < 0.05$). Treatment response is traditionally summarized by cell survival and plotted against applied drug concentration. The cell count relative to control for each cell line is shown as a function of extracellular doxorubicin concentration and equivalent dose in **(E,F)**, respectively. While a significant difference is observed when comparing these cell lines *via* EC_{50} calculated with the extracellular doxorubicin concentration, no significant difference is observed when comparing the EC_{50} statistic derived from the equivalent dose. The equivalent dose can account for the differing pharmacokinetic properties to reveal similar doxorubicin pharmacodynamics in these cell lines.

repair pathway affected by NU7441 is not directly measured in the uptake studies. Recall from section Equivalent Dose that k_{FB} is a mixed measure of doxorubicin binding and DNA repair and describes the functional net binding rate. Thus, these values cannot be directly compared to the values extracted from the uptake study.

The equivalent dose can summarize all treatment conditions in the SUM-149PT cell line and is predictive of response. Further, this statistic can be leveraged to quantify the specific effect of NU7441 with observed treatment response data.

Comparison of MDA-MB-468$_{MDR1}$ and MDA-MB-468$_{H2B}$

The measured intracellular doxorubicin concentration timecourses with accompanying best-fit models for the MDA-MB-468$_{H2B}$ and MDA-MB-468$_{MDR1}$ cell lines are shown in **Figure 7**. Decreased doxorubicin accumulation was observed in the MDA-MB-468$_{MDR1}$ cell line relative to its parental line, MDA-MB-468$_{H2B}$. Notably, drug efflux was significantly elevated in the MDA-MB-468$_{MDR1}$ line relative to its parental line with k_{FE} values of 1.01 ($\pm$0.08) $\times$ 10^{-1} h^{-1} and 0.52 ($\pm$0.04) $\times$ 10^{-1} h^{-1}, respectively ($p < 0.05$). The mean errors of the best-fit pharmacokinetic models across all timepoints were 44.7 and 58.7 nM for the MDA-MB-468$_{H2B}$ and the MDA-MD-468$_{MDR1}$ lines, respectively.

The survival of each cell line 72 h following treatment is compared as a function of extracellular doxorubicin concentration and equivalent dose in **Figures 7E,F**. The EC_{50} as measured with the extracellular doxorubicin concentration for the MDA-MB-468$_{H2B}$ and the MDA-MB-468$_{MDR1}$ are 101.6 ($\pm$28.9) and 350.6 ($\pm$109) nM, respectively. These measures indicate that there is a statistically significant difference between these cell lines ($p < 0.05$, t-test). The EC_{50} as measured with the equivalent dose for the MDA-MB-468$_{H2B}$ and the MDA-MB-468$_{MDR1}$ lines are 53.3 ($\pm$15.1) and 93.7 ($\pm$29.2) nM, respectively. These values are not different at $p = 0.05$ (t-test), indicating the similarity of these lines. Indeed, the only difference between cell lines is the overexpression of the MDR1 efflux pump. The intrinsic sensitivity of these cell lines to treatment should remain similar, and the equivalent dose reflects this similarity. The response of the MDA-MB-468$_{H2B}$ and MDA-MB-468$_{MDR1}$ cell lines are not significantly different as measured by D_{eq}.

DISCUSSION

We have proposed and demonstrated the utility of a mathematical modeling framework to quantify pharmacologic properties. We further proposed a new metric, the equivalent dose (D_{eq}), which provides a biochemically-based measure of treatment effect. With the data presented here, we show that a mechanistic mathematical model of treatment response can succinctly summarize a range of treatments to allow for more precise comparison of treatment response among cell lines. Further, we have shown how this model provides quantitative biological insight into the biochemical drivers of treatment response. We demonstrate that a mathematical modeling framework allows for quantification of pharmacologic processes through population-scale measurements.

Treatment response is driven by cell-line specific pharmacologic properties. Conventional summary statistics of treatment response data often conflate these pharmacologic properties, limiting their utility. To more effectively advance the study of treatment response, methods that explicitly consider this variability are needed to more precisely quantify biological drivers of treatment response. While previous treatment response assays provide insight in the relative sensitivity of a cell line to therapy (Fallahi-Sichani et al., 2013), the proposed approach quantifies specific drivers of treatment sensitivity. Through the approach proposed in this work, we demonstrate how intracellular pharmacologic properties can be quantified using limited data from population-level observations of treatment response.

This work is limited by its use of doxorubicin, which is intrinsically fluorescent, thereby allowing for the uptake model to be fit with experimental data. However, this approach need not be limited to fluorescent drugs. With appropriate experimental design, the approach summarized by Equation (7) can be leveraged to quantify any of the rates proposed in the model. Indeed, the optimized values of doxorubicin efflux in the MDA-MB-468$_{MDR1}$ line in **Figure 4** are similar to those values measured by the uptake assay in **Figure 2**. Further, the effect of NU7441 in altering pharmacokinetics was quantified using only the treatment response data, as this effect cannot be directly measured in the uptake assay. Importantly, this work demonstrates that all treatment conditions collapse onto a single, smooth trajectory through parameter space as a function of equivalent dose, and this property can be leveraged to provide quantitative insight into the biological drivers of treatment response. While cell lines could not be compared without precise estimates of all model parameters, this approach can nevertheless be used to quantify therapeutic perturbations within a given cell line. It is straightforward to extend the proposed modeling approach as a means to more precisely quantify the effects of other parameters in the experimental microenvironment (e.g., how does pH or a specific nutrient concentration affect treatment response?). In this way, these variables can be mapped onto a unified treatment response framework to more efficiently advance precision medicine approaches. More generally, the approach outlined in this work demonstrates how mathematical modeling can be used as a "filter" to derive more specific measures from experimental data to advance systems biology.

Therapies that target PK/PD pathways offer the potential to sensitize cells to cytotoxic therapies, increasing the efficacy of therapy and allowing for lower doses of such therapeutics. The approach proposed in this work provides a means to quantify the respective contributions of PK/PD pathways, providing mechanistic insight into treatment response. This approach differs from current methods used to assess drug synergism and antagonism (Chou, 2006; Jones et al., 2014; Foucquier and Guedj, 2015; Chen and Lahav, 2016; Lederer et al., 2018). These methods have great utility in discovering and quantifying drug interactions; however, they cannot be leveraged to understand the mechanisms underlying the identified synergy/antagonism. While other methods have leveraged mechanistic data to identify synergy (Al-Lazikani et al., 2012; Gao et al., 2017; Yin et al., 2018), the proposed equivalent dose framework provides quantitative mechanistic insight into intracellular drug effects and allows for predictions of treatment response under a variety of treatment conditions. We posit that this mechanistic approach could facilitate clinical translation of combination therapies. Notably, therapeutic approaches intended to sensitize tumors to doxorubicin have demonstrated great preclinical activity; however, their efficacy has been limited in clinical trials. Specifically, negative results have been seen with TQR due to excess toxicities and inactivity (Pusztai et al., 2005; Fox and Bates, 2007). Similarly, DNA-PK inhibitors such as NU7441 have yet to demonstrate an effect clinically despite their preclinical promise (Zhao et al., 2006; Helleday et al., 2008; Davidson et al., 2013). We posit that the proposed modeling framework can be used to identify more effective strategies for dosing and assessing these therapeutics. In particular, the proposed modeling approach can provide *precise* guidance on the necessary dose adjustments to achieve a desired effect in the context of combination therapy. As we have demonstrated, a target equivalent dose can be achieved in a variety of ways. For example, the extracellular drug concentration timecourse can be tuned to reach a specified equivalent dose. Alternatively, the same equivalent dose can be achieved by altering cell line pharmacologic properties through sensitizers with concomitant changes in the extracellular doxorubicin timecourse. While realizing this goal *in vivo* will require a more complete model of treatment response (i.e., one that incorporates plasma pharmacokinetics and organ system toxicities), we have demonstrated the proposed model to be robust to various doxorubicin treatments and is general to sensitizing agents.

While the results of this study are promising, several limitations exist in the current approach. The first order pharmacokinetics model assumes static kinetic rates throughout the experiment, and the pharmacokinetic rates were investigated at only a single concentration. These rates are calculated as an average over all observed cells, not accounting for intercellular heterogeneity. Further, these kinetics may saturate as a function of doxorubicin concentration. This method remains to be validated in additional cell lines with other pharmacologic targets to address its generalizability. Additional properties of *in vitro* assays not explicitly considered in the current model have been shown to confound observed effects. For example, local cell densities have been found to affect treatment response (Greene et al., 2016). Finally, this model is deterministic and does

not consider either population heterogeneity or cell evolution. Despite these limiting assumptions, we note the accuracy of the equivalent dose in summarizing population-level response to a range of doxorubicin treatment conditions.

In this work, we have demonstrated how mathematical modeling can be leveraged to quantify PK/PD pathways and more precisely compare treatment response among cell lines. It is the ultimate goal of precision cancer therapy to deliver the optimal therapy on the optimal schedule for the individual patient (McKenna et al., 2018). A necessary step toward this goal is to establish a robust functional relationship between applied treatment and subsequent response. The present study demonstrates the utility of the modeling framework and provides additional evidence that the response to therapy is *predictable*. In summary, analysis of treatment response data with mechanistic models can effectively quantify the effects of various biological and pharmaceutical perturbations on treatment response.

AUTHOR CONTRIBUTIONS

MM and TY conceived the experiments. MM conducted the experiments under the guidance of TY. MM analyzed all results. JW aided MM in developing the numerical methods to fit the proposed models. All authors reviewed the manuscript.

FUNDING

We thank the National Institutes of Health for funding through: NCI R01 CA138599, NCI R01 CA186193, NCI U01 CA174706, NIGMS T32 GM007347, NCI F30 CA203220, NCI K25 CA204599, and NIBIB R21 EB022380. We thank the Cancer Prevention Research Institute of Texas (CPRIT) for RR160005; TY is a CPRIT Scholar in Cancer Research.

ACKNOWLEDGMENTS

We thank the funding agencies listed above for their support of this work. We would also like to thank the editors and reviewers for their critical evaluation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2019.00616/full#supplementary-material

REFERENCES

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* 2012:2284. doi: 10.1038/nbt.2284

Anderson, A. R., and Quaranta, V. (2008). Integrative mathematical oncology. *Nat. Rev. Cancer* 8, 227–234. doi: 10.1038/nrc2329

Bouwman, P., and Jonkers, J. (2012). The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. *Nat. Rev. Cancer* 12, 587–598. doi: 10.1038/nrc3342

Chang, B. D., Broude, E. V., Dokmanovic, M., Zhu, H., Ruth, A., Xuan, Y., et al. (1999). A senescence-like phenotype distinguishes tumor cells that undergo terminal proliferation arrest after exposure to anticancer agents. *Cancer Res.* 59, 3761–3767. doi: 10.1038/nrc2961

Chen, S. H., and Lahav, G. (2016). Two is better than one; toward a rational design of combinatorial therapy. *Curr. Opin. Struct. Biol.* 41, 145–150. doi: 10.1016/j.sbi.2016.07.020

Chou, T.-C. (2006). Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol. Rev.* 58, 621–681. doi: 10.1124/pr.58.3.10

Ciszewski, W. M., Tavecchio, M., Dastych, J., and Curtin, N. J. (2014). DNA-PK inhibition by NU7441 sensitizes breast cancer cells to ionizing radiation and doxorubicin. *Breast Cancer Res. Treat.* 143, 47–55. doi: 10.1007/s10549-013-2785-6

Clarke, R., Leonessa, F., and Trock, B. (2005). Multidrug resistance/P-glycoprotein and breast cancer: review and meta-analysis. *Semin. Oncol.* 32, 9–15. doi: 10.1053/j.seminoncol.2005.09.009

Davidson, D., Amrein, L., Panasci, L., and Aloyz, R. (2013). Small molecules, inhibitors of DNA-PK, targeting DNA repair, and beyond. *Front. Pharmacol.* 4:5. doi: 10.3389/fphar.2013.00005

Elstrodt, F., Hollestelle, A., Nagel, J. H., Gorin, M., Wasielewski, M., van den Ouweland, A., et al. (2006). BRCA1 mutation analysis of 41 human breast cancer cell lines reveals three new deleterious mutants. *Cancer Res.* 66, 41–45. doi: 10.1158/0008-5472.CAN-05-2853

Eom, Y.-W., Kim, M. A., Park, S. S., Goo, M. J., Kwon, H. J., Sohn, S., et al. (2005). Two distinct modes of cell death induced by doxorubicin: apoptosis and cell death through mitotic catastrophe accompanied by senescence-like phenotype. *Oncogene* 24, 4765–4777. doi: 10.1038/sj.onc.1208627

Fallahi-Sichani, M., Honarnejad, S., Heiser, L. M., Gray, J. W., and Sorger, P. K. (2013). Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nat. Chem. Biol.* 9, 708–714. doi: 10.1038/nchembio.1337

Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N., Johnson, D. A., Richardson, T. B., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434, 917–921. doi: 10.1038/nature03445

Fink, D., Aebi, S., and Howell, S. B. (1998). The role of DNA mismatch repair in drug resistance. *Clin. Cancer Res.* 4, 1–6.

Foucquier, J., and Guedj, M. (2015). Analysis of drug combinations: current methodological landscape. *Pharmacol. Res. Perspect.* 2015:149. doi: 10.1002/prp2.149

Fowler, J. F. (1992). Brief summary of radiobiological principles in fractionated radiotherapy. *Semin. Radiat. Oncol.* 2, 16–21. doi: 10.1016/S1053-4296(05)80045-1

Fox, E., and Bates, S. E. (2007). Tariquidar (XR9576): a P-glycoprotein drug efflux pump inhibitor. *Expert Rev. Anticancer Ther.* 7, 447–459. doi: 10.1586/14737140.7.4.447

Frick, P. L., Paudel, B. B., Tyson, D. R., and Quaranta, V. (2015). Quantifying heterogeneity and dynamics of clonal fitness in response to perturbation. *J. Cell Physiol.* 230, 1403–1412. doi: 10.1002/jcp.24888

Gao, H., Yin, Z., Cao, Z., and Zhang, L. (2017). Developing an agent-based drug model to investigate the synergistic effects of drug combinations. *Molecules* 22:2209. doi: 10.3390/molecules22122209

Gerlee, P. (2013). The model muddle: in search of tumor growth laws. *Cancer Res.* 73, 2407–2411. doi: 10.1158/0008-5472.CAN-12-4355

Gewirtz, D. A. (1999). A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. *Biochem. Pharmacol.* 57, 727–741. doi: 10.1016/S0006-2952(98)00307-4

Greene, J. M., Levy, D., Herrada, S. P., Gottesman, M. M., and Lavi, O. (2016). Mathematical modeling reveals that changes to local cell density dynamically modulate baseline variations in cell growth and drug response. *Cancer Res.* 76, 2882–2890. doi: 10.1158/0008-5472.CAN-15-3232

Gudmundsdottir, K., and Ashworth, A. (2006). The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* 25, 5864–5874. doi: 10.1038/sj.onc.1209874

Hafner, M., Niepel, M., Chung, M., and Sorger, P. K. (2016). Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods* 13, 521–527. doi: 10.1038/nmeth.3853

Harris, L. A., Frick, P. L., Garbett, S. P., Hardeman, K. N., Paudel, B. B., Lopez, C. F., et al. (2016). An unbiased metric of antiproliferative drug effect *in vitro*. *Nat. Methods* 13, 497–500. doi: 10.1038/nmeth.3852

Helleday, T., Petermann, E., Lundin, C., Hodgson, B., and Sharma, R. A. (2008). DNA repair pathways as targets for cancer therapy. *Nat. Rev. Cancer* 8, 193–204. doi: 10.1038/nrc2342

Jarrett, A. M., Lima, E. A. B. F., Hormuth, D. A., McKenna, M. T., Feng, X., Ekrut, D. A., et al. (2018). Mathematical models of tumor cell proliferation: a review of the literature. *Expert. Rev. Anticancer Ther.* 18, 1271–1286. doi: 10.1080/14737140.2018.1527689

Jones, L. B., Secomb, T. W., Dewhirst, M. W., and El-Kareh, A. W. (2014). The additive damage model: a mathematical model for cellular responses to drug combinations. *J. Theor. Biol.* 2014:32. doi: 10.1016/j.jtbi.2014.04.032

Jones, T. R., Carpenter, A. E., Sabatini, D. M., and Golland, P. (2006). "Methods for high-content, high-throughput image-based cell screening," in *Proceedings of the Workshop on Microscopic Image Analysis with Applications in Biology*, 65–72. Available online at: https://pdfs.semanticscholar.org/d0a7/250de15140526b06ccd3ed8effceb77b04fe.pdf

Lankelma, J., Fernández Luque, R., Dekker, H., and Pinedo, H. M. (2003). Simulation model of doxorubicin activity in islets of human breast cancer cells. *Biochim. Biophys. Acta* 1622, 169–178. doi: 10.1016/S0304-4165(03)00139-9

Lankelma, J., Fernández Luque, R., Dekker, H., van den Berg, J., and Kooi, B. (2013). A new mathematical pharmacodynamic model of clonogenic cancer cell death by doxorubicin. *J. Pharmacokinet. Pharmacodyn.* 40, 513–525. doi: 10.1007/s10928-013-9326-0

Larsen, A. K., Escargueil, A. E., and Skladanowski, A. (2000). Resistance mechanisms associated with altered intracellular distribution of anticancer agents. *Pharmacol. Ther.* 85, 217–229. doi: 10.1016/S0163-7258(99)00073-X

Larsen, A. K., and Skladanowski, A. (1998). Cellular resistance to topoisomerase-targeted drugs: from drug uptake to cell death. *Biochim. Biophys. Acta* 1400, 257–274. doi: 10.1016/S0167-4781(98)00140-7

Lederer, S., Dijkstra, T. M. H., and Heskes, T. (2018). Additive dose response models: explicit formulation and the loewe additivity consistency condition. *Front. Pharmacol.* 2018:31. doi: 10.3389/fphar.2018.00031

Lobo, E. D., and Balthasar, J. P. (2002). Pharmacodynamic modeling of chemotherapeutic effects: application of a transit compartment model to characterize methotrexate effects *in vitro*. *AAPS PharmSci.* 4, 212–222. doi: 10.1208/ps040442

McKenna, M. T., Weis, J. A., Barnes, S. L., Tyson, D. R., Miga, M. I., Quaranta, V., et al. (2017). A predictive mathematical modeling approach for the study of doxorubicin treatment in triple negative breast cancer. *Sci. Rep.* 7:5725. doi: 10.1038/s41598-017-05902-z

McKenna, M. T., Weis, J. A., Brock, A., Quaranta, V., and Yankeelov, T. E. (2018). Precision medicine with imprecise therapy: computational modeling for chemotherapy in breast cancer. *Transl. Oncol.* 11, 732–742. doi: 10.1016/J.TRANON.2018.03.009

Mechetner, E., Kyshtoobayeva, A., Zonis, S., Kim, H., Stroup, R., Garcia, R., et al. (1998). Levels of multidrug resistance (MDR1) P-glycoprotein expression by human breast cancer correlate with *in vitro* resistance to taxol and doxorubicin. *Clin. Cancer Res.* 4, 389–398.

Mistry, P., Stewart, A. J., Dangerfield, W., Okiji, S., Liddle, C., Bootle, D., et al. (2001). *In vitro* and *in vivo* reversal of P-glycoprotein-mediated multidrug resistance by a novel potent modulator, XR9576. *Cancer Res.* 61, 749–758.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics* 32:761. doi: 10.2307/2529262

Pusztai, L., Wagner, P., Ibrahim, N., Rivera, E., Theriault, R., Booser, D., et al. (2005). Phase II study of tariquidar, a selective P-glycoprotein inhibitor, in patients with chemotherapy-resistant, advanced breast carcinoma. *Cancer* 104, 682–691. doi: 10.1002/cncr.21227

Quaranta, V., Tyson, D. R., Garbett, S. P., Weidow, B., Harris, M. P., and Georgescu, W. (2009). Trait variability of cancer cells quantified by high-content automated microscopy of single cells. *Methods Enzymol.* 467, 23–57. doi: 10.1016/S0076-6879(09)67002-6

Sanga, S., Sinek, J. P., Frieboes, H. B., Ferrari, M., Fruehauf, J. P., and Cristini, V. (2006). Mathematical modeling of cancer progression and response to chemotherapy. *Expert Rev. Anticancer Ther.* 6, 1361–1376. doi: 10.1586/14737140.6.10.1361

Shen, H., Schultz, M., Kruh, G. D., and Tew, K. D. (1998). Increased expression of DNA-dependent protein kinase confers resistance to adriamycin. *Biochim. Biophys. Acta* 1381, 131–138. doi: 10.1016/S0304-4165(98)00020-8

Simeoni, M., Magni, P., Cammia, C., De Nicolao, G., Croci, V., Pesenti, E., et al. (2004). Predictive pharmacokinetic-pharmacodynamic modeling of tumor growth kinetics in xenograft models after administration of anticancer agents. *Cancer Res.* 64, 1094–1101. doi: 10.1158/0008-5472.CAN-03-2524

Smith, G. C., and Jackson, S. P. (1999). The DNA-dependent protein kinase. *Genes Dev.* 13, 916–934.

Tacar, O., Sriamornsak, P., and Dass, C. R. (2013). Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems. *J. Pharm. Pharmacol.* 65, 157–170. doi: 10.1111/j.2042-7158.2012.01567.x

Tyson, D. R., Garbett, S. P., Frick, P. L., and Quaranta, V. (2012). Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. *Nat. Methods* 9, 923–928. doi: 10.1038/nmeth.2138

Wang, Z., Butner, J. D., Cristini, V., and Deisboeck, T. S. (2015). Integrated PK-PD and agent-based modeling in oncology. *J. Pharmacokinet. Pharmacodyn.* 42, 179–189. doi: 10.1007/s10928-015-9403-7

Yankeelov, T. E., Atuegwu, N., Hormuth, D., Weis, J. A., Barnes, S. L., Miga, M. I., et al. (2013). Clinically relevant modeling of tumor growth and treatment response. *Sci. Transl. Med.* 5:187ps9. doi: 10.1126/scitranslmed.3005686

Yankeelov, T. E., Quaranta, V., Evans, K. J., and Rericha, E. C. (2015). Toward a science of tumor forecasting for clinical oncology. *Cancer Res.* 75, 918–923. doi: 10.1158/0008-5472.CAN-14-2233

Yin, Z., Deng, Z., Zhao, W., and Cao, Z. (2018). Searching synergistic dose combinations for anticancer drugs. *Front. Pharmacol.* 9:535. doi: 10.3389/fphar.2018.00535

Zhao, Y., Thomas, H. D., Batey, M. A., Cowell, I. G., Richardson, C. J., Griffin, R. J., et al. (2006). Preclinical evaluation of a potent novel DNA-dependent protein kinase inhibitor NU7441. *Cancer Res.* 66, 5354–5362. doi: 10.1158/0008-5472.CAN-05-4275

Zimmermann, T. (2005). "Spectral imaging and linear unmixing in light microscopy," in *Microscopy Techniques*, ed J. Rietdorf (Berlin, Heidelberg: Springer), 245–265. doi: 10.1007/b102216

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 18 July 2019
doi: 10.3389/fphys.2019.00823

Exploring the Extracellular Regulation of the Tumor Angiogenic Interaction Network Using a Systems Biology Model

Ding Li[1] *and Stacey D. Finley*[2]*

[1] Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, United States, [2] Department of Biomedical Engineering, Mork Family Department of Chemical Engineering and Materials Science, and Department of Biological Sciences, University of Southern California, Los Angeles, CA, United States

Tumor angiogenesis is regulated by pro- and anti-angiogenic factors. Anti-angiogenic agents target the interconnected network of angiogenic factors to inhibit neovascularization, which subsequently impedes tumor growth. Due to the complexity of this network, optimizing anti-angiogenic cancer treatments requires detailed knowledge at a systems level. In this study, we constructed a tumor tissue-based model to better understand how the angiogenic network is regulated by opposing mediators at the extracellular level. We consider the network comprised of two pro-angiogenic factors: vascular endothelial growth factor (VEGF) and basic fibroblast growth factor (FGF2), and two anti-angiogenic factors: thrombospondin-1 (TSP1) and platelet factor 4 (PF4). The model's prediction of angiogenic factors' distribution in tumor tissue reveals the localization of different factors and indicates the angiogenic state of the tumor. We explored how the distributions are affected by the secretion of the pro- and anti-angiogenic factors, illustrating how the angiogenic network is regulated in the extracellular space. Interestingly, we identified a counterintuitive result that the secretion of the anti-angiogenic factor PF4 can enhance pro-angiogenic signaling by elevating the levels of the interstitial and surface-level pro-angiogenic species. This counterintuitive situation is pertinent to the clinical setting, such as the release of anti-angiogenic factors in platelet activation or the administration of exogenous PF4 for anti-angiogenic therapy. Our study provides mechanistic insights into this counterintuitive result and highlights the role of heparan sulfate proteoglycans in regulating the interactions between angiogenic factors. This work complements previous studies aimed at understanding the formation of angiogenic complexes in tumor tissue and helps in the development of anti-cancer strategies targeting angiogenesis.

Keywords: systems biology, angiogenesis, anti-angiogenic therapy, compartmental model, mathematical model

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Hermann Frieboes,
University of Louisville, United States
Walter Lee Murfee,
University of Florida, United States

***Correspondence:**
Stacey D. Finley
sfinley@usc.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *01 March 2019*
Accepted: *12 June 2019*
Published: *18 July 2019*

Citation:
Li D and Finley SD (2019) Exploring the Extracellular Regulation of the Tumor Angiogenic Interaction Network Using a Systems Biology Model. Front. Physiol. 10:823. doi: 10.3389/fphys.2019.00823

INTRODUCTION

Angiogenesis, the growth of new blood microvessels from pre-existing microvasculature, plays a crucial role in tumor development (Hanahan and Weinberg, 2011). Tumor growth relies on angiogenesis to enable waste exchange and provide oxygen and nutrients from the surrounding environment. Several angiogenic factors that affect the extent of tumor vascularization have been

identified and are commonly categorized as pro- and anti-angiogenic factors. Pro-angiogenic factors, including vascular endothelial growth factor-A (VEGF) and fibroblast growth factor 2 (FGF2), bind to their respective receptors to induce pro-angiogenic signaling promoting cell proliferation, cell migration and blood vessel formation (Carmeliet, 2005; Korc and Friesel, 2009). On the other side, anti-angiogenic factors, like thrombospodin-1 (TSP1) and platelet factor 4 (PF4), inhibit pro-angiogenic signaling and induce anti-angiogenic signaling to oppose angiogenesis (Bikfalvi, 2004; Ren et al., 2006). Considering the importance of angiogenesis in tumor development, anti-angiogenic therapies are designed to target the signaling of angiogenic factors to inhibit neovascularization and tumor growth (Vasudev and Reynolds, 2014). Single-agent anti-angiogenic therapies that target a particular angiogenic factor in the network were the first angiogenesis-inhibiting therapies studied. These include antibodies or small molecules targeting pro-angiogenic factors (Abdalla et al., 2018) and peptide mimetics of anti-angiogenic factors (Jeanne et al., 2015). However, these single-agent anti-angiogenic therapies showed limited success in the clinic due to toxicity, low efficacy, or the development of resistance (Wehland et al., 2012; Vasudev and Reynolds, 2014). These drawbacks have promoted efforts to develop combination therapies administering multiple anti-angiogenic agents that simultaneously target various angiogenic species in the network (Alessi et al., 2009; Kim et al., 2009; Uronis et al., 2013; van Beijnum et al., 2015).

Due to the intrinsic complexity of the network regulating tumor angiogenesis, optimizing anti-angiogenic cancer treatment, specifically combination anti-angiogenic therapy, requires detailed knowledge and a holistic view at a systems level. Computational systems biology models offer powerful tools to systematically study tumor angiogenesis and optimize anti-angiogenic tumor therapy. Various types of systems biology models have been constructed to investigate new anti-angiogenic therapies (Finley et al., 2015). Models of intracellular signaling of angiogenic factors characterize the biochemical events inside the cell initiated by ligand binding to signaling receptors on the cell surface. These models help in the identification of new intracellular drug targets. At the extracellular level, models of the extracellular species' reaction network are used to understand the distribution of angiogenic factors in tumor tissue (Mac Gabhann and Popel, 2006) and in the whole body (Stefanini et al., 2008). By linking to the kinetics of anti-angiogenic drugs, models that capture extracellular interactions can be used to study therapeutics that modulate the distribution of angiogenic factors, which directly affects angiogenic signaling (Stefanini et al., 2010; Li and Finley, 2018). To better understand the effects of targeting angiogenic factors in the tumor, we built a new tissue-based systems biology model characterizing the extracellular network that involves four main angiogenic factors regulating tumor angiogenesis, including VEGF, FGF2, TSP1, and PF4.

Our modeling work expanded previous models by incorporating angiogenic factors that were previously omitted from the models, as well as other significant mediators. Thus, our model enables a systematic study of the extracellular regulation of multiple angiogenic factors. The extracellular distribution of VEGF alone was firstly investigated in a computational setting with a tissue-based model (Mac Gabhann and Popel, 2006). Later, this physiologically relevant and molecularly detailed model was extended to include TSP1, a potent endogenous anti-angiogenic factor, to explore the balance of pro- and anti-angiogenic factor in tumor tissue (Rohrs et al., 2016). In the present work, we further expand the model to include the pro-angiogenic factor, FGF2, and an additional anti-angiogenic factor, PF4. These species are reported to interact with VEGF and TSP1 and significantly impact tumor angiogenesis. FGF2 is reported to synergistically enhance the pro-angiogenic signal with VEGF (Seghezzi et al., 1998; Kano, 2005). On the other hand, upregulation of the FGF2 pathway can result in resistance to anti-VEGF therapy (Alessi et al., 2009; van Beijnum et al., 2015). PF4, like the other anti-angiogenic factor TSP1, binds to VEGF and FGF2 to reduce pro-angiogenic signaling (Vandercappellen et al., 2011; Wang and Huang, 2013). Therefore, incorporating FGF2 and PF4 provides a more complete view of the angiogenic interaction network and a more comprehensive understanding of tumor angiogenic state, as compared to previous models. In addition, PF4, TSP1, VEGF, and FGF2 each bind to heparin, competing for the heparan sulfate (HS) binding sites in heparan sulfate proteoglycans (HSPG) on the cell surface and in the extracellular matrix and basement membrane (Sarrazin et al., 2011). The secretion of PF4 and TSP1 leading to displacement of VEGF and FGF2 from HS binding sites is an important mechanism of tumor angiogenesis regulation. Specifically, PF4 is known to interrupt the HSPG-mediated formation of pro-angiogenic complexes to inhibit VEGF and FGF2 signaling (Perollet et al., 1998; Jouan et al., 1999). To account for the regulation of HSPG, our model includes two distinct species with HS binding sites, one of which, the surface-level HSPG, is not explicitly accounted for in previous tumor tissue-based models (Mac Gabhann and Popel, 2006; Rohrs et al., 2016). The previous tissue-based model of VEGF and TSP1 has 120 species and was generated with 27 seed species and 78 reaction rules (Rohrs et al., 2016). After incorporating FGF2, PF4, HSPGs, and their binding partners, the novel model presented in this study is comprised of 168 species, generated with 40 seed species and 127 reaction rules.

With the newly constructed model, we firstly profiled the distribution of these four angiogenic factors in tumor tissue and systematically investigated how the secretion of different angiogenic factors affects the balance of pro- and anti-angiogenic signaling. Furthermore, we generate insights explaining two specific counterintuitive phenomena: (1) the secretion of PF4 increases the levels of free VEGF and FGF2 in tumor tissue and (2) the secretion of PF4 promotes the formation of VEGF signaling complexes. We found HSPG's level directly affects these counterintuitive results in different ways, emphasizing the important role of HSPGs in the regulation of angiogenic factor signaling. Lastly, we apply the model to simulate a controlled release of PF4 in tumor tissue, and our results indicate that the HSPG level in the tumor microenvironment might affect the response to platelet activation and recombinant PF4 anti-angiogenic therapy. Overall, we establish a new computational framework to understand the extracellular distribution of angiogenic factors in tumor tissue and generate new insights

into the regulation of the angiogenic factors' interaction network, which are difficult to examine through experimental study alone.

MATERIALS AND METHODS

The Tumor Tissue Model of Angiogenic Factors

We constructed a molecularly detailed model that describes the extracellular network of four main angiogenic factors in tumor tissue (**Figure 1**). The modeling approach is consistent with previous works (Mac Gabhann and Popel, 2006; Rohrs et al., 2016). The system is represented by a set of coupled non-linear ordinary differential equations (ODEs) to characterize a well-mixed tumor tissue. For the model structure, the extracellular spaces in tumor tissue are divided into three regions: the surface of endothelial cells, the surface of tumor cells and interstitial space. The interstitial space between tumor and endothelial cells is comprised of extracellular matrix (ECM) and the basement membranes surrounding the tumor cells (TBM) and the endothelial cells (EBM). The soluble species are secreted by both cell types and can be removed from the system through degradation in the interstitial space or internalization with receptors at the cell surface.

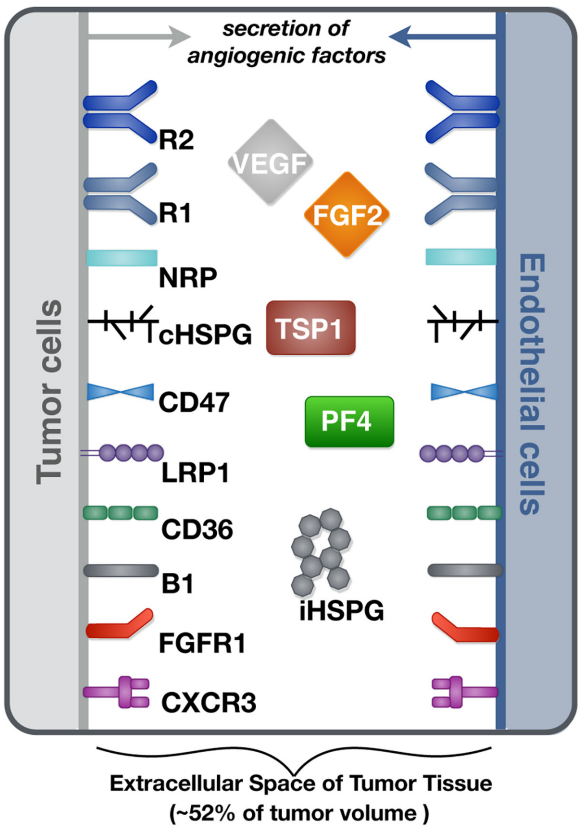

FIGURE 1 | Schematic of the tumor tissue based model. The compartmental model describes the reaction network of four major angiogenic factors (VEGF, FGF2, TSP1, and PF4) in the extracellular space of tumor tissue. Angiogenic receptors are expressed on the cell surfaces. Soluble angiogenic factors exist in the interstitial space, and they bind to cell surface receptors to form signaling complexes. Two different heparan sulfate proteoglycans are included in the model, including cell surface HSPG (cHSPG) and interstitial HSPG (iHSPG).

Ten soluble species are present in the model (**Figure 2**, Legend I). Physiologically, in tumor tissue, $VEGF_{121}$, $VEGF_{165}$, FGF2, TSP1, MMP3, and proMMP9 are mainly produced through the secretion from tumor cells and endothelial cells, while PF4 is stored in the α-granules of the platelets and is released through platelet activation. Thus, in the model, the source of PF4 in tumor tissue is represented by a generic production rate. In addition, $VEGF_{114}$, inactive TSP1 and active MMP9 are formed through cleavage. Nine relevant receptors are present on the cell surface (**Figure 2**, Legend II), including VEGF receptors (VEGFR1, VEGFR2, Neuropilin-1), TSP1 receptors (CD47, CD36, LRP1, $\alpha_x\beta_1$ integrins), FGF2 receptor (FGFR1) and PF4 receptor (CXCR3). Receptors are assumed to be uniformly distributed on the cell surface and are recycled back to the surface to maintain a constant total number for each type of receptor. In addition, we include the heparan sulfate proteoglycans (HSPGs) in the model, which are important modulators of angiogenic signaling. HSPGs are glycoproteins, which have a protein core and one or more covalently attached heparan sulfate (HS) chains. Two types of HSPGs are present in the model (**Figure 2**, Legend III). One is the interstitial heparan sulfate proteoglycans (iHSPG) that are present in the ECM, EBM and TBM. The other one is the cell-surface heparan sulfate proteoglycans (cHSPG). The iHSPG serves as a reservoir for angiogenic factors, while the cHSPG mainly functions as a co-receptor participating in the formation of complexes to modulate angiogenic signaling.

Network of Reactions

The principles of mass action kinetics are used to characterize the species' dynamics. The defined rules that govern the molecular interactions are shown in **Figure 2**, and the detailed reactions are given in **Supplementary File S1.**

VEGF-Receptor Axis (Figure 2A)

Previous work modeling VEGF ligand-receptor interactions did not explicitly include the surface-level HSPGs (cHSPG) (Mac Gabhann and Popel, 2005), assuming the presence of abundant HSPGs on the cell surface. To investigate the impact of HSPG on VEGF signaling, we extended previous VEGF-VEGFR modeling to incorporate the cHSPG-facilitated VEGF binding reactions. Previous works have detailed documentation of estimating the kinetic for two VEGF isoforms ($VEGF_{165}$ and $VEGF_{121}$) binding to VEGF receptors (Mac Gabhann and Popel, 2005, 2006), and we use those parameter values in our model. Below, we present how we have adapted previous works to include cHSPG regulation.

VEGFR2 and co-receptors (first two rows of **Figure 2A**): According its structure, $VEGF_{165}$ binds to VEGFR2 via the exon 4 encoded domain and to NRP1 and HSPG via the exon 7 encoded domain to form a ternary complex (Whitaker et al., 2001; Mac Gabhann and Popel, 2005). It is commonly assumed that VEGFR2 does not directly interact with NRP1, but is bridged

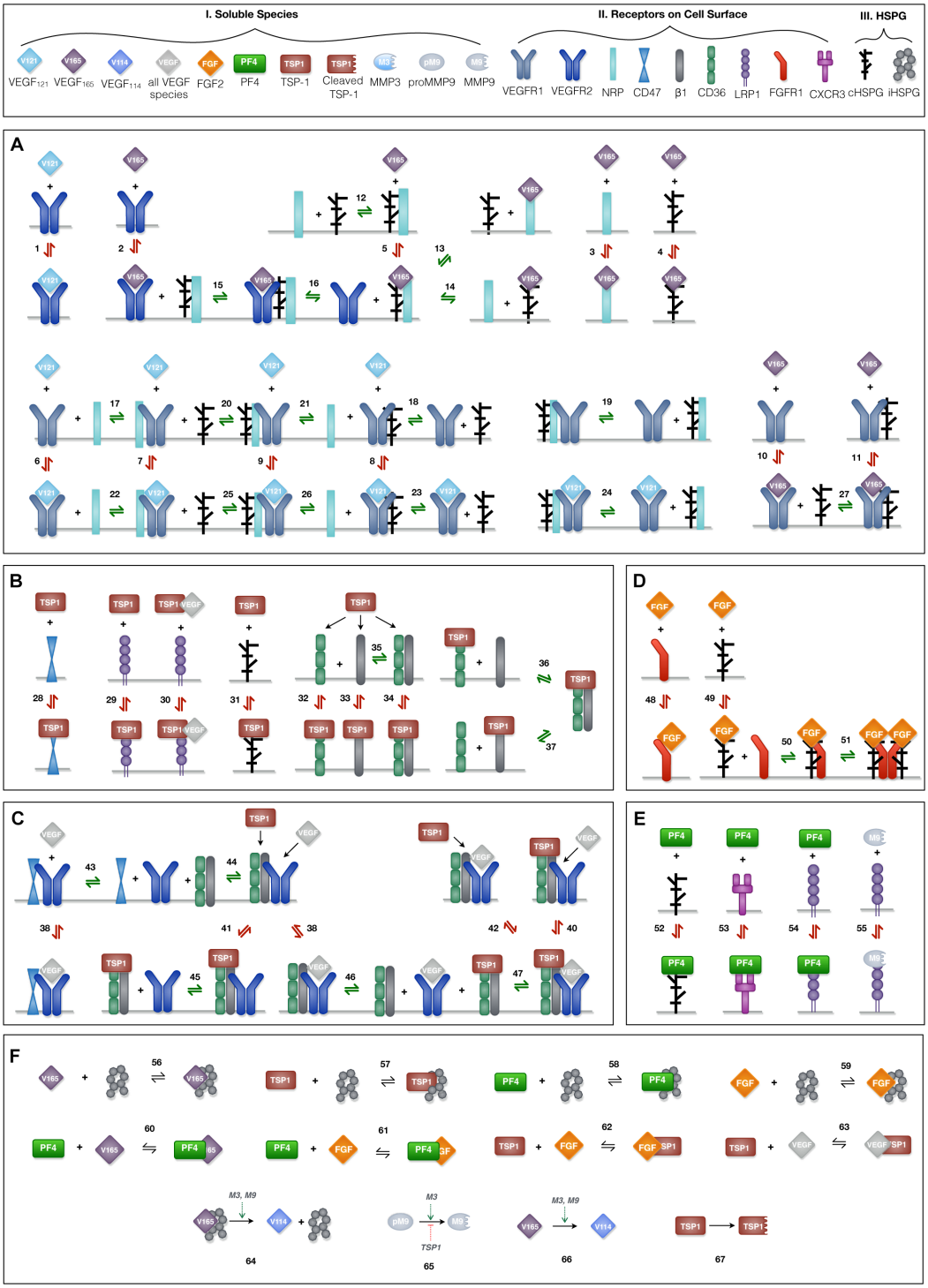

FIGURE 2 | Schematic of the extracellular network of VEGF, FGF2, TSP1, and PF4. **(A)** Molecular interactions of two active VEGF isoforms ($VEGF_{165}$ and $VEGF_{121}$), receptors (VEGFR1, VEGFR2, and NRP1) and heparan sulfate proteoglycans on the cell surface (cHSPG). **(B)** Molecular interactions of TSP1 binding to its receptors (CD36, CD47, LRP1, and $\alpha_x\beta_1$ integrins) and cHSPG. **(C)** Molecular interactions of the coupling between VEGFR2 and TSP1 receptors. **(D)** Molecular interactions of FGF2 binding to FGFR1 and cHSPG and the formation of the full signaling complex through dimerization. **(E)** Molecular interactions of PF4 binding to receptors (CXCR3 and LRP1) and cHSPG, as well as MMP9 binding to LRP1. **(F)** The molecular interactions of angiogenic factors binding to one another and heparan sulfate proteoglycans in the interstitial space (iHSPG), as well as the proteolysis and degradation of soluble species. Numbers for each interaction correspond to the list of reactions in **Supplementary File S1**. One interaction on the schematic may represent multiple reactions (i.e., the same species can bind through different binding sites). In total, the details of the 89 reactions are listed in **Supplementary File S1.**

by the $VEGF_{165}$ (Soker et al., 2002). For the HSPG, in a recent study, VEGFR2 was shown not to interact with heparin directly, and that $VEGF_{165}$ also mediates the interactions between VEGR2 and heparin (Teran and Nugent, 2015). Therefore, in our model, we assume HSPG does not directly interact with VEGFR2, and the impact of HSPG on VEGFR2 signaling is mediated through supporting the $VEGF_{165}$-mediated bridging of VEGFR2 with NRP1. For the interactions between HSPG and NRP1, it is reported that heparin could bind to the b1b2 domain of NRP1 directly, greatly enhancing the binding of $VEGF_{165}$ to NRP1 (Mamluk et al., 2002). This suggests that $VEGF_{165}$ binds to NRP1 in an HSPG-dependent way. To include this knowledge, we allow HSPG to pre-couple with NRP1 before interacting with VEGFR2. The other isoform of VEGF, $VEGF_{121}$, lacks the exon 7 coded region; thus, it does not bind to NRP1 or HSPG (Whitaker et al., 2001; Teran and Nugent, 2015).

VEGFR1 and co-receptors (second two rows of **Figure 2A**): Following our previous modeling (Mac Gabhann and Popel, 2005, 2006), VEGFR1 can couple with NRP1, while the binding with $VEGF_{121}$ is not affected by the coupling. Since VEGR1 was shown to bind to heparin directly and VEGFR1 does not show a heparin-aided VEGF binding as VEGR2 does (Teran and Nugent, 2015), we assume HSPG can couple with VEGR1 and does not affect its binding to VEGF. In addition, we assume VEGFR1 can couple with the NRP1 pre-coupled with HSPG to form a ternary complex and that then binds with VEGF. Following our previous model, $VEGF_{165}$ does not bind to VEGFR1 pre-coupled with NRP1 (Mac Gabhann and Popel, 2005, 2006). In addition, since it is reported that the presence of heparin does not significantly change the binding of $VEGF_{165}$ to VEGFR1 (Teran and Nugent, 2015), we assume the pre-coupling of VEGFR1 with HSPG does not affect VEGFR1's binding with $VEGF_{165}$.

TSP1-Receptor Axis (Figures 2B,C)

The reactions involving interactions between TSP1 and its receptors are taken from previous works (Rohrs et al., 2016), in which TSP1 regulates angiogenic signaling in different ways. TSP1 binds to its own receptors to induce anti-angiogenic signaling (**Figure 2B**). Ligated TSP1 receptors can also couple with VEGFR2 to inhibit the signaling of VEGF (**Figure 2C**). These interactions are included in the model.

FGF2-Receptor Axis (Figure 2D)

The reactions for the FGF2-receptor axis are from the extracellular part of an *in vitro* whole cell FGF2 signaling model (Kanodia et al., 2014), which defines the formation of FGF2 signaling trimeric complexes that then dimerize. FGF2 binds to HSPG to form a complex, which binds to the FGFR1 monomer to form a trimeric complex. Then, dimerization of the trimeric complex leads to the formation of the full FGF2 signaling complex. The choice of this ordering is based on several observations from experimental studies (Ibrahimi et al., 2004; Kanodia et al., 2014): FGF2 shows a lower affinity to FGFR1 than to heparin; the interaction of FGFR1 and heparin has a very weak affinity; and FGF2 dramatically increases the association of FGFR1 with heparin. Alternative orders of the binding reactions are possible; however, they are reported to not conform well with the experimental data (Ibrahimi et al., 2004).

PF4-Receptor Axis (Figure 2E)

PF4 regulates angiogenesis through various mechanisms. On the cell surface, PF4 binds to cell surface receptors (CXCR3 and LRP1) to induce anti-angiogenic signaling (Lasagni et al., 2003; Lambert et al., 2009) and binds to cHSPG to control pro-angiogenic signaling (Perollet et al., 1998; Vandercappellen et al., 2011). We include these interactions in the model. PF4 and TSP1 both are reported to bind to LRP1 (Mikhailenko et al., 2002; Lambert et al., 2009), and we have TSP1 and PF4 compete for LRP1 with different affinities.

Interactions Between Angiogenic Factors (Figure 2F)

The angiogenic factors also interact either by direct binding or through HSPGs in the interstitial space (iHSPG). TSP1 associates with VEGF and FGF2 to reduce pro-angiogenic signaling (Margosio et al., 2003; Rohrs et al., 2016), and it mediates VEGF cleavage through MMP activity (Rohrs et al., 2016). In addition, TSP1 can compete for the HS binding sites on cHSPG and iHSPG to release HS-bound angiogenic factors. Similarly, FGF2 can be trapped by TSP1, PF4 and iHSPG. Additionally, PF4 directly binds to $VEGF_{165}$ and FGF2, reducing the available pro-angiogenic factors (Vandercappellen et al., 2011; Wang and Huang, 2013). Lastly, PF4 competes for the HS binding sites on the iHSPG.

Parameterization

The model parameter values are reported in **Supplementary File S2** with literature references. Here, we describe the derivation of inherited values and the rationales for the parameterization of newly introduced values.

Geometric Parameters

The tumor tissue is parameterized as a 33 cm^3 (Korc and Friesel, 2009) breast tumor, which is modeled as a spatially averaged compartment in the model (**Figure 1**). The geometric parameters define the volume of the compartment, the interstitial space volume fraction, and the tissue surface areas of endothelial cells and tumor cells. These geometric parameters enable the conversion of concentration from moles per cm (Korc and Friesel, 2009) tissue to standard units (pmol/l), where the derivations are thoroughly documented in previous works (Mac Gabhann and Popel, 2006; Stefanini et al., 2008).

Production and Degradation of Soluble Species

The production and degradation rates of VEGF, TSP1, MMP3 and proMMP9 are estimated in our previous work (Rohrs et al., 2016). The baseline production rates of PF4 and FGF2 are set to match an intermediate level within the range of experimental measurements (**Table 1**). The degradation rates of PF4 and FGF2 are set according to their half-life ($t_{1/2}$): the rate of degradation is $ln\ (2)/t_{1/2}$. Since a wide range of reported values for the FGF2 half-life is found in literature (Shiba et al., 2003; Beenken and Mohammadi, 2009), we assume it has a half-life of 60 min, similar to VEGF, which is within the reported range. PF4 is reported to

TABLE 1 | Comparison of the baseline predictions and the experimental measurements of VEGF, FGF2, TSP1, PF4, and MMPs.

Species	Range of experimental measurements†	Predicted concentration	Source and references
VEGF unbound	8.0 – 389 pM	180.2 pM	Multiple (Finley and Popel, 2013)
TSP1 total ‡	1.0 – 6.2 nM (2.0)	3.0 nM	Breast cancer patient (Byrne et al., 2007)
PF4 unbound	1.0 – 11.3 nM	4.7 nM	Multiple (Leitzel et al., 1991; Kurimoto et al., 1995; Peterson et al., 2012; Sabrkhany et al., 2017)
FGF2 total	0.2 – 11.1 nM	3.9 nM	Prostate cancer patient (Giri et al., 1999)
MMP3 total	1.8 – 65.1 nM (5.1)	5.0 nM	Oral squamous cell carcinoma patient (Baker et al., 2006)
MMP9 total	1.0 – 287.8 nM (9.0)	9.2 nM	Oral squamous cell carcinoma patient (Baker et al., 2006)
MMP9 active	0 – 22.4 nM (0.8)	0.2 nM	Oral squamous cell carcinoma patient (Baker et al., 2006)

†*Median value is shown in parentheses, if provided in literature. If the experimental data reflects the total concentration in tissue, we assume 50% of the total protein amount is in the extracellular space.* ‡*TSP1 concentration includes both active TSP1 and cleaved TSP1.*

be rapidly cleared in human body, where the half-life is assumed to be 5 min (Dawes et al., 1978).

Receptor Numbers

The receptor densities for VEGF receptors, TSP1 receptors, FGFR1 and HSPGs are taken from previous modeling works (Kanodia et al., 2014; Rohrs et al., 2016). There is a paucity of measurements for the PF4 receptor, CXCR3. Thus, we referred to the qualitative measurements in Human Protein Atlas (Uhlen, 2005), assuming "low," "medium," and "high" expression levels correspond to 2500, 5000, and 10,000 receptors per cell. CXCR3 has a low expression, which is set to be to be 2,500 receptors per cell accordingly.

Kinetic Parameters

For the VEGF axis, the kinetic parameters have been estimated in previous work, based on experimental measurements (Mac Gabhann and Popel, 2005) and assuming an abundant level of HSPGs. We adopted these values in our current model by incorporating several experimentally observed synergistic interactions in the presence of heparin. Since the previous model is calibrated in conditions with abundant HSPGs, we assume the NRP1 in the previous model is already coupled with HSPGs. Therefore, the parameters of $VEGF_{165}$ binding to NRP1 in the previous model is used for $VEGF_{165}$ binding to the NRP1:HSPG complex in our model. Then, we assume the $VEGF_{165}$ binding to NRP1 alone is 20-fold weaker than binding to the NRP1:HSPG complex, according to a study showing that the presence of heparin significantly increases VEGF binding to NRP1 (Mamluk et al., 2002). Likewise, the rates for VEGFR2 coupling to $VEGF_{165}$-bound NRP in the previous model are used for the VEGFR2 coupling to the $VEGF_{165}$-bound NRP:HSPG complex in our model, while the previous rates for $VEGF_{165}$-bound VEGFR2 coupling to NRP1 are used for the $VEGF_{165}$-bound VEGFR2 coupling to NRP:HSPG complex in our model. To our knowledge, there are no available measurements to estimate the coupling rates of NRP1 to HSPG. Therefore, we assume the rates of NRP1 coupling to HSPG are the same as rates of VEGFR2 coupling to NRP1, which are taken from previous modeling (Mac Gabhann and Popel, 2006). Previous experimental study shows a $VEGF_{165}$-mediated synergistic binding between NRP1 and heparin (Teran and Nugent, 2015), and we accordingly assume the coupling of $VEGF_{165}$:NRP to HSPG and the coupling of $VEGF_{165}$:HSPG to NRP are an order of magnitude stronger than the coupling between NRP and HSPG. Following previous works (Mac Gabhann and Popel, 2006; Li and Finley, 2018), the coupling of VEGFR1 to NRP is set to be an order of magnitude weaker than VEGFR2-NRP coupling. According to the measured binding constants (Teran and Nugent, 2015), the coupling of VEGFR1 to HSPG is assumed to be 5-fold stronger than NRP-HSPG coupling.

For the TSP1 axis, we followed the values used in our previous works (Rohrs et al., 2016). For the kinetic rates governing the FGF2 axis, we used the values estimated from experimental data in a previous study (Kanodia et al., 2014). For the PF4 axis, the K_d values of PF4 binding to CXCR3 and LRP1 are estimated to be 1.85 nmol/l (Lasagni et al., 2003) and 238 nmol/l (Sachais et al., 2002), respectively. These are used to set the dissociation rate, with the association rate held at 5×10^5 $M^{-1}s^{-1}$, based on molecular dynamics studies of biomolecular reaction kinetics (Schlosshauer and Baker, 2004; Northrup and Erickson, 2006). In the model, VEGF, TSP1, PF4, and FGF2 each have a different affinity to HSPG. Their affinities to iHSPG are set according to their binding constants (K_d values) measured with heparin. The K_d values of heparin binding to $VEGF_{165}$, FGF2, TSP1 and PF4 are 80, 39, 41, and 20 nmol/l (Stringer and Gallagher, 1997; Ibrahimi et al., 2004; Zhao et al., 2012; Resovi et al., 2014; Lord et al., 2017), respectively. The rates for FGF2 binding to cHSPG (association rate, K_{on}, and dissociation rate, K_{off}) are estimated in previous work (Kanodia et al., 2014), and this provides the FGF2 affinity to cHSPG. We derive the affinities of $VEGF_{165}$, TSP1 and PF4 binding to cHSPG by scaling the FGF2-cHSPG affinity according to their relative affinity to heparin, assuming the measured heparin affinity reflects their relative binding affinity to cHSPG. We first make VEGF, TSP1, PF4 all have the same association rates (K_{on}) as FGF2, and set the dissociation rates (K_{off}) according to their corresponding affinities. For the associations between pro- and anti-angiogenic factors, PF4 binds to $VEGF_{165}$ and FGF2 with K_d values of 5 and 37 nmol/l (Vandercappellen et al., 2011), respectively. The K_d values of TSP1 binding to VEGF and FGF2 are 10 and 10.8 nmol/ (Perollet et al., 1998; Kaur et al., 2010). The parameters for the protease activity are taken from our previous works (Rohrs et al., 2016; Li and Finley, 2018).

Model Implementation and Simulation

The model ODEs are generated using BioNetGen (Faeder et al., 2009), a rule-based modeling framework. BioNetGen produces all possible molecular species and the corresponding ODEs by specifying a set of starting molecular species and defining reaction rules. Given 40 seed species and 127 reaction rules, the model produced by BioNetGen consists of 168 species. The set of 168 ODEs is implemented in MATLAB (The MathWorks, Natick, MA, United States), which we used to generate the dynamic results, as well as steady state predictions (i.e., when the model outputs change less than 0.01%). The MATLAB model file is provided in **Supplementary File S3.**

RESULTS

Baseline Prediction of the Angiogenic Factors' Distribution in Tumor Tissue

The baseline secretion rates of angiogenic factors were tuned in order to obtain concentrations within the range of available experimental measurements. We report the predicted species' concentrations (for VEGF, TSP1, PF4, FGF2, MMP3, and MMP9) and compare with experimental measurements in **Table 1**.

With the baseline secretion rates, the model predicts that the pro-angiogenic factors (VEGF and FGF2) and anti-angiogenic factors (TSP1 and PF4) have significantly different distribution patterns in tumor tissue (**Figure 3**). The majority of each pro-angiogenic factor in the tumor (∼81% of VEGF and ∼50% of FGF2) is bound to the cell surface, while only a small percentage of the anti-angiogenic factors (∼16% of PF4 and ∼12% of TSP1) exists on the cell surface. The cell surface bound ligands can be further categorized into non-signaling and signaling forms. The non-signaling forms include complexes with cHSPG and non-signaling receptors, and signaling forms include ligated receptors that promote intracellular signaling. Most of the cell-surface bound VEGF is in a signaling form, where VEGFR1-, VEGFR2-and NRP1-bound VEGF comprise 35, 17, and 29% of the total VEGF in the tumor, respectively. Only 0.4% of total VEGF is bound to cHSPG. The model predicts that 23% of total FGF2 is in a signaling form bound to FGFR1:cHSPG dimers. The balance of the cell-surface FGF2 is non-signaling, bound to either cHSPG or FGFR1 monomers, which comprise 6 and 22% of total FGF2, respectively. In comparison to the distributions of the pro-angiogenic factors, most of the cell-surface bound anti-angiogenic factors are in non-signaling forms. The model predicts that 11% of total TSP1 and 16% of total PF4 are bound to cHSPG. Only 1% of total TSP1 is bound to signaling receptors, including CD47, CD36, LRP1 and β_1. In the case of PF4, an even smaller fraction (0.4%) is bound to anti-angiogenic receptors, including CXCR3 and LRP1. However, it is worth noting that the ratio of the number of VEGF and FGF2 signaling complexes to the number of TSP1 and PF4 signaling complexes is approximately 1.7. This means that the number of anti-angiogenic complexes is still in the same order of magnitude as the number of pro-angiogenic complexes.

In the interstitial space, there are three forms of angiogenic factors, including the unbound form, iHSPG-bound form, and the form bound to other angiogenic factors. Approximately 12% of total VEGF is in an unbound active form, including $VEGF_{121}$ and $VEGF_{165}$, and 0.1% of VEGF is present as the inactive isoform $VEGF_{114}$. The percentages of VEGF bound to iHSPG or other angiogenic factors are 6 and 0.7%, respectively. Unlike VEGF, most FGF2 in the interstitial space is trapped by iHSPG. That is, 46% of total FGF2 is bound to iHSPG, while the unbound and angiogenic factor-bound forms only comprise 3 and 0.3% of the total FGF2, respectively. In contrast, the two anti-angiogenic factors, PF4 and TSP1, both have a larger portion in the interstitial space. In the case of PF4, 81% is bound to iHSPG, while the unbound and angiogenic factor-bound forms comprise only 2 and 0.01% of the total PF4, respectively. Finally, most of TSP1 in the interstitial space is bound to iHSPG (48%) or in the cleaved, inactive form (35%). The balance of TSP1 is unbound or bound to other angiogenic factors, comprising 3 and 0.6% of the total TSP1, respectively.

To summarize these results, the model predicts that most of VEGF and FGF2 is bound to the cell surface and in signaling forms, while most of TSP1 and PF4 is in the interstitial space and in non-signaling forms that are trapped by HSPGs or inactive due to proteolysis. It is worth noting that the fraction of the anti-angiogenic factors that is bound to pro-angiogenic factors only comprises a small percentage, which implies that direct binding between pro- and anti-angiogenic factors is not a major mechanism of the extracellular inhibition of pro-angiogenic signaling. Overall, this predicted distribution indicates a tumor state favoring pro-angiogenic signaling and neovascularization. In addition to the prediction under baseline secretion rates, we performed Monte Carlo simulations by sampling the secretion rates of VEGF, TSP1, PF4, FGF2, MMP3, and proMMP9 from a range of 100-fold below and 10-fold above the baseline values. The results (**Supplementary Figure S1**) show that, even with potential uncertainty in the secretion rates, the main conclusions of the tumor distribution remain unchanged.

Secretion of Anti-angiogenic Factors Modulates Both Pro- and Anti-angiogenic Signaling

To characterize the angiogenic state of the tumor, we defined the *angiogenic ratio*: the ratio of the concentrations of the pro-angiogenic signaling complexes to the anti-angiogenic signaling complexes. This ratio captures the activation level of pro-angiogenic receptors relative to the activation level of anti-angiogenic receptors. We examined how different angiogenic factors shift the angiogenic ratio (**Figure 4**, column I) by varying the secretion rates of VEGF, FGF2, TSP1, and PF4 in a range of 100-fold below and 10-fold above the baseline values. We also predict how the concentrations of pro- and anti-angiogenic signaling complexes change in response to varying the secretion rates of the angiogenic factors (**Figure 4**, columns II–V). Varying the secretion rates explores how targeting angiogenic factors changes the tumor angiogenic state, assuming 100-fold below represents strong inhibition and 10-fold above

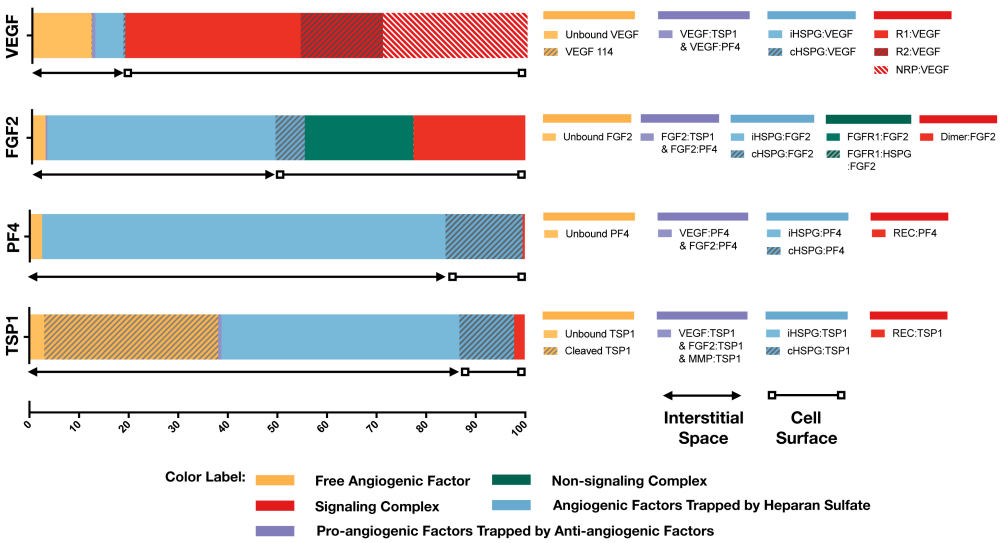

FIGURE 3 | Distribution of VEGF, FGF2, TSP1, and PF4 in tumor tissue at steady state. The percentages of each angiogenic species in its various forms are shown. Species are grouped and labeled with different colors. The sum of the forms bound to the cell surface or in the interstitial space is also indicated.

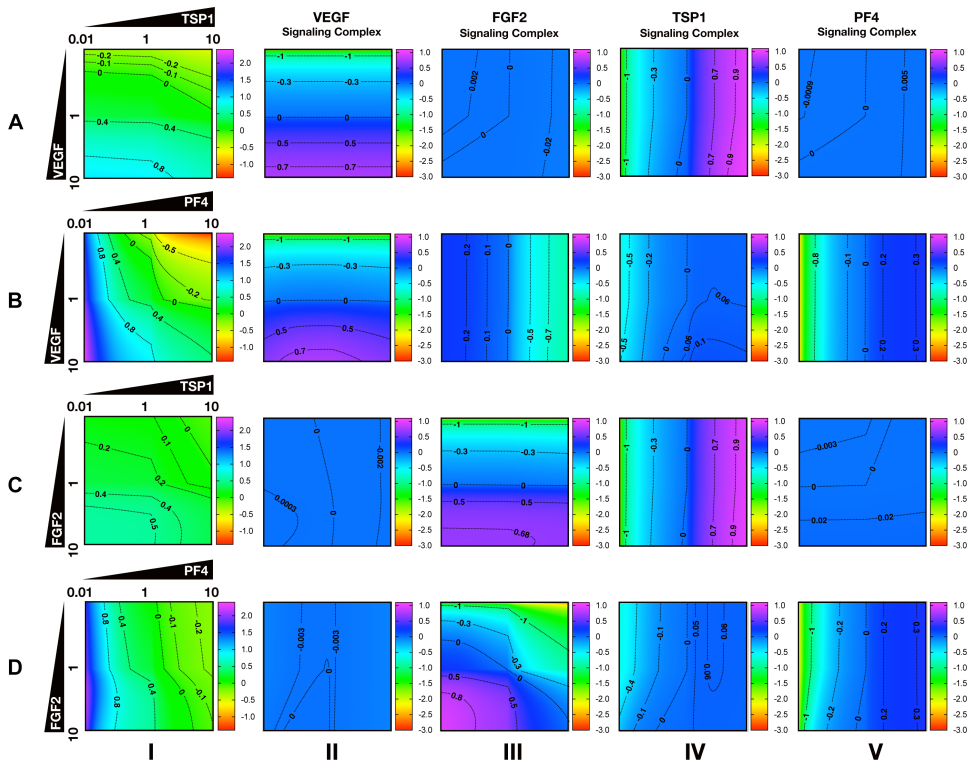

FIGURE 4 | Effects of secretion of angiogenic factors on the angiogenic state of the tumor tissue. Column I shows the angiogenic ratio in the $\log_{10}$ scale. Column II to V show the signaling complex levels (normalized to the baseline prediction) in the $\log_{10}$ scale. The value range is given by the colorbar. The horizontal and vertical axes of each subplot show the fold-change of the corresponding secretion rates, relative to their baseline values. The different rows show the effects of varying the secretion rates of different angiogenic factors: **(A)** VEGF and TSP1 secretion rates vary. **(B)** VEGF and PF4 secretion rates vary. **(C)** FGF2 and TSP1 secretion rates vary. **(D)** FGF2 and PF4 secretion rates vary. The predictions shown in the figures are based on the steady state of the system.

represents upregulation. We plot the angiogenic ratio and the concentrations of the angiogenic complexes normalized to the baseline secretion rates.

Higher secretion of the two pro-angiogenic factors shifts the angiogenic ratio by increasing the level of their corresponding pro-angiogenic complexes (**Figure 4**). The gradient along the vertical axis in **Figures 4A,B**, column I shows the angiogenic ratio will significantly increase with increasing VEGF secretion, which indicates that the tumor moves to a more pro-angiogenic state. The normalized level of the VEGF signaling complexes (**Figures 4A,B**, column II) shows evident color changes along the vertical axis, which indicates the VEGF signaling is strongly enhanced with higher VEGF secretion. Meanwhile, the normalized level of FGF2, TSP1, and PF4 signaling complexes (**Figures 4A,B**, columns III–V) shows no pronounced gradient along the vertical axis, implying that these signaling pathways are not affected by changing VEGF secretion.

Similarly, upregulating FGF2 secretion shifts the angiogenic ratio mainly through enhancing the formation of FGF2 pro-angiogenic complexes. The angiogenic ratio change along vertical axis in **Figure 4C**, column I implies that the upregulation of FGF2 secretion increases the angiogenic ratio and promotes angiogenesis. The normalized level of the FGF2 complex (**Figures 4C,D**, column III) significantly increases with increasing FGF2 secretion, while the concentration of the VEGF signaling complexes (**Figures 4C,D**, column II) is highly stable when the FGF2 secretion is changed. The TSP1 and PF4 signaling complexes (**Figure 4D**, columns IV–V) slightly change in response to FGF2, where increasing FGF2 secretion to a high level slightly promotes the formation of TSP1 bound and PF4 bound anti-angiogenic complexes.

Increasing the secretion of anti-angiogenic factors, particularly PF4, modulates the angiogenic ratio both by upregulating the levels of anti-angiogenic complexes and downregulating the pro-angiogenic complexes levels (**Figure 4**). The gradient along the horizontal axis in **Figures 4A,C**, column I indicates that increasing the secretion of TSP1 can decrease the angiogenic ratio. We also examined the change in the normalized levels of the angiogenic complexes. We found only TSP1 signaling complexes (**Figures 4A,C**, column IV) show an evident color change along the horizontal axis in response to changing TSP1 secretion rates, which indicates that TSP1 secretion decreases the angiogenic ratio mainly through promoting the formation of TSP1-bound anti-angiogenic complexes. Model predictions show that changing PF4 secretion can strongly shift the angiogenic ratio (**Figures 4B,D**, column I). In addition, increasing PF4 secretion promotes the formation of both TSP1- and PF4-bound anti-angiogenic complexes (**Figures 4B,D**, columns IV, V). However, there also appears to be a limit to the effect of PF4, where PF4 does not continue to significantly promote the formation of TSP1 anti-angiogenic complexes when its secretion rate is higher than a certain level. Although varying PF4 secretion only slightly affects the formation of VEGF signaling complexes (**Figure 4B**, column II), the color change along the horizontal axis in **Figures 4B,D**, column III shows that increasing PF4 secretion can strongly inhibit the formation of FGF2 signaling complexes. Furthermore, the secretion of PF4 can nearly neutralize the effect of FGF2 secretion on the formation of FGF2 signaling complex (**Figure 4D**, column III).

Overall, the model predicts that VEGF, FGF2, and TSP1 mainly bind to their own receptors to form more anti-angiogenic complexes and shift the angiogenic ratio, while PF4 affects the formation of signaling complexes of various angiogenic factors to change the angiogenic ratio.

Platelet Factor 4 Secretion Can Increase the Levels of Unbound Pro-angiogenic Factors in Tumor

Increased secretion of PF4 is predicted to affect the formation of both anti- and pro-angiogenic signaling complexes. To get detailed insight into how PF4 modulates the distribution of other angiogenic factors, we report the change of specific signaling species upon varying the PF4 secretion rate (**Figure 5** and **Supplementary Figure S2**). For these simulations, the PF4 secretion rate is again varied in a range of 100-fold below and 10-fold above the baseline value. In the figures, the fold-change of the species on the vertical axis is the species' concentration normalized to its concentration when PF4 secretion is 100-fold below the baseline value (the lower bound of the range over which the secretion rate was varied). Since PF4 mainly influences the other angiogenic factors by competing for the heparan sulfate binding sites, we investigated how the cHSPG level, a tumor-specific property, also affects the outcome of changing PF4 secretion. When we describe the cHSPG level below, we assume the baseline level as intermediate level. For low cHSPG levels, we ran simulations when cHSPG is 2-, 10-, and 100-fold lower than the baseline level. For high cHSPG levels, we considered cHSPH levels 2- and 10-fold higher than the baseline level.

Although anti-angiogenic factors, PF4 and TSP1, can bind to pro-angiogenic factors, VEGF and FGF2, to sequester pro-angiogenic factors, our model predicts that the levels of unbound pro-angiogenic factors do not necessarily decrease in the presence of more anti-angiogenic factors (**Figure 5**). Interestingly, varying PF4 secretion can significantly elevate the levels of unbound FGF2 and unbound VEGF in tumor tissue. For low cHSPG levels, increasing PF4 secretion may only slightly affect the levels of unbound FGF2 (**Figure 5A**, column I; gray and blue lines), while unbound VEGF levels can decrease with increasing PF4 secretion (**Figure 5B**, column I; blue lines). However, the secretion of PF4 strongly increases the level of unbound FGF2 if the tumor has intermediate to high cHSPG level (**Figure 5A**, column I; red, orange, and black lines). Similarly, PF4 secretion can increase unbound VEGF when the cHSPG level is high (**Figure 5B**, column I; red and orange lines). Examining the levels of specific isoforms of VEGF, we find that both unbound $VEGF_{165}$ and unbound $VEGF_{121}$ are affected. The change of $VEGF_{165}$ is more pronounced (**Figure 5C**, column I). Since the majority of unbound VEGF is $VEGF_{121}$, the fold-change of VEGF highly resembles the change of $VEGF_{121}$ (**Figure 5D**, column I).

The counterintuitive increase of unbound the pro-angiogenic factors with increasing PF4 secretion is caused by PF4 displacing pro-angiogenic factors from the cell surface heparan sulfate binding sites. PF4 preferentially competes for the HSPG on the

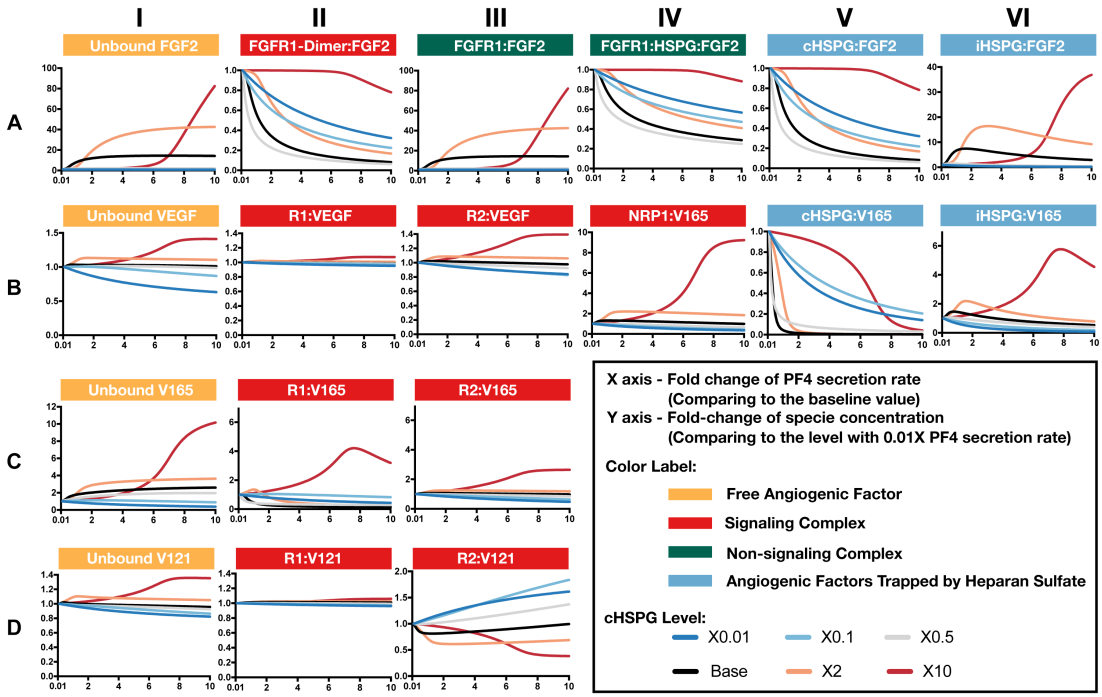

FIGURE 5 | Effects of PF4 secretion on the formation of specific angiogenic complexes. **(A)** The change of species in the FGF2 axis. **(B)** The change of species in VEGF axis. **(C)** Change of $VEGF_{165}$. **(D)** Change of $VEGF_{121}$. The predictions shown in the figures are based on the steady state of the system.

cell surface first, causing cHSPG-bound VEGF and FGF2 to decrease with increasing PF4 secretion (**Figures 5A,B**, column V). The decrement of cHSPG:FGF2 leads to a reduction of the trimeric complex FGFR1:HSPG:FGF2 (**Figure 5A**, column IV) and the FGF2 signaling dimer (**Figure 5A**, column II), which are only formed using cHSPG-bound FGF2. At the same time, the binding of PF4 to cHSPG reduces the availability of HSPG to bind to VEGF and VEGF receptors. Since the cHSPG affects VEGF binding to VEGFR2 and NRP1, the levels of VEGF-bound VEGFR2 and NRP1 change with increasing PF4 secretion. At high cHSPG level (**Figure 5B**, columns III–IV; red and orange lines), increasing PF4 secretion promotes the formation of VEGF-bound VEGFR2 and NRP1. At low to intermediate cHSPG levels (**Figure 5B**, columns III–IV; black, gray, light blue, and dark blue lines), increasing PF4 secretion inhibits the formation of VEGF-bound VEGFR2 and NRP1. This switch is because of the biphasic response to cHSPG level, which will be explored in next section. Since the two isoforms of VEGF have different binding property to receptors, $VEGF_{165}$ and $VEGF_{121}$ bound to VEGFR2 show very different fold-changes in response to increasing PF4 secretion (**Figures 5C,D**, column III). However, for both isoforms, varying PF4 secretion has differential effects on the levels of the pro-angiogenic ligated receptor complexes, depending on the cHSPG level.

iHSPG serves as a reservoir of angiogenic factors that can store and release pro-angiogenic factors. With increasing PF4 secretion, the pro-angiogenic factors displaced from cHSPG bind to iHSPG and form more FGF2- and VEGF-bound iHSPG (black, orange, and red lines in **Figures 5A,B**, column VI). After the depletion of the available cHSPG, secreted PF4 competes for iHSPG binding sites, and iHSPG-bound PF4 significantly increases as the PF4 secretion rate increases. At high PF4 secretion rates, PF4 is even able to displace FGF2 and VEGF from iHSPG and reduce the iHSPG-bound pro-angiogenic factors (**Figures 5A,B**, column VI; black, orange, and red lines).

In summary, the predictions show that the HSPG is an important mediator in how PF4 regulates pro-angiogenic factors. We found that, depending on the HSPG level, the secreted PF4 can displace more pro-angiogenic factors from the HS binding sites than the amount being sequestered, which eventually increases the level of unbound pro-angiogenic factors in the tumor interstitium.

VEGF Signaling Shows a Biphasic Response to the HSPG Level and PF4 Secretion Rate

As presented above, the model predicts that the secretion of PF4 can increase the level of pro-angiogenic complexes on the cell surface. In the case of FGF2, the pro-angiogenic signaling complexes involving FGFR1 dimers decrease with increasing PF4 (**Figure 5A**, column II) and the non-signaling complexes of ligated FGFR1 monomers increase with increasing PF4 secretion (**Figure 5A**, column III). These signaling and non-signaling forms of cell-surface FGF2 are also affected by cHSPG levels. Interestingly, for the VEGF axis, all three VEGF signaling complexes increase with increasing PF4 secretion, particularly for the high cHSPG condition (**Figure 5B**, columns II to IV;

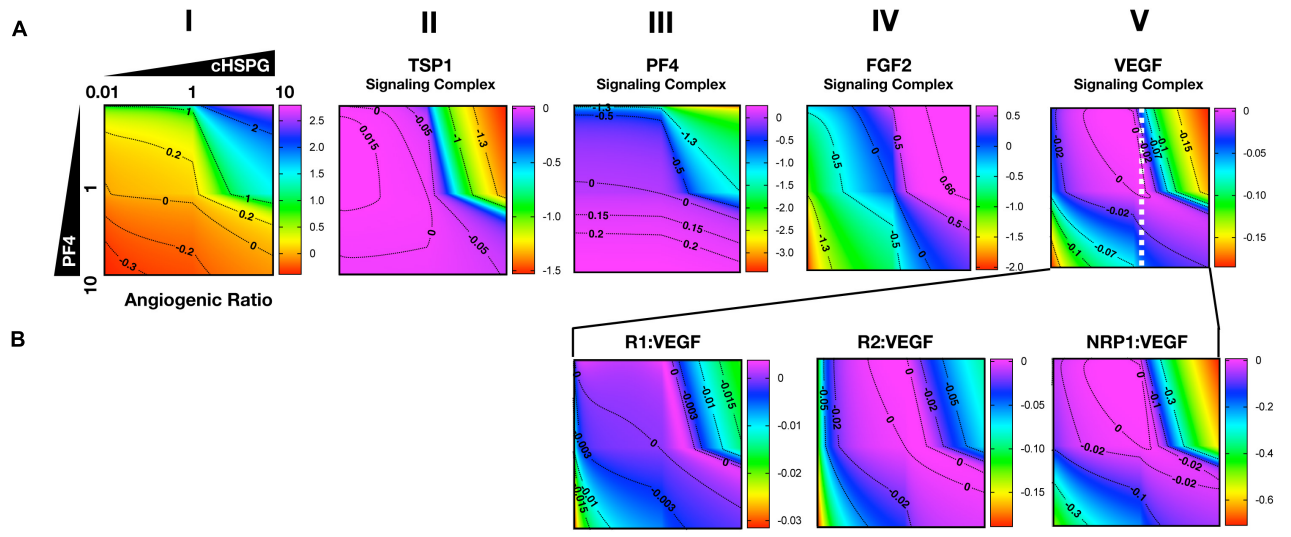

FIGURE 6 | Impact of PF4 and cHSPG on the formation of specific angiogenic complexes. The horizontal axis of each subplot shows the fold-change of cHSPG level, relative the baseline value, and the vertical axes show the fold-change of PF4 secretion rate, relative to its baseline value. The value indicated by the colorbar is in the $\log_{10}$ scale. **(A)** The change of the angiogenic ratio and normalized levels of signaling complexes: Column I shows the change of the angiogenic ratio and column II–V show the changes of the normalized signaling complex levels (the values are normalized to the baseline prediction). **(B)** The change of the normalized levels of specific VEGF signaling complexes. The predictions shown in the figures are based on the steady state of the system.

red and orange lines). This indicates an activation of VEGF pro-angiogenic signaling caused by the anti-angiogenic factor PF4. To explain these results, we further investigate how cHSPG level and PF4 secretion modulate the signaling complexes for each of the angiogenic factors modeled in the tumor tissue (**Figure 6**).

As shown in **Figure 6A**, column I, the model predicts that increasing cHSPG increases the angiogenic ratio (the tumor tissue is shifting to a more pro-angiogenic state) and increasing PF4 decreases the angiogenic ratio (shifting the tumor tissue to a less pro-angiogenic state). Together, these results indicate that HSPG promotes angiogenesis in tumor tissue, and the secretion of PF4 counteracts the pro-angiogenic effect of HSPG. Given the molecular detail of the model, we can explain these results. HSPG traps the two anti-angiogenic factors TSP1 and PF4. Thus, by increasing the HSPG level, the levels of TSP1 and PF4 signaling complexes are reduced (**Figure 6A**, columns II and III). However, HSPG is needed for the formation of the pro-angiogenic FGF2 signaling dimers. Although the HSPG traps FGF2 as well, the predictions show that increasing HSPG increases the FGF2 signaling complexes (**Figure 6A**, column IV). Additionally, increasing PF4 decreases FGF2 signaling complexes (**Figure 6A**, column IV) by displacing FGF2 from cHSPG, as explained in the previous section.

In contrast, VEGF shows biphasic response to HSPG. The gradient along the horizontal axis in **Figure 6A**, column V shows that the concentrations of the VEGF signaling complexes increase and then decrease with increasing HSPG level. Since PF4 competes for HSPG, increasing PF4 secretion decreases the HSPG availability to VEGF. Therefore, VEGF signaling also shows a biphasic response to PF4 secretion. For instance, at a medium HSPG level (dashed white line in **Figure 6A**, column V), along the vertical axis, the color changes from blue to purple then back to blue, which means the concentrations of the VEGF signaling complexes go up and then back down with increasing PF4 secretion. The VEGF signaling complexes include of VEGFR1-, VEGFR2-, and NRP1-bound VEGF. In addition, different types of VEGF signaling complexes, including VEGFR1-, VEGFR2- and NRP1-bound VEGF complexes, show a biphasic response to HSPG and PF4 secretion (**Figure 6B**, columns III and IV).

In summary, although PF4 secretion increases the unbound FGF2 level, the PF4 secretion strongly inhibits the formation of FGF2 signaling dimers that need HSPG to be formed. However, the VEGF signaling complexes can be formed through HSPG-dependent and HSPG-independent ways. Therefore, the VEGF signaling shows a biphasic response to the HSPG level. A low level HSPG limits the formation of VEGF signaling complexes through HSPG-dependent way. At the intermediate HSPG level, the VEGF signaling complexes reaches a peak level, while HSPG mainly traps VEGF and decreases VEGF signaling when it is present at a high level. Given the fact that PF4 secretion can efficiently limit HSPG availability, VEGF signaling shows a biphasic response to PF4 secretion as well. Therefore, at certain HSPG levels, the secretion of PF4 can enhance the VEGF pro-angiogenic signaling in tumor tissue.

The HSPG Level Affects the Response to Platelets Activation and Exogeneous PF4 Therapy

Building on the simulations in which we vary the secretion rate of PF4, we apply the model to predict the effects of a local release of PF4 at the tumor site, mimicking PF4 release following platelet activation (where angiogenic factors are released) or a

bolus injection of exogeneous PF4 as an anti-tumor therapy. The system is first allowed to reach steady state, which occurs in the first day. We then simulate two pulses of 5 mg PF4 per week, injected into the tumor interstitial space. This leads to a peak PF4 concentration of approximately 800 nM. The two pulses of PF4 occur at days 1 and 3.5. The release of PF4 follows an exponential decay with rate constant 2.8×10^{-5} s^{-1}, assuming the PF4 are encapsulated in a biomaterial delivery vehicle (Rohrs et al., 2016). We also perform the simulation at three different cHSPG levels to represent different tumor microenvironments: low (10-fold below the baseline value), medium (baseline value), and high (10-fold above the baseline). In this way, we examined how the tumor-specific property affects the response. Consistent with the results presented above, the model predictions reveal that depending on the HSPG level, platelet activation and recombinant PF4 can impact the pro-angiogenic signaling pathways in different ways.

The model predicts that cHSPG level significantly changes the response of VEGF signaling to the PF4 release (**Figure 7**). In a tumor microenvironment with high HSPG, the release of PF4 in the tumor leads to an activation of VEGF signaling pathway. Specifically, the concentration of unbound VEGF increases from 128 to 177 pM (a 1.9-fold increase) after the release of PF4, and it goes back down due to the degradation of PF4 (**Figure 7A**, red line). The levels of VEGFR1-, VEGFR2-, and NRP1-bound VEGF increase by 1.1-, 1.5-, and 8.1-fold, respectively (**Figures 7B–D**, red line). However, in a microenvironment with medium HSPG level, the release of PF4 inhibits VEGF signaling (**Figure 7**, black line). The concentrations of unbound VEGF and ligated VEGFR1 and VEGFR2 slightly decrease following the release of PF4, while NRP1-bound VEGF decreases 1.4-fold. In the tumor with low HSPG level, the release of PF4 shows a stronger inhibition, particularly for unbound VEGF and the VEGFR2 and NRP1 complexes (**Figure 7**, blue line). The concentrations of unbound VEGF and VEGFR2-bound VEGF each decreased 1.2-fold, and NRP1-bound VEGF significantly decreased, by 2.3-fold.

In addition to affecting the VEGF signaling complexes, release of PF4 influences the FGF2 signaling complexes to different extents, depending on the tumor microenvironment (**Figure 8**). Both unbound FGF2 and FGF2-bound FGFR1 complexes increase upon release of PF4 (**Figures 8A,B**). However, the concentration of the trimeric complex HSPG:FGFR1:FGF2 significantly decreased following each PF4 pulse (**Figure 8C**), which results in the reduction of FGF2-bound dimers. In a tumor with medium HSPG expression, the concentration of FGF2-bound dimers shows the largest decrease (6.5-fold). For low and high HSPG level, the FGF2-bound dimer concentration exhibits a 3.1- and 1.3-fold reduction, respectively.

To summarize, we simulated relevant tumor scenarios in which the PF4 concentration would suddenly increase, such as following platelet activation or administration of exogenous PF4 as an anti-angiogenic treatment strategy. The model predicts that PF4 has differential effects on the concentrations of pro-angiogenic signaling complexes involving VEGF and FGF2, depending on the cell-surface level of HSPG. Particularly, at a high cHSPG level, PF4 is shown to have a counterintuitive effect of promoting the formation of pro-angiogenic VEGF complexes. Overall, these simulations demonstrate the utility of the modeling framework in understanding the possible outcomes of events that are physiologically relevant to tumor angiogenesis.

DISCUSSION

We present a novel systems biology model describing the distribution of two potent pro-angiogenic factors and two important anti-angiogenic factors in tumor tissue. This model significantly expanded previous works to enable a study of four relevant angiogenic factors. Our model considers their interactions with each other in the extracellular space of tumor tissue, which was missing in previous models. In addition, the model expansion allows us to investigate the impact of heparan sulfate proteoglycans (HSPG) on the angiogenic factors' distribution. HSPG is an important modulator of tumor angiogenesis that is present on the cell surface, in the extracellular matrix, and in the cellular basement membranes. HSPG binds to and stores the angiogenic factors, facilitates the angiogenic factors' signaling and mediates the extracellular interactions of pro- and anti-angiogenic factors. Thus, HSPGs are a vital part of the extracellular network of angiogenic factors. Although the role of HSPGs in FGF2 signaling has been modeled in several studies (Ibrahimi et al., 2004; Zhao et al., 2010; Kanodia et al., 2014), the impact of HSPGs on VEGF ligand binding has not been modeled explicitly before. We addressed this gap by incorporating knowledge reported in experimental studies of the synergistic binding of VEGF, its receptors and heparin (Teran and Nugent, 2015). With the expansions upon previous models, our work reports a new computational framework for a comprehensive study of the angiogenic regulation in the extracellular space of tumor tissue.

Given the molecular detail of the model, we gain mechanistic insight into the extracellular regulation of tumor angiogenic signaling. In the tumor extracellular space, TSP1 and PF4 are thought to regulate the formation of pro-angiogenic signaling complexes involving VEGF and FGF2 through two different mechanisms: sequestration – binding directly to VEGF and FGF2 to prevent binding to their pro-angiogenic receptors, and competition – competing for cell-surface HSPG to inhibit the formation of pro-angiogenic complexes. Our study shows that PF4 significantly inhibits pro-angiogenic signaling, mainly by competing for cell-surface HSPG binding sites, not through direct binding. Our model predicts that the majority of TSP1 is in a cleaved form owing to the action of proteases, and this cleaved form is inactive and unable to compete for cell-surface HSPG. Therefore, our predictions show that TSP1 does not strongly inhibit the formation of VEGF and FGF2 signaling complexes. Moreover, the measured binding affinities between the anti-angiogenic factors (TSP1 and PF4) and the pro-angiogenic factors (VEGF and FGF2) are much weaker than their affinities to the receptors, which explains that the binding between them cannot efficiently sequester the pro-angiogenic ligands.

Our model predicts possible counterintuitive outcomes for the angiogenic state of following the release of anti-angiogenic factors. The secretion of anti-angiogenic factors, PF4 and TSP1,

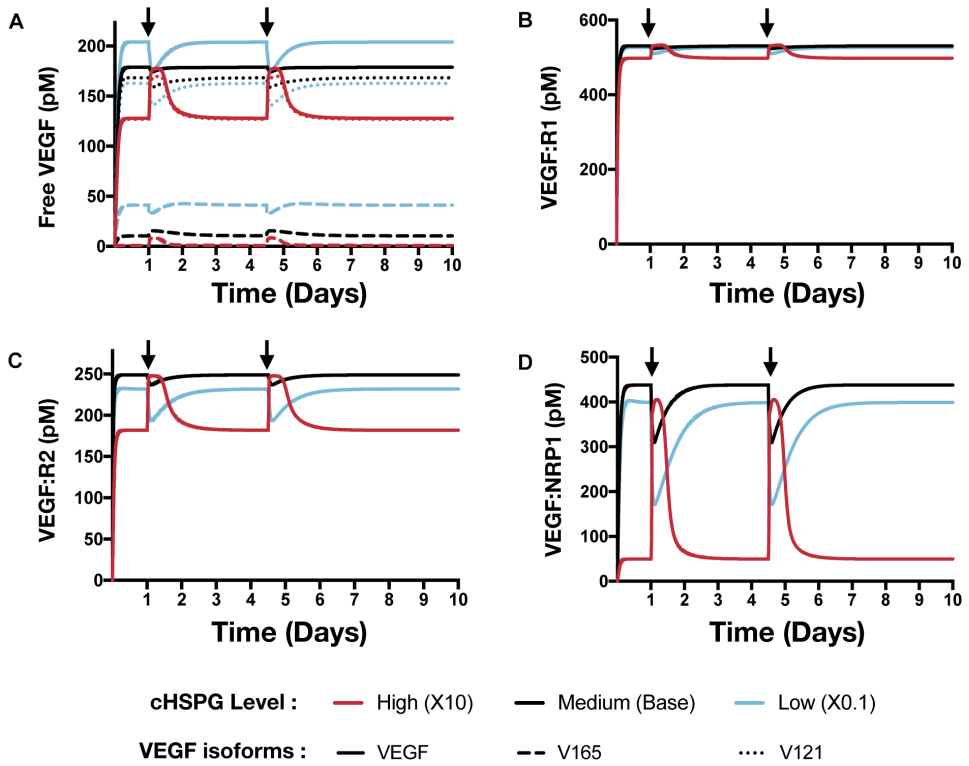

FIGURE 7 | Effect of PF4 pulse release on VEGF signaling pathway. The system is allowed to reach a steady state, followed by two pulses of PF4. The start time of PF4 release is indicated by the arrows: at day 1 and day 3.5. The concentrations of species in the VEGF signaling pathway are predicted: **(A)** Concentration of free VEGF: Solid line, VEGF; Dashed line, V165; Dotted line, V121). **(B)** VEGF:VEGFR1 complexes. **(C)** VEGF:VEGFR2 complexes. **(D)** VEGF:NRP1 complexes.

is generally assumed to reduce the concentrations of the free pro-angiogenic factors and inhibit the formation of pro-angiogenic signaling complexes. However, our model predicts that increasing the secretion of PF4 in tumor tissue can lead to two counterintuitive results: an increase in interstitial FGF2 and VEGF levels (see Platelet Factor 4 Secretion Can Increase the Levels of Unbound Pro-angiogenic Factors in Tumor in Section "Results") and greater formation of pro-angiogenic signaling complexes, particularly in the VEGF signaling pathway (see VEGF Signaling Shows a Biphasic Response to the HSPG Level and PF4 Secretion Rate in Section "Results"). The reason for the increased VEGF and FGF2 levels in the tumor interstitium following PF4 secretion is that PF4 competes for the HSPG binding sites in the cell surface, basement membrane and extracellular matrix, thereby releasing the pro-angiogenic factors from those sites and increasing the level of free pro-angiogenic ligands. When this effect is stronger than the sequestration that occurs when PF4 binds directly to pro-angiogenic factors, the level of unbound VEGF and FGF2 will be higher compared to the tumor microenvironmental condition with lower PF4 secretion (**Figure 5**, Column I). In addition, this counterintuitive outcome depends on the HSPG level. The schematic shown in **Figure 9** illustrates this point. In a low HSPG microenvironment, the pro-angiogenic factors (VEGF and FGF2) primarily bind to their corresponding cell surface receptors, and the rest are mostly in the free form (**Figure 9A**, Column I). In a high HSPG microenvironment, the free pro-angiogenic factors are trapped and stored as an HSPG-bound form (**Figure 9A**, Column II). As secretion of anti-angiogenic factors (TSP1 and PF4) increases, the secreted anti-angiogenic factors bind directly to pro-angiogenic factors and reduce the free pro-angiogenic factors level in a low HSPG condition (**Figure 9B**, Column I), while they mainly replace the pro-angiogenic factors from the HSPG in a high HSPG environment due to their stronger affinity to heparan sulfate (**Figure 9B**, Column II).

The greater formation of VEGF signaling complexes (which presumably will activate intracellular signaling) caused by PF4 is because of the intrinsic biphasic response to HSPG level (**Figure 6A**, Column V). At low levels, HSPG limits the formation of VEGF signaling complexes. When HSPG is present at an intermediate level, it promotes VEGF signaling by facilitating VEGF binding to receptors. However, when HSPG is at an even higher level, it traps VEGF and reduces the formation of VEGF signaling complexes, which leads to a low VEGF signaling again. Higher secretion of PF4 allows PF4 to more strongly compete for HSPG, which can alleviate the HSPG sequestration of VEGF and promote VEGF signaling in certain conditions. Unlike the VEGF signaling complexes, which can be formed through HSPG dependent and independent ways, we assume the formation of FGF2 signaling complexes requires HSPG as a co-receptor in the model (Kanodia et al., 2014). Therefore, the biphasic response is not seen in FGF2 signaling complexes (**Figure 5A**, Column II),

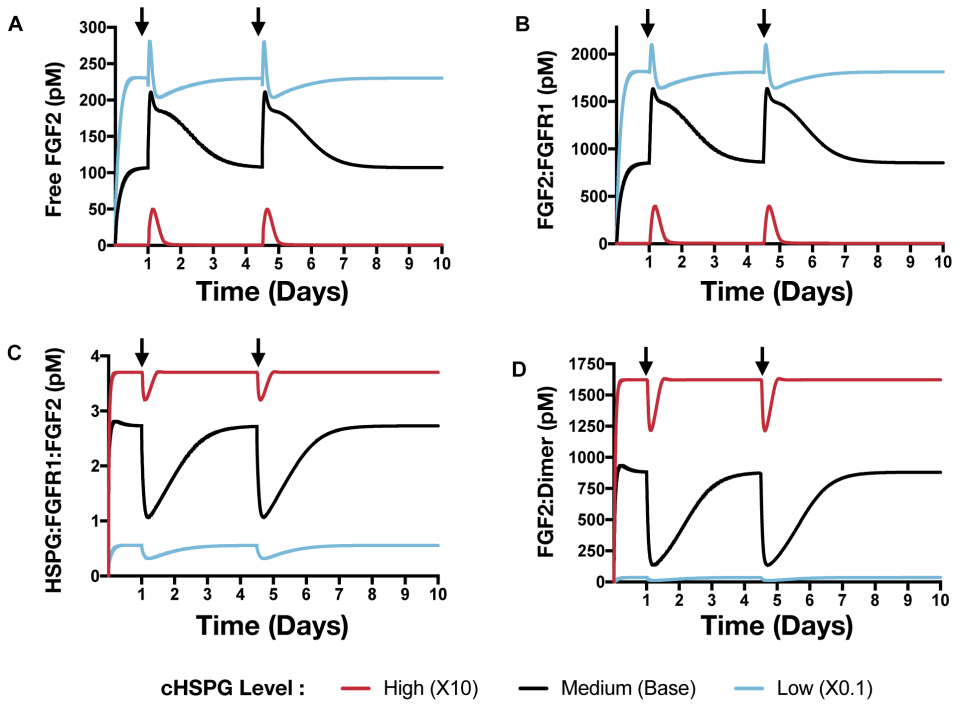

FIGURE 8 | The effect of PF4 pulse release on FGF2 signaling pathway. The system is allowed to reach a steady state, followed by two pulses of PF4. The start time of PF4 release is indicated by the arrows: at day 1 and day 3.5. The concentrations of species in the FGF2 signaling pathway are predicted: **(A)** Concentration of free FGF2. **(B)** FGF2:FGFR1 complexes. **(C)** HSPG:FGFR1:FGF2 complexes. **(D)** HSPG:FGFR1:FGF2 dimers.

in which the increasing of PF4 secretion always decreases the formation of FGF2 signaling complexes.

These predicted counterintuitive results are clinically relevant for understanding the outcome of platelet activation and anti-angiogenic therapy. In the human body, PF4 is stored in platelet α-granules and released upon platelet activation. It is reported that the serum concentrations of PF4 exceeds 8 μg/mL (276 nM) during platelet activation (Chesterman et al., 1978; Leitzel et al., 1991; Kurimoto et al., 1995; Peterson et al., 2012; Sabrkhany et al., 2017). Given the fact that platelets are attracted to and accumulate at tumor sites (Stakiw et al., 2014), it is possible that even higher concentrations of PF4 may be present in the local tumor microenvironment when platelet activation occurs. Besides the release of endogenous PF4 from platelets, recombinant PF4 (rPF4) has been studied as an anti-tumor therapeutic to prevent angiogenesis, showing efficacy in both *in vitro* and *in vivo* settings (Gengrinovitch et al., 1995; Struyf et al., 2007). rPF4 was tested in a mouse model to inhibit tumor growth with a dosage at 0.1 μg/μL for 5 μg in total (Struyf et al., 2007), and rPF4 has been tested in patients with advanced colorectal carcinoma at a dosage of 3 mg/kg in 30-min infusions (Belman et al., 1996). Thus, the administration of rPF4 as a therapeutic agent will greatly increase the total PF4 level in the tumor. To mimic the local increase of PF4 concentration due to either platelet activation or a bolus injection of rPF4, we simulated a situation of controlled release of PF4 in the tumor interstitium. The simulation results highlight the impact of HSPGs level on the outcome of platelet activation and anti-angiogenic therapy (see The HSPG Level Affects the Response to Platelets Activation and Exogeneous PF4 Therapy in Section "Results").

In addition, our model also has other practical applications, complementing pre-clinical and clinical studies. The model can be further expanded to a whole body model to study the clinically tested anti-angiogenic therapy and patient response, as we have done in previous work with VEGF and TSP1 modeling (Li and Finley, 2018). Therefore, our model provides a basis to study the anti-angiogenic therapy targeting multiple angiogenic factors, including VEGF, FGF2, TSP1, and PF4. In our study, we used the angiogenic ratio to characterize the overall angiogenic state of the tumor tissue. This is based on the assumption that the different types of signaling complex have the same contribution to angiogenic signaling, which may not represent the real case. Linking the signaling complexes with corresponding downstream signaling network can help address this issue and enable a better understanding of tumor angiogenesis. The signaling complexes in our model are also the species initiating downstream signaling in previous published downstream signaling models (Kanodia et al., 2014; Wu and Finley, 2017; Song and Finley, 2018); therefore, our model can be connected with models of intracellular signaling to characterize the downstream signaling changes.

We acknowledge that the predictions from the model are sensitive to the values of parameters. In our study, the experimental data we are comparing to are the measured levels

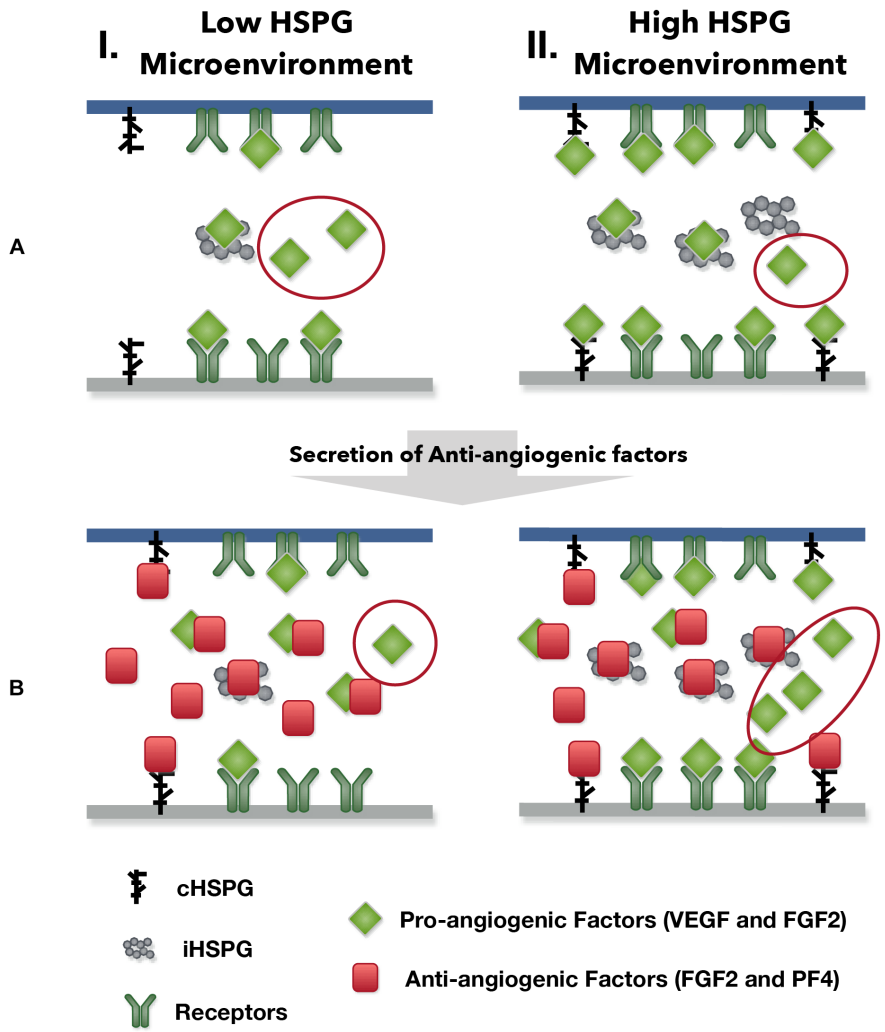

FIGURE 9 | The outcomes of anti-angiogenic factor secretion in different microenvironments. Column I shows the condition of low HSPG level and Column II shows the condition of high HSPG level. **(A)** Before the secretion of anti-angiogenic factors. **(B)** After the secretion of anti-angiogenic factors.

of angiogenic factors in tumor tissue samples (**Table 1**). Since the predicted level of angiogenic factors is highly sensitive to the secretion rates of angiogenic factors, we explicitly performed simulations to vary the secretion rates of angiogenic factors in this study. In addition, we explored the effect of the HSPG level, another influential parameter. We specifically varied the cHSPG level in the model because cHSPG serves both as the reservoir and the co-receptor of angiogenic factors. It is important to notice that there are other unexplored parameters that could affect the model predictions. For example, changing the secretion rates of MMPs will affect the cleavage of VEGF, which can subsequently change the amount of VEGF bound to the extracellular matrix and receptors. Additionally, the affinities of angiogenic factors to HSPG could affect the regulatory role of HSPG. We did not explore all possible parameters in this study. Instead, we focus primarily on the effects of angiogenic factor secretion and HSPG level, parameters that account for key aspects of tumor heterogeneity. In the future, the model can be used to investigate the effects of many more parameters.

There are some more limitations of our model that can be addressed in future work. Given the scarcity of the quantitative data, we used the measurements from tumor types other than breast cancer to tune the baseline value of the angiogenic factors secretion rates. Since there are no available measurements of the PF4 level directly from tumor tissue sample, we use the measured blood PF4 level in breast cancer patients as an estimation of tumor interstitial PF4 level. Additionally, HSPG includes various types, each with different masses and number and types of heparan sulfate chains (Sarrazin et al., 2011). This great complexity is difficult to fully characterize mathematically and warrants its own highly detailed mechanistic model. To make the model more useful, we made a simplification to only explicitly define two generic species of HSPGs that capture the two key HSPG classes with distinct functions, rather than a detailed description of all HSPG species. One of the types

of HSPGs in the model is on the cell surface (cHSPG) that can bind to ligand, couple with receptors, and is subject to internalization. The other type is the interstitial HSPG (iHSPG) in the extracellular matrix and basement membranes, which only traps free angiogenic ligands and is not subject to degradation and internalization. We acknowledge that the soluble form of HSPG, such as heparin, is also important to consider in the context of the tumor (Borsig, 2010). However, we do not explicitly model this class of HSPGs, because its binding to ligands and receptors, as well as its degradation, makes it very similar to the cHSPG in the model. If needed, our model can be extended to include more types of HSPGs. Despite these limitations, our model provides relevant mechanistic insight into interactions between angiogenic factors, their receptors, and HSPGs.

CONCLUSION

In this study, we present a novel model to characterize the extracellular distribution of four important angiogenic factors: VEGF, FGF2, TSP1, and PF4. The model provides mechanistic insights into the regulation of the angiogenic interaction network in the extracellular space of tumor tissue. We expect that the insights generated by our model will enable a better understanding of tumor angiogenesis interactions and aid the development of new anti-angiogenic therapy.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

SF designed the research. DL constructed the model and generated model simulations and analyses. DL and SF contributed to writing of the manuscript and have read and approved the final manuscript.

FUNDING

The authors acknowledge the support of the US National Science Foundation (CAREER Award 1552065 to SF), the American Cancer Society (130432-RSG-17-133-01-CSM to SF), and the USC Provost's Ph.D. Fellowship (DL).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2019.00823/full#supplementary-material

FIGURE S1 | The variations of predicted tumor tissue distribution of VEGF **(A)**, FGF2 **(B)**, TSP1 **(C)**, and PF4 **(D)**. The secretion rates of VEGF, FGF2, TSP1, PF4, and MMPs are sampled within a range of 100-fold below and 10-fold above the baseline values. The mean value and the standard deviations of the predictions of 5000 Monte Carlo simulations are shown in plots.

FIGURE S2 | Effects of PF4 secretion on the angiogenic distribution. The predicted change of **(A)** TSP1 and **(B)** PF4 species with increasing PF4 secretion.

FILE S1 | List of the reactions in the model.

FILE S2 | List of the parameter values used in the model.

FILE S3 | MATLAB model file.

REFERENCES

Abdalla, A. M. E., Xiao, L., Ullah, M. W., Yu, M., Ouyang, C., and Yang, G. (2018). Current challenges of cancer anti-angiogenic therapy and the promise of nanotherapeutics. *Theranostics* 8, 533–548. doi: 10.7150/thno.21674

Alessi, P., Leali, D., Camozzi, M., Cantelmo, A. R., Albini, A., and Presta, M. (2009). Anti-FGF2 approaches as a strategy to compensate resistance to anti-VEGF therapy: long-pentraxin 3 as a novel antiangiogenic FGF2-antagonist. *Eur. Cytokine Netw.* 20, 225–234. doi: 10.1684/ecn.2009.0175

Baker, E. A., Leaper, D. J., Hayter, J. P., and Dickenson, A. J. (2006). The matrix metalloproteinase system in oral squamous cell carcinoma. *Br. J. Oral Maxillofac. Surg.* 44, 482–486. doi: 10.1016/j.bjoms.2005.10.005

Beenken, A., and Mohammadi, M. (2009). The FGF family: biology, pathophysiology and therapy. *Nat. Rev. Drug Discov.* 8, 235–253. doi: 10.1038/nrd2792

Belman, N., Bonnem, E. M., Harvey, H. A., and Lipton, A. (1996). Phase I trial of recombinant platelet factor 4 (rPF4) in patients with advanced colorectal carcinoma. *Invest. New Drugs* 14, 387–389. doi: 10.1007/BF00180815

Bikfalvi, A. (2004). Platelet factor 4: an inhibitor of angiogenesis. *Semin. Thromb. Hemost.* 30, 379–385. doi: 10.1055/s-2004-831051

Borsig, L. (2010). Heparin as an inhibitor of cancer progression. *Prog. Mol. Biol. Transl. Sci.* 93, 335–349. doi: 10.1016/S1877-1173(10)93014-7

Byrne, G. J., Hayden, K. E., Mcdowell, G., Lang, H., and Kirwan, C. C. (2007). Angiogenic characteristics of circulating and tumoural thrombospondin-1 in breast cancer. *Int. J. Oncol.* 31, 1127–1132.

Carmeliet, P. (2005). VEGF as a key mediator of angiogenesis in cancer. *Oncology* 69(Suppl. 3), 4–10. doi: 10.1159/000088478

Chesterman, C. N., McGready, J. R., Doyle, D. J., and Morgan, F. J. (1978). Plasma levels of platelet factor 4 measured by radioimmunoassay. *Br. J. Haematol.* 40, 489–500. doi: 10.1111/j.1365-2141.1978.tb05819.x

Dawes, J., Smith, R. C., and Pepper, D. S. (1978). The release distribution and clearance of human beta-thromboglobulin and platelet factor 4. *Thromb. Res.* 12, 851–861.

Faeder, J. R., Blinov, M. L., and Hlavacek, W. S. (2009). Rule-based modeling of biochemical systems with BioNetGen. *Methods Mol. Biol.* 500, 113–167. doi: 10.1007/978-1-59745-525-1_5

Finley, S. D., Chu, L.-H., and Popel, A. S. (2015). Computational systems biology approaches to anti-angiogenic cancer therapeutics. *Drug Discov. Today* 20, 187–197. doi: 10.1016/j.drudis.2014.09.026

Finley, S. D., and Popel, A. S. (2013). Effect of tumor microenvironment on tumor VEGF during anti-VEGF treatment: systems biology predictions. *J. Natl. Cancer Inst.* 105, 802–811. doi: 10.1093/jnci/djt093

Gengrinovitch, S., Greenberg, S. M., and Cohen, T. (1995). Platelet factor-4 inhibits the mitogenic activity of $VEGF_{121}$ and $VEGF_{165}$ using several concurrent mechanisms. *J. Biol. Chem.* 270, 15059–15065. doi: 10.1074/jbc.270.25.15059

Giri, D., Ropiquet, F., and Ittmann, M. (1999). Alterations in expression of basic fibroblast growth factor (FGF) 2 and its receptor FGFR-1 in human prostate cancer. *Clin. Cancer Res.* 5, 1063–1071.

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Ibrahimi, O. A., Zhang, F., Hrstka, S. C. L., Mohammadi, M., and Linhardt, R. J. (2004). Kinetic model for FGF, FGFR, and proteoglycan signal

transduction complex assembly. *Biochemistry* 43, 4724–4730. doi: 10.1021/bi0352320

Jeanne, A., Schneider, C., Martiny, L., and Dedieu, S. (2015). Original insights on thrombospondin-1-related antireceptor strategies in cancer. *Front. Pharmacol.* 6:252. doi: 10.3389/fphar.2015.00252

Jouan, V., Canron, X., and Alemany, M. (1999). Inhibition of in vitro angiogenesis by platelet factor-4-derived peptides and mechanism of action. *Blood* 94, 984–993.

Kano, M. R. (2005). VEGF-A and FGF-2 synergistically promote neoangiogenesis through enhancement of endogenous PDGF-B-PDGFR signaling. *J. Cell Sci.* 118(Pt 16), 3759–3768. doi: 10.1242/jcs.02483

Kanodia, J., Chai, D., and Vollmer, J. (2014). Deciphering the mechanism behind fibroblast growth factor (FGF) induced biphasic signal-response profiles. *Cell Commun. Signal.* 12:34. doi: 10.1186/1478-811X-12-34

Kaur, S., Martin-Manso, G., Pendrak, M. L., Garfield, S. H., Isenberg, J. S., and Roberts, D. D. (2010). Thrombospondin-1 inhibits VEGF receptor-2 signaling by disrupting its association with CD47. *J. Biol. Chem.* 285, 38923–38932. doi: 10.1074/jbc.M110.172304

Kim, T. J., Landen, C. N., and Lin, Y. G. (2009). Combined anti-angiogenic therapy against VEGF and integrin alphaVbeta3in an orthotopic model of ovarian cancer. *Cancer Biol. Ther.* 8, 2263–2272

Korc, M., and Friesel, R. E. (2009). The role of fibroblast growth factors in tumor growth. *Curr. Cancer Drug Targets* 9, 639–651.

Kurimoto, M., Nishijima, M., Hirashima, Y., Endo, S., and Takaku, A. (1995). Plasma platelet-derived growth factor-B chain is elevated in patients with extensively large brain tumour. *Acta Neurochir.*137, 182–187. doi: 10.1007/BF02187191

Lambert, M. P., Wang, Y., Bdeir, K. H., Nguyen, Y., Kowalska, M. A., and Poncz, M. (2009). Platelet factor 4 regulates megakaryopoiesis through low-density lipoprotein receptor-related protein 1 (LRP1) on megakaryocytes. *Blood* 114, 2290–2298. doi: 10.1182/blood-2009-04-216473

Lasagni, L., Francalanci, M., and Annunziato, F. (2003). An alternatively spliced variant of CXCR3 mediates the inhibition of endothelial cell growth induced by IP-10, Mig, and I-TAC, and acts as functional receptor for platelet factor 4. *J. Exp. Med.* 197, 1537–1549. doi: 10.1084/jem.20021897

Leitzel, K., Bryce, W., and Tomita, J. (1991). Elevated plasma platelet-derived growth factor B-chain levels in cancer patients. *Cancer Res.* 51, 4149–4154.

Li, D., and Finley, S. D. (2018). The impact of tumor receptor heterogeneity on the response to anti-angiogenic cancer treatment. *Integr. Biol.* 10, 844–860. doi: 10.1039/c6ib00093b

Lord, M. S., Cheng, B., Farrugia, B. L., McCarthy, S., and Whitelock, J. M. (2017). Platelet factor 4 binds to vascular proteoglycans and controls both growth factor activities and platelet activation. *J. Biol. Chem.* 292, 4054–4063. doi: 10.1074/jbc.M116.760660

Mac Gabhann, F., and Popel, A. S. (2005). Differential binding of VEGF isoforms to VEGF receptor 2 in the presence of neuropilin-1: a computational model. *Am. J. Physiol. Hear. Circ. Physiol.* 288, H2851–H2860. doi: 10.1152/ajpheart.01218.2004

Mac Gabhann, F., and Popel, A. S. (2006). Targeting neuropilin-1 to inhibit VEGF signaling in cancer: comparison of therapeutic approaches. *PLoS Comput. Biol.* 2:e180. doi: 10.1371/journal.pcbi.0020180

Mamluk, R., Gechtman, Z., Kutcher, M. E., Gasiunas, N., Gallagher, J., and Klagsbrun, M. (2002). Neuropilin-1 binds vascular endothelial growth factor 165, placenta growth factor-2, and heparin via its b1b2 domain. *J. Biol. Chem.* 277, 24818–24825. doi: 10.1074/jbc.M200730200

Margosio, B., Marchetti, D., and Vergani, V. (2003). Thrombospondin 1 as a scavenger for matrix-associated fibroblast growth factor 2. *Blood* 102, 4399–4406. doi: 10.1182/blood-2003-03-0893

Mikhailenko, I., Krylov, D., Argraves, K. M., Roberts, D. D., Liau, G., and Strickland, D. K. (2002). Cellular internalization and degradation of thrombospondin-1 is mediated by the amino-terminal heparin binding domain (HBD). High affinity interaction of dimeric HBD with the low density lipoprotein receptor-related protein. *J. Biol. Chem.* 272, 6784–6791. doi: 10.1074/jbc.272.10.6784

Northrup, S. H., and Erickson, H. P. (2006). Kinetics of protein-protein association explained by brownian dynamics computer simulation. *Proc. Natl. Acad. Sci. U.S.A.* 89, 3338–3342. doi: 10.1073/pnas.89.8.3338

Perollet, C., Han, Z. C., Savona, C., Caen, J. P., and Bikfalvi, A. (1998). Platelet factor 4 modulates fibroblast growth factor 2 (FGF-2) activity and inhibits FGF-2 dimerization. *Blood* 91, 3289–3299.

Peterson, J. E., Zurakowski, D., and Italiano, J. E. (2012). VEGF, PF4 and PDGF are elevated in platelets of colorectal cancer patients. *Angiogenesis* 15, 265–273. doi: 10.1007/s10456-012-9259-z

Ren, B., Yee, K. O., Lawler, J., and Khosravi-Far, R. (2006). Regulation of tumor angiogenesis by thrombospondin-1. *Biochim. Biophys. Acta* 1765, 178–188. doi: 10.1016/j.bbcan.2005.11.002

Resovi, A., Pinessi, D., Chiorino, G., and Taraboletti, G. (2014). Current understanding of the thrombospondin-1 interactome. *Matrix Biol.* 37, 83–91. doi: 10.1016/j.matbio.2014.01.012

Rohrs, J. A., Sulistio, C. D., and Finley, S. D. (2016). Predictive model of thrombospondin-1 and vascular endothelial growth factor in breast tumor tissue. *NPJ Syst. Biol. Appl.* 2:16030. doi: 10.1038/npjsba.2016.30

Sabrkhany, S., Kuijpers, M. J. E., and van Kuijk, S. M. J. (2017). A combination of platelet features allows detection of early-stage cancer. *Eur. J. Cancer* 80, 5–13. doi: 10.1016/j.ejca.2017.04.010

Sachais, B. S., Kuo, A., and Nassar, T. (2002). Platelet factor 4 binds to low-density lipoprotein receptors and disrupts the endocytic itinerary, resulting in retention of low-density lipoprotein on the cell surface. *Blood* 99, 3613–3622. doi: 10.1182/blood.V99.10.3613

Sarrazin, S., Lamanna, W. C., and Esko, J. D. (2011). Heparan sulfate proteoglycans. *Cold Spring Harb. Perspect. Biol.* 3:a004952. doi: 10.1101/cshperspect.a004952

Schlosshauer, M., and Baker, D. (2004). Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Sci.* 13, 1660–1669. doi: 10.1110/ps.03517304

Seghezzi, G., Patel, S., and Ren, C. J. (1998). Fibroblast growth factor-2 (FGF-2) induces vascular endothelial growth factor (VEGF) expression in the endothelial cells of forming capillaries: an autocrine mechanism contributing to angiogenesis. *J. Cell Biol.* 141, 1659–1673. doi: 10.1083/jcb.141.7.1659

Shiba, T., Nishimura, D., and Kawazoe, Y. (2003). Modulation of mitogenic activity of fibroblast growth factors by inorganic polyphosphate. *J. Biol. Chem.* 278, 26788–26792. doi: 10.1074/jbc.M303468200

Soker, S., Miao, H. Q., Nomi, M., Takashima, S., and Klagsbrun, M. (2002). VEGF165 mediates formation of complexes containing VEGFR-2 and neuropilin-1 that enhance VEGF165-receptor binding. *J. Cell Biochem.* 85, 357–368. doi: 10.1002/jcb.10140

Song, M., and Finley, S. D. (2018). Mechanistic insight into activation of MAPK signaling by pro-angiogenic factors. *BMC Syst. Biol.* 12:145. doi: 10.1186/s12918-018-0668-665

Stakiw, J., Burnouf, T., Kotb, R. R., Emara, M. E., and Goubran, H. A. (2014). Regulation of tumor growth and metastasis: the role of tumor microenvironment. *Cancer Growth Metastasis* 7, 9–18. doi: 10.4137/cgm.s11285

Stefanini, M. O., Wu, F. T. H., Mac Gabhann, F., and Popel, A. S. (2008). A compartment model of VEGF distribution in blood, healthy and diseased tissues. *BMC Syst. Biol.* 2:77. doi: 10.1186/1752-0509-2-77

Stefanini, M. O., Wu, F. T. H., Mac Gabhann, F., and Popel, A. S. (2010). Increase of plasma VEGF after intravenous administration of bevacizumab is predicted by a pharmacokinetic model. *Cancer Res.* 70, 9886–9894. doi: 10.1158/0008-5472.CAN-10-1419

Stringer, S. E., and Gallagher, J. T. (1997). Specific binding of the chemokine platelet factor 4 to heparan sulfate. *J. Biol. Chem.* 272, 20508–20514. doi: 10.1074/jbc.272.33.20508

Struyf, S., Burdick, M. D., and Peeters, E. (2007). Platelet factor-4 variant chemokine CXCL4L1 inhibits melanoma and lung carcinoma growth and metastasis by preventing angiogenesis. *Cancer Res.* 67, 5940–5948. doi: 10.1158/0008-5472.CAN-06-4682

Teran, M., and Nugent, M. A. (2015). Synergistic binding of vascular endothelial growth factor-a and its receptors to heparin selectively modulates complex affinity. *J. Biol. Chem.* 290, 16451–16462. doi: 10.1074/jbc.M114.627372

Uhlen, M. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics* 4, 1920–1932. doi: 10.1074/mcp.M500279-MCP200

Uronis, H. E., Cushman, S. M., and Bendell, J. C. (2013). A phase I study of ABT-510 plus bevacizumab in advanced solid tumors. *Cancer Med.* 2, 316–324. doi: 10.1002/cam4.65

van Beijnum, J. R., Griffioen, A. W., Huijbers, E. J. M., Thijssen, V. L., and Nowak-Sliwinska, P. (2015). The great escape; the hallmarks of resistance to antiangiogenic therapy. *Pharmacol. Rev.* 67, 441–461. doi: 10.1124/pr.114.010215

Vandercappellen, J., Van Damme, J., and Struyf, S. (2011). The role of the CXC chemokines platelet factor-4 (CXCL4/PF-4) and its variant (CXCL4L1/PF-4var) in inflammation, angiogenesis and cancer. *Cytokine Growth Factor Rev.* 22, 1–18. doi: 10.1016/J.CYTOGFR.2010.10.011

Vasudev, N. S., and Reynolds, A. R. (2014). Anti-angiogenic therapy for cancer: current progress, unresolved questions and future directions. *Angiogenesis* 17, 471–494. doi: 10.1007/s10456-014-9420-y

Wang, Z., and Huang, H. (2013). Platelet factor-4 (CXCL4/PF-4): an angiostatic chemokine for cancer therapy. *Cancer Lett.* 331, 147–153. doi: 10.1016/j.canlet.2013.01.006

Wehland, M., Bauer, J., Infanger, M., and Grimm, D. (2012). Target-based anti-angiogenic therapy in breast cancer. *Curr. Pharm. Des.* 18, 4244–4257. doi: 10.2174/138161212802430468

Whitaker, G. B., Limberg, B. J., and Rosenbaum, J. S. (2001). Vascular endothelial growth factor receptor-2 and neuropilin-1 form a receptor complex that is responsible for the differential signaling potency of VEGF165 and VEGF121. *J. Biol. Chem.* 276, 25520–25531. doi: 10.1074/jbc.M102315200

Wu, Q., and Finley, S. D. (2017). Predictive model identifies strategies to enhance TSP1-mediated apoptosis signaling. *Cell Commun. Signal.* 15, 6–9. doi: 10.1186/s12964-017-0207-209

Zhao, B., Zhang, C., Forsten-Williams, K., Zhang, J., and Fannon, M. (2010). Endothelial cell capture of heparin-binding growth factors under flow. *PLoS Comput. Biol.* 6:e1000971. doi: 10.1371/journal.pcbi.1000971

Zhao, W., McCallum, S. A., Xiao, Z., Zhang, F., and Linhardt, R. J. (2012). Binding affinities of vascular endothelial growth factor (VEGF) for heparin-derived oligosaccharides. *Biosci. Rep.* 32, 71–81. doi: 10.1042/BSR20110077

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 September 2019
doi: 10.3389/fgene.2019.00821

Screening and Identification of Potential Prognostic Biomarkers in Adrenocortical Carcinoma

Wen-Hao Xu[1,2†], Junlong Wu[1,2†], Jun Wang[1,2†], Fang-Ning Wan[1,2], Hong-Kai Wang[1,2], Da-Long Cao[1,2], Yuan-Yuan Qu[1,2], Hai-Liang Zhang[1,2*] and Ding-Wei Ye[1,2*]*

[1] Department of Urology, Fudan University Shanghai Cancer Center, Shanghai, China, [2] Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada,
United States

Reviewed by:
Pawel Buczkowicz,
Gene42, Inc., Canada
Maria Candida Barisson Villares Fragoso,
University of São Paulo, Brazil
Reju Korah,
Yale University,
United States

***Correspondence:**
Yuan-Yuan Qu
quyy1987@163.com
Hai-Liang Zhang
zhanghl918@163.com
Ding-Wei Ye
dwyeli@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: 21 December 2018
Accepted: 08 August 2019
Published: 11 September 2019

Citation:
Xu W-H, Wu J, Wang J, Wan F-N, Wang H-K, Cao D-L, Qu Y-Y, Zhang H-L and Ye D-W (2019) Screening and Identification of Potential Prognostic Biomarkers in Adrenocortical Carcinoma. Front. Genet. 10:821. doi: 10.3389/fgene.2019.00821

Objective: Adrenocortical carcinoma (ACC) is a rare but aggressive malignant cancer that has been attracting growing attention over recent decades. This study aims to integrate protein interaction networks with gene expression profiles to identify potential biomarkers with prognostic value *in silico*.

Methods: Three microarray data sets were downloaded from the Gene Expression Omnibus (GEO) database to identify differentially expressed genes (DEGs) according to the normalization annotation information. Enrichment analyses were utilized to describe biological functions. A protein–protein interaction network (PPI) of the DEGs was developed, and the modules were analyzed using STRING and Cytoscape. LASSO Cox regression was used to identify independent prognostic factors. The Kaplan–Meier method for the integrated expression score was applied to analyze survival outcomes. A receiver operating characteristic (ROC) curve was constructed with area under curve (AUC) analysis to determine the diagnostic ability of the candidate biomarkers.

Results: A total of 150 DEGs and 24 significant hub genes with functional enrichment were identified as candidate prognostic biomarkers. LASSO Cox regression suggested that *ZWINT*, *PRC1*, *CDKN3, CDK1* and *CCNA2* were independent prognostic factors in ACC. In multivariate Cox analysis, the integrated expression scores of the modules showed statistical significance in predicting **disease-free survival (DFS, $P = 0.019$) and overall survival (OS, $P < 0.001$)**. Meanwhile, ROC curves were generated to validate the ability of the Cox model to predict prognosis. The AUC index for the integrated genes scores was 0.861 ($P < 0.0001$).

Conclusion: In conclusion, the present study identifies DEGs and hub genes that may be involved in poor prognosis and early recurrence of ACC. The expression levels of *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2* are of high prognostic value, and may help us understand better the underlying carcinogenesis or progression of ACC. Further studies are required to elucidate molecular pathogenesis and alteration in signaling pathways for these genes in ACC.

Keywords: adrenocortical carcinoma, bioinformatics analysis, biomarker, prognosis, network module

INTRODUCTION

Adrenocortical carcinoma (ACC) is a rare endocrine malignancy with an annual incidence of 0.7–2.0 per million people, accounting for an estimated 0.02% of all cancers (Wajchenberg et al., 2000; Kebebew et al., 2006; Kerkhofs et al., 2013). Although comparatively uncommon, ACC patients often face aggressive progression, with merely less than 35% of patients surviving 5 years after initial diagnosis (Else et al., 2014). Currently, the preferred treatment regimen for ACC is surgical resection of the primary tumor (Fassnacht et al., 2013). However, almost half of ACC patients have disseminated metastasis, and approximately one-third of patients have locoregional metastases after surgery (Else et al., 2014). The first–line treatment, and the only ACC-specific medical therapy approved by the US Food and Drug Administration, is Mitotane, which is regularly used as an adjuvant agent in these patients (Else et al., 2014). Mitotane disrupts mitochondria and activates an apoptotic process (Poli et al., 2013). A major concern of the therapeutic management with Mitotane is the risk of toxicity, which may lead to severe adrenal insufficiency (Paragliola et al., 2018).

Accumulating evidence has demonstrated that gene expression levels and related pathways are involved in the carcinogenesis and progression of ACC. For example, the most frequent alterations observed in ACC are overexpression of insulin-like growth factor 2 (*IGF-2*) (Gicquel et al., 2001; Giordano et al., 2003; de Fraipont et al., 2005) and constitutive activation of the Wnt/β-Catenin pathway (Gaujoux et al., 2011). Despite these encouraging advances in ACC clinical strategies, only a minority of patients receive any significant survival benefit because of a lack of effective therapeutic strategies (Mohan et al., 2018). Therefore, it is crucial to understand the underlying molecular mechanisms involved in the carcinogenesis, proliferation and recurrence of ACC and thus develop effective diagnostic and therapeutic strategies.

Over the last decade, microarray technologies and bioinformatic analysis have been widely used to detect comprehensive mRNA expression levels, which have assisted in identifying the differentially expressed genes (DEGs) and functional pathways involved in the tumorigenesis and progression of ACC. However, because of the rarity of this tumor, there has been a problem in identifying potential markers to differentiate ACC from other renal neoplasms, and thus guiding potential treatment strategy. In the present study, three mRNA microarray datasets were downloaded from GEO database and analyzed to obtain DEGs between cancer tissues and adjacent normal tissues. Subsequently, functional pathway enrichment analyses were implemented to further understand the molecular mechanisms underlying carcinogenesis. The protein–protein interaction (PPI) network reveals the functions of all proteins and the importance of these interactions with regards to biological processes, molecular functions, and signal transduction (Sharan et al., 2007; Wu et al., 2009; Bapat et al., 2010). This may provide insights into the mechanisms of generation or development of diseases.

To investigate candidate biomarkers in tumor tissue and to define their value in ACC patients, this work focuses on analyzing the gene expression profiles, revealing the underlying biological interaction networks and assessing their prognostic value. We hypothesize that the oncogenic activity of significant hub genes correlates with poor prognosis, and might reveal potential prognostic markers and therapeutic targets for ACC.

Abbreviations: ACC, adrenocortical carcinoma; DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein-protein interaction; TCGA, the Cancer Genome Atlas; HPF, high power field; DFS, disease-free survival; OS, overall survival; HR, hazard ratio; CI, confidence interval; ROC, receiver operating characteristic curve; AUC, area under curve; GSEA, Gene set enrichment analysis.

MATERIALS AND METHODS

Raw Biological Microarray Data

The raw DNA microarray data were obtained from GEO (http://www.ncbi.nlm.nih.gov/geo) (Edgar et al., 2002) for patients with ACC. Corresponding genes converted into the probes were converted into symbols according to the annotation information in the platform. Three chip data sets GSE14922, GSE19750 and GSE90713 (4 normal and 4 ACC samples in GSE14922, 4 normal and 44 ACC samples in GSE19750, and 5 normal and 58 ACC samples in GSE90713) were downloaded from GEO (Agilent GPL6480 platform, Affymetrix GPL570 platform and Affymetrix GPL15270 platform, respectively).

Normalization and Elucidation of DEGs

DNA microarray analysis begins with preprocessing and normalization of raw biological data. This process removes noise from the biological data and ensures its integrity. Next, background correction of probe data, normalization, and summarization were executed by robust multi-array average analysis algorithm17 in affy package of R.

The DEGs between ACC and non-cancerous samples were screened and identified across experimental conditions. Delineating parameters such as adjusted P-values (adj. P), Benjamini and Hochberg false discovery rate (FDR) and fold change were utilized for filtering of DEGs and applied to provide a balance between discovery of statistically significant genes and limitations of false-positives. Probe sets without corresponding gene symbols or genes with more than one probe set were removed or averaged. Log_2FC (fold change) > 1 and adj. P-value <0.01 were considered statistically significant.

Functional Enrichment of DEGs

Discerning the role of DEGs in ACC, biological attributes including biological processes (BP), molecular functions (MF), and cellular components (CC) were extracted from Gene Ontology (GO) enrichment analysis (Ashburner et al., 2000). Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016) is a database resource for understanding high-level functions and biological systems from large-scale molecular datasets generated by high-throughput experimental technologies. The online Database for Annotation, Visualization

and Integrated Discovery (DAVID; https://david-d.ncifcrf.gov/summary.jsp Version 6.8) was used to explore the role of development-related signaling pathways in ACC (Huang et al., 2007). P-value < 0.05 was considered statistically significant. GO enrichment was analyzed and displayed using a bubble chart.

PPI Network Construction and Module Analysis

In the present study, the Search Tool for the Retrieval of Interacting Genes (STRING; http://string-db.org) (version 10.0) online database was used to predict PPI network of DEGs and analyze the functional interactions between proteins (Franceschini et al., 2013). An interaction with a combined score >0.4 was considered statistically significant.

Cytoscape (version 3.5), an open source bioinformatics software platform, was used to visualize molecular interaction networks (Smoot et al., 2011). Molecular Complex Detection (MCODE) (version 1.4.2) is a plug-in for Cytoscape used for clustering a given network based on topology to find densely connected regions (Bandettini et al., 2012). MCODE could identify the most significant module in the PPI networks with selection as follows: MCODE scores >5, degree cut-off = 2, node score cut-off = 0.2, Max depth = 100 and k-score = 2. Subsequently, the KEGG and GO analyses for genes in this module were performed using DAVID.

Hub Genes Selection and Analysis

The hub nodes of network with connectivity degrees >10 were identified. A network of the 24 genes and their co-expression genes was analyzed using cBioPortal (http://www.cbioportal.org) online platform (Cerami et al., 2012). ClueGO is a Cytoscape plug-in that visualizes the non-redundant biological terms for large clusters of genes in a functionally grouped network (Bindea et al., 2009). The biological process from GO and KEGG pathway analysis of hub genes was performed and visualized using ClueGO (version 2.5.3) and CluePedia (version 1.5.3), a functional extension of ClueGO, plug-in of Cytoscape (Bindea et al., 2013). Potential coexpression relationship between the 24 hub genes and possible prognostic value are shown in a heat map.

Statistical Analysis

Phenotype and expression profiles of hub genes in 76 ACC patients from TCGA were analyzed and displayed. Clinical and pathological parameters of the cohort were summarized. Expression of hub genes was respectively identified as binary variables (high vs. low) referring to median expression of each hub gene in the TCGA cohort. Then, a LASSO Cox regression model was constructed to find independent prognostic factors. The significant hub gene expression profiles of common neoplasm were analyzed and displayed using Oncomine online database (http://www.oncomine.com) (Giordano et al., 2003; Giordano et al., 2009).

The Kaplan–Meier method was applied to analyze survival differences between groups. The primary end point was overall survival (OS) for patients, which was evaluated from the date of first therapy to the date of death or last follow-up. Disease-free survival (DFS), as the secondary end point, was the length of time from the initiation of curative treatment to the date of progression or the start date of a second-line treatment or the date of death, whichever occurred first. The follow-up duration was estimated using the Kaplan-Meier method with 95% confidence intervals (95%CI) and log-rank test in separate curves. Univariate analyses were performed with Cox logistic regression models to find independent variables, including age at diagnosis, gender, laterality, TNM stage, pathologic stage, mitotic rate, invasion of tumor capsule, sinusoid invasion, necrosis, Weiss score, new tumor event after first treatment and integrated expression score. Parameters with P-value less than 0.1 were enrolled in multivariate Cox regression analyses of DFS and OS in "Back-LR" method. Integrated score was identified as the sum of the weight of each significant hub gene. X-tile software was utilized to take the cut-off value. All hypothetical tests were two-sided and P-values less than 0.05 were considered significant in all tests. The receiver operating characteristic curve (ROC) was constructed by predicting the probability of a diagnosis being of high or low integrated score of significant hub gene expression. Area under curve (AUC) analysis was performed to determine the diagnostic ability.

Sensitivity Analysis of Chip Datasets

In this study, GSE14922 only contains 4 ACC patients and 4 normal patients, while datasets 2 and 3 have 44 and 58 ACC samples respectively. To avoid penitential bias and see what other significant genes may have been missed, the analysis was re-run without GSE14922. Prognostic values of DEGs were then also tested against the TCGA validation cohort.

Data Processing of Gene Set Enrichment Analysis (GSEA)

TCGA database was implemented with the GSEA method using the Category version 2.10.1 package. For each separate analysis, Student's-t-test statistical score was performed in consistent pathways, and the mean of the differential expression genes was calculated. A permutation test of 1000 times was used to identify the significantly changed pathways. The adjusted P values (adj. P) using Benjamini and Hochberg (BH) false discovery rate (FDR) method by default were applied to correct the occurrence of false positive results (Subramanian et al., 2005). The significant related genes were defined with an adj. P less than 0.01 and FDR less than 0.25. Statistical analysis and graphical plotting were conducted using R software (Version 3.3.2).

RESULTS

This study consisted of three stages. In the first stage, we assessed DEGs using three datasets hosted on the GEO platform. In the second stage, coexpression, functional annotation of hub genes and patient survival analysis were carried out. In the third stage,

the most significant hub genes were selected, evaluated and integrated to predict their prognostic value.

Identification of DEGs in ACC

After standardization and identification of the microarray results, the DEGs (1,804 probe samples with 1,539 DEGs in GSE14922, 2,454 probe samples with 2,040 DEGs in GSE19750 and 1,216 probe samples with 806 DEGs in GSE90713) were determined to be significant based on the analysis and the statistical parameters of the data processing steps. The overlap among the three datasets included 150 significant genes and is displayed in the Venn diagram in **Figure 1A**.

GO and KEGG Enrichment Assessment of DEGs

To analyze the biological classification of the DEGs, functional and pathway enrichment analyses were performed using DAVID. As shown in **Supplementary Figure 1**, gene ontology (GO) analysis indicated that changes in the biological processes of the DEGs were significantly associated with the mitotic cell cycle, cell cycle process, movement of cells or subcellular components and cell locomotion activity. Changes in molecular function were mostly enriched in growth factor binding, kinase activity, extracellular matrix structure constituents and insulin-like growth factor binding. Changes in

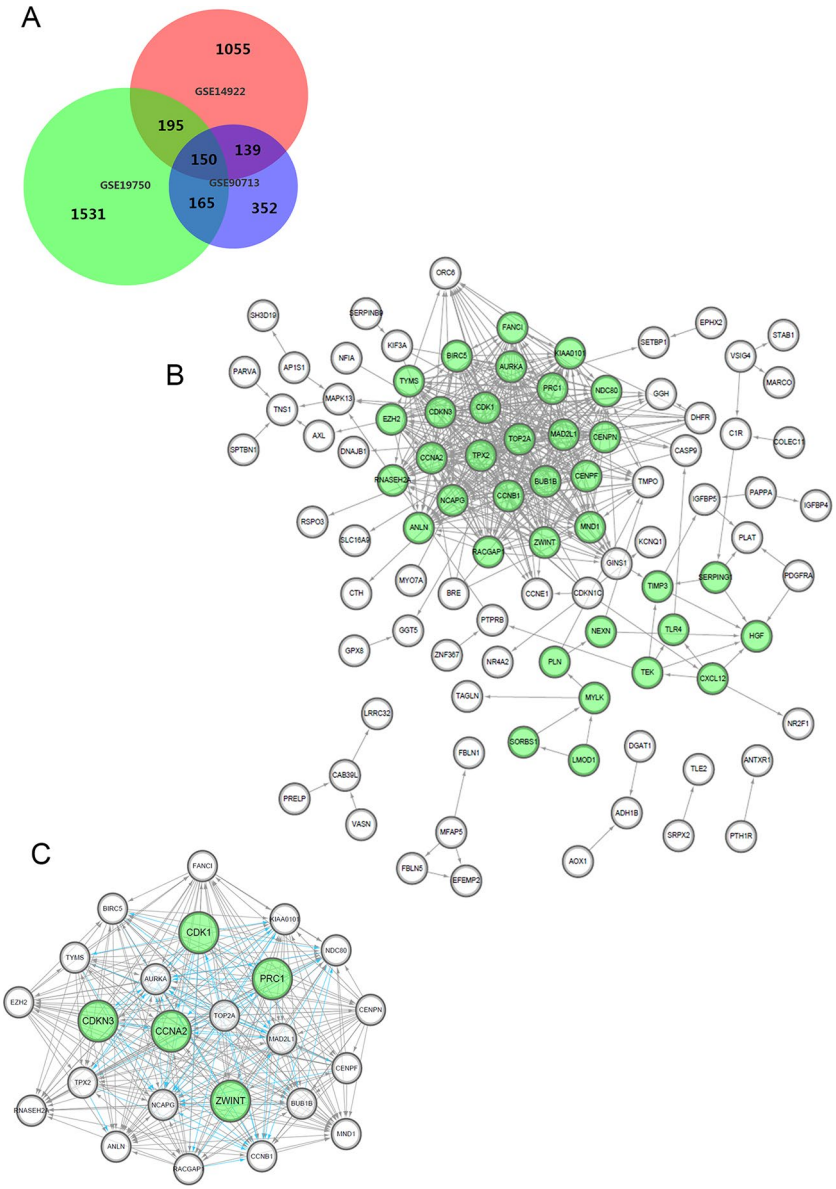

FIGURE 1 | Venn diagram, PPI network and the most significant module of DEGs. **(A)** DEGs were selected with a fold change >2 and P-value <0.01 among the mRNA expression profiling chip datasets GSE14922, GSE19750 and GSE90713. The 3 datasets show an overlap of 150 genes in the Venn diagram. **(B)** The PPI network of DEGs was constructed using Cytoscape. **(C)** The most significant module was obtained from PPI network with 24 nodes. Significant edges are marked in light blue with a K-score >0.800.

cellular components were mainly enriched in the chromosome, centromeric region, extracellular region, actin cytoskeleton and mitotic spindle. KEGG pathway analysis revealed that the DEGs were mainly enriched in cell cycle, progesterone-mediated oocyte maturation, oocyte meiosis, arachidonic acid metabolism and the p53 signaling pathway, summarized in **Table 1**.

PPI Network Establishment and Module Analysis

We constructed the PPI network of the DEGs (**Figure 1B**) and subsequently found the most significant module penal using a Cytoscape plugin (**Figure 1C**). The enrichment profiles from DAVID functional analyses of the 24 hub genes suggested that the hub genes in this module were primarily enriched in cell cycle phase, M phase, the mitotic cell cycle and mitosis (**Table 2**).

Hub Gene Selection and Analysis

After statistical selection, the significant hub nodes of the network included RACGAP1, AURKA, KIAA0101, MAD2L1, ZEH2, CCNB1, BIRC5, ZWINT, NDC80, NCAPG, TOP2A, PRC1, CENPF, CENPN, FANCI, CDKN3, MND1, RNASEH2A, TYMS, CDK1, BUB1B, CCNA2, TPX2 and ANLN. A visual network of the 24 genes and their coexpressed genes was set up (**Figure 2A**). The GO biological processes and KEGG functional annotation analysis of the hub genes are shown in **Figure 2B**. The detailed functional notes and classification pie charts are provided in the **Supplementary Figure 2**. Of the GO biological processes, 66.67% of terms belonged to the mitotic cell cycle checkpoint, 15.79% to mitotic spindle organization, 12.28% to anaphase-promoting complex-dependent catabolic processes, 3.51% to protein localization to kinetochore, and 1.75% to chromosome condensation. A heat map shows that a potential coexpression relationship may exist between the 24 hub genes, which could suggest they have value for prognostic prediction (**Figure 2C**).

TABLE 1 | KEGG pathways enrichment analysis of DEGs in ACC samples.

Term	Description	Count in gene set	P value
Has04110	Cell cycle	7	7.01E-04
Has04914	Progesterone-mediated oocyte maturation	5	6.05E-03
Has04114	Oocyte meiosis	5	0.01757
Has00590	Arachidonic acid metabolism	4	0.01758
Has04115	p53 signaling pathway	4	0.02312
Has05133	Pertussis	4	0.02667
Has06161	Hepatitis B	5	0.04010
Has00380	Tryptophan metabolism	3	0.04228

KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes; ACC, adrenocortical carcinoma.

TABLE 2 | GO and KEGG pathways enrichment analysis of DEGs in the most significant module.

Term	Description	Count in gene set	P value
GO:0022403	Cell cycle phase	16	6.836E-19
GO:0022402	Cell cycle process	17	1.154E-18
GO:0000279	M phase	15	1.898E-18
GO:0000278	Mitotic cell cycle	15	9.924E-18
GO:0007067	Mitosis	13	6.478E-17
GO:0000280	Nuclear division	13	6.477E-17
GO:0000087	M phase of mitotic cell cycle	13	8.063E-17
GO:0005819	Spindle	9	4.531E-11
GO:0015630	Microtubule cytoskeleton	11	4.532E-9
GO:0005694	Chromosome	10	3.922E-8
hsa04110	Cell cycle	5	2.268E-5
hsa04914	Progesterone-mediated oocyte maturation	4	2.461E-4
hsa04114	Oocyte meiosis	4	5.097E-4

GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes.

Clinicopathological Statistical Analysis

The clinical and pathological parameters from phenotype and expression profiles of the hub genes in 76 ACC patients from The Cancer Genome Atlas (TCGA) are summarized in **Table 3.** Each hub gene was classified into dichotomous variables according to the median expression in the analysis. Subsequently, the univariate survival analysis of the hub genes was performed using a Kaplan–Meier curve. Apart from *MND1*, ACC patients with elevated expression of the other 23 hub genes showed significantly worse OS and DFS (**Supplementary Figure 3**). LASSO Cox regression suggested that *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2* are significant weighted prognostic factors, and that an integrated gene panel may serve as an independent penal in ACC samples. The five significant hub gene expression profiles showed significantly elevated expression in tumor tissues compared with the corresponding normal tissues (**Figures 3A–E**). In addition, differential analysis from the ONCOMINE online database of tumor and normal tissue in two cohorts indicated that *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2* were highly expressed in ACC samples (**Figure 3F**). Elevated expression patterns were significantly associated with distant metastasis, necrosis, Weiss score and mitotic rate of >5 mitoses per 50 high power fields (HPF), plotted in **Figure 4**.

Cox Regression Analyses and Survival Outcomes of the Cohorts

In this study, the integrated expression score was identified as the sum of the weight of each binary gene expression. In univariate models, traditional prognostic factors, specifically T stage, M stage and pathologic stage, were significantly correlated with DFS ($P < 0.001$) and OS ($P < 0.001$) in ACC patients. Importantly, in univariate Cox regression analyses of DFS, subgroups of integrated expression score (High vs. Low) showed that integrated gene expression amplification significantly correlated with poor DFS ($P < 0.001$) for ACC patients. In addition, mitotic rate (≤5/50 HPF vs. >5/50 HPF) ($P = 0.022$), necrosis (**Present** vs. **Absent**) ($P = 0.011$), Weiss score(≤3 vs. >3) ($P = 0.008$) and new tumor event (**Present** vs. **Absent**) ($P < 0.001$) were correlated with poor DFS. In univariate Cox regression analyses of OS, invasion of tumor capsule (**Present** vs. **Absent**) ($P = 0.013$), sinusoid invasion (**Present** vs. **Absent**) ($P = 0.011$), necrosis (**Present** vs. **Absent**) ($P = 0.026$) and new tumor event

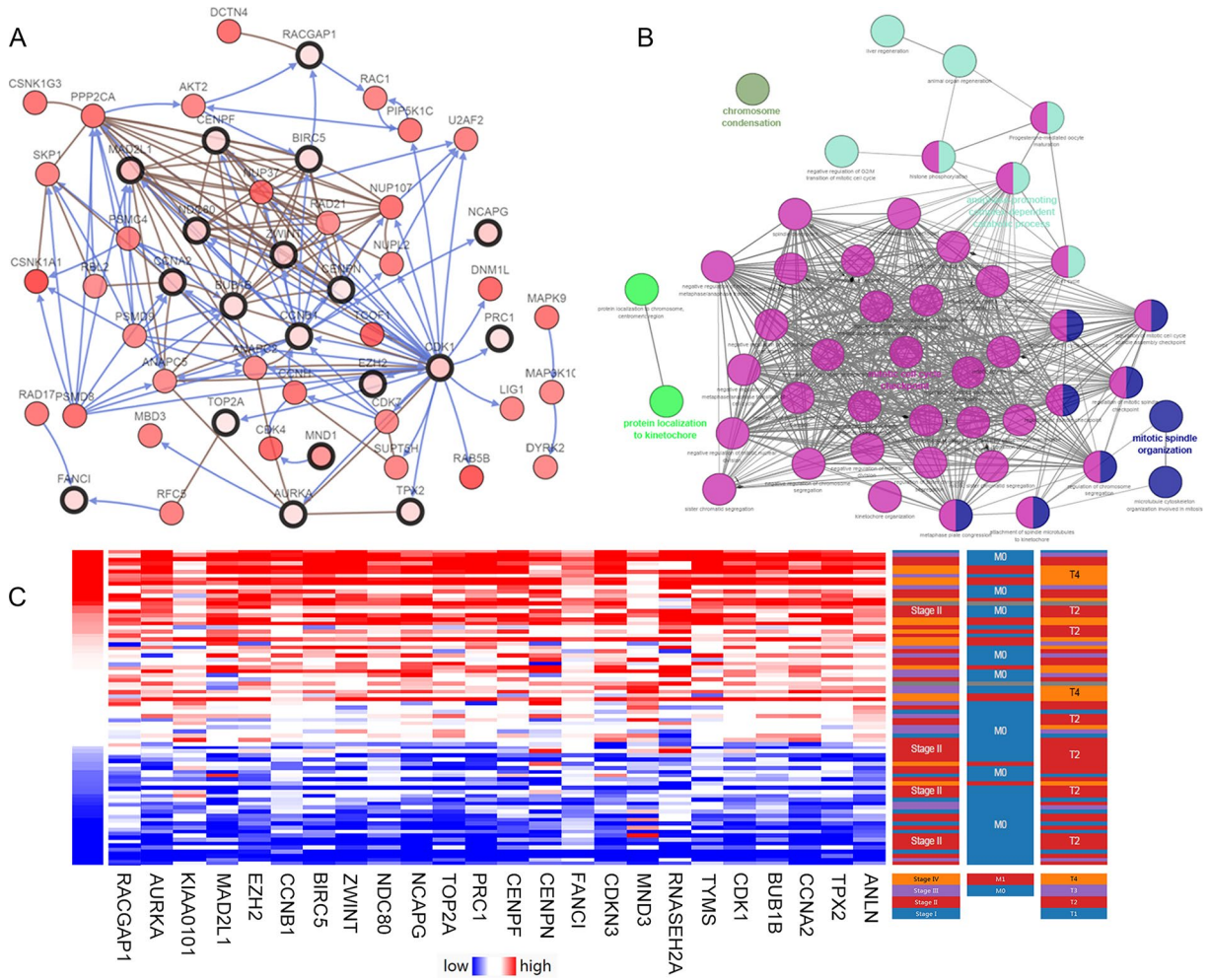

FIGURE 2 | Interaction network and biological process analysis of the hub genes. **(A)** Hub genes and their co-expression network were analyzed using cBioPortal. Nodes with bold black outline represent hub genes. Nodes with thin black outline represent the co-expression genes. **(B)** The biological process analysis of hub genes was constructed using ClueGO. Different color of nodes refers to the functional annotation of ontologies. Corrected P-value < 0.01 was considered statistically significant. **(C)** Hierarchical partitioning of 24 hub genes was obtained from DNA microarrays. It represent the level of expression of 24 genes across a number of comparable samples with high expression samples marked in red and low in blue.

(**Present** vs. **Absent**) ($P < 0.001$) were associated with shorter OS. However, in multivariate prognostic analysis, new tumor event after first treatment ($P < 0.001$) and integrated expression score ($P = 0.019$) were statistically significant parameters in predicting DFS (**Table 4**). Age at diagnosis ($P = 0.003$), M stage ($P = 0.033$), new tumor event after first treatment ($P = 0.013$) and integrated expression score ($P < 0.001$) were significantly associated with shorter OS (**Table 5**).

*Multivariate Cox regression analyses of OS in 76 enrolled ACC patients was run in "Back-LR" method. After integrating all the significant gene expression profiles in the Cox regression models, the Kaplan–Meier method was used to determine the significant survival outcomes (DFS: $P < 0.0001$; OS: $P < 0.0001$), shown in **Figures 5A, B**. Meanwhile, ROC curves were generated to validate the ability of the logistic model to predict prognosis. The AUC index for the integrated gene scores was 0.861 ($P < 0.0001$) (**Figure 5C**).

Sensitivity Analysis of Chip Datasets

To avoid penitential bias and see what other significant genes may have been missed, two datasets GSE19750 and GSE90713 were enrolled to re-run the analysis. The overlap among the two datasets, which includes 315 differential expressed genes (DEGs), is displayed in the Venn diagram (**Supplementary Figure 4A**). A PPI network of the new DEGs was constructed in **Supplementary Figure 4B**. Subsequently, we selected the most significant module penal using M-CODE, a plug-in of Cytoscape, and found 31 hub genes including *NDC80, MND1, MAD2L1, UBE2C, NCAPG, GINS1, CENPN, CDKN3, CCNA2, ZWINT, BIRC5, KIAA0101, TOP2A, BUB1B, CCNB1, AURKA, SMC2, ATAD2, PRC1, TPX2, CDK1, RACGAP1, TYMS, ANLN, PRIM1, NUSAP1, CENPF, SPAG5, SMC4, EZH2, FANCI*. Interestingly, five significant DEGs we have focused on (*ZWINT, PRC1, CDKN3, CDK1, CCNA2*) still consist of this new module penal (**Supplementary Figure 4C**), indicating a good stability of our molecular model. Eight different

TABLE 3 | Clinicopathologic characteristics of 76 ACC patients from TCGA database.

Characteristics	Entire cohort (N = 76)
N (%)	
Age, years	
≤57	54 (71.1)
>58	22 (28.9)
Gender	
Male	30 (39.5)
Female	46 (60.5)
Germline testing performed	
Present	12 (18.2)
Absent	54 (81.1)
Laterality	
Left	42 (55.3)
Right	34 (44.7)
pT stage	
T1 – T2	50 (65.8)
T3 – T4	26 (34.2)
pN stage	
N0	68 (89.5)
N1	8 (10.5)
M stage	
M0	62 (81.6)
M1	14 (18.4)
Pathologic stage	
I – II	46 (60.5)
III – IV	30 (39.5)
Histological type	
Myxiod	1 (1.3)
Oncocytic	3 (3.9)
Usual	72 (94.7)
Mitotic rate	
≤5/50 HPF	26 (38.8)
>5/50 HPF	41 (61.2)
Invasion of tumor capsule	
Absent	30 (42.9)
Present	40 (57.1)
Sinusoid invasion	
Absent	34 (58.6)
Present	24 (41.4)
Necrosis	
Absent	17 (23.9)
Present	54 (76.1)
Weiss score	
≤3	17 (22.0)
>4	47 (78.0)
Persistent distant metastasis	
Absent	56 (73.7)
Present	20 (26.3)

ACC, adrenocortical carcinoma; TCGA, the Cancer Genome Atlas; HPF, high power field.

DEGs are found different from these in three-chipset study, including *UBE2C, GINS1, SMC2, ATAD2, PRIM1, NUSAP1, SPAG5, SMC4*. Kaplan-Meier method was used to analyze mRNA expression level of 8 hub genes in TCGA cohort, which also showed statistically significant correlation with progressive progression and poor prognosis (**Supplementary Figure 5**).

Significant Genes and Pathways Obtained by GSEA

A total of 100 significant genes were obtained by gene set enrichment analysis (GSEA) with positive and negative correlation. Importantly, GSEA was used to perform hallmark analysis for *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2*. This suggested that the most involved significant pathways included mitotic spindle, G2M checkpoint and E2F targets. The details are shown in **Figure 6**.

DISCUSSION

Adrenocortical carcinoma (ACC) is a rare but aggressive cancer, with a typically high incidence in children with a *TP53* germline mutation (Fassnacht et al., 2013). The Wnt/β-catenin pathway and IGF-2 signaling have been confirmed as altered signaling pathways in ACC patients, while increasing data indicate that the available evidence is inadequate for malignant phenotype and poor prognosis (Berthon et al., 2010; Heaton et al., 2012), especially for the diagnosis of low-grade ACC confined to the adrenal gland (Mete et al., 2018). Although there are diagnostic and prognostic molecular tests for ACC such as the IGF-2, Ki-67, p53, BUB1B, PBK, HURP, NEK2, DAX, Wnt/β-catenin and PI3K signaling pathways, they remain largely unutilized in morphologic assessment coupled with ancillary diagnostic and prognostic modeling of ACC (Mete et al., 2018). Therefore, the major molecular mechanisms in the pathogenesis and progression are poorly understood. In 2003 and 2009, Giordano et al. performed unsupervised cluster analyses of transcriptome data to identify subgroups with different prognoses (Giordano et al., 2003; Giordano et al., 2009). These two studies laid the foundation for the molecular classification and prognostication of adrenocortical tumors and also provided a rich source of potential diagnostic and prognostic markers. Still, most cases of ACC were initially diagnosed with highly aggressive progression but were not candidates for curative therapies. Hence, potential biomarkers for diagnosis and treatment with high efficiency are urgently demanded.

Currently, microarray technology enables comprehensive mRNA expression profiling in ACC and can identify and investigate new biomarkers involved in tumorigenesis. A total of 150 DEGs and 24 hub genes were identified by microarray data analysis. GO and KEGG enrichment analysis showed association to the cell cycle, especially mitotic cycle checkpoint, mitotic spindle and oocyte meiosis, which was the most significant annotated function. Furthermore, among the 24 hub genes, the most significant molecular prognostic model integrated *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2*. Importantly, after reintegrating the weight of each gene, the new score was statistically the most significant parameter in both univariate and multivariate regression analysis. The gene set enrichment analysis (GSEA) method was used to visualize the significant signaling pathway analysis of *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2*.

ZW10 interactor (*ZWINT*), an interactor with ZW10, plays a vital role in rectifying incorrect centromere-microtubule attachment and regulating the miotic spindle checkpoint (Starr et al., 2000). Increased expression of *ZWINT* correlates with poor outcomes in human malignancies, including prostate, ovarian, bladder and lung cancers (Bhattacharjee et al., 2001; Endoh et al., 2004; Urbanucci et al., 2012; Xu et al., 2016). These new findings encourage further investigation of the potential clinical

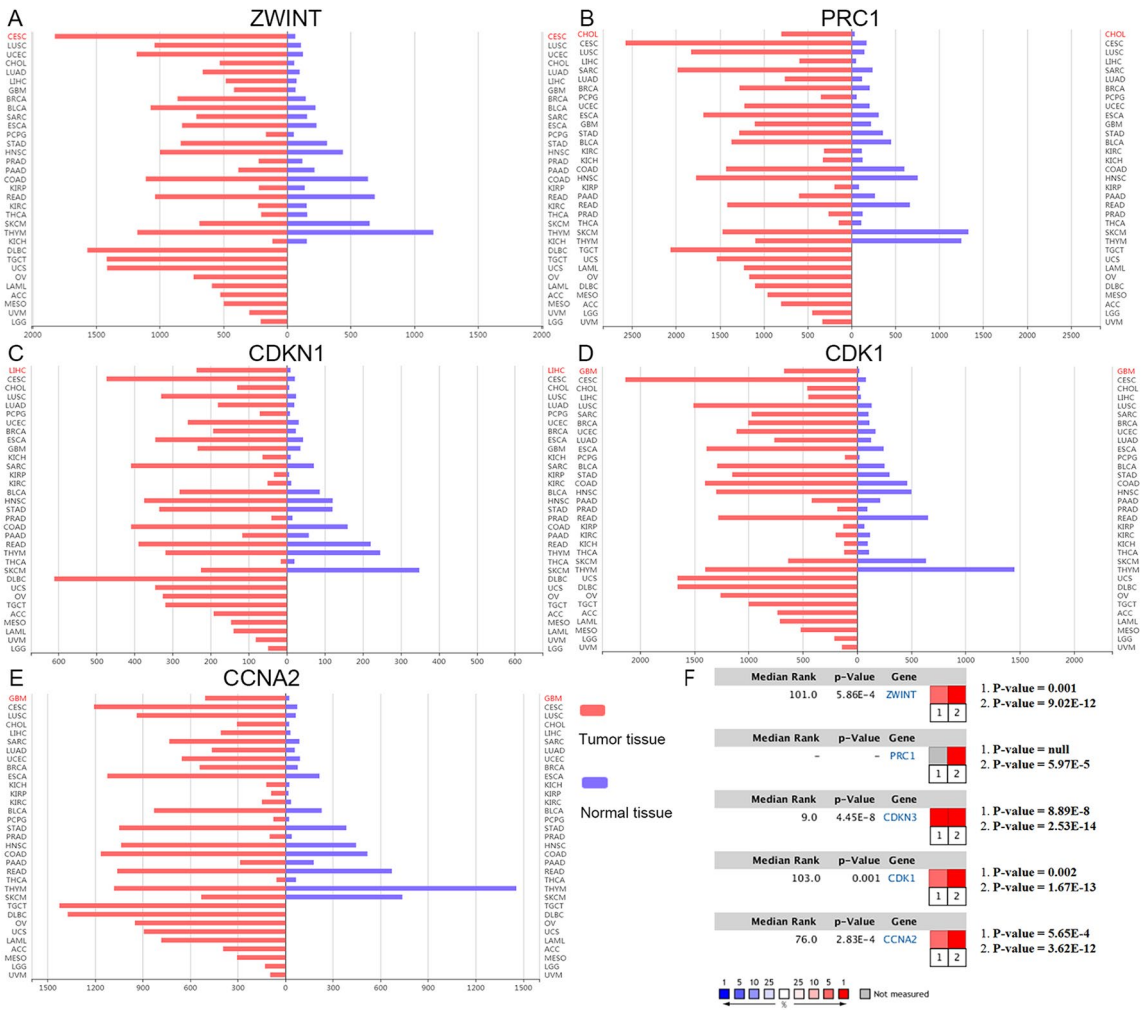

FIGURE 3 | Oncomine expression analysis of cancer vs. normal tissue. **(A–E)** Histogram of *ZWINT, PRC1, CDKN1, CDK1* and *CCNA2* gene expression in total tumor spectrum samples vs. normal tissues. **(F)** Documented expression of significant molecular score in ACC samples. 1. Giordano, T.J., et al., Distinct transcriptional profiles of adrenocortical tumors uncovered by DNA microarray analysis. Am J Pathol, 2003 (Giordano et al., 2003). 2. Giordano, T.J., et al., Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. Clin Cancer Res, 2009 (Giordano et al., 2009).

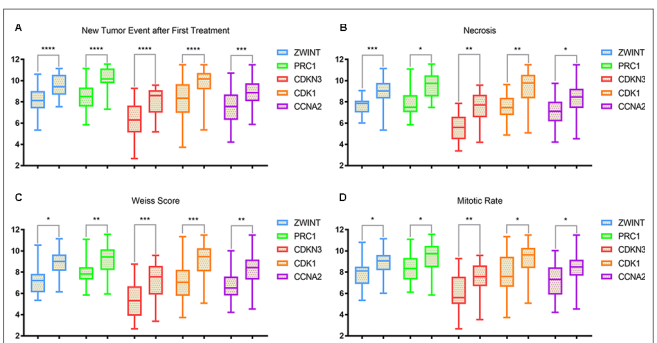

FIGURE 4 | Elevated expression patterns (*ZWINT* presents in light blue, PRC1 in light green, *CDKN3* in maroon, *CDK1* in orange and *CCNA2* in purple) were significantly associated with **(A)** new tumor events after first treatments, **(B)** necrosis, **(C)** Weiss score and **(D)** mitotic rate >50 high power field (HPF).

significance in human malignancies, yet the prognostic value of *ZWINT* in ACC has rarely been reported.

Protein Regulator of cytokinesis 1 (*PRC1*) protein is located in the nucleus. It is highly expressed in S and G2/M phases and shows an obvious drop in the G1 phase of the cell cycle (Freedland et al., 2013). During anaphase, it dynamically locates with the mitotic spindle and localizes to the cell midbody (Wu et al., 2018). Increasing evidence suggests that PRC1 may be involved in a cancer-specific manner, because of its negative correlation with p53 and overexpression in p53-defective cells *in vitro* (Li et al., 2004). In addition, Chen et al. and Zhan et al. demonstrated that *PRC1* contributes to tumorigenesis by regulating the Wnt/β-catenin signaling pathway in a positive feedback loop (Chen et al., 2016; Zhan et al., 2017), in which carcinogenesis and progression may feasibly be mediated in ACC.

TABLE 4 | Univariate and multivariate Cox regression analyses of DFS in 76 enrolled ACC patients.

	Univariate analysis		Multivariate analysis	
Covariates	**HR (95%CI)**	**P value**	**HR (95%CI)**	**P value**
Age at diagnosis (≤57 years vs. >58 years)	1.703 (0.819 – 3.541)	0.154		
Gender (male vs. female)	0.977 (0.477 – 2.002)	0.950		
Laterality (left vs. right)	0.771 (0.381 – 1.563)	0.471		
T stage (T1-T2 vs. T3-T4)	3.846 (1.841 – 8.034)	**<0.001**		
N stage (N0 vs. N1)	2.151 (0.820 – 5.641)	0.119		
M stage (M0 vs. M1)	3.104 (1.471 – 6.546)	**0.003**	2.193 (0.977 – 4.921)	0.057
Pathologic stage (I - II vs. III - IV)	3.937 (1.853 – 8.364)	**<0.001**		
Mitotic rate (≤5/50 HPF vs. >5/50 HPF)	2.851 (1.164 – 6.984)	**0.022**		
Invasion of tumor capsule (Present vs. Absent)	2.074 (0.965 – 4.455)	0.062		
Sinusoid invasion (Present vs. Absent)	1.516 (0.678 – 3.389)	0.311		
Necrosis (Present vs. Absent)	6.501 (1.542 – 27.404)	**0.011**		
Weiss score (≤3 vs. >3)	2.816 (1.303 – 6.085)	**0.008**		
New tumor event (Present vs. Absent)	16.642 (5.673 – 48.822)	**<0.001**	9.041 (2.983 – 27.234)	**<0.001**
Integrated expression score (High vs. Low)	7.819 (3.569 – 17.114)	**<0.001**	2.767 (1.185 – 6.460)	**0.019**

DFS, disease-free survival; ACC, clear cell renal cell carcinoma; HR, hazard ratio; CI, confidence interval; HPF, high power field.
**Multivariate Cox regression analyses of DFS in 76 enrolled ACC patients was run in "Back-LR" method. Statistically significant is considered as P value less than 0.05, indicated in bold.*

TABLE 5 | Univariate and multivariate Cox regression analyses of OS in 76 enrolled ACC patients.

	Univariate analysis		Multivariate analysis*	
Covariates	**HR (95%CI)**	**P value**	**HR (95%CI)**	**P value**
Age at diagnosis (≤57 years vs. >58 years)	1.957 (0.913 – 4.196)	0.085	4.959 (1.744 – 14.098)	**0.003**
Gender (male vs. female)	0.996 (0.466 – 2.127)	0.991		
Laterality (left vs. right)	1.262 (0.591 – 2.695)	0.548		
T stage (T1-T2 vs. T3-T4)	10.693 (4.276 – 28.110)	**<0.001**		
N stage (N0 vs. N1)	0.451 (0.171 – 1.191)	0.108		
M stage (M0 vs. M1)	7.340 (3.300 – 16.327)	**<0.001**	3.045 (1.094 – 8.477)	**0.033**
Pathologic stage (I - II vs. III - IV)	7.157 (3.023 – 16.941)	**<0.001**		
Mitotic rate (≤5/50 HPF vs. >5/50 HPF)	1.708 (0.743 – 3.926)	0.208		
Invasion of tumor capsule (Present vs. Absent)	3.015 (1.257 – 7.231)	**0.013**		
Sinusoid invasion (Present vs. Absent)	3.069 (1.297 – 7.262)	**0.011**		
Necrosis (Present vs. Absent)	5.135 (1.214 – 21.725)	**0.026**		
Weiss score (≤3 vs. >3)	3.467 (1.136 – 10.577)	0.307		
New tumor event (Present vs. Absent)	5.833 (2.351 – 14.473)	**<0.001**	3.609 (1.305 – 9.980)	**0.013**
Integrated expression score (High vs. Low)	18.892 (6.830 – 52.255)	**<0.001**	18.719 (5.122 – 68.415)	**<0.001**

OS, overall survival; ACC, adrenocortical carcinoma; HR, hazard ratio; CI, confidence interval; HPF, high power field.
**Multivariate cox regression analyses of DFS in 76 enrolled ACC patients was run in "Back-LR" method. Statistically significant is considered as P value less than 0.05, indicated in bold.*

Cyclin-dependent kinase inhibitor 3 (*CDKN3*) is part of the dual-specificity protein phosphatase family that dephosphorylates CDK2/CDK1 kinase and other cytokines (Hannon et al., 1994). Interestingly, a relationship between elevated *CDKN3* expression and poor prognosis has been reported in many cancers by modulation of the cell cycle, mitotic spindle or p53 pathways (Berumen et al., 2014; Fan et al., 2015). A previous study has distinguished five genes modeling ACC using *TOP2A, NDC80, CEP55, CDKN3* and *CDK1*, which may be utilized to form a board of progressive and predictive biomarkers for ACC for clinical purpose (Xiao et al., 2018). Thus, it is inferred that *CDKN3* may be an oncogene in human ACC.

Cyclin dependent kinase 1 (*CDK1*) is a catalytic subunit of a highly conserved protein and is involved in many biological processes including cell cycle control, DNA damage repair, and checkpoint transcription (Skotheim et al., 2008; Enserink and Kolodner, 2010). *CDK1* plays an important regulatory role in the control of the eukaryotic cell cycle by modulating the centrosome cycle (Asghar et al., 2015). It has been previously reported that inhibition of *CDK1* could serve as a therapeutic target *via* microRNA-7 for ACC samples *in vivo* (Glover et al., 2015). Meanwhile, CDC2, sharing approximately 63% amino-acid homology with CDK1, was found to be dysregulated in the cell cycle or retinoic acid signaling pathway by meta-analysis of genomic profiling data of adrenocortical tumors (Szabo et al., 2010).

Cyclin-A2 (*CCNA2*) belongs to a highly conserved cyclin family whose members function as regulators of the cell cycle. This protein interacts with CDK2 during G1/S and in G2/M phase, therefore promoting cell cycle transition (Pagano et al., 1992). There is accumulating evidence suggesting a role for CCNA2 in tumorigenesis of human malignancies. Kim et al. identified an SNP (rs769236) at the *CCNA2* promoter

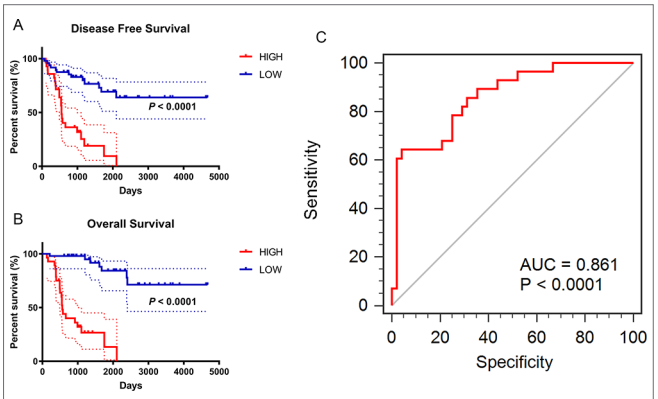

FIGURE 5 | Prognostic and diagnostic value of integrated significant molecular score in ACC samples after LASSO Cox regression. **(A–B)** Kaplan–Meier method was used to perform the significant survival outcomes (DFS: $P < 0.0001$; OS: $P < 0.0001$). **(C)** ROC curves of the integrated models were synchronously plotted to predict diagnosis probability. Red line represents integrated expression score with AUC of 0.861 ($P < 0.0001$).

that may be significantly associated with an increased risk of colon, liver and lung cancers (Kim et al., 2011). In addition, a significant delay in liver tumor formation was observed in mice with CCNA2-deficient hepatocytes (Gopinathan et al., 2014). As well as a prognostic value for *CDK1* in ACC (Xiao et al., 2018), the mitotic checkpoint regulator *CCNA2* may combine with other cell-cycle coding genes and be involved in aberrant regulation of the cell cycle network. It has not been evaluated whether this could be an effective approach to ACC treatment.

Our study represents the first attempt to construct a gene regulatory network incorporating DEGs and functional annotation of hub genes in ACC. An additional strength is that the alteration of *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2* is significantly associated with worse OS and DFS, indicating that these genes may play important roles in the aggressive malignant phenotypes of ACC. At the same time, several limitations of this study are as follows. First, the data utilized in the study consisted of unbalanced ACC and normal control samples, which were restricted in quantity and downloaded from the GEO database, not generated by new

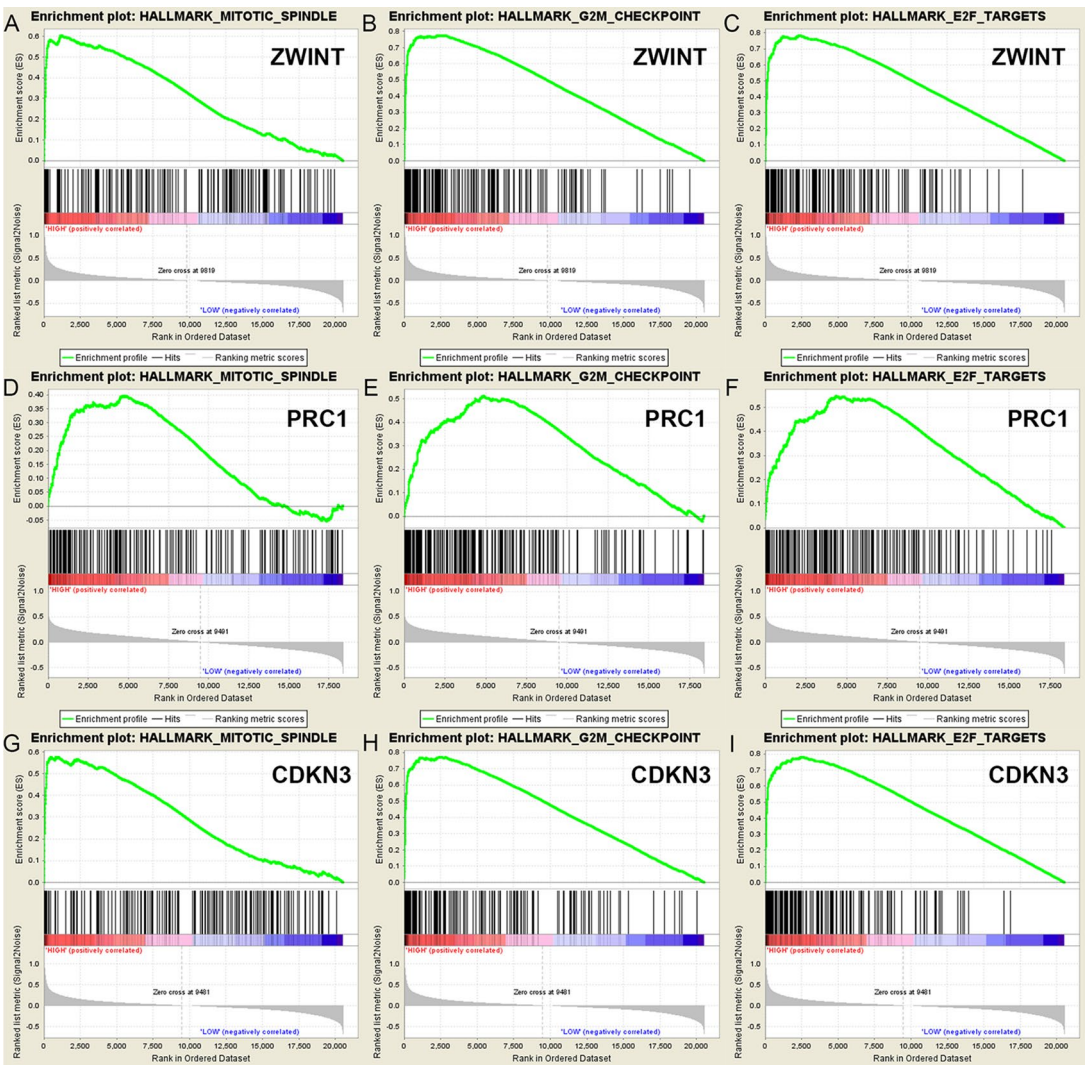

FIGURE 6 | Continued

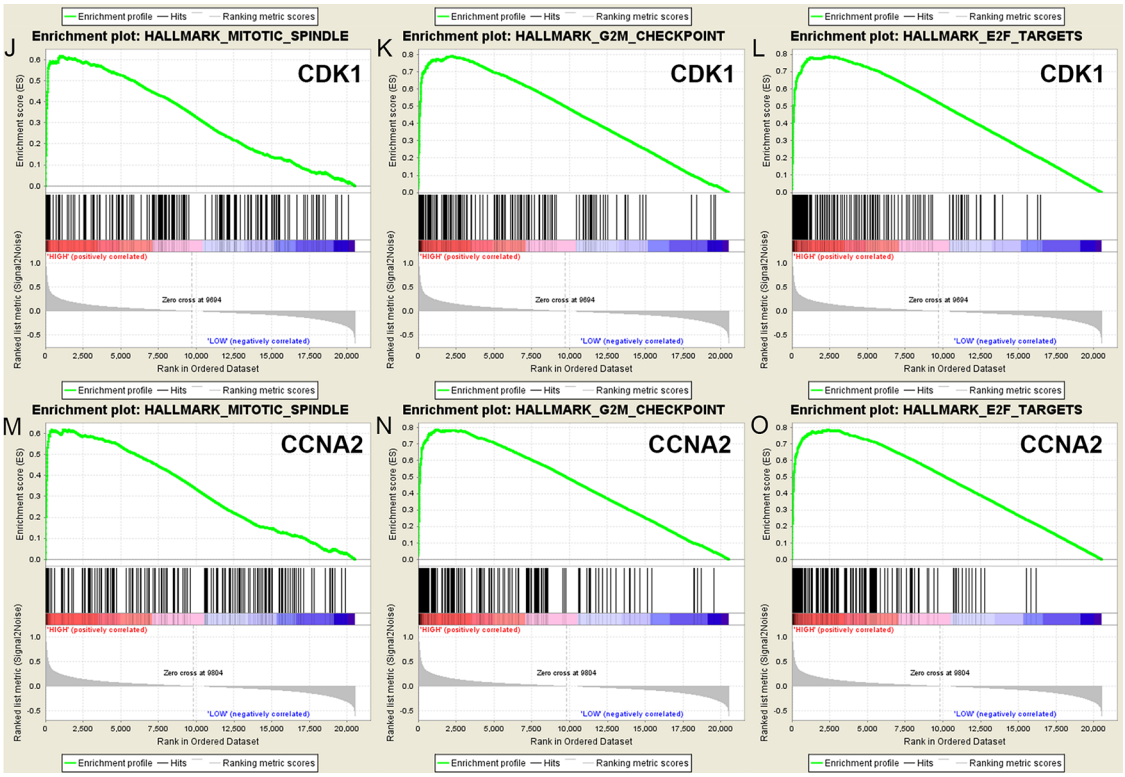

FIGURE 6 | A total of 100 significant genes were obtained from GSEA with positive and negative correlation. GSEA was used to perform hallmark analysis in *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2*, respectively. Results of GESA suggested that **(A–C)** *ZWINT*, **(D–F)** *PRC1*, **(G–I)** *CDKN3*, **(J–L)** *CDK1*, **(M–O)** *CCNA2* significantly involved in the same hallmarks pathways including mitotic spindle, G2M checkpoint and E2F targets.

DNA microarrays. Second, the microarray data contained relatively few ACC samples in the public database, and only 76 patients were enrolled from the TCGA cohort with corresponding transcriptome data. Third, prospective cohort was not used in this study. In addition, only the mRNA levels of hub genes are shown in this study, thus further functional works and validated cohorts are needed to verify these findings.

CONCLUSION

In conclusion, the present study identifies DEGs and hub genes that may be involved in poor prognosis and recurrence of ACC *in silico*. The transcriptional profiles of *ZWINT, PRC1, CDKN3, CDK1* and *CCNA2* are of prognostic value, and may assist in better understanding the underlying carcinogenesis or progression of ACC. Further studies are required to elucidate the molecular pathogenesis and alterations in signaling pathways of these genes in ACC.

ETHICS STATEMEMT

The ethics approval and consent to participate of the current study was approved and consented by the ethics committee of Fudan University Shanghai Cancer Center.

AUTHOR CONTRIBUTIONS

The work presented here was carried out in collaboration among all authors. D-WY, H-LZ, and Y-YQ defined the research theme, discussed analyses, interpretation and presentation. W-HX and JlW drafted the manuscript, analyzed the data, developed the algorithm and interpreted the results. JW co-worked on associated data collection, cohort validation and helped to draft the manuscript. F-NW, H-KW, and D-LC helped to perform the statistical analysis and reference collection. All authors read and approved the final manuscript.

FUNDING

This work is supported by Grants from the National Natural Science Foundation of China (No. 81202004, 81802525), Natural Science Foundation of Shanghai (No. 16ZR1406400), and Shanghai Sailing Program of China (No. 17YF1402700).

ACKNOWLEDGMENTS

We thank Catherine Perfect, MA (Cantab), from Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00821/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Functional and pathway enrichment analyses were performed using DAVID in bubble chart. **(A)** Changes in cellular components of DEGs were mainly enriched in the chromosome, centromeric region, extracellular region, actin cytoskeleton and mitotic spindle. **(B)** Changes in molecular funtions were mostly enriched in growth factor binding, kinase activity, extracellular matrix structure constituent and insulin-like grouth factor binding. **(C)** GO analysis results showed that changes in biological processes of DEGs were significantly enriched in mitotic cell cycle, cell cycle process, movement of cell or subcellular component and cell locomotion activity.

SUPPLEMENTARY FIGURE 2 | A network of the 24 genes and their co-expression genes was visualized and displayed in detail. **(A)** The biologic process and KEGG enrichment analysis of the hub genes were shown in different color. **(B)** The detailed functional notes and classification pie charts are listed as follows. 66.67% terms belong to mitotic cell cycle checkpoint, 15.79% to mitotic spindle organization, 12.28% to anaphase-promoting complex-dependent catabolic process, 3.51% to protein localization to kinetochore and 1.75% to chromosome consederation.

SUPPLEMENTARY FIGURE 3 | Univariate survival analysis of the hub genes was performed using Kaplan-Meier curve. Besides MND1, each elevated expression in 24 hub gene showed markedly significant worse OS and DFS in ACC samples ($P < 0.05$).

SUPPLEMENTARY FIGURE 4 | Sensitivity analyze of GSE19750 and GSE90713 with Venn diagram, PPI network and the most significant module of DEGs. **(A)** DEGs were selected with a fold change >2 and P-value <0.01 among the mRNA expression profiling chip datasets GSE19750 and GSE90713. The 2 datasets showed an overlap of 315 genes in Venn diagram. **(B)** The PPI network of DEGs was constructed using Cytoscape. **(C)** The most significant module was obtained from PPI network with 31 nodes including ZWINT, PRC1, CDKN3, CDK1, CCNA2. Significant edges are marked in light blue with a K-score >0.800.

SUPPLEMENTARY FIGURE 5 | Univariate survival analysis of hub genes from sensitivity validated datasets was performed using Kaplan-Meier curve. Eight different DEGs are found different from these in three-chipset study, including *UBE2C, GINS1, SMC2, ATAD2, PRIM1, NUSAP1, SPAG5, SMC4*. Kaplan-Meier method was used to analysis mRNA expression level of 8 hub genes in TCGA cohort, which also showed significant correlation between elevated expression and progressive progression or poor prognosis ($P < 0.05$).

REFERENCES

Asghar U., Witkiewicz A. K., Turner, N. C., and Knudsen, E. S. (2015). The history and future of targeting cyclin-dependent kinases in cancer therapy. *Nat. Rev. Drug Discov.* 14 (2), 130–146. doi: 10.1038/nrd4504

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Gene Ontol. Consort. Nat. Genet.* 25 (1), 25–29. doi: 10.1038/75556

Bandettini, W. P., Kellman, P., Mancini, C., Booker, O. J., Vasu, S., Leung, S. W., et al. (2012). MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. *J. Cardiovasc. Magn. Reson.* 14, 83. doi: 10.1186/1532-429X-14-83

Bapat, S. A., Krishnan, A., Ghanate, A. D., Kusumbe, A. P., and Kalra, R. S. (2010). Gene expression: protein interaction systems network modeling identifies transformation-associated molecules and pathways in ovarian cancer. *Cancer Res.* 70 (12), 4809–4819. doi: 10.1158/0008-5472.CAN-10-0447

Berthon, A., Sahut-Barnola, I., Lambert-Langlais, S., de Joussineau, C., Damon-Soubeyrand, C., Louiset, E., et al. (2010). Constitutive beta-catenin activation induces adrenal hyperplasia and promotes adrenal cancer development. *Hum. Mol. Genet.* 19 (8), 1561–1576. doi: 10.1093/hmg/ddq029

Berumen, J., Espinosa, A. M., and Medina, I. (2014). Targeting CDKN3 in cervical cancer. *Expert Opin. Ther. Targets* 18 (10), 1149–1162. doi: 10.1517/14728222.2014.941808

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U. S. A.* 98 (24), 13790–13795. doi: 10.1073/pnas.191502998

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25 (8), 1091–1093. doi: 10.1093/bioinformatics/btp101

Bindea, G., Galon, J., and Mlecnik, B. (2013). CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics* 29 (5), 661–663. doi: 10.1093/bioinformatics/btt019

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2 (5), 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chen, J., Rajasekaran, M., Xia, H., Zhang, X., Kong, S. N., Sekar, K., et al. (2016). The microtubule-associated protein PRC1 promotes early recurrence of hepatocellular carcinoma in association with the Wnt/beta-catenin signalling pathway. *Gut* 65 (9), 1522–1534. doi: 10.1136/gutjnl-2015-310625

de Fraipont, F., El Atifi, M., Cherradi, N., Le Moigne, G., Defaye, G., Houlgatte, R., et al. (2005). Gene expression profiling of human adrenocortical tumors using complementary deoxyribonucleic Acid microarrays identifies several candidate genes as markers of malignancy. *J. Clin. Endocrinol. Metab.* 90 (3), 1819–1829. doi: 10.1210/jc.2004-1075

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30 (1), 207–210. doi: 10.1093/nar/30.1.207

Else, T., Williams, A. R., Sabolch, A., Jolly, S., Miller, B. S., and Hammer, G. D. (2014). Adjuvant therapies and patient and tumor characteristics associated with survival of adult patients with adrenocortical carcinoma. *J. Clin. Endocrinol. Metab.* 99 (2), 455–461. doi: 10.1210/jc.2013-2856

Endoh, H., Tomida, S., Yatabe, Y., Konishi, H., Osada, H., Tajima, K., et al. (2004). Prognostic model of pulmonary adenocarcinoma by expression profiling of eight genes as determined by quantitative real-time reverse transcriptase polymerase chain reaction. *J. Clin. Oncol.* 22 (5), 811–819. doi: 10.1200/JCO.2004.04.109

Enserink, J. M., and Kolodner, R. D. (2010). An overview of Cdk1-controlled targets and processes. *Cell Div.* 5, 11. doi: 10.1186/1747-1028-5-11

Fan, C., Chen, L., Huang, Q., Shen, T., Welsh, E. A., Teer, J. K., et al. (2015). Overexpression of major CDKN3 transcripts is associated with poor survival in lung adenocarcinoma. *Br. J. Cancer* 113 (12), 1735–1743. doi: 10.1038/bjc.2015.378

Fassnacht, M., Kroiss, M., and Allolio, B. (2013). Update in adrenocortical carcinoma. *J. Clin. Endocrinol. Metab.* 98 (12), 4551–4564. doi: 10.1210/jc.2013-3020

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41 (Database issue), D808–D815. doi: 10.1093/nar/gks1094

Freedland, S. J., Gerber, L., Reid, J., Welbourn, W., Tikishvili, E., Park, J., et al. (2013). Prognostic utility of cell cycle progression score in men with prostate cancer after primary external beam radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* 86 (5), 848–853. doi: 10.1016/j.ijrobp.2013.04.043

Gaujoux, S., Grabar, S., Fassnacht, M., Ragazzon, B., Launay, P., Libé, R., et al. (2011). β-catenin activation is associated with specific clinical and pathologic characteristics and a poor outcome in adrenocortical carcinoma. *Clin. Cancer Res.* 17 (2), 328–336. doi: 10.1158/1078-0432.CCR-10-2006

Gicquel, C., Bertagna, X., Gaston, V., Coste, J., Louvel, A., Baudin, E., et al. (2001). Molecular markers and long-term recurrences in a large cohort of patients with sporadic adrenocortical tumors. *Cancer Res.* 61 (18), 6762–6767.

Giordano, T. J., Thomas, D. G., Kuick, R., Lizyness, M., Misek, D. E., Smith, A. L., et al. (2003). Distinct transcriptional profiles of adrenocortical tumors uncovered by DNA microarray analysis. *Am. J. Pathol.* 162 (2), 521–531. doi: 10.1016/S0002-9440(10)63846-1

Giordano, T. J., Kuick, R., Else, T., Gauger, P. G., Vinco, M., Bauersfeld, J., et al. (2009). Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. *Clin. Cancer Res.* 15 (2), 668–676. doi: 10.1158/1078-0432.CCR-08-1067

Glover, A. R., Zhao, J. T., Gill, A. J., Weiss, J., Mugridge, N., Kim, E., et al. (2015). MicroRNA-7 as a tumor suppressor and novel therapeutic for adrenocortical carcinoma. *Oncotarget* 6 (34), 36675–36688. doi: 10.18632/oncotarget.5383

Gopinathan, L., Tan, S. L., Padmakumar, V. C., Coppola, V., Tessarollo, L., and Kaldis, P. (2014). Loss of Cdk2 and cyclin A2 impairs cell proliferation and tumorigenesis. *Cancer Res.* 74 (14), 3870–3879. doi: 10.1158/0008-5472.CAN-13-3440

Hannon, G. J., Casso, D., and Beach, D. (1994). KAP: a dual specificity phosphatase that interacts with cyclin-dependent kinases. *Proc. Natl. Acad. Sci. U. S. A.* 91 (5), 1731–1735. doi: 10.1073/pnas.91.5.1731

Heaton, J. H., Wood, M. A., Kim, A. C., Lima, L. O., Barlaskar, F. M., Almeida, M. Q., et al. (2012). Progression to adrenocortical tumorigenesis in mice and humans through insulin-like growth factor 2 and beta-catenin. *Am. J. Pathol.* 181 (3), 1017–1033. doi: 10.1016/j.ajpath.2012.05.026

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 8 (9), R183. doi: 10.1186/gb-2007-8-9-r183

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 04, 44. doi: 10.1093/nar/gkv1070

Kebebew, E., Reiff, E., Duh, Q. Y., Clark, O. H., and McMillan, A. (2006). Extent of disease at presentation and outcome for adrenocortical carcinoma: have we made progress? *World J. Surg.* 30 (5), 872–878. doi: 10.1007/s00268-005-0329-x

Kerkhofs, T. M., Verhoeven, R. H., Van der Zwan, J. M., Dieleman, J., Kerstens, M. N., Links, T. P., et al. (2013). Adrenocortical carcinoma: a population-based study on incidence and survival in the Netherlands since 1993. *Eur. J. Cancer* 49 (11), 2579–2586. doi: 10.1016/j.ejca.2013.02.034

Kim, D. H., Park, S. E., Kim, M., Ji, Y. I., Kang, M. Y., Jung, E. H., et al. (2011). A functional single nucleotide polymorphism at the promoter region of cyclin A2 is associated with increased risk of colon, liver, and lung cancers. *Cancer* 117 (17), 4080–4091. doi: 10.1002/cncr.25930

Li, C., Lin, M., and Liu, J. (2004). Identification of PRC1 as the p53 target gene uncovers a novel function of p53 in the regulation of cytokinesis. *Oncogene* 23 (58), 9336–9347. doi: 10.1038/sj.onc.1208114

Mete, O., Gucer, H., Kefeli, M., and Asa, S. L. (2018). Diagnostic and prognostic biomarkers of Adrenal cortical carcinoma. *Am. J. Surg. Pathol.* 42 (2), 201–213. doi: 10.1097/PAS.0000000000000943

Mohan, D. R., Lerario, A. M., and Hammer, G. D. (2018). Therapeutic targets for adrenocortical carcinoma in the Genomics era. *J. Endocr. Soc.* 2 (11), 1259–1274. doi: 10.1210/js.2018-00197

Pagano, M., Pepperkok, R., Verde, F., Ansorge, W., and Draetta, G. (1992). Cyclin A is required at two points in the human cell cycle. *EMBO J.* 11 (3), 961–971. doi: 10.1002/j.1460-2075.1992.tb05135.x

Paragliola, R. M., Torino F., Papi, G., Locantore, P., Pontecorvi, A., and Corsello, S. M. (2018). Role of mitotane in Adrenocortical Carcinoma - Review and state of the art. *Eur. Endocrinol.* 14 (2), 62–66. doi: 10.17925/EE.2018.14.2.62

Poli, G., Guasti, D., Rapizzi, E., Fucci, R., Canu, L., Bandini, A., et al. (2013). Morphofunctional effects of mitotane on mitochondria in human adrenocortical cancer cells. *Endocr. Relat. Cancer* 20 (4), 537–550. doi: 10.1530/ERC-13-0150

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.* 3, 88. doi: 10.1038/msb4100129

Skotheim, J. M., Di Talia, S., Siggia, E. D., and Cross, F. R. (2008). Positive feedback of G1 cyclins ensures coherent cell cycle entry. *Nature* 454 (7202), 291–296. doi: 10.1038/nature07118

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27 (3), 431–432. doi: 10.1093/bioinformatics/btq675

Starr, D. A., Saffery, R., Li, Z., Simpson, A. E., Choo, K. H., Yen, T. J., et al. (2000). HZwint-1, a novel human kinetochore component that interacts with HZW10. *J. Cell Sci.* 113 (Pt 11), 1939–1950.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi: 10.1073/pnas.0506580102

Szabó, P. M., Tamási, V., Molnár, V., Andrásfalvy, M., Tömböl, Z., Farkas, R., et al. (2010). Meta-analysis of adrenocortical tumour genomics data: novel pathogenic pathways revealed. *Oncogene* 29 (21), 3163–3172. doi: 10.1038/onc.2010.80

Urbanucci, A., Sahu, B., Seppälä, J., Larjo, A., Latonen, L. M., Waltering, K. K., et al. (2012). Overexpression of androgen receptor enhances the binding of the receptor to the chromatin in prostate cancer. *Oncogene* 31 (17), 2153–2163. doi: 10.1038/onc.2011.401

Wajchenberg, B. L., Albergaria Pereira, M. A., Medonca, B. B., Latronico, A. C., Campos Carneiro, P., Alves, V. A., et al. (2000). Adrenocortical carcinoma: clinical and laboratory observations. *Cancer* 88 (4), 711–736. doi: 10.1002/(SICI)1097-0142(20000215)88:4<711::AID-CNCR1>3.0.CO;2-W

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat. Methods* 6 (1), 75–77. doi: 10.1038/nmeth.1282

Wu, F., Shi, X., Zhang, R., Tian, Y., Wang, X., Wei, C., et al. (2018). Regulation of proliferation and cell cycle by protein regulator of cytokinesis 1 in oral squamous cell carcinoma. *Cell Death Dis.* 9 (5), 564. doi: 10.1038/s41419-018-0618-6

Xiao, H., Xu, D., Chen, P., Zeng, G., Wang, X., and Zhang, X. (2018). Identification of five genes as a potential biomarker for predicting progress and prognosis in adrenocortical carcinoma. *J. Cancer* 9 (23), 4484–4495. doi: 10.7150/jca.26698

Xu, Z., Zhou, Y., Cao, Y., Dinh, T. L., Wan, J., and Zhao, M. (2016). Identification of candidate biomarkers and analysis of prognostic values in ovarian cancer by integrated bioinformatics analysis. *Med. Oncol.* 33 (11), 130. doi: 10.1007/s12032-016-0840-y

Zhan, P., Zhang, B., Xi, G. M., Wu, Y., Liu, H. B., Liu, Y. F., et al. (2017). PRC1 contributes to tumorigenesis of lung adenocarcinoma in association with the Wnt/beta-catenin signaling pathway. *Mol. Cancer.* 16 (1), 108 doi: 10.1186/s12943-017-0682-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 12 November 2019
doi: 10.3389/fams.2019.00056

Metastases Growth Patterns *in vivo*—A Unique Test Case of a Metastatic Colorectal Cancer Patient

Gili Hochman[1†], *Einat Shacham-Shmueli*[2†], *Tchia Heymann*[1], *Stephen Raskin*[2] and *Svetlana Bunimovich-Mendrazitsky*[1*]

[1] Department of Mathematics, Ariel University, Ariel, Israel, [2] Sheba Medical Center, Tel Hashomer, Israel

Colorectal cancer (CRC) is one of the most common causes of cancer-related mortality worldwide. Most cases of deaths result from metastases, assumed to be shed, in many cases, before disease detection. Providing reliable predictions of the metastases' growth pattern may help planning treatment. Available mathematical tumor growth models rely mainly on primary tumor data, and rarely relate to metastases growth. The aim of this work was to explore CRC lung metastases growth patterns. We used data of a metastatic CRC patient, for whom 10 lung metastases were measured while untreated by seven serial computed tomography (CT) scans, during almost 3 years. Three mathematical growth models—Exponential, logistic, and Gompertzian—were fitted to the actual measurements. Goodness of fit of each of the models to actual growth was estimated using different scores. Factors affecting growth pattern were explored: size, location, and primary tumor resection. Exponential growth model demonstrated good fit to data of all metastases. Logistic and Gompertzian growth models, in most cases, were overfitted and hence unreliable. Metastases inception time, calculated by backwards extrapolation of the fitted growth models, was 8–19 years before primary tumor diagnosis date. Three out of ten metastases demonstrated enhanced growth rate shortly after primary tumor resection. Our unique data provide evidence that exponential growth of CRC lung metastases is a legitimate approximation, and encourage focusing research on short-term effects of surgery on metastases growth rate.

SIGNIFICANCE

Providing reliable predictions of the metastases' growth pattern using mathematical models may help determining the optimal treatment plan that fits a given patient best and maximizes the probability of cure.

Keywords: lung metastases, mathematical growth models, exponential growth, logistic growth, gompertzian growth, primary tumor resection

OPEN ACCESS

Edited by:
Doron Levy, University of Maryland, College Park, United States

Reviewed by:
Spyros K. Stamatelos, Sanofi, United States
Elisa Domínguez-Hüttinger, National Autonomous University of Mexico, Mexico

***Correspondence:**
Svetlana Bunimovich-Mendrazitsky
svetlanabu@ariel.ac.il

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *03 July 2019*
Accepted: *22 October 2019*
Published: *12 November 2019*

Citation:
Hochman G, Shacham-Shmueli E, Heymann T, Raskin S and Bunimovich-Mendrazitsky S (2019) Metastases Growth Patterns in vivo—A Unique Test Case of a Metastatic Colorectal Cancer Patient. Front. Appl. Math. Stat. 5:56. doi: 10.3389/fams.2019.00056

INTRODUCTION

Colorectal cancer (CRC) is one of the most common causes of cancer-related morbidity and mortality worldwide. Most cases of deaths result from development of metastatic disease [1]. CRC has a slow natural history (i.e., development of disease) that provides a great opportunity for early detection and prevention strategies. Surgery is the main curative treatment, but despite

complete resection of the primary tumor, metastatic disease might develop in a significant number of patients [2].

The exact dynamics of tumor and metastasis formation is not well-established. It is assumed that many (if not most) of metastases are shed before primary tumor is even detectable [1, 3, 4]. Hence, preventing metastases growth by adjuvant or perioperative treatments is indicated in many cases after resection of primary tumor [5, 6]. Providing reliable predictions of the metastases' growth pattern using mathematical models may help determining the optimal treatment plan.

Growth laws of primary tumors are thoroughly investigated [7, 8], however, not many mathematical models are dealing with metastases growth dynamics in humans. Many of the mathematical models for primary tumor growth are based on fitting *in-vivo* data to relatively simple growth models, such as exponential, logistic, Gompertzian, or power law [9, 10], and models for metastases dynamics rely on the same laws. The Gompertzian law is considered most reliable, because it was found that generally, doubling time of tumors usually decreases with time. Nevertheless, the assumption of exponential growth is preferred over logistic or Gompertzian because it includes one less parameter, which reduces the degree of freedom in the model, consequently reducing the difficulty in getting numerical convergence with limited amount of data. Hence, exponential law is often assumed, at least for the first period of growth [10–12]. However, this assumption is hard to prove *in vivo*, since there are very few available data of untreated metastases growth in humans. Moreover, diversity between patients, and between metastases of the same patient, further increases the challenge when trying to find growth patterns that can be used as predictors.

Here, we describe a CRC patient with 10 lung metastases, for which uncommon data of *in vivo* growth over time is available. The metastases were followed and measured—while untreated—for over 2 years. Our aims were:

- To describe metastases growth pattern and decide which of the three models—Gompertzian, exponential, logistic—fits best.
- To determine whether factors such as location and size of metastases have an effect on growth pattern and rate.
- To estimate natural history of the disease (i.e., time of onset of metastases).

MATERIALS AND METHODS

Data

A 65 years old patient was diagnosed with rectal cancer TNM stage [13] T3N0 (and colon polyp containing superficial cancer TNM stage T1N0). A CT scan at the time of first diagnosis showed also 8 mm nodules in the lungs. A PET-CT scan did not show FDG uptake in these nodules, which may have implied that these nodules are not malignant. Fifty-four days after first diagnosis, the primary tumor in rectum (and colonic polyp) were resected. On post-surgery follow up, six additional CT scans were conducted, roughly every 6 months, in which 10 lung metastases were evidently growing. During this time period systemic treatment (chemotherapy, targeted treatment) was offered, but not administered, because of personal preference of the patient. The measured volumes of metastases at these seven timepoints (marked as W1–W7) are reported in **Table 1**. See **Figure S1** for examples of CT tomographic images and **Figure S2** for illustration of the locations of all diagnosed metastases (marked #1–#10) in the lungs.

Modeling

Based on the data available, we wanted to set a growth model (exponential, Gompertzian, or logistic) for each of the 10 metastases, and assess the values of growth rate parameters.

Exponential growth was modeled by the equation:

$$\Psi\,(\mathrm{t}) = \mathrm{N}_0^{exp} e^{\lambda t}, \tag{1}$$

where Ψ (t) is the metastasis volume at time t, counted from the day of primary tumor resection, $N_0^{\exp}$ is the size of metastasis at $t = 0$, and λ is the growth rate parameter.

Logistic growth was modeled by the equation:

$$\Theta\,(\mathrm{t}) = \frac{K^{logistic}}{1 + \left(\frac{K^{logistic}}{\mathrm{N}_0^{logistic}} - 1\right)\mathrm{e}^{-rt}}, \tag{2}$$

where $\Theta\,(t)$ is metastasis volume at time t, $N_0^{logistic}$ is the size of metastasis at $t = 0$, $K^{logistic}$ is the limiting tumor size—carrying capacity, and r is a rate parameter.

Gompertzian growth was modeled by:

$$\Phi\,(\mathrm{t}) = K^{gomp} e^{\ln\left(\frac{\mathrm{N}_0^{gomp}}{K^{gomp}}\right)e^{-\beta t}}, \tag{3}$$

where Φ (t) is metastasis volume at time t, N_0^{gomp} is the size of metastasis at t = 0, K^{gomp} is the limiting tumor size and β is a rate parameter.

Direct fit of the data, by numerical minimization of the sum of squared errors (SSE) was done for each of the metastases separately, to optimize the parameter values for each of the three equations: $N_0^{\exp}$ and λ in Equation (1), $N_0^{logistic}$, $K^{logistic}$, and r in Equation (2) and N_0^{gomp}, K^{gomp} and β in Equation (3). Specifically, the minimization was done for errors of the model predictions of the log-volume of tumor size:

$$SSE = \sum\nolimits_{i=1}^{n} \left(ln\left(f\left(t_i, p\right)\right) - ln\left(Y_i\right)\right)^2, \tag{4}$$

where Y_i is the observed metastasis volume at time t_i and $f\left(t_i, p\right)$ is predicted metastasis volume at the same time, as calculated by each of the model Equations (1)–(3), depending on the estimated parameters vector p. The minimization procedure was performed using the Matlab functions *lsqnonlin* and *nlinfit*.

TABLE 1 | Metastases sizes measured by CT scans of the patient, at different times marked W1–W7.

	W1	W2	W3	W4	W5	W6	W7
Date	17/10/2012	08/05/2013	06/10/2013	22/04/2014	05/10/2014	27/04/2015	29/09/2015
MET1	0.014	0.016	0.050	0.099	0.177	0.326	0.776
MET2	0.178	0.236	0.309	0.754	1.466	3.613	6.589
MET3	0.004	0.101	0.197	0.356	0.544	0.940	1.371
MET4	0.128	0.330	0.506	0.921	2.384	5.370	9.292
MET5	0.108	0.205	0.349	0.674	1.039	3.933	14.547
MET6	–	0.077	0.347	0.479	0.887	3.031	4.475
MET7	–	0.058	0.197	0.410	0.675	1.565	2.138
MET8	0.292	0.807	4.944	8.548	12.718	32.654	66.693
MET9	0.108	0.209	0.361	0.543	0.954	1.338	1.897
MET10	0.175	0.429	2.719	8.045	19.250	55.708	91.538

Primary tumor was resected on 10/12/2012, 54 days after W1. First row is date of the CT scan, and other rows are metastases volumes in cm^3.

The fit was done for each metastasis using data of all available measurements in time, including at time W1, conducted 54 days before resection. Indeed, the growth law and rate may change between W1 and W2 due to the resection, however we assumed that the time between W1 and resection time was short enough that it would change the measure only slightly, within the measurement error.

Goodness of Fit Analysis

Different criteria for the goodness of fit were compared, in order to determine the best growth model for each of the metastases, and the reliability of the estimated parameter values [10, 14]. For this purpose, the root of mean square of errors (RMSE) was calculated for each of the three models that were fitted to each of the 10 metastases.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{(n-P)}}, \tag{5}$$

where *SSE* is defined by Equation (4). The MSE is normalized to the number of measurements (n) available for the specific metastasis, and to the number of model parameters (P), to enable fair comparison between exponential model (where $P = 2$) and the other models (where $P = 3$).

Another criterion used for the goodness of fit of predicted curves to the data was the adjusted coefficient of determination:

$$R^2 = 1 - \frac{n-1}{(n-P)}\frac{SSE}{SST}, \tag{6}$$

where $SST = \sum_{i=1}^{n}\left(\ln(Y_i) - \overline{\ln(Y_i)}\right)^2$, and $\overline{\ln(Y_i)}$ is the time average of the observed log-volumes measured at all n time-points. This metric quantifies how much of the variability in the data is described by the model, as the denominator is proportional to the data variance. In this case, $R^2$is adjusted to the number of measurements (n) and number of model parameters (P).

To quantify the reliability of the estimated parameter values, the variance-covariance matrix of parameters was calculated, in the context of non-linear least squares regression:

$$Cov = MSE \cdot \left(J^T J\right)^{-1}, \tag{7}$$

where J is the Jacobian of the model as a function of the parameters vector p: $J_{ij} = \frac{\partial f(t_i,p)}{\partial p_j}$. p_j is the jth element of p. The variance of an estimated parameter p_j is defined by the diagonal element of the covariance matrix, Cov_{jj}. This is a measure of the sensitivity of model prediction to the estimated value of the parameter p_j.

Evaluation of Metastases' Natural History

After the best fitted models are chosen, and their reliability is established, the fitted models can be used to estimate the time of onset of metastasis. For this purpose, the fitted curve with estimated parameters for each metastasis k was extrapolated backwards to determine the time of onset of metastasis (T_k), defined as time of appearance of the first malignant cell, adopting the evaluation of $10^{-9}\ cm^3$ for the volume of a single tumor cell. For example, in case of an exponential model the value for T_k was derived from $\Psi\,(t = T_k) = 10^{-9}\ cm^3$. This method was also used to assess the time of metastasis' size reaching to the threshold enabling detection by CT scan (D_k), approximated as $0.002\ cm^3$.

RESULTS

Fitting and Comparing Growth Models

For every one of the metastases, values for the parameters of each of the three growth models examined were fitted to the dataset of all available measurements at times W1–W7. Metastases #6 and #7 were not detectable at W1 timepoint (see **Table 1**). For metastasis #3, the measure at W1 was omitted from the fit since it was very small—close to the limit of detection. The parameters' optimal values, as well as different scores for goodness of fit

(see section Materials and Methods), are presented in **Tables 2**, **3**. In general, all growth models provided good fit for most of metastases, as their predicted curves are within or close to the measurement error bounds (**Figure 1**). This accuracy reflects in the adjusted R squared values (**Table 3**) which are almost all >0.94. unlike R^2, the SSE and RMSE values are not normalized to variability of observed values. Hence, comparing SSE or RMSE values of different metastases would reflect the variability in absolute values of metastases volumes: metastases that grow to high volumes would have higher SEE and RMSE values. However, their values can be used to compare between goodness of fit of different growth models for the same metastasis, as detailed bellow.

For metastases numbered 1, 2, 4, 5, comparing goodness of fit of the three models shows that the exponential model demonstrated the closet prediction to actual growth measurements for these metastases in all three scores (see **Table 3**). Logistic and Gompertzian models converged with extremely high values of carrying capacity parameter K (marked orange in **Table 2**), which means that they essentially degenerate into exponent. The variance of K could not be calculated in those cases, because the Jacobian was singular or close to singular, i.e., curve fit does not depend on the value of K. This may point on redundancy in these models.

For metastases numbered 6, 8, 10 the exponential fit scores were inferior than those of the other two models. However, Gompertzian fit has the same problem of parameter redundancy, where curve fit does not depend on the value of K (marked orange in **Table 2**). The logistic fit is also not reliable in these cases, since the variance of the parameter K is very large, 10–100 times its value (marked pink in **Table 2**). Hence, in these metastases, exponential model is also the preferable one.

For metastases 3, 7, 9 the logistic curve seems reliable, and it has better scores than exponential. However, these metastases are very small in size hence measurement error is relatively large, and exponential curve is also within this error. Gompertzian fit is not reliable from the same reasons mentioned above for other metastases.

For all metastases described above, it seems that using exponential approximation for the growth law is a good enough approximation, at least for a range of 2 years from the time of primary tumor detection and resection (for the first timepoints of measure W1–W5, or W2–W5 for metastases #3, #6, and #7).

Variability of Growth Rates of Metastases

Looking at the fitted exponential model parameters, the value of exponent of the growth rate λ (see Equation 1) is in the same order of magnitude for all metastases, and its value is estimated to be the average of their fitted values: 1.48 years^{-1}, with standard deviation of 0.34 years^{-1} (**Table 2**). Their distribution (assumed to be normal) is presented in **Figure 2A**. Note, that the metastases most distant from this mean value are #9 and #10, which are both located in the left lung, while all other metastases are in the right lung. Other than that, no relation was found between fitted growth rates to the metastasis location in the lungs, nor

TABLE 2 | Values of estimated optimal parameters, for each of the 10 observed metastases, for the three fitted models (see Equations 1–3), along with their variance (see Equation 7) presented in parentheses.

	Exponential			Logistic				Gompertz				
	N_0^{exp} [cm^3] (var)		λ [years^{-1}] (var)	$N_0^{logistic}$ [cm^3] (var)		$K^{logistic}$ [cm^3] (var)		r [years^{-1}] (var)	N_0^{gomp} [cm^3] (var)		K^{gomp} [cm^3] (var)		β [years^{-1}] (var)	
MET #1	0.0136	(4.1E-06)	1.4016	(2.2E-05)	0.0136	(4.1E-06)	2.33E+06	(–)	1.4008	(2.2E-05)	0.0132	(4.4E-06)	5.44E+08	(–)	0.0621	(4.7E-08)
MET #2	0.1495	(5.8E-04)	1.2945	(2.6E-05)	0.1495	(5.8E-04)	2.75E+07	(–)	1.2943	(2.6E-05)	0.1468	(7.7E-04)	6.14E+08	(–)	0.0629	(8.1E-08)
MET #3	0.0758	(5.2E-05)	1.0604	(7.6E-06)	0.0638	(4.0E-05)	2.6565	(0.8103)	1.3243	(3.8E-05)	0.0562	(3.3E-05)	22.31	(593.1)	0.2692	(1.5E-05)
MET #4	0.1595	(1.7E-04)	1.4504	(6.4E-06)	0.1594	(1.7E-04)	1.37E+08	(–)	1.4496	(6.4E-06)	0.1536	(2.3E-04)	6.46E+08	(–)	0.0717	(2.2E-08)
MET #5	0.1012	(6.4E-04)	1.5730	(6.1E-05)	0.1012	(6.4E-04)	3.13E+06	(–)	1.5725	(6.1E-05)	0.0992	(8.2E-04)	1.08E+09	(–)	0.0741	(1.8E-07)
MET #6	0.0565	(2.7E-04)	1.6017	(7.3E-05)	0.0506	(5.3E-04)	15.00	(1495)	1.7406	(5.1E-04)	0.0411	(1.9E-04)	2.82E+03	(–)	0.1954	(1.0E-06)
MET #7	0.0465	(1.3E-04)	1.4452	(5.0E-05)	0.0318	(7.9E-05)	2.8551	(1.4E+00)	1.9492	(2.3E-04)	0.0207	(5.6E-05)	9.8627	(104)	0.4930	(8.2E-05)
MET #8	0.5395	(2.3E-02)	1.7831	(7.7E-05)	0.4525	(1.9E-02)	76.82	(2822)	2.1618	(3.3E-04)	0.4187	(9.2E-03)	1245.44	(–)	0.3410	(1.7E-06)
MET #9	0.1426	(1.3E-04)	0.9631	(6.1E-06)	0.1309	(4.7E-05)	3.3190	(6.5E-01)	1.2171	(1.8E-05)	0.1301	(4.0E-05)	38.1023	(1940)	0.2266	(1.0E-05)
MET #10	0.2821	(5.1E-03)	2.2069	(6.3E-05)	0.2282	2.3E-03	118.51	2637)	2.6072	(1.3E-04)	0.2249	(1.9E-03)	5.37E+04	(–)	0.2430	(4.2E-07)

Colored in pink are parameters for which calculated variance is very high (> 10 times the parameter value). Colored in orange are parameters for which the variance could not be calculated, because the Jacobian in Equation (7) was singular or close to singular.

TABLE 3 | Different measures of the goodness of fit, for each of the 10 observed metastases, for the three fitted models.

	SSE			RMSE			Adjusted R^2		
	Exp	Logistic	Gomp	Exp	Logistic	Gomp	Exp	Logistic	Gomp
MET #1	0.267	0.267	0.297	0.231	0.258	0.272	0.977	0.976	0.967
MET #2	0.317	0.317	0.415	0.252	0.281	0.322	0.968	0.968	0.947
MET #3	0.046	0.016	0.009	0.107	0.074	0.054	0.988	0.996	0.997
MET #4	0.080	0.080	0.112	0.126	0.141	0.167	0.993	0.993	0.988
MET #5	0.761	0.762	0.973	0.390	0.436	0.493	0.948	0.948	0.917
MET #6	0.443	0.424	0.401	0.333	0.376	0.366	0.950	0.953	0.940
MET #7	0.305	0.137	0.083	0.276	0.214	0.167	0.958	0.981	0.985
MET #8	0.959	0.703	0.525	0.438	0.419	0.362	0.949	0.963	0.965
MET #9	0.076	0.020	0.017	0.123	0.071	0.066	0.986	0.996	0.996
MET #10	0.782	0.340	0.402	0.396	0.292	0.317	0.972	0.988	0.982

See Equations (4)–(6).

to its size at the time of detection. The distribution of initial metastases sizes (at time W1), and the lack of correlation to growth rate λ can be seen in **Figure 2B**. The similarity of exponential growth rates of metastases can be also seen in **Figure 3A**, where fitted models for all metastases are presented on the same graph (note that the figure is presented in log-scale, and volumes are also normalized to the initially detected volume, at time W1).

Metastases' Natural History

If we assume each metastasis has followed the same growth law since its formation, then for each metastasis k the onset time (i.e., time of emergence of the first malignant clonogenic cell,) T_K, could be estimated. The earliest possible detection time (i.e., time of metastasis' size reaching to the threshold enabling detection by CT scan,) D_k, could also be evaluated. These evaluations were obtained by extrapolating backwards of the fitted exponential growth model, assuming growth rate was the same through all the time of metastasis' existence. Results, presented in **Figure 3**, show that according to the model, all metastases were formed 8–19 years before primary tumor was detected (**Figure 3B**), however, earliest possible time on which they could be detected, assuming detection limit is $0.002\,cm^3$, was years later -1–5 years before primary tumor detection (This can be seen in **Figure 3B**, and more clearly in **Figure 3A**).

For metastases #3, #6, and #7, the fitted models imply that they have reached detection limit 3–4 years before primary detection time (see **Figure 3A**), and were far larger than this threshold at the day of disease detection (see smooth lines compared to the asterisks in corresponding subplots in **Figure 1**). This result stands in contrast with the fact that they were not observed at the CT done on primary tumor detection (timepoint W1). We assume that these metastases were either undetectable because their size was bellow detection limit, or small enough to be missed at the scan, i.e., their size was above detection limit but close to it. Either way, for these metastases it seems that growth law was *not* the same all the time; it was dramatically changed in the 6 months between W1 and W2, the beginning of the period for which exponential curve was well-fitted. The enhanced growth rate for these metastases between W1 and W2 was evaluated by assuming exponential growth and fitting it to these two timepoints (taking maximal possible metastasis size at W1, when it was not detectable, as $0.002\,cm^3$). Results showed its minimal possible value was 5.85, 6.55, and 6.05 years^{-1} for metastases #3, #6, and #7, respectively. This is at least four times higher than the growth rate at the following period of time, between W2 and W7 (**Table 2**). For all other metastases, the exponential curve is well-fitted to the measure at W1, which means that the exponential growth rate remained the same.

DISCUSSION

Understanding metastases growth is crucial for treating cancer patients. However, little is known about the dynamics of untreated metastases, because such dynamical data in humans are rare. In this paper, we used rare data of a metastatic CRC patient, for which CT measurements of growth of 10 untreated lung metastases during 3 years are available. We aimed to examine the common hypothesis that metastases growth rate can be approximated as exponential. Our results showed that all the metastases could be regarded as growing exponentially, at least for the first 2 years after disease detection and primary tumor resection. Logistic and Gompertzian growth models were also examined, but in most cases, they are overfitted and could not be used.

In addition, we found evidence that the exponential growth curve does not always demonstrate the closet prediction to actual growth measurements throughout all the follow-up time period. That was true for the first time period after primary tumor resection. Our results imply that some of the metastases (#3, #6, and #7) grew more rapidly between the time of first diagnosis and shortly after primary tumor surgery, while during the period of the next 2 years the growth rate was exponential with a constant, slower rate. This result is supported by literature describing implications of surgery on metastases growth rate,

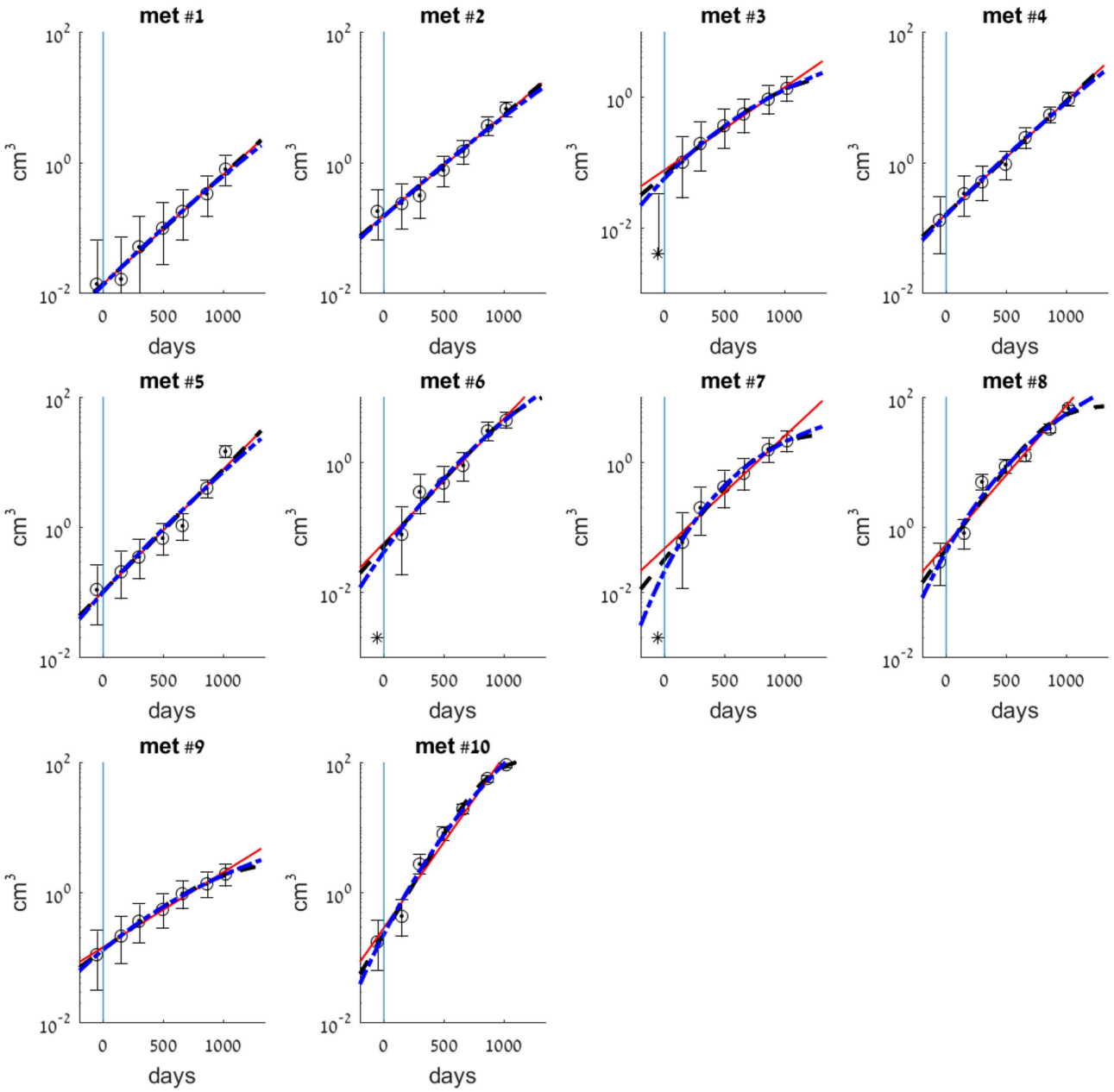

FIGURE 1 | Exponential (smooth red line), logistic (dashed black line), and Gompertzian (dashed-dotted blue line) growth law fits for each of the 10 metastases. Model parameters were fitted to observed volumes of metastases as measured by CT scans at times W1–W7 (see **Table 1**). Circles are clinical measurements that models were fitted to. Asterisk at time W1 ($t = 54$ days) in met #3 is detected size, that was not used for the fit. Asterisks in mets #6 and #7, which were not detected at time W1, mark detection limit 0.002 cm^3. Vertical line at t = 0 marks the day of primary tumor resection.

especially in the short term. There is emerging evidence that the stress response caused by surgery as well as anesthesia and analgesia may promote growth of pre-existing micro-metastasis [5, 15–18]. Such post-surgery metastatic acceleration (PSMA) might be related to surgical stress through several mechanisms, such as suppression of anti-tumor immune response, stimulatory effects on tumor cells, and activation of the coagulation system [5, 16–19]. There are mathematical models that assume PSMA is caused by removal of the suppression that the primary tumor induces on metastases, through systemic inhibition of angiogenesis [15, 20, 21]. This mechanism may explain changes in growth rate between the time before primary resection and the time after it (which was not examined in this work), but it does not explain the change in the growth rate during the period after resection, i.e., in the first month after resection compared to the next following years. Our results show that the most

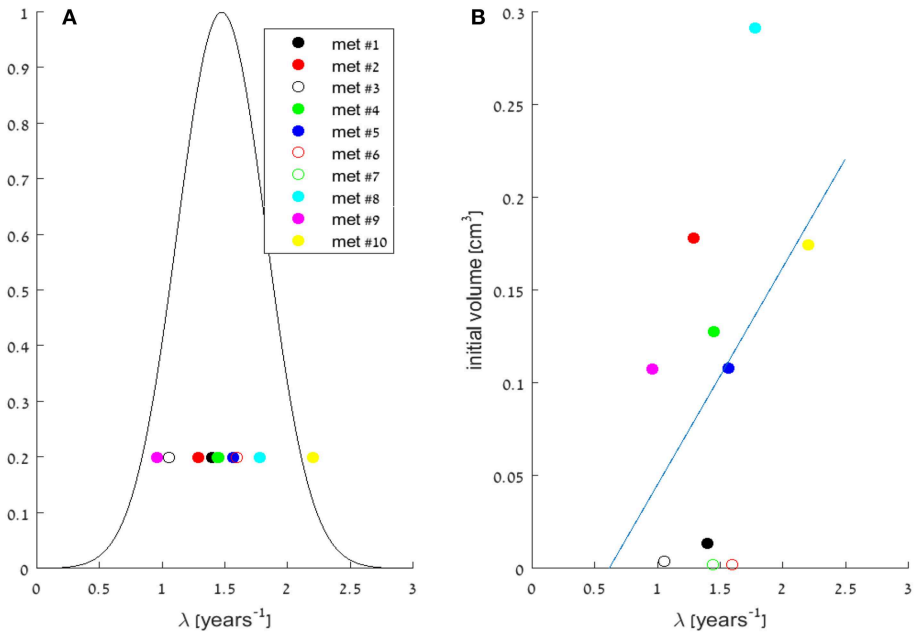

FIGURE 2 | (A) Distribution of the growth rate parameter λ in the fitted exponential models for the 10 metastases. Circles are the specific values of λ, for all metastases (as detailed in **Table 2**), and a gaussian curve is presented, assuming normal distribution around the mean value: 1.48 years^{-1}, with standard deviation of 0.34 years^{-1}. **(B)** Scattering of the growth rate parameter λ vs. initial metastases' volumes, measured at time W1. Circles are the specific values, for all metastases, and a linear model trendline of their correlation is presented. No significant correlation was observed (correlation coefficient 0.43).

significant impact of surgery is in the short term, at least for some of the metastases. It implies that other mechanisms, which decay shortly after surgery, for example—increased angiogenesis factors production due to wound healing [22], are dominant and should be investigated.

For the period of 6 months up to 2–3 years after surgery, all the metastases could be modeled as exponentially growing. The value of the growth rate parameter λ (see Equation 1) is quite similar for all of them, with a mean value of 1.5 years^{-1}. Based on these results, we can assume that exponential growth assumption is legitimate for most of metastases. Moreover, it is reasonable to assume a single value for the growth rate parameter that would fit all metastases found at the same site, as done in some mathematical models [11, 23]. No relation was found between fitted growth laws or rates to the metastases sizes or locations within the lungs.

A timeline of disease progression was constructed and estimated that onset of metastases occurred 8–19 years before primary tumor was detected (**Figure 3B**), and that they grew slowly and became detectable several years later. This was done by backwards extrapolation of the exponential fitted curves, assuming that growth rates before and after resection were the same. However, if we assume that the growth rate was faster—or at least not slowed—after surgery, then the real inception time was *no later* (and possibly earlier) than at the estimated times shown in **Figure 3B**. This result reinforces the notion that metastases were formed many years before detection of primary tumor [1, 5, 12]. The growth dynamics of metastases before primary tumor resection may be further investigated by applying a natural history model on this patient's data, developed by Hanin et al. [23]. As validated here, we can use this model under the assumption of exponential growth after resection, with one value for the growth rate parameter for all metastases.

It is well-known that great variability exists between different primary tumors, between different patients with the same primary tumor and even between metastases at different locations in the same patient [24–26]. The main limitation of this work is that it is based on data of a single patient with rectal cancer metastatic to lungs. Different growth patterns might apply to other sites of metastases or to other primary tumors. Also, all measurements are prone to minute deviation errors especially when millimetric lesions are measured on a bidimensional CT scan. It would be interesting to analyze in the same way data of other patients with measurable metastases either in the lungs or other sites, either from CRC or other primary tumors. Another limitation is the fact that metastases were measured along 3 years period only. The dynamics of metastases growth before first diagnosis and more than 3 years after primary tumor resection are lacking. Hence other factors that could influence growth patterns in time are beyond the scope of this case.

In summary, our unique and uncommon data provide firm evidence that exponential growth model demonstrated precise prediction to actual growth measurements of CRC lung metastases, at least for a limited time period, starting half a year after surgery until about 2 years afterwards. In addition, the results imply that growth rate of some metastases might

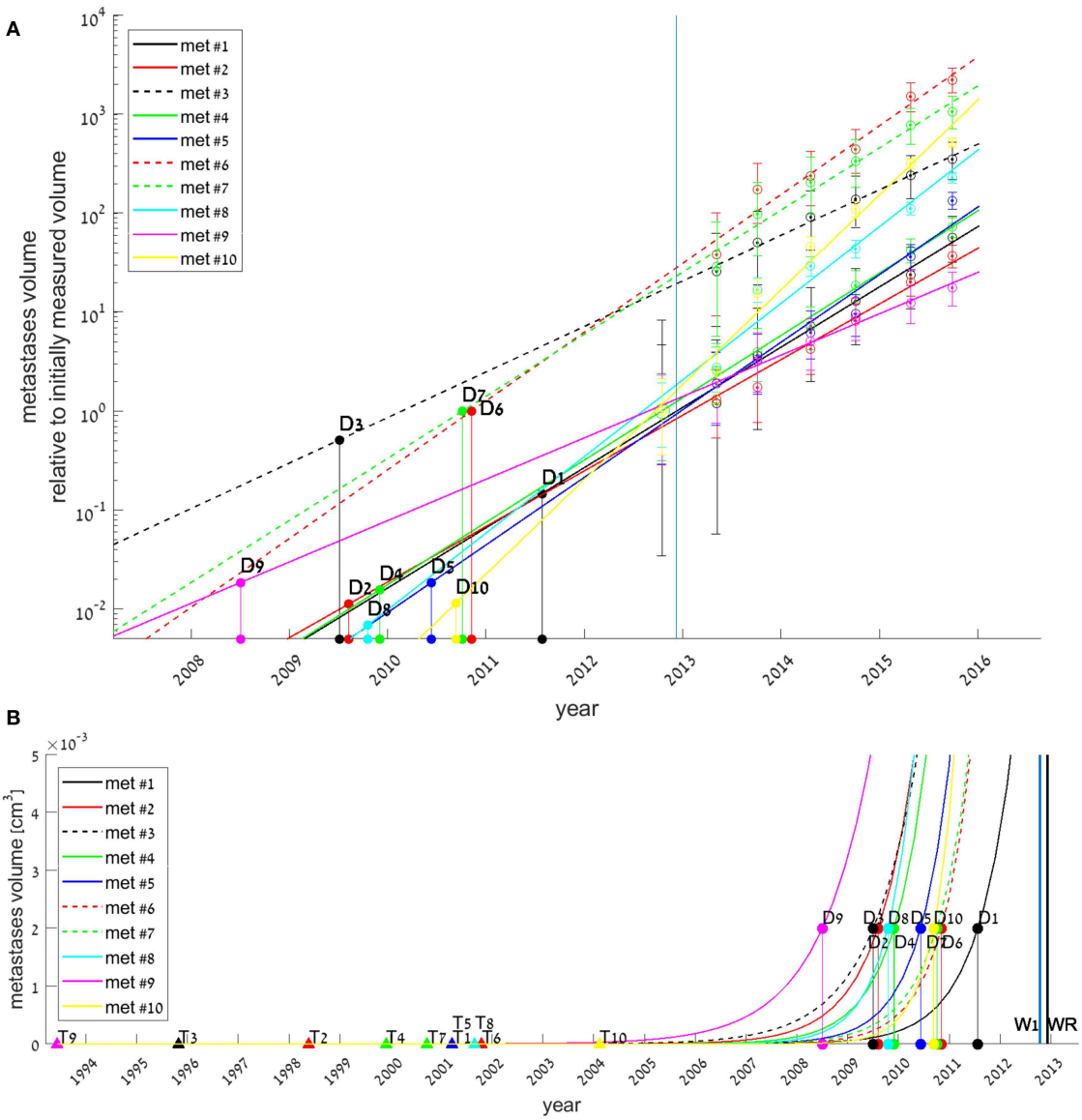

FIGURE 3 | Exponential growth law fits for all 10 metastases (colored lines), with estimated date for each metastasis onset (filled triangles, marked by T_k for each metastasis #*k*), and estimated date for its earliest possible detection (filled circles, D_k). T_k and D_k values were extrapolated from the fitted models, assuming growth rates did not change since metastasis' onset. In **(A)**, measured volumes are normalized to each metastasis initial volume at time W1. Early growth of metastasis (<0.005 cm^3) is not presented, hence T_k values are not shown. In **(B)**, early growth of metastasis is presented, and the whole timeline of the disease natural history is shown. Blue and black vertical lines represent the days of disease detection (W1) and of primary tumor resection (WR), respectively.

accelerate shortly after primary tumor surgery, getting more moderated later. These results encourage further research of the suggested mechanisms for metastases growth acceleration caused by short-term effects of surgery, and of the effects of adjuvant treatment in this period of time.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SB-M designed research. GH and TH performed research. SB-M and GH contributed analytic tools. SB-M, GH, and ES-S analyzed data. SR measured metastases. ES-S and GH wrote the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2019.00056/full#supplementary-material

Figure S1 | CT images from dates W2–5. The metastases are marked by arrows.

Figure S2 | A scheme based on the CT scans, showing locations of all 10 diagnosed metastases in the lungs.

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* (2011) **144**:646–74. doi: 10.1016/j.cell.2011.02.013
2. Stein U, Schlag PM. Clinical, biological, and molecular aspects of metastasis in colorectal cancer. In: Dietel M, Berlin E, editors. *Targeted Therapies in Cancer*. Heidelberg: Springer (2007). p. 61–80.
3. Fisher B, Montague E, Redmond C, Barton B, Borland D, Fisher ER, et al. Comparison of radical mastectomy with alternative treatments for primary breast cancer: a first report of results from a prospective randomized clinical trial. *Cancer.* (1977) **39**:2827–39. doi: 10.1002/1097-0142(197706)39:6<2827::AID-CNCR2820390671>3.0.CO;2-I
4. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A, et al. Colorectal cancer statistics, 2017. *CA Cancer J Clin.* (2017) **67**:177–93. doi: 10.3322/caac.21395
5. Retsky M, Demicheli R, Hrushesky W, Baum M, Gukas I. Surgery triggers outgrowth of latent distant disease in breast cancer: an inconvenient truth? *Cancers.* (2010) **2**:305–37. doi: 10.3390/cancers2020305
6. Horowitz M, Neeman E, Sharon E, Ben-Eliyahu S. Exploiting the critical perioperative period to improve long-term cancer outcomes. *Nat Rev Clin Oncol.* (2015) **12**:213–26. doi: 10.1038/nrclinonc.2014.224
7. Kozusko F, Bajzer Z. Combining gompertzian growth and cell population dynamics. *Math Biosci.* (2003) **185**:153–67. doi: 10.1016/S0025-5564(03)00094-4
8. Brú A, Albertos S, Subiza JL, García-Asenjo JL, Brú I. The universal dynamics of tumor growth. *Biophys J.* (2003) **85**:2948–61. doi: 10.1016/S0006-3495(03)74715-8
9. Rodriguez-Brenes IA, Komarova NL, Wodarz D. Tumor growth dynamics: insights into evolutionary processes. *Trends Ecol Evol.* (2013) **28**:597–604. doi: 10.1016/j.tree.2013.05.020
10. Benzekry S, Lamont C, Beheshti A, Tracz A, Ebos JML, Hlatky L, et al. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol.* (2014) **10**:e1003800. doi: 10.1371/journal.pcbi.1003800
11. Hanin L, Bunimovich-Mendrazitsky S. Reconstruction of the natural history of metastatic cancer and assessment of the effects of surgery: gompertzian growth of the primary tumor. *Math Biosci.* (2014) **247**:47–58. doi: 10.1016/j.mbs.2013.10.010
12. Hanin L, Seidel K, Stoevesandt D. A 'universal' model of metastatic cancer, its parametric forms and their identification: what can be learned from site-specific volumes of metastases. *J Math Biol.* (2016) **72**:1633–62. doi: 10.1007/s00285-015-0928-6
13. Mahul LRM, Amin B, Edge S, Greene FL, Byrd DR, Brookland RK, et al. *AJCC Cancer Staging Manual*. New York, NY: Springer Science+Business Media (2017).
14. Richter PH. Estimating errors in least-squares fitting. *Telecommun Data Acquis Prog Rep.* (1995) **42**:107–36.
15. Benzekry S, Gandolfi A, Hahnfeldt P. Global dormancy of metastases due to systemic inhibition of angiogenesis. *PLoS ONE.* (2014) **9**:26–30. doi: 10.1371/journal.pone.0084249
16. Pinson H, Cosyns S, Ceelen WP. The impact of surgical resection of the primary tumor on the development of synchronous colorectal liver metastasis: a systematic review. *Acta Chir Belg.* (2018) **118**:203–11. doi: 10.1080/00015458.2018.1446602
17. Behrenbruch C, Shembrey C, Paquet-Fifield S, Mølck C, Cho HJ, Michael M, et al. Surgical stress response and promotion of metastasis in colorectal cancer: a complex and heterogeneous process. *Clin Exp Metastasis.* (2018) **35**:333–45. doi: 10.1007/s10585-018-9873-2
18. Demicheli R, Retsky MW, Hrushesky WJM, Baum M, Gukas ID. The effects of surgery on tumor growth: a century of investigations. *Ann Oncol.* (2008) **19**:1821–28. doi: 10.1093/annonc/mdn386
19. Zheng J, Jia L, Mori S, Kodama T. Evaluation of metastatic niches in distant organs after surgical removal of tumor-bearing lymph nodes. *BMC Cancer.* (2018) **18**:1–13. doi: 10.1186/s12885-018-4538-8
20. Benzekry S, Lamont C, Barbolosi D, Hlatky L, Hahnfeldt P. Mathematical modeling of tumor–tumor distant interactions supports a systemic control of tumor growth. *Cancer Res.* (2017) **77**:5183–93. doi: 10.1158/0008-5472.CAN-17-0564
21. Hanin L, Rose J. Suppression of metastasis by primary tumor and acceleration of metastasis following primary tumor resection: a natural law? *Bull Math Biol.* (2018) **80**:519–39. doi: 10.1007/s11538-017-0388-9
22. Maida V, Ennis M, Kuziemsky C, Corban J. Wounds and survival in cancer patients. *Eur J Cancer.* (2009) **45**:3237–44. doi: 10.1016/j.ejca.2009.05.014
23. Hanin L, Rose J, Zaider M. A stochastic model for the sizes of detectable metastases. *J Theor Biol.* (2006) **243**:407–17. doi: 10.1016/j.jtbi.2006.07.005
24. Franko J, Shi Q, Meyers JP, Maughan TS, Adams RA, Seymour MT. Prognosis of patients with peritoneal metastatic colorectal cancer given systemic therapy: an analysis of individual patient data from prospective randomised trials from the Analysis and Research in Cancers of the Digestive System (ARCAD) database. *Lancet Oncol.* (2016) **17**:1709–19. doi: 10.1016/S1470-2045(16)30500-9
25. Riihimaki M, Hemminki A, Sundquist J, Hemminki K. Patterns of metastasis in colon and rectal cancer. *Sci Rep.* (2016) **6**:1–9. doi: 10.1038/srep29765
26. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* (2012) **366**:883–92. doi: 10.1056/NEJMoa1113205

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

BRIEF RESEARCH REPORT
published: 10 December 2019
doi: 10.3389/fphys.2019.01501

Identifying SNAREs by Incorporating Deep Learning Architecture and Amino Acid Embedding Representation

Nguyen Quoc Khanh Le[1*†] *and Tuan-Tu Huynh*[2,3*]

[1] Professional Master Program in Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan, [2] Department of Electrical Electronic and Mechanical Engineering, Lac Hong University, Bien Hoa, Vietnam, [3] Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

OPEN ACCESS

Edited by:
Panayiotis V. Benos, University of Pittsburgh, United States

Reviewed by:
Litao Sun, The Scripps Research Institute, United States
Alexey Goltsov, Abertay University, United Kingdom

***Correspondence:**
Nguyen Quoc Khanh Le
khanhlee@tmu.edu.tw; khanhlee87@gmail.com
Tuan-Tu Huynh
huynhtuantu@lhu.edu.vn

†Present address:
Nguyen Quoc Khanh Le, Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *28 March 2019*
Accepted: *26 November 2019*
Published: *10 December 2019*

Citation:
Le NQK and Huynh T-T (2019) Identifying SNAREs by Incorporating Deep Learning Architecture and Amino Acid Embedding Representation. Front. Physiol. 10:1501. doi: 10.3389/fphys.2019.01501

SNAREs (soluble N-ethylmaleimide-sensitive factor activating protein receptors) are a group of proteins that are crucial for membrane fusion and exocytosis of neurotransmitters from the cell. They play an important role in a broad range of cell processes, including cell growth, cytokinesis, and synaptic transmission, to promote cell membrane integration in eukaryotes. Many studies determined that SNARE proteins have been associated with a lot of human diseases, especially in cancer. Therefore, identifying their functions is a challenging problem for scientists to better understand the cancer disease as well as design the drug targets for treatment. We described each protein sequence based on the amino acid embeddings using fastText, which is a natural language processing model performing well in its field. Because each protein sequence is similar to a sentence with different words, applying language model into protein sequence is challenging and promising. After generating, the amino acid embedding features were fed into a deep learning algorithm for prediction. Our model which combines fastText model and deep convolutional neural networks could identify SNARE proteins with an independent test accuracy of 92.8%, sensitivity of 88.5%, specificity of 97%, and Matthews correlation coefficient (MCC) of 0.86. Our performance results were superior to the state-of-the-art predictor (SNARE-CNN). We suggest this study as a reliable method for biologists for SNARE identification and it serves a basis for applying fastText word embedding model into bioinformatics, especially in protein sequencing prediction.

Keywords: SNARE proteins, deep learning, convolutional neural networks, word embedding, skip-gram

INTRODUCTION

Soluble N-ethylmaleimide-sensitive factor activating protein receptors (SNAREs) are the most important and broadly studied proteins in membrane fusion, trafficking, and docking. They are membrane-associated proteins that consist of distinguishing SNARE domains: heptad restates ∼60 amino acids in length that are predicted to assemble coiled-coils (Duman and Forte, 2003). Most SNAREs consist of only one SNARE motif adjacent to a single C-terminal membrane (e.g., syntaxin 1 and synaptobrevin 2). **Figure 1** shows the domain architecture of some example

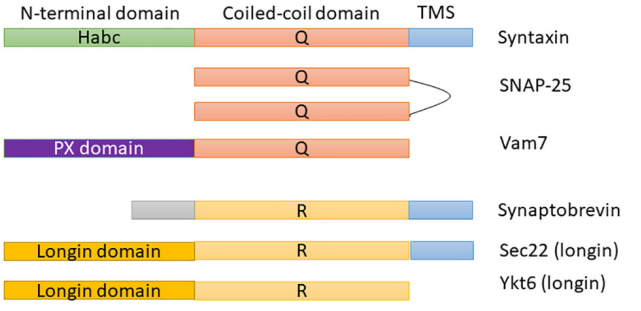

FIGURE 1 | Domain architecture model of SNARE proteins.

SNAREs (e.g., syntaxin, SNAP-25, or Vam 7). As shown in these proteins, SNAREs generally consist of a central "SNARE domain" that is flanked by a variable N-terminal domain and a C-terminal single α-helical transmembrane anchor (Ungermann and Langosch, 2005). SNARE proteins are crucial for a broad range of cell processes, e.g., cytokinesis, synaptic transmission, and cell growth, to promote cell membrane integration in eukaryotes (Jahn and Scheller, 2006; Wickner and Schekman, 2008). There are two categories of SNARE: v-SNAREs incorporated into the membranes of transport vesicles during budding, and t-SNAREs associated with nerve terminal membranes. Researchers have recently identified a lot of SNARE proteins in human and they demonstrated that there is a crucial link between SNARE proteins and numerous diseases [e.g., neurodegenerative (Hou et al., 2017), mental illness (Dwork et al., 2002), and especially cancer (Meng and Wang, 2015; Sun et al., 2016)]. As a detail, a 1 bp deletion in SNAP-29 causes a novel neurocutaneous syndrome (Sprecher et al., 2005), mutation in the b-isoform of neuronal SNARE synaptosomal-associated protein of 25 kDa (SNAP-25) results in both diabetes and psychiatric disease (Jeans et al., 2007), mutations in VPS33B cause arthrogryposis–renal dysfunction–cholestasis (ARC) syndrome (Gissen et al., 2004), and so on.

Because SNARE proteins play an essential molecular function in cell biology, a wide variety of techniques were presented and used to investigate them. One of the best studies on SNAREs is molecular docking of synaptic vesicles with the presynaptic membrane in neurons. Another solution is to identify SNAREs from unknown sequence according to their motif information. In order to address it, Kloepper team is a first group that used bioinformatics techniques in this kind of problem. In their research, they have already built a database for retrieving and classifying SNARE proteins (Kloepper et al., 2007, 2008; Kienle et al., 2009). Furthermore, SNARE functions in sub-Golgi localization had also been predicted using bioinformatics techniques (van Dijk et al., 2008). Yoshizawa et al. (2006) identified SNAREs in membrane trafficking via extracting sequence motifs and the phylogenetic features. In the latest work, Le and Nguyen (2019) identified SNAREs by treating position-specific scoring matrices as images to feed into 2D convolutional neural network (CNN).

To our knowledge, only the study from Le and Nguyen (2019) conducted the SNARE protein prediction in membrane fusion by using machine learning techniques. However, their performance results need a lot of improvements, and we therefore motivate to create a better model for this. To address this, we transform the protein sequences into a continuous bag of nucleobases using fastText model (Bojanowski et al., 2017) and then carry out to identify them with the use of deep neural networks. Releasing by Facebook Research, fastText is a natural language processing (NLP) model for word embedding and text classification. It uses neural network for learning text representations and since its discovery, it has been used in a lot of different NLP problems (Joulin et al., 2017). It has been also used in interpreting biological sequences such as DNA sequences (Le, 2019; Le et al., 2019b) and protein sequences (Asgari et al., 2019), and here we provide a different application with a more in-depth analysis.

The idea is to treat protein sequence as a sentence and amino acids as words, we used fastText to train the language model on all sequences. Subsequently, this language model will be used to generate vectors for protein sequences. At the latest stage, we used a deep neural network to learn these vectors as features and perform supervised learning for classification. The rest of this paper is organized as follows: our materials and methods are introduced in the section "Methods"; some of our relevant experiments and results are introduced in the section "Results"; discussions of the model performance as well as limitations are given in the section "Discussion."

METHODS

Figure 2 illustrates our flowchart which consists of three major processes: data collection, training fastText model and 1D CNN model. We describe the detailed description of our approach in the following paragraphs.

Data Collection

The dataset retrieved from the National Center for Biotechnology Information (NCBI) (by 4-2-2019) (Coordinators, 2015), which is a large suite of online resources for biological information and data. Moreover, on-line resource conserved domain database (CDD) (Zheng et al., 2014) suggested that "SNARE superfamily" members could be identified using the SNARE motif "cl22856," therefore, we used this information to generate non-redundant (annotated) SNARE proteins. This step ensures that we collected all corrected SNARE proteins including SNARE motif. There are many protein sources in NCBI, and we chose to collect all protein sequences from RefSeq (Pruitt et al., 2006). Next, to prevent overfitting problem, we used CD-HIT (Fu et al., 2012) to eliminate the redundant sequences with similarity greater than 30%, and the rest of proteins reaches 26,789 SNAREs. We used full sequences of proteins, thus it includes typical coiled coil as well as other motifs.

In the next step, we collected a negative set to treat our problem as a binary classification between positive (SNAREs) and negative set. To perform this, we retrieved all general proteins without the SNARE motif and with similarity more

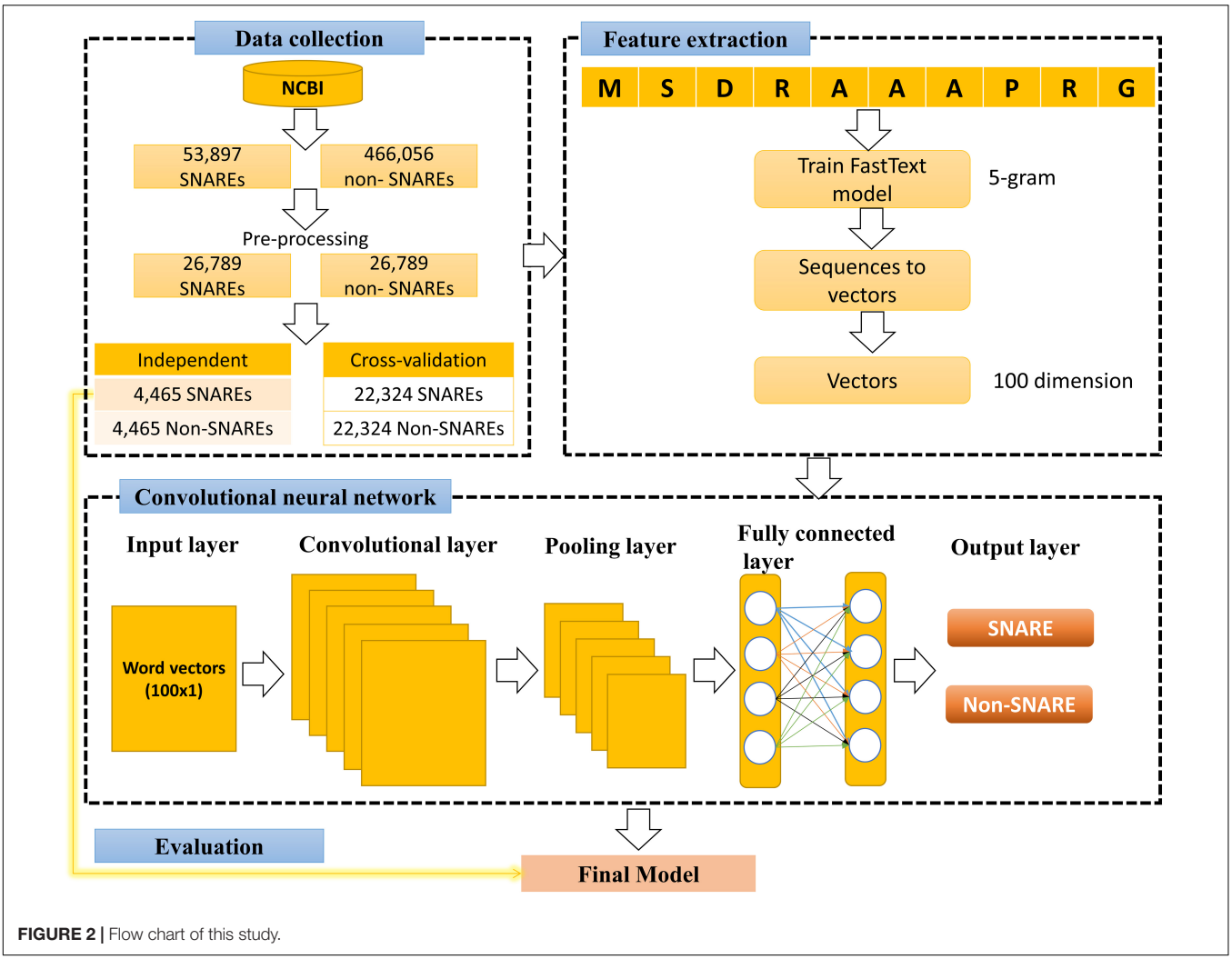

FIGURE 2 | Flow chart of this study.

than 30%. Because the number of negative data was much higher than the number of positive data, it will cause difficulties in machine learning problem. Therefore, we randomly selected 26,789 negative samples to give balance training in our problem.

Amino Acid Embedding Representation

Encouraged by the high performance of word embedding in many NLP tasks, we presented a similar feature set called "amino acid embedding." The objective is to apply recent NLP models into biological sequences. It was first proposed by Asgari and Mofrad (2015) and successfully used to solve the latter biological problems related to sequence information (Habibi et al., 2017; Vang and Xie, 2017; Öztürk et al., 2018). Nevertheless, with the use of Word2Vector to describe the biological sequences, these findings had some disadvantages such as out-of-vocabulary cases for unknown words as well as not taking care of the inner structure of words. Accordingly, a critical issue therefore needs to be resolved is that instead of using an single specific vector representation for the protein word, the internal structure of each word needs to be taken into account. Facebook suggested fastText, which is a Word2vec extension that can handle the word as a continuous bag of character n-grams (Bojanowski et al., 2017), to perform this task. The vector for a word therefore consists of the number of n-grams of this type. It has been shown that fastText was more accurate than using Word2vec in a variety of fields (Joulin et al., 2017). Inspired by its accomplishments, previous researchers used it to describe biological sequences such as DNA enhancer sequence (Le et al., 2019b), DNA N6-methyladenine sites (Le, 2019) and protein sequence (Asgari et al., 2019).

The goal of this step is to encode nucleotides by establishing their vector space distribution, enabling them to be adopted by supervised learning algorithms. To perform a supervised learning classification, we need a set of features having the same dimension. Nonetheless, our protein sequences are of different lengths, so to address this issue, we set the embedding vector dimension to 100. This means that each protein sequence is represented as real numerical values of 100 and can be fed directly without pre-processing into any machine learning classifier. We have more special features for a good prediction by bringing this information into the dataset.

Convolutional Neural Network

Convolutional neural network generally consists of multiple layers with each layer performing a particular function of translating its output into a functional representation. All layers are combined to form the architecture of our CNN system using a specific order. Similar to many published works in this field (Le et al., 2018, 2019a,c; Nguyen et al., 2019), different layers used in CNN for the current study include:

(1) Input layer of our CNN is a 1D vector, which is a vector of size 1×100 (created by fastText model).
(2) Convolutional layers were used as convolution operations to extract features embedded in the 1D input vector. These layers took a sliding window with specific stride shifting across all the input shapes. After sliding, the input shapes will be transformed into representative values. The spatial relationship between numeric values in the vectors has been preserved in this convolutional process. It will help this layer learn the important features using small slides of input data. Since the input of our CNN model is a vector of small size, we used the kernel size of 3 to deduce more information. This number of kernel has been used in previous works on CNN (Le et al., 2017, 2018, 2019a).
(3) Activation layer was performed after convolutional layers. It is an additional non-linear operation, called ReLU (Rectified Linear Unit) and is calculated as follows:

$$f(x) = \max(0, x) \tag{1}$$

Where x is the number of inputs in a neural network. The purpose of ReLU is to introduce non-linearity in our CNN and help our model learn better from the data.
(4) Pooling layer was applied in convolutional layers to reduce the computational size for the next layers. There are three types of pooling layers, and we selected max pooling in our architecture to select the maximum value over a window of 2.
(5) Dropout layer was applied aiming to reduce the overfitting of our model and also to improve the performance results in some cases (Srivastava et al., 2014).
(6) Flatten layer was used to transform the input matrix into a vector. It always stand before fully connected layers.
(7) Fully connected layer was usually applied in the last stages of neural network architectures. In this layer, each node is fully connected with all the nodes of the previous layers. Two fully connected layers have been included in the current model. The first one connected all the input nodes to the flatten layer to help our model to gain more knowledge and perform better. This one was then connected to the output layer by the second layer. The number of nodes in the output layer is equal to 2 as identifying SNARE proteins was as a binary classification problem.
(8) Softmax was an evaluation function standing at the output of the model to determine the probability of each possible output. Its function could be calculated by the formula:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \tag{2}$$

where z indicates the input vector with K-dimensional vector, $\sigma(z)$i is real values in the range (0, 1) and ith class is the predicted probability from sample vector x.

Assessment of Predictive Ability

We firstly trained the model on the entire training set using 5-fold cross-validation technique. Since every 5-fold cross-validation produces different results each time, we performed ten times 5-fold cross-validation to achieve more reliable results. Thereafter, we reported the cross-validation performance by averaging all the ten times cross-validation tests. In the training process, hyper-parameter optimization has been used to identify the best parameters for each dataset. Finally, an independent test was applied to evaluate the performance and to ensure preventing any systematic bias in the cross-validation set.

Moreover, to evaluate the performance of our method, we applied Chou's criterion (Chou, 2001) used in many bioinformatics studies. With this criterion, some standard metrics sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC) are as follows:

$$Sensitivity = 1 - \frac{N_-^+}{N^+}, \quad 0 \le Sen \le 1 \tag{3}$$

$$Specificity = 1 - \frac{N_+^-}{N^-}, \quad 0 \le Spec \le 1 \tag{4}$$

$$Accuracy = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}, \quad 0 \le Acc \le 1 \tag{5}$$

$$MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N^+}{N^+}\right)\left(1 + \frac{N_-^+ - N^-}{N^-}\right)}},$$

$$-1 \le MCC \le 1 \tag{6}$$

The relations between these symbols and the symbols in Eqs. (3–6) are given by:

$$\begin{cases} N_+^- = FP \\ N_-^+ = FN \\ N^+ = TP + N_-^+ \\ N^- = TN + N_+^- \end{cases} \tag{7}$$

Where TP, FP, TN, FN are true positive, false positive, true negative, and false negative values, respectively.

RESULTS

Composition of Amino Acid Representation in SNAREs and Non-SNAREs

In this section, we would like to analyze the differences between SNARE and non-SNARE sequences in our dataset by computing the composition of amino acid representation between them. The amino acids which had the highest frequency in the positive and negative set are shown in **Figure 3**. It is easy to point out some of the differences between the two types of dataset. For instance, we were aware of the higher frequency of amino acid L, and F, and R in the SNARE proteins but lower in the non-SNAREs. Otherwise, the amino acids that appeared a lot in non-SNARE sequences are G, T, N, and D. Besides, we plotted the standard error bars at each column to statistically see the differences among amino acid compositions. These error bars aim to calculate confidence intervals, or margins of error to quantify uncertainty. As shown in **Figure 3**, there are some amino acids had significantly differences (with no overlap error bars) such as N, D, G, L, F, and T. Therefore, these amino acids might play a crucial role in identifying SNARE sequences and they can be special features that help our model predict SNAREs with high accuracy. This finding also plays an important role in further research that aims to analyze the motif information in SNARE proteins.

Hyperparameters Optimization

Hyper-parameters are architecture-level parameters and are different from parameters of a model trained via backpropagation. To tune hyperparameters, we used the approach to choose a set of hyperparameters for speeding up the training process as well as preventing overfitting. As suggested by Chollet (2015), each step of the above hyper-parameter-tuning approach was integrated into the hyper-parameter-tuning process as follows:

- Selecting a specific set of hyper-parameters.
- Creating the model according to the specific set.
- Evaluating the performance results using testing dataset.
- Moving to the next set of hyper-parameters.
- Repeating.
- Measuring performance results on an independent dataset.

Keras framework library (Chollet, 2015) with a TensorFlow backend (Abadi et al., 2016) was used as a deep learning framework to build the 1D CNN architecture. We performed grid search on training set and used accuracy to select the next set of hyperparameters. Furthermore among the six optimizers

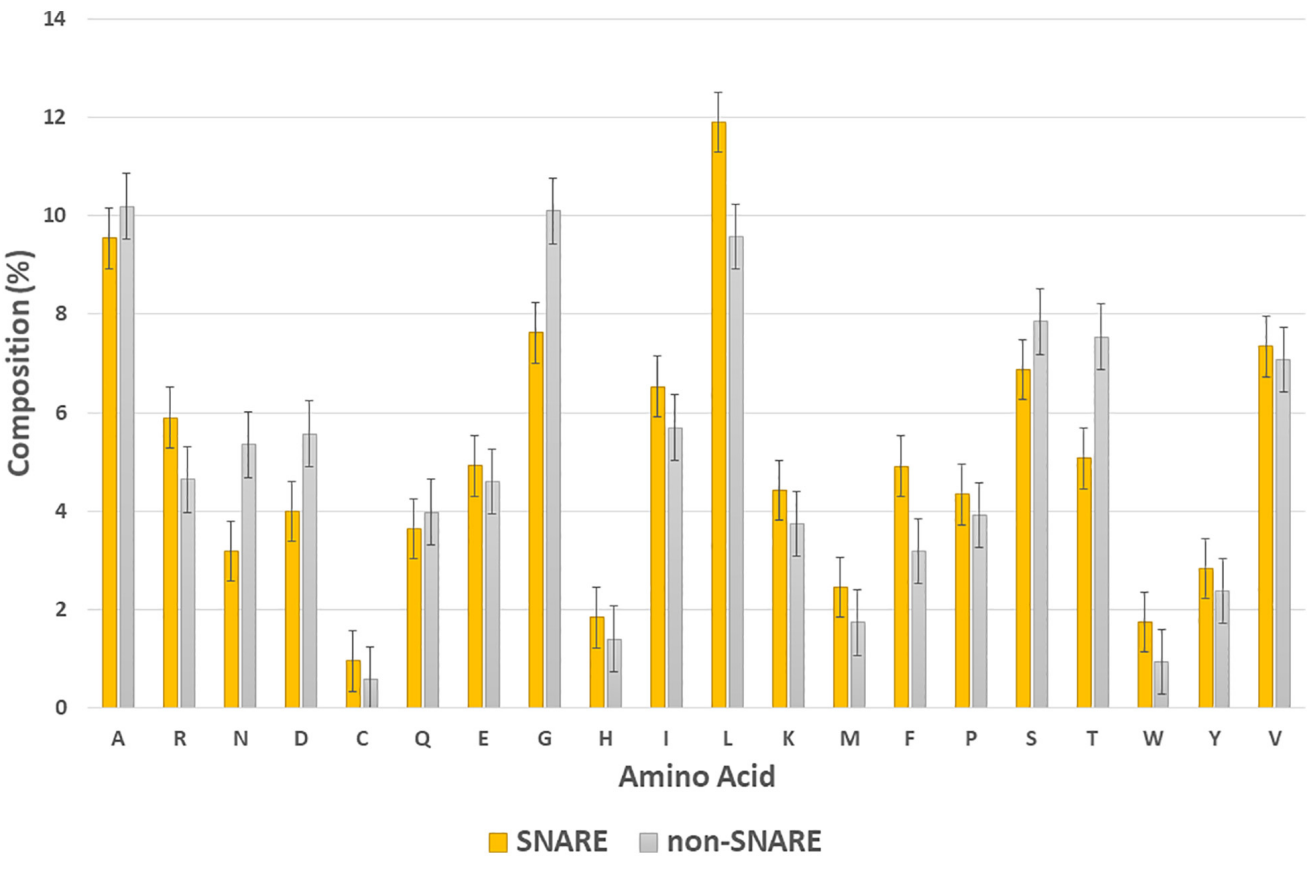

FIGURE 3 | Composition of amino acid in SNAREs and non-SNAREs.

in Keras [e.g., Adam, Adadelta, Adagrad, Stochastic Gradient Descent (SGD), RMSprop, and Adamax], Adadelta has given a superior performance. Therefore, we used Adadelta in our model to achieve an optimal result. This point is also proven in the previous protein function prediction using CNN (Le et al., 2017; Nguyen et al., 2019).

SNARE Identification With Different n-Gram Levels

After tuning the optimal parameters for 1D CNN model, we evaluated the performance of this architecture on the datasets of different n-gram levels (from 1 to 5). In this step, all the measurement metrics were used to evaluate the comparative performance in both cross-validation and independent test. The result is displayed in **Table 1**. **Table 1** shows that the performance results of n-gram levels are proportional. We were not able to achieve the best performance unless we used high levels of n-gram values. To maximize the performance of our models, we should choose the n-gram levels from 4 (accuracy of more than 97%). This means that the model only captures the special information in a high level of n-gram, increasing high level of n-gram will help to increase much in the results. In this study, we chose n-gram = 5 with the best metrics (accuracy of 97.5 and 92.8% in the cross-validation and independent test, respectively) to perform further experiments.

In most of the supervised learning problems, our model can perform well during training test, but worse in another invisible data. This is called overfitting and our study, no exception also included in this issue. Therefore, an independent test was used in our study to ensure that our model also works well in a blind dataset with unseen data. As described in the previous part, our independent dataset contained 4,465 SNAREs and 4,465 non-SNAREs. None of these samples occur in the training set. As shown in **Table 1**, our independent testing results also comply with cross-validation results in most metrics. To detail, our independent testing performance achieved the accuracy of 92.8%, sensitivity of 88.5%, specificity of 97%, and MCC of 0.86. There is a very few overfitting in our model and it can demonstrate that our model has been well done in this type of dataset. Another reason is the use of dropout inside CNN structure and it helps us prevent overfitting.

Comparative Performance Between Proposed Method and the Existing Methods

From the previous section, we chose the combination of 1D CNN and 5-gram as our optimal model for SNARE identification. In this section, we aim to compare the effectiveness of our proposed features with other research groups studying the same problem. As mentioned in the literature review, there have been some published works on identifying SNARE proteins using computational techniques. However, among of them, there is only one predictor to propose the machine learning techniques on predicting SNARE (Le and Nguyen, 2019). Therefore, we compared our performance with them in both cross-validation and independent test. **Table 2** shows the performance results by highlighting the higher values for each metrics. It is clear that on average, our method outperforms the previous model in all measurement metrics. Therefore, we are able to generate effective features for identifying SNAREs with a better performance than PSSM profiles which had been used in the previous work.

DISCUSSION

Based on the outstanding results of word embeddings in NLP, applying it to protein function prediction is an essential concern for biological researchers. In this study, we have approached a method using word embedding and deep learning for identifying SNARE proteins. Our structure is a combination between fastText (to train vectors model) and 1D CNN (to train deep learning model from the generated vectors). By using fastText, the protein sequences have been interpreted via different

TABLE 1 | Performance results on identifying SNAREs with different n-gram levels.

	Cross validation				Independent			
n-gram	**Sens**	**Spec**	**Acc**	**MCC**	**Sens**	**Spec**	**Acc**	**MCC**
1	83.8	88.7	86.3	0.73	39.4	94.6	67	0.41
2	93.7	91.6	92.6	0.85	83.1	87.4	85.2	0.71
3	95.8	97.6	96.7	0.93	87.4	95	91.2	0.83
4	96.7	98.1	97.4	0.95	88.7	96.4	92.6	0.85
5	96.6	98.4	97.5	0.95	88.5	97	92.8	0.86

TABLE 2 | Comparative performance of predicting SNAREs between the proposed method and the previous published work.

	Cross validation				Independent			
Predictor	**Sens**	**Spec**	**Acc**	**MCC**	**Sens**	**Spec**	**Acc**	**MCC**
SNARE-CNN	76.6	93.5	89.7	0.7	65.8	90.3	87.9	0.46
Ours	**96.6**	**98.4**	**97.5**	**0.95**	**88.5**	**97**	**92.8**	**0.86**

The bold values is to show the significant values for each metric.

representations and we could generate the hidden information of them. While the other NLP models do not have sub-word information, it is an advantage of fastText that can help to improve this problem. Benefits of fastText when comparing to the other features have been also proven in the previous works based on their results (Do and Khanh Le, 2019; Le, 2019; Le et al., 2019b). We used 5-fold cross-validation set to train our model and an independent set to examine the performance results. Compared to the state-of-the-art predictor, our method produced superior performance in all the typical measurement metrics. Through this study, biologists can use our model to identify SNARE proteins with high accuracy and use them as necessary information for drug development. In addition, we contribute a method to interpret the information of protein sequences and further research is able to apply in bioinformatics research, especially in protein function prediction.

Furthermore, we provided our source codes and datasets at https://github.com/khanhlee/fastSNARE. The readers and biologists are able to reproduce our results as well as perform their classifications according our method. We also hope that our future research would be able to provide a web-server for the method of prediction as presented in this paper. Moreover, a limitation of using language model is that it could not consider mutations and SNPs in SNARE sequence. Therefore, further studies could integrate these information into fastText model to improve the predictive performance.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

Both authors conceived the ideas, designed the study, participated in the discussion of the results and writing of the manuscript, and read and approved the final version of the manuscript. NL conducted the experiments and analyzed the results.

FUNDING

This work was supported in part by the Ministry of Science and Technology of the Republic of China under Grant MOST 109-2811-E-155-501.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of Nvidia Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (eds) (2016). "Tensorflow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, Savannah, GA.

Asgari, E., McHardy, A. C., and Mofrad, M. R. K. (2019). Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Sci. Rep.* 9:3577. doi: 10.1038/s41598-019-38746-w

Asgari, E., and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 10:e0141287. doi: 10.1371/journal.pone.0141287

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5, 135–146. doi: 10.1162/tacl_a_00051

Chollet, F., (2015). *Keras*. Available at: https://github.com/fchollet/keras (accessed November 20, 2018).

Chou, K.-C. (2001). Using subsite coupling to predict signal peptides. *Protein Eng.* 14, 75–79. doi: 10.1093/protein/14.2.75

Coordinators, N. R. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 44, D7–D19. doi: 10.1093/nar/gkv1290

Do, D. T., and Khanh Le, N. Q. (2019). A sequence-based approach for identifying recombination spots in *Saccharomyces cerevisiae* by using hyper-parameter optimization in fastText and support vector machine. *Chemometr. Intell. Lab. Syst.* 194:103855. doi: 10.1016/j.chemolab.2019.103855

Duman, J. G., and Forte, J. G. (2003). What is the role of SNARE proteins in membrane fusion? *Am. J. Physiol. Cell Physiol.* 285, C237–C249. doi: 10.1152/ajpcell.00091.2003

Dwork, A. J., Li, H.-Y., Mann, J. J., Xie, J., Hu, L., Falkai, P., et al. (2002). Abnormalities of SNARE mechanism proteins in anterior frontal cortex in severe mental illness. *Cereb. Cortex* 12, 349–356. doi: 10.1093/cercor/12.4.349

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gissen, P., Johnson, C. A., Morgan, N. V., Stapelbroek, J. M., Forshew, T., Cooper, W. N., et al. (2004). Mutations in VPS33B, encoding a regulator of SNARE-dependent membrane fusion, cause arthrogryposis–renal dysfunction–cholestasis (ARC) syndrome. *Nat. Genet.* 36, 400–404. doi: 10.1038/ng1325

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., and Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 33, i37–i48. doi: 10.1093/bioinformatics/btx228

Hou, C., Wang, Y., Liu, J., Wang, C., and Long, J. (2017). Neurodegenerative disease related proteins have negative effects on SNARE-Mediated membrane fusion in pathological confirmation. *Front. Mol. Neurosci.* 10:66. doi: 10.3389/fnmol.2017.00066

Jahn, R., and Scheller, R. H. (2006). SNAREs — engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* 7, 631–643. doi: 10.1038/nrm2002

Jeans, A. F., Oliver, P. L., Johnson, R., Capogna, M., Vikman, J., Molnár, Z., et al. (2007). A dominant mutation in Snap25 causes impaired vesicle trafficking, sensorimotor gating, and ataxia in the blind-drunk mouse. *Proc. Natl. Acad. Sci.U.S.A.* 104, 2431–2436. doi: 10.1073/pnas.0610222104

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (eds) (2017). "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2*, Valencia.

Kienle, N., Kloepper, T. H., and Fasshauer, D. (2009). Phylogeny of the SNARE vesicle fusion machinery yields insights into the conservation of the secretory pathway in fungi. *BMC Evol. Biol.* 9:19. doi: 10.1186/1471-2148-9-19

Kloepper, T. H., Kienle, C. N., and Fasshauer, D. (2008). SNAREing the basis of multicellularity: consequences of protein family expansion during evolution. *Mol. Biol. Evol.* 25, 2055–2068. doi: 10.1093/molbev/msn151

Kloepper, T. H., Kienle, C. N., Fasshauer, D., and Munro, S. (2007). An elaborate classification of SNARE proteins sheds light on the conservation of the eukaryotic endomembrane system. *Mol. Biol. Cell* 18, 3463–3471. doi: 10.1091/mbc.e07-03-0193

Le, N. Q. K. (2019). iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. *Mol. Genet. Genomics* 294, 1173–1182. doi: 10.1007/s00438-019-01570-y

Le, N. Q. K., Ho, Q. T., and Ou, Y. Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Comput. Chem.* 38, 2000–2006. doi: 10.1002/jcc.24842

Le, N. Q. K., Ho, Q.-T., and Ou, Y.-Y. (2018). Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. *Anal. Biochem.* 555, 33–41. doi: 10.1016/j.ab.2018.06.011

Le, N. Q. K., Huynh, T.-T., Yapp, E. K. Y., and Yeh, H.-Y. (2019a). Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. *Comput. Methods Prog. Biomed.* 177, 81–88. doi: 10.1016/j.cmpb.2019.05.016

Le, N. Q. K., Yapp, E. K. Y., Ho, Q.-T., Nagasundaram, N., Ou, Y.-Y., and Yeh, H.-Y. (2019b). iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* 571, 53–61. doi: 10.1016/j.ab.2019.02.017

Le, N. Q. K., Yapp, E. K. Y., Ou, Y.-Y., and Yeh, H.-Y. (2019c). iMotor-CNN: identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. *Anal. Biochem.* 575, 17–26. doi: 10.1016/j.ab.2019.03.017

Le, N. Q. K., and Nguyen, V. N. (2019). SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data. *PeerJ Comput. Sci.* 5:e177. doi: 10.7717/peerj-cs.177

Meng, J., and Wang, J. (2015). Role of SNARE proteins in tumourigenesis and their potential as targets for novel anti-cancer therapeutics. *Biochim. Biophys. Acta* 1856, 1–12. doi: 10.1016/j.bbcan.2015.04.002

Nguyen, T.-T.-D., Le, N.-Q.-K., Kusuma, R. M. I., and Ou, Y.-Y. (2019). Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network. *J. Mol. Graph. Model.* 92, 86–93. doi: 10.1016/j.jmgm.2019.07.003

Öztürk, H., Ozkirimli, E., and Özgür, A. (2018). A novel methodology on distributed representations of proteins using their interacting ligands. *Bioinformatics* 34, i295–i303. doi: 10.1093/bioinformatics/bty287

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35(Suppl._1), D61–D65. doi: 10.1093/nar/gkl842

Sprecher, E., Ishida-Yamamoto, A., Mizrahi-Koren, M., Rapaport, D., Goldsher, D., Indelman, M., et al. (2005). A mutation in SNAP29, coding for a SNARE protein involved in intracellular trafficking, causes a novel neurocutaneous syndrome characterized by cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma. *Am. J. Hum. Genet.* 77, 242–251. doi: 10.1086/432556

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Sun, Q., Huang, X., Zhang, Q., Qu, J., Shen, Y., Wang, X., et al. (2016). SNAP23 promotes the malignant process of ovarian cancer. *J. Ovarian Res.* 9:80. doi: 10.1186/s13048-016-0289-289

Ungermann, C., and Langosch, D. (2005). Functions of SNAREs in intracellular membrane fusion and lipid bilayer mixing. *J. Cell Sci.* 118, 3819–3828. doi: 10.1242/jcs.02561

van Dijk, A. D. J., van der Krol, A. R., ter Braak, C. J. F., Bosch, D., and van Ham, R. C. H. J. (2008). Predicting sub-Golgi localization of type II membrane proteins. *Bioinformatics* 24, 1779–1786. doi: 10.1093/bioinformatics/btn309

Vang, Y. S., and Xie, X. (2017). HLA class I binding prediction via convolutional neural networks. *Bioinformatics* 33, 2658–2665. doi: 10.1093/bioinformatics/btx264

Wickner, W., and Schekman, R. (2008). Membrane fusion. *Nat. Struct. Mol. Biol.* 15, 658–664. doi: 10.1038/nsmb.1451

Yoshizawa, A. C., Kawashima, S., Okuda, S., Fujita, M., Itoh, M., Moriya, Y., et al. (2006). Extracting sequence motifs and the phylogenetic features of SNARE-Dependent membrane traffic. *Traffic* 7, 1104–1118. doi: 10.1111/j.1600-0854.2006.00451.x

Zheng, C., Lanczycki, C. J., Zhang, D., Hurwitz, D. I., Chitsaz, F., Lu, F., et al. (2014). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, D222–D226. doi: 10.1093/nar/gku1221

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 08 January 2020
doi: 10.3389/fphys.2019.01551

Image-Based Network Analysis of DNp73 Expression by Immunohistochemistry in Rectal Cancer Patients

Tuan D. Pham[1,2], Chuanwen Fan[3,4], Daniella Pfeifer[3], Hong Zhang[5] and Xiao-Feng Sun[3]**

[1] Department of Biomedical Engineering, Linkoping University, Linkoping, Sweden, [2] The Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia, [3] Department of Oncology, Clinical and Experimental Medicine, Linkoping University, Linkoping, Sweden, [4] Institute of Digestive Surgery, West China Hospital, Sichuan University, Chengdu, China, [5] Department of Medical Sciences, Orebro University, Orebro, Sweden

Background: Rectal cancer is a disease characterized with tumor heterogeneity. The combination of surgery, radiotherapy, and chemotherapy can reduce the risk of local recurrence. However, there is a significant difference in the response to radiotherapy among rectal cancer patients even they have the same tumor stage. Despite rapid advances in knowledge of cellular functions affecting radiosensitivity, there is still a lack of predictive factors for local recurrence and normal tissue damage. The tumor protein DNp73 is thought as a biomarker in colorectal cancer, but its clinical significance is still not sufficiently investigated, mainly due to the limitation of human-based pathology analysis. In this study, we investigated the predictive value of DNp73 in patients with rectal adenocarcinoma using image-based network analysis.

Methods: The fuzzy weighted recurrence network of time series was extended to handle multi-channel image data, and applied to the analysis of immunohistochemistry images of DNp73 expression obtained from a cohort of 25 rectal cancer patients who underwent radiotherapy before surgery. Two mathematical weighted network properties, which are the clustering coefficient and characteristic path length, were computed for the image-based networks of the primary tumor (obtained after operation) and biopsy (obtained before operation) of each cancer patient.

Results: The ratios of two weighted recurrence network properties of the primary tumors to biopsies reveal the correlation of DNp73 expression and long survival time, and discover the non-effective radiotherapy to a cohort of rectal cancer patients who had short survival time.

Conclusion: Our work contributes to the elucidation of the predictive value of DNp73 expression in rectal cancer patients who were given preoperative radiotherapy. Mathematical properties of fuzzy weighted recurrence networks of immunohistochemistry images are not only able to show the predictive factor of DNp73 expression in the patients, but also reveal the identification of non-effective application of radiotherapy to those who had poor overall survival outcome.

Keywords: fuzzy weighted recurrence networks, network properties, multi-channel images, DNp73, immunohistochemistry, predictive biomarker, rectal cancer, survival outcome

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Neda Slade,
Rudjer Boskovic Institute, Croatia
Sanjay Ram Kharche,
University of Western Ontario, Canada
Stella Logotheti,
University of Rostock, Germany

****Correspondence:***
Tuan D. Pham
tuan.pham@liu.se;
tpham@pmu.edu.sa
Xiao-Feng Sun
xiao-feng.sun@liu.se

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *21 December 2019*
Accepted: *09 December 2019*
Published: *08 January 2020*

Citation:
Pham TD, Fan C, Pfeifer D, Zhang H and Sun X-F (2020) Image-Based Network Analysis of DNp73 Expression by Immunohistochemistry in Rectal Cancer Patients. Front. Physiol. 10:1551. doi: 10.3389/fphys.2019.01551

1. INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer in the world (Arnold et al., 2017). There are around 6,500 new diagnosed cases of CRC yearly among the population of Sweden. The causes for CRC are considered to be associated with gene mutations, gene variants, and changed expression of proteins. The combination of surgery and radio-chemotherapy is the most beneficial regimens in current treatment of advanced rectal cancer. Preoperative radiotherapy (RT) is often given to rectal cancer patients as a complement to surgery to improve treatment outcome. However, tumor recurrence plays a major cause of death for progressive rectal patients after surgery. As a result, a significant proportion of patients did not benefit from preoperative RT (Fan et al., 2013).

It remains up to date that the clinical testing of specific mutations in KRAS, BRAF, RAS, and RAF genes along with mismatch repair gene deficiency assists either as prognostic or predictive biomarkers in CRC (Sinicrope et al., 2016; Zarkavelis et al., 2017). Other methods for identifying biomarkers in the treatment of CRC include molecular subtype classification (Cuyle and Prenen, 2017), identifying molecular signatures at protein and RNA levels by microarray analysis (Rahman et al., 2019), protein identification in cell proliferation and new blood vessels (Chatterjee et al., 2019), changes in the amounts of certain proteins (Letellier et al., 2017), and proteomic strategies (Lee et al., 2018). A recent review of methods for discovering prognostic and predictive biomarkers in CRC for personalized therapy can be found in Patel et al. (2019).

The role of predictive biomarkers is known to be essential for the field of radiation oncology (Yaromina et al., 2012). While most efforts aim to improve cancer treatment with respect to physical conditions and technology, such as precision in treatment plans and dose administration (Sonke and Belderbos, 2010), the inclusion of patient-specific biological characteristics into cancer treatment decision would be very useful for personalized treatment. However, such important biological information of individual patients is not well explored. To achieve this purpose, predictive biomarkers are needed to guide radiation oncologists to determine optimal dose prescription, select patient-specific schemes, and treatments for individual cancer patients (Yaromina et al., 2012).

However, it is a big challenge to find predictive biomarkers that can select patients who can benefit from RT, although our and other groups have spent much effort to identify potential predictors for the RT response (Ryan et al., 2016; Ye and Guo, 2019). A previous study of our group suggested that p73 independently predicted poor prognosis in colorectal cancer and p73-negative tumors tended to have a lower local recurrence after RT compared with unirradiated case (Ye and Guo, 2019).

One of important reasons is that the TP73 gene expresses isoforms with divergent and/or opposing roles in cancer. These are mainly categorized in two classes, the anti-oncogenic TAp73 isoforms, which contain an intact N-terminal, transactivation domain, and the oncogenic DNp73 isoforms, which lack part or whole of the transactivation domain and act as dominant negative forms of TAp73 proteins (Logotheti et al., 2013). The TAp73 isoforms are generated from an external P1 promoter. The DNp73 proteins are transcribed (a) by the P1 promoter, followed post-transcriptionally by alternative splicing in exons 2 and/or 3 at the 5′ end (Stiewe et al., 2002), or (b) by an alternative, internal P2 promoter which generates variants lacking exons 2 and 3, but instead containing an exon 3′ that encodes for a unique 13-amino acid domain (Irwin, 2006). Additional complexity is created by alternative splicing in the 3′ end, which gives rise to a large number of C-terminal variants of the abovementioned isoforms (Logotheti et al., 2013). Altogether, the TP73 gene expresses at least 35 mRNA variants, which can encode theoretically 29 different p73 protein isoforms (Murray-Zmijewski et al., 2006). Notably, the ratio between TAp73 and DNp73 isoforms has essential effects on the cellular response (Dulloo et al., 2010; Rufini et al., 2011).

The imbalance between TAp73 and DNp73 isoforms may be useful to predict response to chemotherapy and prognosis (Muller et al., 2005; Lucena-Araujo et al., 2015). High DNp73 expression has strong correlation with unfavorable prognosis in several types of cancer patients, and DNp73-positive tumors show a reduced response to chemotherapy and irradiation (Uramoto et al., 2004; Di et al., 2013; Zhu et al., 2015). The upregulation of DNp73 was frequently detected in radioresistant cervical cancers (Liu et al., 2006). Our previous findings indicated that DNp73 is increased in colon cancer cell line that is resistant to γ-irradiation (Pfeifer et al., 2009). Thus, these findings suggested that DNp73 expression may play an important role in the regulation of radiosensitivity. However, the prognostic and preditive role of DNp73 in rectal cancer patients with radiation still remains unclear.

This study aimed to elucidate the role of DNp73 as a predictive biomarker by investigating if DNp73 was related to the survival time of rectal cancer patients who were administered with RT before surgery. To overcome the subjective and time-consuming task of pathologist-based analysis of immunohistochemistry (IHC) images stained for DNp73 expression, we carried out a study by means of a novel image-based recurrence network approach. The motivation for developing this new image-based network analysis was based on the recurrence of image attributes inherently existing in the complex nature of IHC images of rectal cancer tissue arrays.

In fact, network analysis in graph theory has been increasingly recognized as a useful tool for studying cancer. Such studies include the prediction of outcomes of ovarian cancer treatment (Zhang et al., 2013), analysis of breast cancer progression and reversal (Parikh et al., 2014), drug response prediction in cancer cell lines (Zhang et al., 2018), identification of novel cancer gene candidates (Josef Gladitz et al., 2018), tumor biology for precision cancer medicine (Ozturk et al., 2018), and prediction of cancer recurrence (Ruan et al., 2019).

In this present study, we introduce a new method of fuzzy weighted recurrence networks of multi-channel images for computing useful properties of the complex networks of the expression patterns of the DNp73 IHC. The ratios of these network properties discover the predictive value of DNp73 in rectal cancer patients in the Swedish Rectal Cancer Trial.

TABLE 1 | Demographic information of the rectal cancer patients who had a median age of 68 years (range: 39–78 years), were followed for a median period of 81 months (range: 0–129 months), and had the median time to disease free of 101 months after surgery (range: 15–288 months).

	Number of patients
Male	16 (64%)
Female	9 (36%)
Shorter survival time (15–75 months)	11 (44%)
Longer survival time (101–288 months)	14 (56%)

TABLE 2 | Clinico-pathological characteristics of the rectal cancer patients.

Parameters		Number of cases
Age	<60	6
	>60	19
Gender	Male	9
	Female	16
Growth pattern	Expansion	11
	Infiltration	13
	Null	1
Grade	Well	2
	Moderate	14
	Poor	9
Pathological stages	I	8
	II	6
	III	8
	IV	3

2. MATERIALS AND METHODS

2.1. Rectal Cancer Patients

This study included the patients with rectal adenocarcinoma from the Southeast Swedish Health Care region who participated in a clinical trial of preoperative RT for rectal cancer (Swedish Rectal Cancer Trial et al., 1997). Samples of biopsy and primary tumor from the same patients were selected for the analysis. In this Swedish Rectal Cancer Trial study, we collected samples from both pre-radiotherapy and non-radiotherapy rectal cancer patients. The biopsy samples were taken from the rectal cancer before the RT and went through the routine pathological process, and eventually embedded in paraffin blocks. The primary tumor samples were taken from the primary rectal cancer after the RT.

There were 25 patients with RT whose demographic information is given in **Table 1**. This study was carried out in accordance with the recommendations of Good Clinical Practice, the Research Ethics Committee in Linkoping, Sweden with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Research Ethics Committee in Linkoping, Sweden. The clinico-pathologic characteristics of the patients are listed in **Table 2**.

2.2. Immunochemistry and Image Extraction

The five-micrometer paraffin-embedded tissue micro-array (TMA) sections were deparaffinized in xylene and rehydrated with a series of gradient ethanol to water. The sections were heated to boiling point in citrate buffer (pH 6.0) for 30 min to unmasked antigen, followed by a washing in phosphate-buffered saline (PBS). Endogenous peroxidase activity was blocked with 3% H_2O_2 in methanol followed by washing three-times in PBS. The sections were incubated with protein block (Dako, Carpinteria, CA) for 10 min and then incubated with anti-DNp73 antibody (clone 38C674.2, Novus Biologicals, 1:200), which specifically recognized DNp73 isoforms, but not TAp73.

After that, the sections were washed in PBS and then incubated with goat anti-mouse secondary antibody (Dako) at room temperature for 25 min. Next, the sections were subjected to 3,3′-diaminobenzidine tetrahydrochloride for 8 min and then counterstained with hematoxylin. Negative and positive controls were added in each staining run. All slides were scored by two independent investigators. Whole-slide images of entire sections were captured with an Aperio CS2 slide scanner system (Leica Biosystems, Wetzlar, Germany) using a 40x magnification.

All sections were reviewed to remove images containing tissue-processing artifacts, including bubbles, section folds and poor staining. A total of 46 whole-slide images from the 25 unique patients were extracted from the TMA slides using ObjectiveViewer (https://www.objectivepathology.com/objectiveview) with the original resolution.

2.3. Multi-Channel Fuzzy Weighted Recurrence Networks

The term "channel" is a conventional expression used to refer to a certain component of an image. For example, an RGB image has 3 channels that are red (R), green (G) and blue (B) components. A grayscale image has only one channel. Let $\mathbf{I} = [f_{ijk}]$ be a multi-channel image of size $M \times N \times K$, where $i = 1, \ldots, M$, $j = 1, \ldots, N$, and $k = 1, \ldots, K$. Let $m \geq 1$ be an integer, a local image window $\mathbf{W}_{ij}^k \in \mathbf{I}$ of size $(2m+1) \times (2m+1)$ is constructed for each pixel located at ij in each of the k components of the multi-channel image, where ij is the center of the window. This window can be considered as embedding dimensions in two-dimensional space, which considers the local spatial distribution around f_{ij} of the k-th image channel. The Frobenius norm can be used to transform each local window into a scalar measure that has the useful property of invariance under rotations as

$$\|\mathbf{W}_{ij}^k\|_F = \sqrt{\sum_{i-m}^{i+m}\sum_{j-m}^{j+m} |f_{ijk}|^2}, \quad (1)$$

where $(i-m), (j-m) > 0, (i+m) \leq M, (j+m) \leq N$, and any pixel at the center of the window that requires values from beyond the image boundaries is skipped.

We can then obtain a set of feature vectors $\mathbf{y}_{ij}, (i-m), (j-m) > 0$, by joining the Frobenius norms computed for each window of

the k-th image channel at the same location, for example, a color image of 3 channels:

$$\mathbf{y}_{ij} = \left(\|\mathbf{W}_{ij}^1\|_F, \|\mathbf{W}_{ij}^2\|_F, \|\mathbf{W}_{ij}^3\|_F\right), \quad (2)$$

where $(i-m),(j-m) > 0, (i+m) \leq M, (j+m) \leq N$.

Since the Frobenius norm induced feature vector set $\mathbf{y}_{ij}$ can be computed for the multi-channel image $\mathbf{I}$, the multi-channel fuzzy weighted recurrence network (MC-FWRN), which is an extension of the FWRN of time series (Pham, 2019), can be constructed as follows. To simplify the notation in subsequent mathematical presentation, $\mathbf{y}_{ij}$ is now denoted as $\mathbf{x}_n$, $n = 1, \ldots, L$, where L is the total number of feature vectors, and some same indices are used but defined differently.

Let $\mathbf{X} = \{\mathbf{x}_n\}, n = 1, \ldots, L$, c a given number of clusters of the feature space, and a set of c fuzzy clusters, $\mathbf{V} = \{\mathbf{v}_i : i = 1, \ldots, c\}$. Fuzzy clusters are groups that contain data points, where every data point has a degree of fuzzy membership of belonging to each group. A fuzzy relation $\tilde{\mathbf{R}}$ between $\mathbf{v}_i$ and $\mathbf{v}_j$, $i, j = 1, \ldots, c$, is characterized by a fuzzy membership function $\mu \in [0, 1]$, which expresses the degree of similarity of each pair $(\mathbf{v}_i, \mathbf{v}_j)$ in $\tilde{\mathbf{R}}$. This fuzzy relation has the following three properties (Zadeh, 1971):

1. Reflexivity: $\mu(\mathbf{v}_i, \mathbf{v}_i) = 1, \forall \mathbf{v}_i \in \mathbf{V}$.
2. Symmetry: $\mu(\mathbf{v}_i, \mathbf{x}_n) = \mu(\mathbf{x}_n, \mathbf{v}_i), \forall \mathbf{x}_n \in \mathbf{X}, \forall \mathbf{v}_i \in \mathbf{V}$.
3. Transitivity: $\mu(\mathbf{v}_i, \mathbf{v}_j) = \vee_{\mathbf{x}_n} [\mu(\mathbf{v}_i, \mathbf{x}_n) \wedge \mu(\mathbf{v}_j, \mathbf{x}_n)], \forall \mathbf{x}_n \in \mathbf{X}, \forall \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}$, where the symbols $\vee$ and $\wedge$ stand for max and min, respectively.

The computation of $\mu(\mathbf{v}_i, \mathbf{x}_n), i = 1, \ldots, c, n = 1, \ldots, L$, which are necessary for the construction of the fuzzy relation $\tilde{\mathbf{R}}$ can be carried out by means of the fuzzy c-means (FCM) algorithm (Bezdek, 1981) as follows.

Let μ_{nj} denote a fuzzy membership grade of $\mathbf{x}_n, n = 1, \ldots, L$, which belongs to a cluster $j, j = 1, \ldots, c$, whose center is $\mathbf{v}_j$. This fuzzy membership is calculated by the FCM as

$$\mu_{nj} = \frac{1}{\sum_{i=1}^{c} \left[\frac{d(\mathbf{x}_n, \mathbf{v}_j)}{d(\mathbf{x}_n, \mathbf{v}_i)}\right]^{2/(\alpha-1)}}, \quad (3)$$

where $1 \leq \alpha < \infty$ is the weighting exponent, and $d(\mathbf{x}_n, \mathbf{v}_j)$ is used as a Euclidean distance between $\mathbf{x}_n$ and $\mathbf{v}_j$.

Using the fuzzy membership grades, each cluster center $\mathbf{v}_j$ is computed as

$$\mathbf{v}_j = \frac{\sum_{n=1}^{L} (\mu_{nj})^{\alpha} \mathbf{x}_n}{\sum_{n=1}^{L} (\mu_{nj})^{\alpha}}, \forall j. \quad (4)$$

The iterative procedure of the FCM is outlined as follows.

1. Given c, α, step t, $t = 0, \ldots, T$, initialize matrix $\mathbf{U}^{(t=0)} = [\mu_{nj}]^{(t=0)}$
2. Compute $\mathbf{v}_j^{(t)}, j = 1, \ldots, c$, using Equation (4).
3. Update $\mathbf{U}^{(t+1)}$ using Equation (3).
4. If $\|\mathbf{U}^{(t+1)} - \mathbf{U}^{(t)}\| < \epsilon$ or $t = T$, stop. Otherwise, set $\mathbf{U}^{(t)} = \mathbf{U}^{(t+1)}$ and return to step 2.

The predefined FCM parameters α, T and ϵ usually take the values of 2, 100, and 0.00001, respectively. The number of clusters can be estimated using a cluster validity measure such as the partition entropy, denoted by H, which is defined as (Bezdek, 1981)

$$H = \frac{1}{L} \sum_{j=1}^{c} \sum_{n=1}^{L} \mu_{nj} \log(\mu_{nj}). \quad (5)$$

This cluster validity works by computing the partition entropy H for a range of a given number of clusters, $c \geq 2$, and the number of clusters that has the minimum value of H is considered as an optimal c for the FCM algorithm.

Finally, an $N \times N$ MC-FWRN can be constructed with the fuzzy relation $\tilde{\mathbf{R}}$ as

$$\mathbf{W} = \tilde{\mathbf{R}} - \mathbf{I}, \quad (6)$$

where $\mathbf{W}$ is an $N \times N$ adjacency matrix of edge weights, and $\mathbf{I}$ is the $N \times N$ identity matrix. The interested reader is referred to the work described in Pham (2019) to obtain more detailed information about the concept of fuzzy weighted recurrence networks originally developed for time series.

2.4. Network Properties

Two most well-known measures of the statistical characterization of a complex network are the average clustering coefficient and characteristic path length (Watts and Strogatz, 1998; Albert and Barabasi, 2002; Barrat et al., 2004). The clustering coefficient of a node in a network is a numerical indicator of a node that tends to cluster with other neighboring nodes. The average clustering coefficient expresses the average amount of connectivity around individual nodes of a network, whereas the characteristic path length is considered as a measure of the efficiency of transfer of information in a network.

The average clustering coefficient for an unweighted network represented with an $N \times N$ (binary) adjacency matrix $\mathbf{A} = [a_{ij}]$, $i, j = 1, \ldots, N$, is defined as

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i, \quad (7)$$

where C_i is the local unweighted clustering coefficient for node i, and defined as

$$C_i = \frac{\sum_{j,k} a_{ij} a_{jk} a_{ki}}{k_i(k_i - 1)}, \; k_i \neq 0, 1, \quad (8)$$

where k_i is the degree of node i, which is the number of links of node i.

The average clustering coefficient for a weighted network is defined as

$$CC = \frac{1}{N} \sum_{i=1}^{N} C_i^w, \quad (9)$$

where C_i^w is the local weighted clustering coefficient for node i, and defined as (Fagiolo, 2007)

$$C_i^w = \frac{\sum_{j,k}[w_{ij}w_{ik}w_{jk}]^{1/3}}{k_i(k_i-1)}, k_i \neq 0, 1, \quad (10)$$

where $w_{ij}, w_{ik}, w_{jk} \in W$.

In general, the clustering coefficient of a node is the ratio of existing links connecting a node's neighbors to each other to the maximum possible number of such links. The clustering coefficient for the entire network is the average of the clustering coefficients of all the nodes.

The characteristic path length of a network is defined as the average of all shortest path lengths:

$$CP = \frac{1}{N(N-1)} \sum_{i \neq j, i,j=1}^{N} d_{ij}, \quad (11)$$

where d_{ij} is the length of the shortest path between nodes i and j. The Dijkstra's algorithm (Newman, 2010) was used for computing the shortest weighted path in this study. The characteristic path length is calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs. This operation shows on average the number of steps it takes to get from one node of the network to another.

2.5. Algorithm for Computing Network Properties From MC-FWRN

1. Given a multi-channel image **I**, window parameter m, number of clusters c, and FCM parameters.
2. Using Equation (1) to compute the Frobenius norm for each window $(2m+1) \times (2m+1)$ of each image channel, and using Equation (2) to form a matrix of vectors of length 3 with the number of pixels that can be used to construct the windows.
3. Compute the fuzzy weighted adjacency matrix **W** using Equation (6) via the FCM.
4. Using **W** to calculate the clustering coefficient with Equation (9), and the characteristic path length with Equation (11).

3. RESULTS AND DISCUSSION

Table 3 shows the screening results of the 25 rectal cancer patients. Patient numbers 1-11 are those who had shorter survival time, and patient numbers 12–25 are those who had longer survival time. The evaluation of the IHC-stained color intensity of the whole slide of a tissue core with brown antibody stain and blue counter-stain were assessed as being positive and negative, respectively. The positive stain is subjectively classified as weak = 1 (light brown), moderate = 2 (moderate brown), and strong = 3 (dark brown), whereas the negative stain = 0 (blue). **Figure 1** shows representative IHC staining for DNp73 expression on the biopsy and primary tumor tissue images obtained from a rectal cancer patient survived 40 months after radiotherapy, and biopsy and primary tumor tissue images obtained from a rectal cancer patient who survived 255 months after radiotherapy at the censoring date.

TABLE 3 | Screening results of rectal cancer patients.

Patient #	Disease-free time	Recurrence status	Survival time	IHC score	
				Primary tumor	Biopsy
1	0	Yes	15	1	1
2	6	Yes	19	1	3
3	20	Yes	25	2	1
4	13	Yes	40	3	3
5	37	Yes	60	3	3
6	44	Yes	62	2	3
7	0	Yes	15	3	2
8	63	Yes	75	3	3
9	26	Yes	27	3	3
10	12	No	26	3	3
11	34	Yes	43	2	2
12	100	Yes	101	1	1
13	0	Yes	180	1	3
14	114	Yes	114	2	2
15	122	Yes	255	3	2
16	167	Yes	167	2	2
17	81	No	101	3	2
18	129	Yes	129	2	3
19	129	Yes	288	2	3
20	126	Yes	126	3	2
21	186	Yes	238	3	3
22	122	Yes	122	2	2
23	168	Yes	288	2	3
24	151	Yes	151	3	3
25	168	Yes	168	2	2

Time is in months. For IHC score, 0 = negative, 1 = weak, 2 = moderate, and 3 = strong.

To capture the local information of the DNp73 expression over the whole IHC-stained slides, images of biopsy and primary tumor of each of the 25 rectal cancer patients were divided into subimages of 150×150 pixels. The subimages that contain either the background or a large portion of the background were excluded in the analysis. To construct the FWRNs of the IHC-stained subimages, we selected the FWRN parameters m=3 to establish a reasonable local window size of 7 × 7, c = 20 that was approximately based on the partition entropy, and the FCM parameters α = 2, T = 100, and ϵ = 0.00001, which are widely adopted for the FCM analysis. The clustering coefficient and characteristic path length were calculated for each subimage of each patient, and the total average values of the clustering coefficients and characteristic path lengths of all subimages represent the reported values.

Figures 2, **3** show the clustering coefficients and characteristic path lengths of the FWRNs of the biopsy and primary tumor images obtained from the 25 rectal cancer patients, respectively. The scatter plot of the survival time against the ratios of the clustering coefficients of the primary tumors to those of the biopsies, and the ratios of the characteristic path lengths of the primary tumors to those of the biopsies are shown in **Figure 4**.

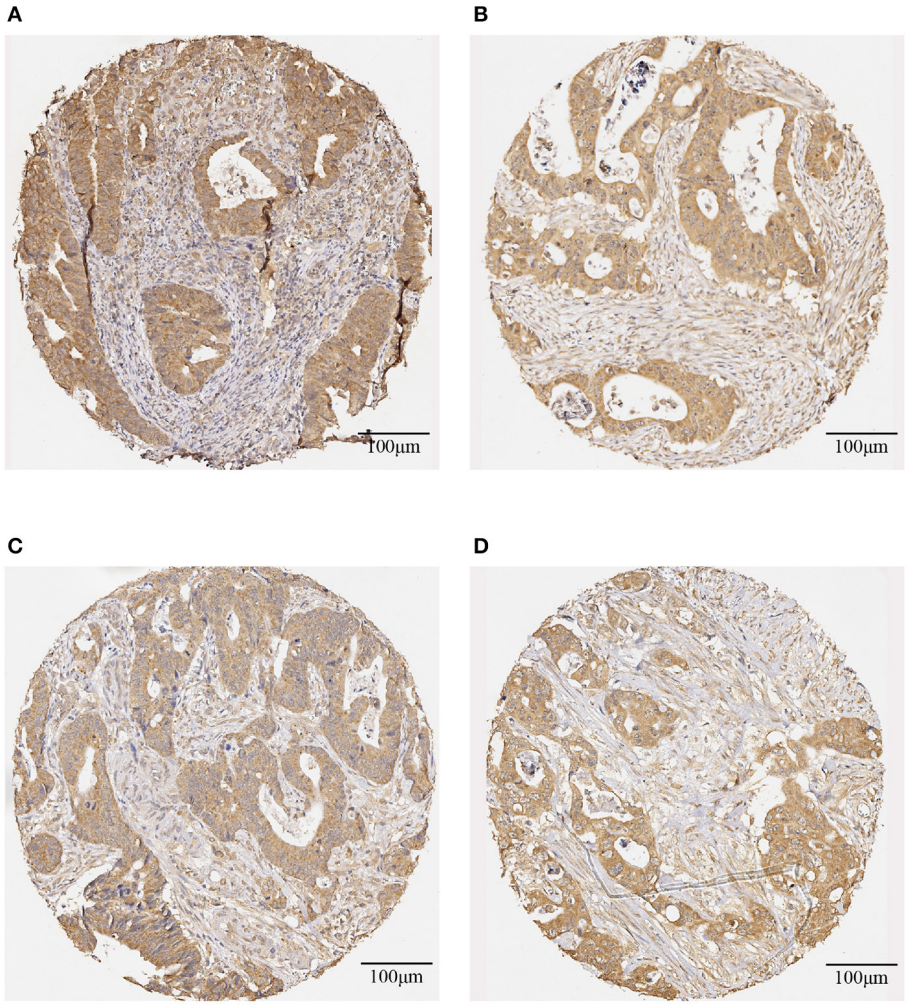

FIGURE 1 | Representative IHC-stained images of DNp73 expression: **(A)** a biopsy image and **(B)** a primary tumor image obtained from a rectal cancer patient survived 40 months after radiotherapy; and **(C)** a biopsy image and **(D)** a primary tumor image obtained from a rectal cancer patient survived 255 months after radiotherapy.

Based on the visualization of the scatter plot, we discovered the predictive value of DNp73 in the rectal cancer patients in terms of the clustering-coefficient and characteristic-path-length ratios, which are shown in **Figure 5**. The probability (p) for the predicted survival time based on the clustering-coefficient ratio was computed as the number of patients who lived between 101 and 288 months divided by the total number of patients whose clustering-coefficient ratios are within the ratio range (p = 11/15 = 0.7333). The probability (p) for the predicted survival time based on the characteristic-path-length ratio was computed as the number of patients who lived between 126 and 288 months divided by the total number of patients whose characteristic-path-length ratios are within the ratio range (p = 7/9 = 0.7778).

Both intensity and percentage of the IHC staining have to be considered when we score the slides. We have been working with such a classic scoring system for many years. We have realized that even two experienced pathologists score the slides, there is still difficulty to make clear decisions for about 10% of the cases. In this study, a new image-based network analysis was developed to analyze the immunostaining array slides and to extract patterns of the IHC staining, including both intensity and percentage in the whole arrays. We further analyzed the associations of the immunostaining patterns with our clinical data to provide more precise information for rectal cancer.

The mean values of both clustering coefficients and characteristic path lengths of the rectal cancer patients of shorter survival are lower than those of longer survival. There is no correlation between the ratios of the clustering coefficients of the tumor to those of the biopsy and the survival time (correlation coefficient R = 0.0120, p-value = 9.7656e-04) among the shorter-surviving rectal cancer patients whose maximum survival time was about over 6 years (75 months). This can be observed from **Figure 4**. There is also no correlation between

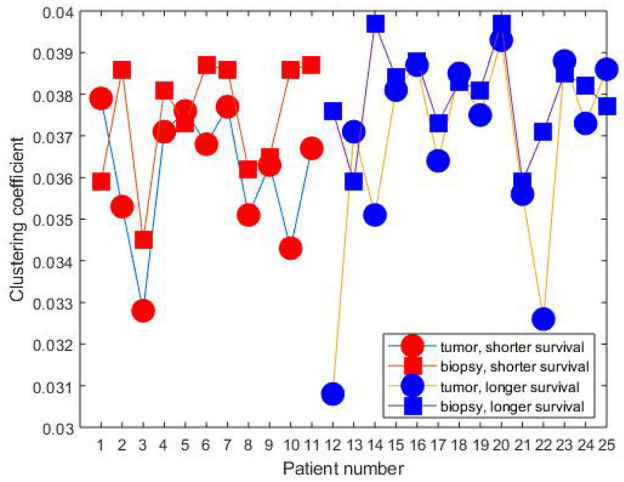

FIGURE 2 | Clustering coefficients of rectal cancer patients.

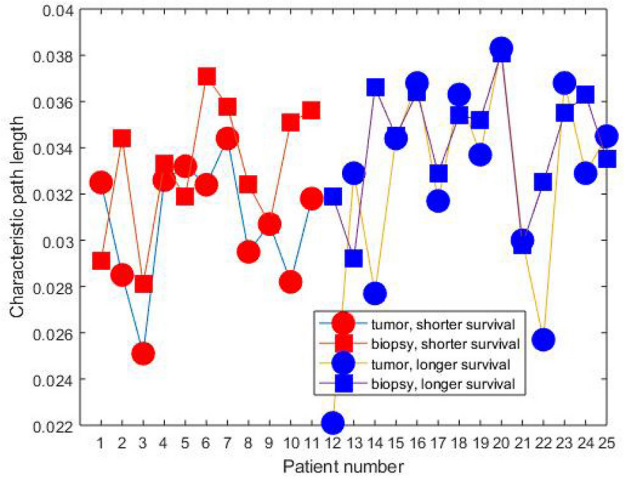

FIGURE 3 | Characteristic path length of rectal cancer patients.

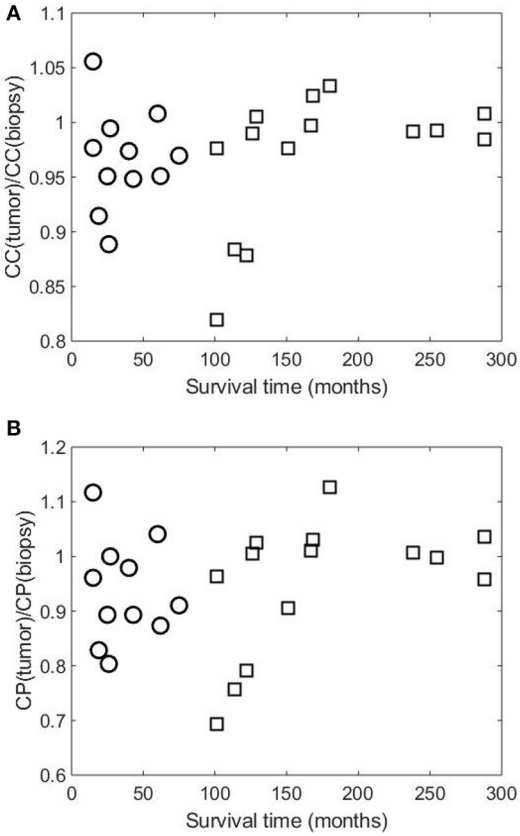

FIGURE 4 | Scatter plots of survival time of 25 rectal cancer patient against **(A)** ratios of clustering coefficients of primary tumor [CC(tumor)] to those of biopsy [CC(biopsy)], and **(B)** ratios of characteristic path lengths of primary tumor [CP(tumor)] to those of biopsy [CP(biopsy)]. Symbols "○" and "□" indicate patient groups with shorter and longer times of survival, respectively.

the ratios of the characteristic path lengths of the tumor to those of the biopsy and the survival time (R = −0.0780, p-value = 9.7656e-04) among the shorter-surviving patients. This can also be observed from **Figure 4**. There is an indication of correlation between the ratios of the clustering coefficients of the tumor to those of the biopsy and the survival time (R = 0.4924, p-value = 1.2207e-04) among the longer-surviving rectal patients whose maximum survival time was 24 years (288 months). This can be observed from **Figure 4**. There is also evidence of correlation between the ratios of the characteristic path lengths of the tumor to those of the biopsy and the survival time (R = 0.4778, p-value = 1.2207e-04) among the longer-surviving rectal cancer patients. This can also be observed from **Figure 4**.

Figure 4 shows similar plots of the ratios of the two network-property parameters of the tumor to biopsy against the survival time, suggesting the consistency of the results. It is reported that rectal patients who survive at least 5 years (60 months) are likely to die from causes that are common in the general population (London, 2017). This finding highlights the predictive value of DNp73 revealed by the image-based FWRN analysis among the cohort of rectal cancer patients whose survival time was between 8.4 years (101 months) and 24 years (288 months) correlated with the clustering-coefficient ratios, and 10.5 years (126 months) and 24 years (288 months) correlated with the characteristic-path-length ratios.

The lack of correlation of the ratios of the MC-FWRN parameters and the (shorter) survival time may suggest an implication of poor responses or non-effective treatment of the RT provided to the rectal cancer patients. Meanwhile, those patients who have positive correlation between the ratios of the FWRN parameters and the (longer) survival time were very likely to have a good or better response to the RT. General findings are that higher values of the ratios of the MC-FWRN parameters indicate longer survival time. The longest survival time is found with the values of the MC-FWRN parameter ratios being about 1. Based on the MC-FWRN parameters of the 25 rectal cancer patients and their survival months, we can predict

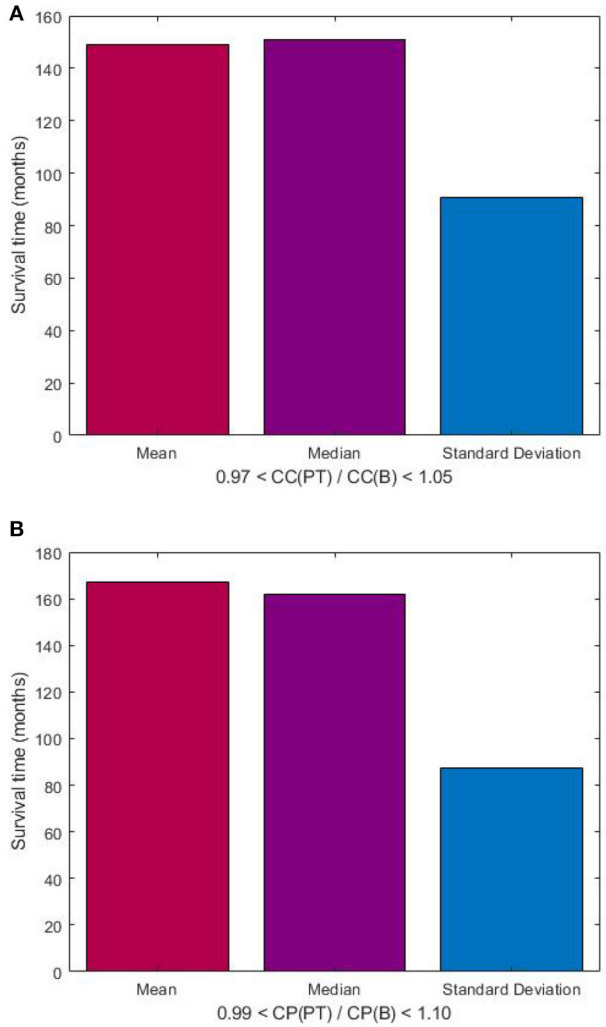

FIGURE 5 | Predictive value of DNp73 in rectal cancer patients, who are deemed to have longer survival time as being between 101 and 288 months, in terms of ratios of clustering coefficient **(A)**, and characteristic path length **(B)** of image-based FWRNs. CC(PT) and CC(B) denote clustering coefficients of primary tumor and biopsy, respectively; and CP(PT) and CP(B) denote characteristic path lengths of primary tumor and biopsy, respectively.

the survival time between 101 months (8.42 years) and 288 months (24 years) with a probability of 73% for those patients whose clustering-coefficient ratio is within the range between 0.97 and 1.05. Similarly, the survival time between 126 months (10.5 years) and 288 months (24 years) with a probability of 78% for those patients whose characteristic-path-length ratio is within the range between 0.99 and 1.10.

The application of a novel image-based network analysis presented in this study was able to discover the predictive factor of DNp73 biomarker in rectal cancer patients having preoperative RT. Predictive biomarkers provide useful information on the probability of obtaining a response to treatment (Walther et al., 2009) and support the process of therapeutic decision for personalized cancer treatment (Voon and Kong, 2011). Such a discovery of DNp73 expression as a predictive biomarker in rectal cancer patients is expected to provide early assessment of the patient outcome, clinical value in the diagnostics of the disease, identification of targeted postoperative therapy.

Regarding the MC-FWRN introduced in this study, this new method appears to be the first of its kind mathematically formulated to capture the recurrence features of multi-channel data inherently existing in complex histology images in a way that is both effective and easily implemented for practical use. Complex networks consist of certain attributes that can be computed to analyze the properties and characteristics of the networks. Mathematical properties of these networks are utilized to define network models and to elucidate how certain models different to each other. The proposed MC-FWRN allows the calculation of the clustering coefficients and characteristic path lengths of DNp73 expression in the primary tumors and biopsies. These values can used to predict the survival time of a cohort of rectal cancer patients who were deemed to be positively influenced by preoperative RT.

The fuzzy weighted recurrence network analysis proposed herein is not supposed to be the study of the complexity of DNp73-controlled networks, but the derivation of structural properties of DNp73 expression from complex microscopy images that can be difficult to understand by pathologists. The results suggest that there are relationships between the graph properties of fuzzy weighted recurrence networks and the color distribution of the stained images. Hence, the network analysis yields new quantitative characteristics of the complexity of the IHC detection of the protein in tissue sections. From a molecular biology perspective, the average clustering coefficient and characteristic path length of the image-based fuzzy weighted recurrence network provide a mathematical measure of the heterogeneity of DNp73 in IHC staining, in correlation with clinicopathological characteristics. This heterogeneity may reflect diverse cell populations in expressing different levels of DNp73.

In this study, we have shown significant results concerning the DNp73 protein expression in predicting the outcome for the rectal cancer patients with the proposed mathematical approach. A limitation of this study is a relatively small number of the rectal cancer patients selected in the analysis. Therefore, future studies with more subgroups of rectal patients will be considered. It should be pointed out that although the total samples of the RT clinical trail from the Southeast Swedish Health Care region included 216 cases, only 102 cases randomly received preoperative RT. Given the aim of this study, only the paired samples of biopsy and primary tumors that are from the same patients were selected for the analysis. Many of the biopsy samples are too tiny to be used for IHC staining, constituting to the limitation of the sample size carried out in this pilot study, which still can provide some representative indication due to the paired samples from the same patient and all the samples derived from the random clinical trial.

Furthermore, results from rectal cancer patients with and without preoperative RT will be obtained and compared. Images of biopsies, primary cancers and metastatic cancers should be further investigated. Eventually,

we will analyze the associations of the reactions from tumor invasive margins and stroma with the patients' prognosis.

Another limitation in this study is that the TAp73 expression was not performed in the present 25 pairs of rectal cancer samples. It is known that TAp73 acts as a tumor suppressor, while DNp73 exerts as an oncogene that is opposite to TAp73 (Amelio et al., 2014; Stantic et al., 2015). Therefore, it is necessary to expand the sample size and simultaneously evaluate TAp73 and DNp73 in the future, based on the methodology we have developed in the current study. DNp73 links to the ability to act as dominant-negative of the TAp73 isoforms and p53. This negative regulation by DNp73 forms an autoregulatory feedback loop, since both TAp73 and p53 can induce expression of DNp73 isoforms by direct binding to the P2 promoter (Irwin, 2006; Rufini et al., 2011; Di et al., 2013). A newest evidence showed that DNp73 isoforms has higher applicant potential in colorectal cancer patients than the canonical p73 protein (Garranzo-Asensio et al., 2019). Thus, it is reasonable to focus on DNp73. In addition, in the present paper, we mainly focused on the automated quantification of IHC expression using an image-based complex network model. The expression of TAp73 and the relationship between TAp73 and DNp73 will be investigated in our future study.

The highlights of the technical development and findings addressed in this paper are summarized as follows. First, the proposed MC-FWRN analysis of DNp73 expression by IHC in rectal cancer is the first of its kind. Second, a new mathematical analysis of IHC-stained biopsy and tumor images reveals the predictive power of DNp73 in rectal cancer patients who received RT. Third, a new method of multi-channel fuzzy weighted recurrence networks is developed for extracting two useful complex network properties of IHC images that can be used as prognostic indicators of rectal cancer. Fourth, the proposed approach for quantifying the expression of IHC is not limited to the study of DNp73, but can also be generally applied to discovering image patterns of other tumor proteins. Fifth, the proposed approach can be utilized as a computerized tool for extracting features from whole slide images in digital pathology.

4. CONCLUSION

The findings presented herein show the useful application of complex network analysis of images for studying the predictive factor of DNp73 biomarker expression in rectal cancer patients. The use of DNp73 biomarker can give insight into preoperative RT that has been considered as an important companion in the treatment of rectal cancer. A larger sample size when being available in future clinical trial will further confirm the current findings. Moreover, the proposed approach is not only found useful to rectal cancer but also can be adopted for the analysis of other biomarkers as well as other types of cancer, where human-based pathology practice is of limited capacity. In fact, there are many reports on the computerized image analysis of H&E (Haemotoxylin and Eosin) staining, much less effort has been made to apply computational methods for the automated analysis of IHC staining. The MC-FWRN presented in this paper can be generally applied for studying the expression of other potential biomarkers.

Although there are many studies reported about the association between DNp73 protein biomarker expression and malignant potential, the function of DNp73 still remains unclear. Our work contributes to the elucidation of the predictive value of DNp73 expression in rectal cancer patients who were given preoperative RT. We developed an original method for constructing weighted recurrence networks of multi-channel images. These networks allow the extraction of useful network properties from complex IHC images. The clustering coefficients and characteristic path lengths of the MC-FWRNs are not only able to show the predictive factor of DNp73 expression in the patients, but also reveal the identification of non-effective application of RT to those who had poor overall survival outcome.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Good Clinical Practice, the Research Ethics Committee in Linkoping, Sweden with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the the Research Ethics Committee in Linkoping, Sweden.

AUTHOR CONTRIBUTIONS

TP, HZ, and X-FS conceived the project. TP developed the fuzzy weighted recurrence networks of multi-channel images. TP, CF, HZ, and X-FS analyzed the results. TP, CF, and X-FS wrote the paper. TP carried out the computer implementation of the computational methods. DP contributed to the Lab work and data preparation. All authors edited and approved the manuscript.

REFERENCES

Albert, R., and Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97. doi: 10.1103/RevModPhys.74.47

Amelio, I., Antonov, A. A., Catani, M. V., Massoud, R., Bernassola, F., Knight, R. A., et al. (2014). TAp73 promotes anabolism. *Oncotarget* 5, 12820–12934. doi: 10.18632/oncotarget.2667

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683–691. doi: 10.1136/gutjnl-2015-310912

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3747–3752. doi: 10.1073/pnas.0400087101

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY: Plenum Press.

Chatterjee, S. B., Hou, J., Bandaru, V. V. R., Pezhouh, M. K., Syed Rifat Mannan, A. A., and Sharma, R. (2019). Lactosylceramide synthase β-1,4-GalT-V: a novel target for the diagnosis and therapy of human colorectal cancer. *Biochem. Biophys. Res. Commun.* 508:380. doi: 10.1016/j.bbrc.2018.11.149

Cuyle, P. J., and Prenen, H. (2017). Current and future biomarkers in the treatment of colorectal cancer. *Acta Clin. Bel.* 72, 103–115. doi: 10.1080/17843286.2016.1262996

Di, C., Yang, L., Zhang, H., Ma, X., Zhang, X., and Sun, C. (2013). Mechanisms, function and clinical applications of DNp73. *Cell Cycle* 12, 1861–1867. doi: 10.4161/cc.24967

Dulloo, I., Gopalan, G., Melino, G., and Sabapathy, K. (2010). The antiapoptotic DeltaNp73 is degraded in a c-Jun-dependent manner upon genotoxic stress through the antizyme-mediated pathway. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4902–4907. doi: 10.1073/pnas.0906782107

Fagiolo, G. (2007). Clustering in complex directed networks. *Phys. Rev. E* 76:026107. doi: 10.1103/PhysRevE.76.026107

Fan, C. W., Chen, T., Shang, Y. N., Gu, Y. Z., Zhang, S. L., Lu, R., et al. (2013). Cancer-initiating cells derived from human rectal adenocarcinoma tissues carry mesenchymal phenotypes and resist drug therapies. *Cell Death Dis.* 4:e828. doi: 10.1038/cddis.2013.337

Garranzo-Asensio, M., Guzmán-Aránguez, A., Povés, C., Fernández-Aceñero, M. J., Montero-Calle, A., Ceron, M. Á., et al. (2019). The specific seroreactivity to ΔNp73 isoforms shows higher diagnostic ability in colorectal cancer patients than the canonical p73 protein. *Sci. Rep.* 9:13547. doi: 10.1038/s41598-019-49960-x

Irwin, M. S. (2006). ΔNp73: misunderstood protein? *Cancer Biol. Ther.* 5, 804–807. doi: 10.4161/cbt.5.7.3023

Josef Gladitz, J., Klink, B., and Seifert, M. (2018). Network-based analysis of oligodendrogliomas predicts novel cancer gene candidates within the region of the 1p/19q co-deletion. *Acta Neuropathol. Commun.* 6:49. doi: 10.1186/s40478-018-0544-y

Lee, P. Y., Chin, S. F., Rahman, T. Y. L., and Jamal, R. (2018). Probing the colorectal cancer proteome for biomarkers: current status and perspectives. *J. Proteom.* 118, 93–105. doi: 10.1016/j.jprot.2018.06.014

Letellier, E., Schmitz, M., Ginolhac, A., Rodriguez, F., Ullmann, P., Qureshi-Baig, K., et al. (2017). Loss of Myosin Vb in colorectal cancer is a strong prognostic factor for disease recurrence. *Brit. J. Cancer* 117, 1689–1701. doi: 10.1038/bjc.2017.352

Liu, S. S., Chan, K. Y. K., Cheung, A. N. Y., Liao, X. Y., Leung, T. W., and Ngan, H. Y. S. (2006). Expression of ΔNp73 and TAp73a independently associated with radiosensitivities and prognoses in cervical squamous cell carcinoma. *Clin. Cancer Res.* 12, 3922–3927. doi: 10.1158/1078-0432.CCR-05-2573

Logotheti, S., Pavlopoulou, A., Galtsidis, S., Vojtesek, B., and Zoumpourlis, V. (2013). Functions, divergence and clinical value of TAp73 isoforms in cancer. *Cancer Metastasis Rev.* 32, 511–534. doi: 10.1007/s10555-013-9424-x

London, S. (2017, May 25). Common causes of death predominate among long-term colorectal cancer survivors. *The ASCO Post*.

Lucena-Araujo, A. R., Kim, H. T., Thome, C., Jacomo, R. H., Melo, R. A., and Bittencourt, R. (2015). High ΔNp73/TAp73 ratio is associated with poor prognosis in acute promyelocytic leukemia. *Blood* 126, 2302–2306. doi: 10.1182/blood-2015-01-623330

Muller, M., Schilling, T., Sayan, A. E., Kairat, A., Lorenz, K., and Schulze-Bergkamen, H. (2005). TAp73/Delta Np73 influences apoptotic response, chemosensitivity and prognosis in hepatocellular carcinoma. *Cell Death Differ.* 12, 1564–1577. doi: 10.1038/sj.cdd.4401774

Murray-Zmijewski, F., Lane, D. P., and Bourdon, J. C. (2006). p53/p63/p73 isoforms: an orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death Differ.* 13, 962–972. doi: 10.1038/sj.cdd.4401914

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.

Ozturk, K., Dow, M., Carlin, D. E., Bejar, R., and Carter, H. (2018). The emerging potential for network analysis to inform precision cancer medicine. *J. Mol. Biol.* 430, 2875–2899. doi: 10.1016/j.jmb.2018.06.016

Parikh, A. P., Curtis, R. E., Kuhn, I., Becker-Weimann, S., Bissell, M., Xing, E. P., et al. (2014). Network analysis of breast cancer progression and reversal using a tree-evolving network algorithm. *PLoS Comput. Biol.* 10:e1003713. doi: 10.1371/journal.pcbi.1003713

Patel, J. N., Fong, M. K., and Jagosky, M. (2019). Colorectal cancer biomarkers in the era of personalized medicine. *J. Pers. Med.* 9:E3. doi: 10.3390/jpm9010003

Pfeifer, D., Wallin, A., Holmlund, B., and Sun, X. F. (2009). Protein expression following gamma-irradiation relevant to growth arrest and apoptosis in colon cancer cells. *J. Cancer Res. Clin. Oncol.* 135, 1583–1592. doi: 10.1007/s00432-009-0606-4

Pham, T. D. (2019). Fuzzy weighted recurrence networks of time series. *Physica A* 513, 409–417. doi: 10.1016/j.physa.2018.09.035

Rahman, M. R., Islam, T., Gov, E., Turanli, B., Gulfidan, G., Shahjaman, M., et al. (2019). Identification of prognostic biomarker signatures and candidate drugs in colorectal cancer: insights from aystems biology analysis. *Medicina* 55:20. doi: 10.3390/medicina55010020

Ruan, J., Jahid, M. J., Gu, F., Lei, C., Huang, Y. W., Hsu, Y. T., et al. (2019). A novel algorithm for network-based prediction of cancer recurrence. *Genomics* 111, 17–23. doi: 10.1016/j.ygeno.2016.07.005

Rufini, A., Agostini, M., Grespi, F., Tomasini, R., Sayan, B. S., Niklison-Chirou, M. V., et al. (2011). p73 in cancer. *Genes Cancer* 2, 491–502. doi: 10.1177/1947601911408890

Ryan, J. E., Warrier, S. K., Lynch, A. C., Ramsay, R. G., Phillips, W. A., and Heriot, A. G. (2016). Predicting pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a systematic review. *Colorectal Dis.* 18, 234–246. doi: 10.1111/codi.13207

Sinicrope, F. A., Okamoto, K., Kasi, P. M., and Kawakami, H. (2016). Molecular biomarkers in the personalized treatment of colorectal cancer. *Clin. Gastroenterol. Hepatol.* 14, 651–658. doi: 10.1016/j.cgh.2016.02.008

Sonke, J. J., and Belderbos, J. (2010). Adaptive radiotherapy for lung cancer. *Semin. Radiat. Oncol.* 20:94e106. doi: 10.1016/j.semradonc.2009.11.003

Stantic, M., Sakil, H. A., Zirath, H., Fang, T., Sanz, G., Fernandez-Woodbridge, A., et al. (2015) TAp73 suppresses tumor angiogenesis through repression of proangiogenic cytokines and HIF-1α activity. *Proc. Natl. Acad. Sci. U.S.A.* 112, 220–225. doi: 10.1073/pnas.1421697112

Stiewe, T., Zimmermann S., Frilling A., Esche, H., and Putzer, B. M. (2002). Transactivation-deficient DeltaTA-p73 acts as an oncogene. *Cancer Res.* 62, 3598–3602.

Swedish Rectal Cancer Trial, Cedermark, B., Dahlberg, M., Glimelius, B., Påhlman, L., Rutqvist, L. E., et al. (1997). Improved survival with preoperative radiotherapy in resectable rectal cancer. *N. Engl. J. Med.* 8, 980–987. doi: 10.1056/NEJM199704033361402

Uramoto, H., Sugio, K., Oyama, T., Nakata, S., Ono, K., Morita, M., et al. (2004). Expression of ΔNp73 predicts poor prognosis in lung cancer. *Clin Cancer Res.* 10, 6905–6911. doi: 10.1158/1078-0432.CCR-04-0290

Voon, P. J., and Kong, H. L. (2011). Tumour genetics and genomics to personalise cancer treatment. *Ann. Acad. Med. Singapore* 40, 362–368.

Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I., and Kerr, D. (2009). Genetic prognostic and predictive markers in colorectal cancer. *Nat. Rev. Cancer* 9, 489–499. doi: 10.1038/nrc2645

Watts, D. J., and Strogatz, S. (1998). Collective dynamics of "small-world" networks. *Nature* 393, 440–442. doi: 10.1038/30918

Yaromina, A., Krause, M., and Baumann, M. (2012). Individualization of cancer treatment from radiotherapy perspective. *Mol. Oncol.* 6, 211–221. doi: 10.1016/j.molonc.2012.01.007

Ye, H., and Guo, X. (2019). TP73 is a credible biomarker for predicting clinical progression and prognosis in cervical cancer patients. *Biosci. Rep.* 39:BSR20190095. doi: 10.1042/BSR20190095

Zadeh, L. A. (1971). Similarity relations and fuzzy orderings. *Informat. Sci.* 3, 177–200. doi: 10.1016/S0020-0255(71)80005-1

Zarkavelis, G., Boussios, S., Papadaki, A., Katsanos, K. H., Christodoulouc, D. K., and Pentheroudakis, G. (2017). Current and future biomarkers in colorectal cancer. *Ann. Gastroenterol.* 30, 613–621. doi: 10.20524/aog.2017.0191

Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. (2018). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* 8:3355. doi: 10.1038/s41598-018-21622-4

Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B., and Kuang, R. (2013). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol.* 9:e1002975. doi: 10.1371/journal.pcbi.1002975

Zhu, W., Pan, X., Yang, Z., Xing, P., Zhang, Y., Li, F., et al. (2015). Expression and prognostic significance of TAp73 and ΔNp73 in FIGO stage I-II cervical squamous cell carcinoma. *Oncol. Lett.* 9, 2090–2094. doi: 10.3892/ol.2015.3052

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 03 March 2020
doi: 10.3389/fphys.2020.00151

Modeling Oncolytic Viral Therapy, Immune Checkpoint Inhibition, and the Complex Dynamics of Innate and Adaptive Immunity in Glioblastoma Treatment

Kathleen M. Storey[1]*, *Sean E. Lawler*[2] and *Trachette L. Jackson*[1]

[1] Department of Mathematics, University of Michigan, Ann Arbor, MI, United States, [2] Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA, United States

Oncolytic viruses are of growing interest to cancer researchers and clinicians, due to their selectivity for tumor cells over healthy cells and their immunostimulatory properties. The immune response to an oncolytic virus plays a critical role in treatment efficacy. However, uncertainty remains regarding the circumstances under which the immune system either assists in eliminating tumor cells or inhibits treatment via rapid viral clearance, leading to the cessation of the immune response. In this work, we develop an ordinary differential equation model of treatment for a lethal brain tumor, glioblastoma, using an oncolytic Herpes Simplex Virus. We use a mechanistic approach to model the interactions between distinct populations of immune cells, incorporating both innate and adaptive immune responses to oncolytic viral therapy (OVT), and including a mechanism of adaptive immune suppression via the PD-1/PD-L1 checkpoint pathway. We focus on the tradeoff between viral clearance by innate immune cells and the innate immune cell-mediated recruitment of antiviral and antitumor adaptive immune cells. Our model suggests that when a tumor is treated with OVT alone, the innate immune cells' ability to clear the virus quickly after administration has a much larger impact on the treatment outcome than the adaptive immune cells' antitumor activity. Even in a highly antigenic tumor with a strong innate immune response, the faster recruitment of antitumor adaptive immune cells is not sufficient to offset the rapid viral clearance. This motivates our subsequent incorporation of an immunotherapy that inhibits the PD-1/PD-L1 checkpoint pathway by blocking PD-1, which we combine with OVT within the model. The combination therapy is most effective for a highly antigenic tumor or for intermediate levels of innate immune localization. Extreme levels of innate immune cell activity either clear the virus too quickly or fail to activate a sufficiently strong adaptive response, yielding ineffective combination therapy of GBM. Hence, we show that the innate and adaptive immune interactions significantly influence treatment response and that combining OVT with an immune checkpoint inhibitor expands the range of immune conditions that allow for tumor size reduction or clearance.

Keywords: oncolytic viral therapy, mathematical modeling, glioblastoma, immune checkpoint inhibitor, combination therapy, innate and adaptive immunity

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Hermann Frieboes,
University of Louisville, United States
Jason H. Yang,
New Jersey Medical School, Rutgers, The State University of New Jersey, United States

***Correspondence:**
Kathleen M. Storey
storeyk@umich.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *19 December 2019*
Accepted: *12 February 2020*
Published: *03 March 2020*

Citation:
Storey KM, Lawler SE and Jackson TL (2020) Modeling Oncolytic Viral Therapy, Immune Checkpoint Inhibition, and the Complex Dynamics of Innate and Adaptive Immunity in Glioblastoma Treatment. Front. Physiol. 11:151. doi: 10.3389/fphys.2020.00151

1. INTRODUCTION

Oncolytic viral therapy (OVT) shows promise as a cancer treatment option that selectively targets cancer cells over healthy cells. Viral therapy is also viewed as a type of immunotherapy because the viral presence stimulates an adaptive immune response (Kaufman et al., 2015). However, after decades of development, OVT has yet to become a widely used treatment option. This is likely due to the multifaceted immune response to the virus, surrounding which uncertainty remains.

This work adds to a growing literature developing mathematical models of OVT. In Wodarz (2001), Wodarz developed a model to study the virus-specific and tumor-specific cytotoxic T lymphocyte response to OVT, and determined the viral and host conditions that produce an optimal tumor response. Wodarz and Komarova (2009) and Komarova and Wodarz (2010) focus on the role of the viral infection rate and develop a general framework to study oncolytic viral dynamics. Eftimie et al. (2011) study the phenomena of multi-stability and multi-instability that arise in interactions between an oncolytic virus and adaptive immune cells, and they conclude that the immune response is primarily responsible for multi-stability, while the virus is primarily responsible for multi-instability. Eftimie and Eftimie (2018) investigate the role that two disparate types of macrophages, M1 and M2, can play in enhancing OVT, finding that polarization toward M1 or M2 phenotype can enhance OVT through either anti-tumor immune activation or increased cytotoxic activity, and that the total number of macrophages plays a consequential role in treatment outcomes. Friedman et al. (2006) consider the effect of the immunosuppressive drug, cyclophosphamide, on glioma response to OVT, and find that it decreases the percentage of uninfected tumor cells.

Many of the papers within this body of work focus on either the innate immune response or the adaptive immune response to OVT, and we build on this work by incorporating both of these branches of the immune system and focusing on the interactions between them. The innate immune system plays two major roles in response to OVT: clearance of the virus and recruitment of adaptive immune cells (McDonald and Levy, 2019). The adaptive immune system, and in particular, the $CD8^+$ T cells, target tumor-associated cognate antigens, in order to specifically target and kill tumor cells. Hence, the innate immune cells play a complex role in response to OVT, potentially clearing the virus before the infection takes hold within the tumor microenvironment, while simultaneously recruiting antitumor adaptive immune cells. We investigate the circumstances under which the innate immune system either assists or hinders viral therapy, thereby providing insight regarding the barriers to successful cancer treatment.

In particular, we study the use of an oncolytic Herpes Simplex Virus (HSV) to treat glioblastoma (GBM), the most aggressive primary malignant brain tumor, killing half of all patients within a year of diagnosis, and nearly all patients within 2 years (Alexander and Cloughesy, 2017). The standard treatment of care for GBM is surgical resection, followed by concurrent radiotherapy and chemotherapy, and subsequent cycles of adjuvant chemotherapy until the tumor recurs (Stupp et al., 2005). A major impediment to GBM treatment is the frequent development of resistance to the standard chemotherapy agent, temozolomide (Hegi et al., 2005; Zhang et al., 2012). Thus, novel therapies are frequently being developed and tested for use in conjunction with, or as an alternative to, temozolomide, to effectively treat GBM. In this work, we consider the effectiveness of OVT as an alternative treatment modality, by developing and analyzing an ordinary differential equation model of GBM response to OVT.

Our results from modeling GBM response to OVT suggest that this treatment is frequently ineffective due to the inhibition of T cell activity by the PD-1/PD-L1 immune checkpoint. PD-1 (programmed cell death-1) is a protein expressed on activated T cells, and its ligand, PD-L1, is frequently upregulated on cancer cells, on innate immune cells, and on T cells (Cheng et al., 2013; Shi et al., 2013). When PD-1 on the surface of a T cell is engaged by PD-L1 on neighboring tumor or innate immune cells, the T cell becomes dysfunctional or "exhausted" and loses the ability to kill its target cells. In recent years, monoclonal antibody therapies against PD-1 and PD-L1, known as immune checkpoint inhibitors, have been developed to target the PD-1/PD-L1 pathway (Barber et al., 2006; He et al., 2015; Speranza et al., 2018). Our initial model investigations suggest the necessity of increased T cell activity in response to OVT, so we also present a second model that combines OVT and an anti-PD-1 immunotherapy, known as nivolumab.

Complex interactions frequently arise when combining cancer therapies, so a number of mathematical models have been developed to study combination treatments. To highlight a few examples, de Pillis et al. (2006) develop a model of tumor response to a combination of chemotherapy and immunotherapy. In Lai and Friedman (2017) model the combination of a cancer vaccine that activates dendritic cells with an immune checkpoint inhibitor, finding that these treatments work effectively together, and developing a notion of synergy between the drugs. In Bagheri et al. (2011) model the combination of an oncolytic adenovirus with MEK-inhibitor treatment. Kim et al. (2018, 2019) investigate the effect of combining OVT, natural killer cell treatment, and a proteasome inhibitor known as bortezomib, suggesting dosing strategies that account for factors in the tumor microenvironment. Our work supplements the existing literature by investigating a combination of an oncolytic Herpes simplex virus with anti-PD-1 immunotherapy, while focusing on the crucial role of the innate immune cells in response to this treatment.

The outline of this paper is as follows: in section 2 we describe our mathematical model with OVT alone, followed by a modified version that incorporates the combination of OVT and an immune checkpoint inhibitor. In this section, we also describe the use of experimental murine data to calibrate the parameters used in the model. In section 3, we present our results with OVT alone, in section 3.1, suggesting a need to combine OVT with anti-PD-1 immunotherapy. We proceed in section 3.2 by discussing the increased efficacy of the combination therapy over OVT alone, and the multifaceted role of the innate immune

TABLE 1 | Model variables.

Variable	Description	Units
T_s	Susceptible tumor cells	#
T_I	Infected tumor cells	#
V	Free viral particles	pfu
Z	Activated innate immune cells	#
Y_T	Adaptive tumor-specific immune cells	#
Y_V	Adaptive virus-specific immune cells	#
P	Concentration of PD-1	μM

system in response to the combination therapy. We summarize these results and discuss future directions in section 4.

2. MATERIALS AND METHODS

We have developed a model to investigate the treatment of GBM through OVT and an immune checkpoint inhibitor. We use *in vivo* parameter values to simulate GBM in a murine model. We first present a model including OVT alone, and then we present an additional equation to incorporate the immune checkpoint inhibitor. In the initial set of seven equations, we model the temporal changes in five immune/cancer cell types; the oncolytic virus, which we assume to be HSV; and the molar concentration of PD-1 molecules expressed by the cells within the model. In the second version of the model, we modify the equation for the molar concentration of PD-1, and add an equation describing the molar concentration of an anti-PD-1 immunotherapy drug, which we assume to be nivolumab. The initial set of seven variables are listed in **Table 1**.

Figure 1 provides a visual representation of the model. The labeled connections between cell types in **Figure 1** correspond to specific terms in the model equations described in the next section. Model parameters and their sources are listed in **Table 2**, and the process used to estimate them can be found in section 2.2 and in the **Appendix**.

2.1. Model Equations

The initial model consists of a system of seven non-linear differential equations, listed below. Each equation describes the rate of change in hours of the population of a single cell type or of the virus. We assume that the innate immune cell population includes both macrophages and natural killer cells, due to the positive feedback loop that exists between these two cellular types. Macrophages engulf and destroy viral particles, while natural killer cells primarily target and kill infected cells, so in our model we assume the collective group of innate immune cells are activated by and target both viral particles and infected tumor cells. The innate immune cells release cytokines, which recruit adaptive $CD8^+$ T cells, and we assume that the T cells can be divided into two groups that primarily target either viral antigens or tumor antigens (McDonald and Levy, 2019). Following Lai and Friedman (2017) and Nikolopoulou et al. (2018), we incorporate suppression of these adaptive immune cells via the PD-1/PD-L1 checkpoint with the factor $F(P, L)$ in Equations (5),(6).

We start simulations with the initial conditions $X_s = 10^5$ cells, $V = 10^7$ pfu (plaque-forming units), and all other cell populations beginning at 0. Time $t = 0$ represents the time at which the initial viral dose is administered, and we assume that any pre-treatment antitumor immune activity is factored into the net tumor growth rate, so there are no new immune cells being recruited to target the tumor at the time of the initial viral dose.

2.1.1. Oncolytic Viral Therapy Alone

Equation (1), shown below, models the susceptible tumor population. The term (1a) represents logistic growth of the susceptible tumor cells with intrinsic growth rate r_t, and with a carrying capacity C_T for all tumor cells. We assume a baseline growth rate of 0.0192 per hour, corresponding to a tumor doubling rate of about 35 days, and a carrying capacity of 5.157×10^8 cells. We obtained these values by fitting a logistic growth curve to control data in Linsenmann et al. (2019), displayed in **Figure 2A**. Note that in Friedman et al. (2006), they use a similar value of 0.02 h^{-1}, based on the growth of glioma cells.

The term $\beta T_s V$ represents viral infection of susceptible tumor cells at rate β, which shifts tumor cells from the susceptible population to the infected population, T_I. We assume a baseline viral infection rate of 2.5×10^{-9} virion^{-1}h^{-1}, but vary this parameter in much of the numerical analysis. The final term, (1c), represents the killing of susceptible tumor cells by tumor-specific adaptive immune cells, Y_T. We use Michaelis-Menten kinetics to model saturation in the immune response, assuming that an over-abundance of tumor cells restricts movement within the tumor architecture (Kirschner and Panetta, 1998). The k_{TA} denotes the maximum immune killing rate of tumor cells, for which we assume a baseline value of $\frac{1}{24}$ cell^{-1}h^{-1} from Mahasa et al. (2017), corresponding to each T cell killing one tumor cell each day. The parameter h_T represents the population of T_s at which the immune cells lyse tumor cells at half of their maximum killing rate. We use a baseline value of $h_T = 2.7 \times 10^4$ cells from Banerjee et al. (2015), but we allow for a feasible range that includes much smaller values, as seen in Mahasa et al. (2017).

$$\frac{dT_s}{dt} = \underbrace{r_t T_s \left(1 - \frac{T_s + T_I}{C_T}\right)}_{\text{(1a) Tumor growth}} - \underbrace{\beta T_s V}_{\substack{\text{(1b) Viral} \\ \text{infection}}} - \underbrace{k_{TA} Y_T \frac{T_s}{h_T + T_s}}_{\substack{\text{(1c) Adaptive} \\ \text{immune killing}}} \tag{1}$$

Equation (2) models the infected tumor population, in which term (2a) denotes the addition of cells to the population T_I via viral infection of susceptible tumor cells at rate β. Term (2b) denotes the death of infected cells, induced by the viral infection, at rate δ_T. We assume a baseline viral lysis rate of $\delta_T = 1/18$ h^{-1} from Friedman et al. (2006). The final three terms, (2c)–(2e), denote the killing of infected cells via innate immune cells, antitumor adaptive immune cells, and antiviral adaptive immune cells, respectively. All three types of immune killing are modeled using Michaelis-Menten kinetics, analogously to the immune killing term in Equation (2). We assume the killing rate of infected tumor cells by antiviral T

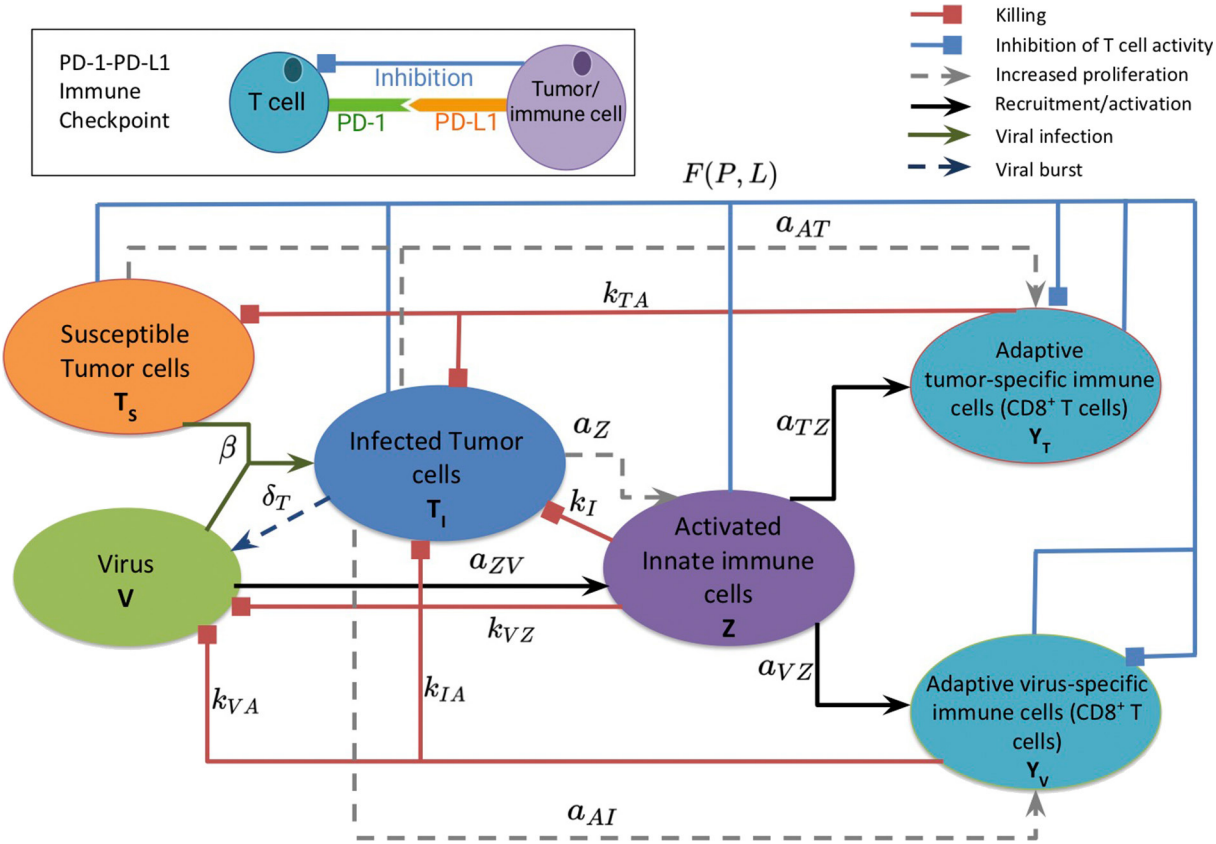

FIGURE 1 | A schematic diagram of the model, depicting the interactions between the susceptible and infected tumor cells, innate and adaptive immune cells, and the virus. Susceptible tumor cells become infected by the oncolytic virus, and following successful viral replication within the infected cells, these cells lyse to give rise to new viral particles. Innate immune cells, representing both macrophages and natural killer cells, become activated when they encounter the virus or viral antigens on the infected cells, and subsequently target and kill viral particles and infected cells. The innate immune cells present viral and tumor antigens to $CD8^+$ T cells, which activate antiviral and antitumor T cells. Antiviral T cells attack and kill infected cells and viral particles, and antitumor T cells target both infected and susceptible tumor cells. The inset diagram displays the inhibition of adaptive immune cell activity by the PD-1/PD-L1 immune checkpoint.

cells is $k_{IA} = \frac{1}{24}$ $\text{cell}^{-1}\text{h}^{-1}$ from Mahasa et al. (2017), and we estimate a killing rate of infected tumor cells by innate immune cells with $k_I = 0.02$ $\text{cell}^{-1}\text{h}^{-1}$. We estimate this value based on the assumption that T cells primarily target infected cells, while innate immune cells primarily target the virus itself, and thus the innate immune-mediated killing rate should be smaller than the adaptive immune-mediated killing rate of infected cells.

$$\frac{dT_I}{dt} = \underbrace{\beta T_s V}_{\text{(2a) Viral Infection}} - \underbrace{\delta_T T_I}_{\text{(2b) Viral lysis}} - \underbrace{k_I T_I \frac{Z}{h_I + Z}}_{\text{(2c) Innate killing}} - \underbrace{k_{TA} Y_T \frac{T_I}{h_I + T_I}}_{\text{(2d) Antitumor adaptive killing}} - \underbrace{k_{IA} Y_V \frac{T_I}{h_I + T_I}}_{\text{(2e) Antiviral adaptive killing}} \quad (2)$$

Equation (3) models the virus population, with term (3a) representing the addition of new viral particles that are released when an infected tumor cell lyses. The parameter b_T denotes the viral burst size released from each infected cell, which we assume to be 50 viral particles per cell, the estimated burst size for HSV, from Friedman et al. (2006). The term $k_{VZ}VZ$ represents consumption of the virus by innate immune cells at rate k_{VZ}. We estimate a baseline value for k_{VZ} and use the values for the rate at which primed innate immune cells consume a pathogen from Reynolds et al. (2006), and the mean rate of phagocytosis by macrophages in the presence of an unlimited supply of targets, from Branwood et al. (1992), to dictate a feasible range for k_{VZ}. The term (3c) represents viral clearance by antiviral adaptive immune cells. We estimate that the adaptive immune-mediated killing rate of the virus $k_{VA} << k_{IA}$, stemming from our assumption that the innate immune cells have a larger impact than adaptive immune cells on the clearance of the virus itself. The final term (3d), corresponds to clearance of the viral particles, resulting from local non-specific immune cells in the tumor region. We use clearance rate $\omega = 0.025\ \text{h}^{-1}$ from Friedman et al. (2006), corresponding to a half-life of about 1.15 days.

$$\frac{dV}{dt} = \underbrace{b_T \delta_T T_I}_{\text{(3a) Viral burst}} - \underbrace{k_{VZ} VZ}_{\text{(3b) Innate killing}} - \underbrace{k_{VA} V Y_V}_{\text{(3c) Adaptive killing}} - \underbrace{\omega V}_{\text{(3d) Natural clearance}} \quad (3)$$

TABLE 2 | Model parameters.

	Parameter	Description	Baseline	Range	References
1	r_t	Tumor cell growth rate	0.0192 h^{-1}	0.005–0.05	Friedman et al., 2006
2	C_T	Tumor cell carrying capacity	5.157×10^8 cells	$10^7 - 10^9$	Fit from Linsenmann et al., 2019
3	β	Viral infection rate	2.5×10^{-9} $pfu^{-1}h^{-1}$	$2.5 \times 10^{-13} - 2.5 \times 10^{-7}$	Okamoto et al., 2014, Est.
4	δ_T	Death rate of infected tumor cells	$\frac{1}{18}$ h^{-1}	1/48 – 1/9	Friedman et al., 2006
5	ω	Viral clearance rate	0.025 h^{-1}	0.001–1	Friedman et al., 2006
6	b_T	Burst size of infected cells	50 pfu/cell	10–1,350	Friedman et al., 2006
7	a_Z	Rate of infected cell-mediated proliferation of innate immune cells	2.4×10^{-6} $cell^{-1}h^{-1}$	$2.4 \times 10^{-8} - 2.4 \times 10^{-4}$	Estim.
8	a_{ZV}	Virus-mediated activation rate of resting innate immune cells	0.1 $pfu^{-1}h^{-1}$	$2.4 \times 10^{-4} - 0.2$	Reynolds et al., 2006
9	δ_Z	Death rate of innate immune cells	0.008 h^{-1}	$5 \times 10^{-4} - 1/12$	Eftimie and Eftimie, 2018
10	δ_{YT}	Death rate of tumor-specific adaptive immune cells	3.75×10^{-4} h^{-1}	0.001 – 0.0074	Banerjee et al., 2015; Mahasa et al., 2017
11	δ_{YV}	Death rate of virus-specific adaptive immune cells	5.54×10^{-3} h^{-1}	0.001–0.01	Mahasa et al., 2017
12	k_I	Killing rate of infected cells by innate immune cells	0.02 $cell^{-1}h^{-1}$	0.001–0.1	Estim.
13	k_{VZ}	Killing rate of virions by innate immune cells	0.005 $cell^{-1}h^{-1}$	0.001 – 2	Est., Reynolds et al., 2006
14	k_{TA}	Killing rate of tumor cells by tumor-specific adaptive immune cells	$\frac{1}{24}$ $cell^{-1}h^{-1}$	0.0004–0.2	Mahasa et al., 2017
15	k_{IA}	Killing rate of infected cells by virus-specific adaptive immune cells	$\frac{1}{24}$ $cell^{-1}h^{-1}$	0.0004–0.2	Mahasa et al., 2017
16	k_{VA}	Killing rate of virions by virus-specific adaptive immune cells	10^{-5} $cell^{-1}h^{-1}$	$10^{-6} - 10^{-3}$	Estim.
17	a_{TZ}	Activation rate of tumor-specific adaptive immune cells via innate immune cells	0.025 h^{-1}	$10^{-3} - 0.1$	Estim.
18	a_{VZ}	Activation rate of virus-specific adaptive immune cells via innate immune cells	0.025 h^{-1}	$10^{-3} - 0.1$	Estim.
19	a_{AT}	Rate of tumor cell-mediated proliferation of tumor-specific adaptive immune cells	0.0016 $cell^{-1}h^{-1}$	$10^{-5} - 10^{-1}$	Mahasa et al., 2017
20	a_{AI}	Rate of infected cell-mediated proliferation of virus-specific adaptive immune cells	0.025 $cell^{-1}h^{-1}$	$10^{-5} - 0.1042$	Mahasa et al., 2017
21	h_T	Half-saturation constant of tumor cells	2.7×10^4 cells	$40–10^5$	Banerjee et al., 2015; Mahasa et al., 2017
22	h_I	Half-saturation constant of infected tumor cells	10^4 cells	$20–5 \times 10^4$	Banerjee et al., 2015, Est.

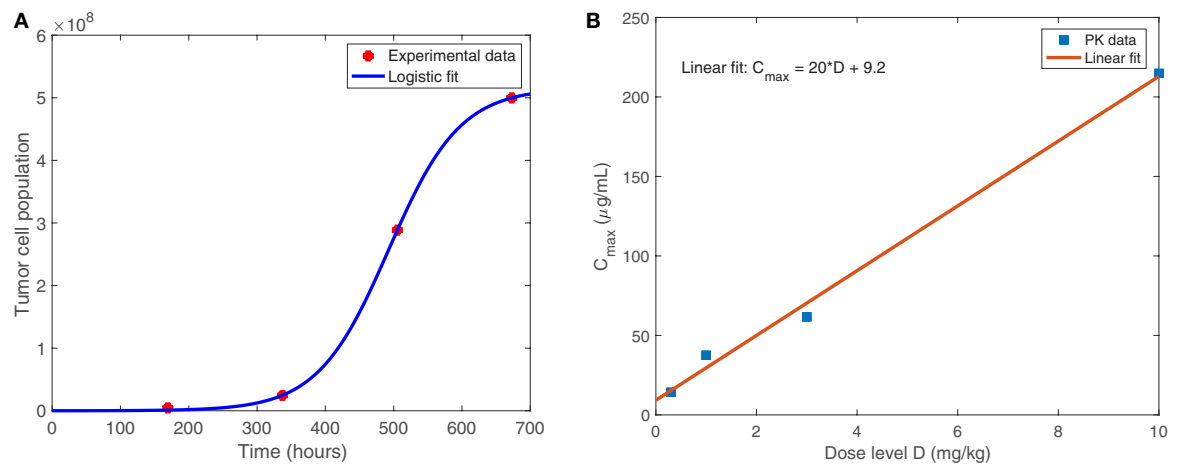

FIGURE 2 | (A) Shows the logistic growth fit to murine GBM size data in the absence of treatment, from Linsenmann et al. (2019). We use this to determine the net growth rate, $r_t = 0.0192$ h^{-1}, and carrying capacity, $C_T = 5.157 \times 10^8$ cells. **(B)** Displays the linear relationship between the administered dose (mg/kg) of nivolumab and the maximum plasma concentration, C_{max} (μg/mL). Data is obtained from Brahmer et al. (2010).

Equation (4) models the activated innate immune cell population, in which term (4a) represents the activation of resting innate immune cells. The parameter s_{ZR} denotes the rate at which new resting innate immune cells arrive in the tumor microenvironment. These resting cells are activated by interactions with the virus at rate a_{ZV}, and previously activated innate immune cells recruit more resting innate immune cells at rate $a_Z Z$, creating a positive feedback loop. We assume that the activation of resting innate immune cells occurs quickly, and thus use a quasi-steady state analysis for the resting innate immune population to obtain term (4a), as shown in Reynolds et al. (2006). We use baseline parameter values $s_{ZR} = 0.08$ cell h^{-1}, $a_{ZV} = 0.1$ pfu^{-1} h^{-1}, and $a_{ZZ} = 0.01$ cell^{-1} h^{-1} from Reynolds et al. (2006). Term (4b) represents increased proliferation of innate immune cells, induced by interactions with infected cells. Here we estimate a_Z to be 2.4×10^{-6} cell^{-1}h^{-1}, or equivalently 5.7×10^{-5} per infected cell per day. This accounts for macrophages and natural killer cells signaling to and recruiting each other, resulting in an innate immune cell positive feedback loop, with the assumption that activation occurs more commonly by encountering the virus itself, rather than by encountering infected cells. The term $\delta_Z Z$ denotes natural innate immune cell death, at rate δ_Z, which we assume to be $\delta_Z = 0.008$ h^{-1} from Eftimie and Eftimie (2018).

$$\frac{dZ}{dt} = \underbrace{\frac{s_{ZR}(a_{ZZ}Z + a_{ZV}V)}{\delta_{ZR} + a_{ZZ}Z + a_{ZV}V}}_{\text{(4a) Activation of resting innate immune cells}} + \underbrace{a_Z T_I Z}_{\text{(4b) Infected cell-mediated proliferation}} - \underbrace{\delta_Z Z}_{\text{(4c) Natural death}} \tag{4}$$

Equation (5) models the tumor-specific adaptive immune response. Term (5a) models the recruitment of T cells by innate immune cells, in which we assume a recruitment rate of $a_{TZ} = 0.025$ h^{-1}. This is an *ad hoc* estimate, as this relationship has not been well-explored previously, and we rely on the parameter sensitivity analysis to consider a range of values for this parameter. Term (5b) represents the proliferation of adaptive T cells due to the presence of tumor antigens on both susceptible and infected tumor cells. We assume a baseline value for the tumor cell-mediated proliferation rate of tumor-specific adaptive immune cells, a_{AT}, of 0.0016 h^{-1}, converted from the rate in Mahasa et al. (2017). We again use Michaelis-Menten kinetics with half-saturation constant h_T to model the saturation of T cell activity, due to the restrictive tumor architecture. The factor $F(P, L)$ factor represents suppression of T cell activation and proliferation via the PD-1/PD-L1 checkpoint. P, L denote the molar concentrations of PD-1 and PD-L1, respectively, expressed by cells within the model. The molar concentrations are obtained by first calculating the PD-1 expression on all T cells and the PD-L1 expression on all T cells, tumor cells, and innate immune cells, as outlined in the **Appendix**. As P and L increase, so does the number of PD-1/PD-L1 complexes within the tumor region. This increase corresponds to a smaller $F(P, L)$ value, modeling the inhibition of T cell activity. Term (5d) represents natural death of antitumor T cells. We use $\delta_{YT} = 3.75 \times 10^{-4}$ h^{-1} from Mahasa et al. (2017), corresponding to a half-life of about 77 days.

$$\frac{dY_T}{dt} = \left(\underbrace{a_{TZ}Z}_{\substack{\text{(5a)} \\ \text{Activation via innate immune cells}}} + \underbrace{a_{AT}Y_T \frac{T_s + T_I}{h_T + T_s + T_I}}_{\text{(5b) Tumor cell-mediated proliferation}} \right) \underbrace{F(P, L)}_{\substack{\text{(5c) PD-1-PD-L1} \\ \text{suppression}}} - \underbrace{\delta_{YT} Y_T}_{\substack{\text{(5d) Natural} \\ \text{death}}} \tag{5}$$

Equation (6) models the adaptive virus-specific immune response. Term (6a) represents the recruitment of CD8$^+$ T cells by innate immune cells. We assume equal activation rates of antitumor and antiviral T cells via innate immune cells, so we use the same estimate, $a_{VZ} = a_{TZ} = 0.025$ h^{-1}. Term (6b) represents the proliferation of virus-specific CD8$^+$ T cells resulting from viral antigens expressed on infected cells, and we use infected cell-mediated proliferation rate $a_{AI} = 0.025$ cell^{-1}h^{-1} from Mahasa et al. (2017). Similarly to Equation (5), the factor $F(P, L)$ represents the PD-1/PD-L1-mediated inhibition of T cell activation and proliferation. The parameter δ_{YV} denotes the rate of natural cell death of antiviral T cells. We use $\delta_{YV} = 5.54 \times 10^{-3}$ h^{-1} from Mahasa et al. (2017), corresponding to a half-life of 5.2 days.

$$\frac{dY_V}{dt} = \left(\underbrace{a_{VZ}Z}_{\substack{\text{(6a)} \\ \text{Activation via innate immune cells}}} + \underbrace{a_{AI}Y_V \frac{T_I}{h_I + T_I}}_{\substack{\text{(6b) Infected} \\ \text{cell-mediated} \\ \text{proliferation}}} \right) \underbrace{F(P, L)}_{\substack{\text{(6c) PD-1-} \\ \text{PD-L1} \\ \text{suppression}}} - \underbrace{\delta_{YV} Y_V}_{\substack{\text{(6d) Natural} \\ \text{death}}} \tag{6}$$

$$\frac{dP}{dt} = \underbrace{\rho_p \left(\frac{dY_T}{dt} + \frac{dY_V}{dt} \right)}_{\substack{\text{(7a) PD-1 expression on} \\ \text{adaptive immune cells}}}, \tag{7}$$

where $\frac{dY_T}{dt}$ and $\frac{dY_V}{dt}$ denote the expressions in Equations (5) and (6), L denotes the molar concentration of PD-L1 within the tumor microenvironment, represented by

$$L = \underbrace{\rho_L (Y_T + Y_V + \epsilon_T (T_s + T_I) + \epsilon_Z Z)}_{\substack{\text{PD-L1 expression on adaptive immune cells,} \\ \text{tumor cells, and innate immune cells}}} \tag{8}$$

and

$$F(P, L) = \frac{1}{1 + PL/K_{YQ}} \tag{9}$$

2.1.2. With Immune Checkpoint Inhibitor

When we incorporate the immune checkpoint inhibitor within the model, the functional forms for Equations (1)–(6) remain the same. For the equation describing PD-1 concentration, we modify Equation (7) and replace with (10) below. We also add an eighth equation, representing the change in molar concentration of an anti-PD-1 immunotherapy drug, A, as follows:

$$\frac{dP}{dt} = \underbrace{\frac{P}{Y_T + Y_V}\left(\frac{dY_T}{dt} + \frac{dY_V}{dt}\right)}_{\substack{\text{(8a) PD-1 expression on} \\ \text{adaptive immune cells}}} - \underbrace{\mu_{PA} PA}_{\substack{\text{(8b) Blocking} \\ \text{by anti-PD-1}}} \tag{10}$$

$$\frac{dA}{dt} = \underbrace{A(t)}_{\substack{\text{(9a)} \\ \text{anti-PD-1} \\ \text{dosing}}} - \underbrace{\mu_{PA} PA}_{\substack{\text{(9b) Depletion by} \\ \text{blocking PD-1}}} - \underbrace{\delta_A A}_{\substack{\text{(9c)} \\ \text{Natural} \\ \text{depletion}}}, \tag{11}$$

In the presence of anti-PD-1 therapy, Equation (10) models the total number of free molecules of PD-1 within the tumor microenvironment. In term (8a), we replace ρ_p from Equation (7) with $P/(Y_T + Y_V)$, since the mass of PD-1 changes when the drug binds to it. Hence the ratio between the mass of a PD-1 molecule and the mass of a T cell does not remain constant in the presence of the anti-PD-1 drug. Term (8b) models the binding of the drug to PD-1 at rate μ_{PA}, thereby blocking the PD-1 from forming a complex with PD-L1.

In Equation (11), $A(t)$ represents the source of the anti-PD-1 drug, which is derived from pharmacokinetic data in section 2.2. Term (9b) models the depletion of the drug as it binds to PD-1. Term (9c) represents the natural depletion of free drug that has not bound to PD-1. We estimate the parameter δ_A, the natural decay rate of anti-PD-1, to be 0.0019 h^{-1}, converted from the half-life of 15 days, published in Brahmer et al. (2010). In our parameter sensitivity analysis, we vary δ_A in the range $1.37 \times 10^{-3} - 0.058$ h^{-1}, converted from the range in Nikolopoulou et al. (2018). To estimate the drug-mediated blocking rate of PD-1, μ_{PA}, we use a similar argument to one used in Lai and Friedman (2017) to obtain $\mu_{PA} = 8.945$ L/μmol/h. See the **Appendix** for a full derivation of this parameter value.

2.2. Immune Checkpoint Parameter Estimation

The function in terms (5c) and (6c) is defined in Equation (9), with L given by (8). In the expression for L, the molar concentration of PD-L1 in the tumor microenvironment, ρ_L, denotes the molar concentration of PD-L1 per T cell. In our simulations, we use $\rho_L = 2.510 \times 10^{-11}\mu$M. See **Appendix** for the full derivation of this parameter value. To complete the derivation of term (6c), we define Q to be the molar concentration of PD-1/PD-L1 complexes formed from the binding of PD-1 and PD-L1, modeled by

$$P + L \underset{\delta_Q}{\overset{\alpha_{PL}}{\rightleftharpoons}} Q,$$

where α_{PL}, δ_Q are the association and dissociation rates of Q. As in Lai and Friedman (2017) and Nikolopoulou et al. (2018), we assume that the association and dissociation of Q are fast (Mautea et al., 2015), so applying a quasi-steady state argument, we can approximate Q using the equation:

$$Q = \frac{\alpha_{PL}}{\delta_Q} PL.$$

In Lai and Friedman (2017), they incorporate T cell inhibition via Q in the T cell differential equation by multiplying the activation terms by the following factor:

$$\frac{1}{1 + Q/K_{TQ}}.$$

They define $K_{TQ} = \frac{1}{2}\bar{Q} = \frac{1}{2}\frac{\alpha_{PL}}{\delta_Q}\bar{P}\bar{L}$, where $\bar{P}, \bar{L}$ denote the steady state quantities for P, L. Thus, we define $K_{YQ} = \frac{1}{2}\bar{P}\bar{L}$ so that we can rewrite the previous factor as

$$F(P, L) = \frac{1}{1 + PL/K_{YQ}}.$$

We use $K_{YQ} = 1.296 \times 10^{-9}\mu\text{M}^2$, as determined by a process outlined in the **Appendix**.

Equation (7) models the micromolar concentration of PD-1 (in μmol/L) within the tumor microenvironment in the absence of anti-PD-1 treatment. PD-1 is expressed on T cells, so we can represent P by

$$P = \rho_p(Y_T + Y_V),$$

where ρ_p denotes the molar concentration of PD-1 per T cell. In our simulations, we use $\rho_p = 1.259 \times 10^{-11}\mu$M. See the **Appendix** for a full derivation of this parameter value. By differentiating this equation with respect to t, we obtain the equation shown for $\frac{dP}{dt}$.

The approved flat dosage regimen for nivolumab is 240 mg every 2 weeks. In Lee et al. (2018), they cite that the flat dosage results in similar exposure to 3 mg/kg. The typical treatment schedule consists of a single intravenous dose of 3 mg/kg nivolumab, administered for 1 h, once every 2 weeks. We use pharmacokinetic data from the Phase I study in Brahmer et al. (2010) to relate the dosage, D, in mg/kg to plasma concentration C_{max}, in μg/mL. As shown in **Figure 2B**, we obtained the following linear relationship:

$$C_{max}(D) = 20D + 9.2.$$

We convert this to μM units using the molar mass of nivolumab, 1.436×10^{-1} g/μmol (Wishart et al., 2017). Hence, $\hat{C}_{max}$, the μM plasma concentration, is given by

$$\hat{C}_{max}(D) = C_{max}(D) \times \frac{10^{-3}\,\frac{\text{g*mL}}{\mu\text{g*L}}}{1.436 \times 10^{-1}\ \text{g}/\mu\text{mol}}$$

Thus,

$$\hat{C}_{max}(D) = 0.139D + 0.064.$$

For simplicity, we use $\hat{C}_{max}(3 \text{ mg/kg}) = 0.481\ \mu$M as our baseline estimate for $A(t)$ during the hour following each anti-PD-1 dose. Hence, for each time t_d at which anti-PD-1 is administered,

$$A(t) = \begin{cases} 0.481 & t_d \leq t < t_d + 1 \\ 0 & \text{otherwise.} \end{cases}$$

3. RESULTS

3.1. Oncolytic Viral Therapy Alone

First, we discuss our results for the model in the absence of an immune checkpoint inhibitor, given by Equations (1)–(7).

3.1.1. Parameter Sensitivity Analysis

We perform a global parameter sensitivity analysis with OVT alone, to simulate a virtual experimental trial with 300 mice, each with distinct tumor and immune characteristics. We use this analysis to identify the parameters that most significantly contribute to treatment efficacy. We first determined a reasonable range of values in which to vary each model parameter using estimates in the literature when available, and otherwise estimating based on available biological information, as shown in **Tables 2**, **3**. We performed the sensitivity analysis using Latin hypercube sampling (LHS) and partial rank correlation coefficient (PRCC) analysis (McKay et al., 1979). See the **Appendix** for details describing this process.

We performed the global sensitivity analysis with four different simulation endpoints, at $t = 100$, $t = 300$, $t = 1,000$, and $t = 3,000$ h. **Figure 3** depicts the PRCC for each parameter and each endpoint, determined through this global sensitivity analysis. In all cases, the parameter with the strongest relationship to the final tumor size was β, the viral infection rate. The tumor cell growth rate, r_t, was another highly significant parameter for $t \leq 300$. However, on the longer time scale, when $t = 1,000$ or $t = 3,000$, the tumor carrying capacity, C_T, had a more significant impact on the final tumor size than the growth rate. The parameter k_{VZ}, representing the innate immune-mediated killing rate of virus, also gains some significance as the simulation end time increases, but on a much smaller scale than the viral infection rate and tumor carrying capacity.

We are particularly interested in the role that the immune system plays in treatment success. In order to isolate this effect, overshadowed by the impact of the viral infection and tumor growth properties in the full sensitivity analysis, we perform another sensitivity analysis, varying only the parameters directly related to the innate immune response and fixing all other parameters. This models a trial of mice with similar tumors, treated by the same virus, but characterized by distinct innate immune responses to the treatment. We found that the most significant innate immune-related parameter on an intermediate time-scale is the innate immune-mediated killing rate of virus, k_{VZ}. Using the notation $\mathcal{P}(x, t)$ to denote the PRCC between the parameter x and the tumor size after t h, the PRCC for k_{VZ} was $\mathcal{P}(k_{VZ}, 300) = 0.6591$, indicating a strong direct correlation between this parameter value and the susceptible tumor population after 300 h. The left plot in **Figure 4A** displays the tumor size for each simulation within the innate immune sensitivity analysis with simulation endpoint $t = 300$ h, as a function of the innate immune-mediated killing rate of virus, k_{VZ}. The second most significant parameter in this analysis was the source of the innate immune cells, s_{ZR}, with PRCC $\mathcal{P}(s_{ZR}, 300) = 0.3241$, indicating a moderate direct relationship to the post-treatment susceptible tumor population, shown in **Figure 4B**. The PRCC between each remaining innate immune-related parameters and the susceptible tumor population was under 0.09.

We continued this investigation by isolating the parameters directly related to the adaptive immune response and fixing all other parameters, simulating an experimental trial of mice with similar tumors and viral treatment, but characterized by distinct adaptive immune responses to the treatment. When varying only parameters related to the adaptive immune response, there was very little variation in tumor size after 300 h for most parameter sets. However, the parameter with the strongest correlation to tumor size was the virus-specific adaptive immune-mediated killing rate of virus, k_{VA}, with a strong direct relationship, indicated by $\mathcal{P}(k_{VA}, 300) = 0.7961$. **Figure 5A** displays the susceptible tumor population as k_{VA} varies, and we observe that most simulations ended at a comparable high tumor size level, but for very large values of k_{VA}, this post-treatment tumor size increases further, due to rapid killing of the virus by the adaptive immune cells. The second most significant parameter is the innate immune-mediated activation rate of virus-specific adaptive immune cells, a_{VZ}, with $\mathcal{P}(a_{VZ}, 300) = 0.3128$, and the third most significant parameter is the rate of tumor cell-mediated proliferation of tumor-specific adaptive immune cells, a_{AT}, which can be interpreted as the level of antigenicity of the tumor. The PRCC between this parameter and the tumor size after 300 h is $\mathcal{P}(a_{AT}, 300) = -0.1228$, indicating an inverse relationship between antigenicity and the tumor size. Although a_{AT} does not have a strong correlation coefficient when compared to the parameter k_{VA}, **Figure 5B** shows that the only simulations resulting in a reduced tumor size had high levels of antigenicity. Thus, there seems to be an important range of a_{AT} that allows for more successful therapy results, making the tumor antigenicity level potentially more interesting than the adaptive immune-mediated killing rate of virus.

3.1.2. Treatment Dependence on Viral Infection Rate

We observed in the global sensitivity analysis that the effectiveness of oncolytic viral therapy to treat GBM is highly dependent on the viral infection rate. The infectivity of an oncolytic virus is not an intrinsic property of the system; this viral characteristic can be genetically modified via gene deletions, so it is undoubtedly a parameter worthy of investigation. We investigate the effect of the viral infectivity by fixing all other parameters at their baseline level while varying only the viral infection rate, β. Due to uncertainty regarding a biologically

TABLE 3 | Parameters used in immune checkpoints.

	Parameter	Description	Baseline	Range	References
23	K_{YQ}	Inhibition of T cells by PD-1/PD-L1	1.296×10^{-9} $(\mu M)^2$	$10^{-10} - 10^{-8}$	Lai and Friedman, 2017, Est.
24	ρ_p	Molar concentration of PD-1 per T cell	1.259×10^{-11} μM	$10^{-12} - 10^{-10}$	Nikolopoulou et al., 2018, Est.
25	ρ_L	Molar concentration of PD-L1 per T cell	2.510×10^{-11} μM	$10^{-12} - 2 \times 10^{-10}$	Nikolopoulou et al., 2018, Est.
26	ϵ_T	Expression of PD-L1 on tumor cells vs. T cells	10	1–50	Estim.
27	ϵ_Z	Expression of PD-L1 on innate immune cells vs. T cells	10	1–50	Estim.
28	s_{ZR}	Source of the resting innate immune cells	0.08 cell h^{-1}	0.005 – 0.2	Reynolds et al., 2006
29	a_{ZZ}	Activation of resting innate immune cells by previously activated innate immune cells	0.01 $cell^{-1}h^{-1}$	0.005 – 0.2	Reynolds et al., 2006
30	δ_{ZR}	Death rate of resting innate immune cells	0.12 $cell^{-1}h^{-1}$	0.069 – 0.12	Reynolds et al., 2006
31	δ_A	Decay rate of anti-PD-1	0.0019 h^{-1}	1.37×10^{-3} – 0.058	Brahmer et al., 2010; Nikolopoulou et al., 2018
32	μ_{PA}	Anti-PD-1 blocking rate of PD-1	8.945 L/μ mol/h	$6.45 - 2.73 \times 10^2$	Lai and Friedman, 2017

achievable upper bound for viral infectivity, we let β vary in a large range, for the purpose of identifying the level of infectivity required for successful treatment. For each distinct β level, we simulate the model until $t = 3,000$, when all populations have settled toward their steady state behavior. **Figure 6** shows in yellow that there is a clear threshold, $\beta \approx 4.9 \times 10^{-8}$, above which the tumor is eliminated through treatment, and below which the tumor reaches its carrying capacity. There is little available information about specific limitations for viable oncolytic viral infection rates, so it may be the case that many oncolytic viruses cannot feasibly reach this high level of infectivity.

We also investigate the degree to which this critical β threshold changes as the immune landscape changes. To model a tumor in a strong innate immune environment, we increase the two most influential innate immune parameters, k_{VZ} and s_{ZR}, to the upper bounds of the ranges over which we vary these parameters in the sensitivity analysis, i.e., to $k_{VZ} = 2$ $cell^{-1}h^{-1}$ and $s_{ZR} = 0.2$ cell h^{-1}. As β varies, the dotted green line in **Figure 6** shows that the strong innate immunity prevents treatment success for all levels of viral infectivity, which we hypothesize is due to the rapid innate immune-mediated clearance of all viral particles.

It may also be the case that the oncolytic viral treatment is administered to a tumor that elicits a strong adaptive immune response. In order to test the benefit that the strong adaptive immune response may confer to treatment response, we increase the antigenicity parameter a_{AT}, whose upper range yielded a reduced tumor size in the sensitivity analysis, to 0.05 $cell^{-1}h^{-1}$. As β varies, the dashed purple curve in **Figure 6** shows that the viral infection threshold shifts downward from the baseline case, suggesting that in an environment with a strong adaptive immune response, treatment can be effective with a less infectious virus, due to the increase in tumor-mediated recruitment of adaptive immune cells.

3.1.3. Innate Immune Suppression of OVT

In the previous subsection, we observe that on its own, a strong innate immune response negatively impacts the tumor response to OVT. However, intuition suggests that when paired with a strong adaptive immune response, for a sufficiently strong innate response, the innate immune cell recruitment of adaptive immune cells could potentially outweigh the rapid clearance of the virus. In **Figure 7**, we consider the tumor size after 300 h as the source of innate immune cells, s_{ZR} varies. The curve in blue shows a monotone increase in tumor size in the baseline case, as s_{ZR} increases. The green curve shows the results in a tumor microenvironment with a strong adaptive immune response, modeled as before, with a high level of tumor antigenicity, $a_{AT} = 0.05$ $cell^{-1}h^{-1}$. In this case, the tumor size again increases monotonically with s_{ZR}, albeit deviating to some extent from the baseline case for large s_{ZR} values. The monotonic behavior suggests that even when paired with a strong adaptive immune response, the strong innate immune system is not beneficial to OVT response. Hence, with a larger innate immune presence, the faster recruitment of adaptive immune cells is not sufficient to offset the rapid viral clearance from the innate immune cells.

However, in the absence of PD-1/PD-L1 immune suppression, i.e., when $F(P, L) = 1$ in Equations (5), (6), we observe the opposite trend for large s_{ZR}. The dashed yellow curve in **Figure 7** shows that eliminating the immune checkpoints in the baseline case has essentially no effect on the treatment response, but when paired with a strong adaptive immune response, displayed in purple, the tumor size decreases for sufficiently large s_{ZR}. Hence, without the PD-1/PD-L1 suppression of T-cell activity, the faster recruitment of adaptive immune cells resulting from a large innate immune presence, can yield more effective treatment results. This suggests that for tumors with strong adaptive immunity, combining OVT with immunotherapies that inhibit the PD-1/PD-L1 checkpoint may improve treatment efficacy. These observations motivated the inclusion of anti-PD-1 immunotherapy in our model. We will discuss the results from the combination therapy model in the following section.

3.2. Combination Therapy With Anti-PD-1

Next, we discuss our results for the model that includes both oncolytic viral therapy and the immune checkpoint inhibitor, anti-PD-1, described by Equations (1)–(6), (10), (11).

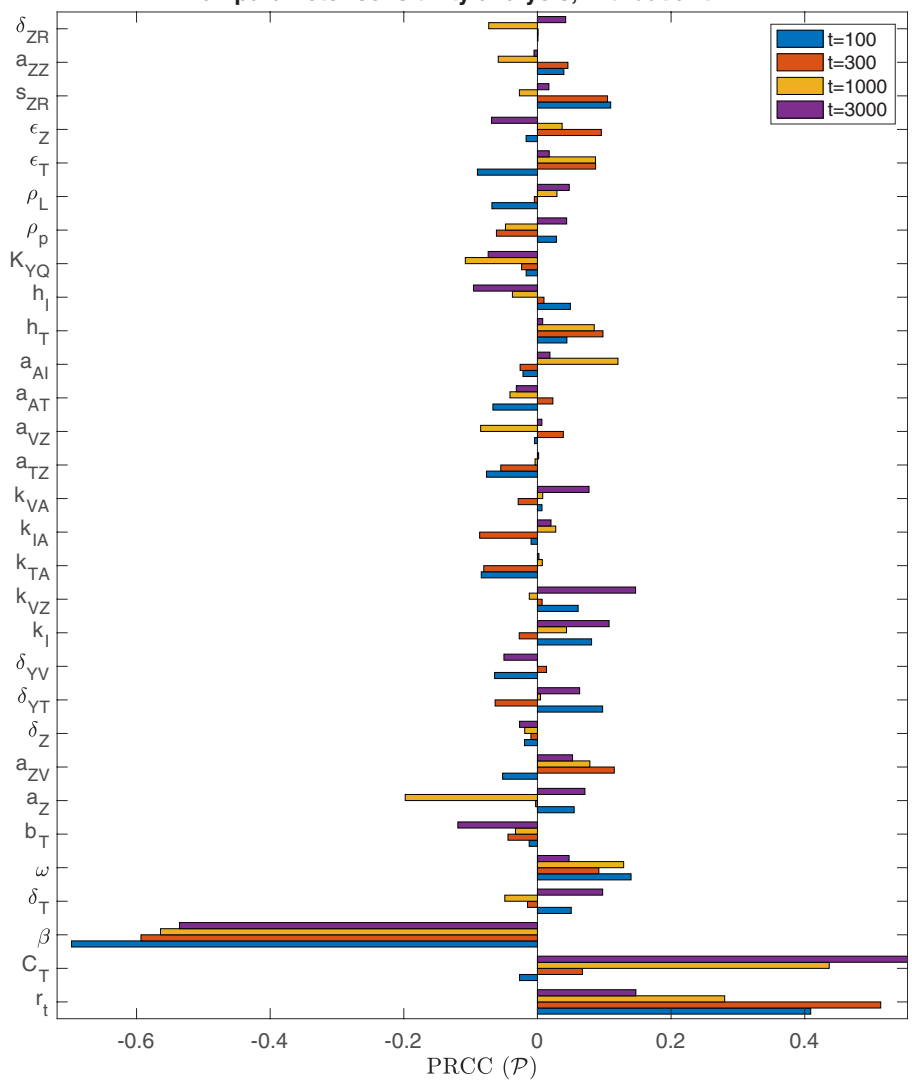

FIGURE 3 | Parameter sensitivity analysis without anti-PD-1. Bar plot of the partial rank correlation coefficient between each model parameter and the susceptible tumor population at the end of the simulation, shown for simulation end points at $t = 100$, $t = 300$, $t = 1,000$, and $t = 3,000$ h.

3.2.1. Parameter Sensitivity Analysis

We also perform a parameter sensitivity analysis with both oncolytic viral therapy and anti-PD-1 immunotherapy, using the method described in section 3.1.1, in order to identify parameters that gain or lose significance with the combination therapy, when compared to the sensitivity analysis with OVT alone. **Figure 8** displays the PRCC for each parameter in this global sensitivity analysis. We will represent this PRCC by $\hat{\mathcal{P}}$ when it refers to the model with anti-PD-1. The most substantial difference between this analysis and the analysis with OVT alone relates to the parameter a_{AT}, representing the level of tumor antigenicity. With anti-PD-1, the PRCC between a_{AT} and tumor size after 1,000 h is $\hat{\mathcal{P}}(a_{AT}, 1,000) = -0.4532$, and after 3,000 h is $\hat{\mathcal{P}}(a_{AT}, 3,000) = -0.4705$, whereas with oncolytic viral therapy alone, the corresponding PRCC values for a_{AT} are $\mathcal{P}(a_{AT}, 1,000) = -0.0411$ and $\mathcal{P}(a_{AT}, 3,000) = -0.0316$. Hence, the parameter a_{AT} has a much stronger correlation with post-treatment tumor size when the tumor is treated with anti-PD-1, suggesting that tumor antigenicity contributes significantly more to the effectiveness of the combination therapy than to the effectiveness of OVT alone. Otherwise, the viral infection rate, β, is still the most significant parameter for simulation end-time $t \leq 1,000$. For $t = 3,000$, the parameter a_{AT} surpasses β, with $\hat{\mathcal{P}}(a_{AT}, 3,000) = -0.4705$ and $\hat{\mathcal{P}}(\beta, 3,000) = -0.3071$. We also note that the carrying capacity, C_T, is much less significant with anti-PD-1 than with OVT alone, suggesting more effective treatment with the combination therapy, leading to more frequent tumor size reduction or clearance.

Analogously to the previous section, we perform additional sensitivity analyses, first varying only the parameters directly

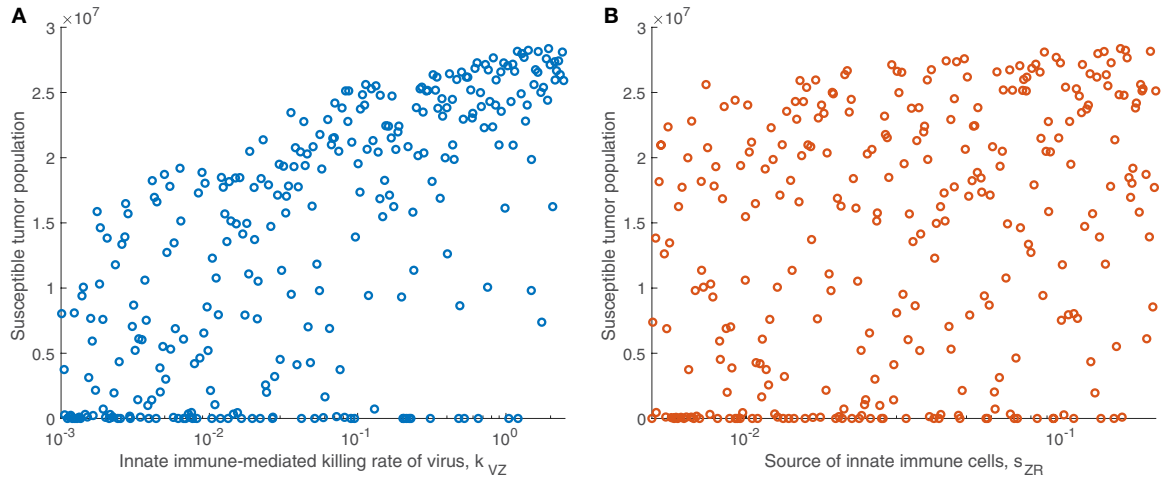

FIGURE 4 | Tumor size in innate immune sensitivity analysis. The tumor size **(A)** as a function of k_{VZ}, the killing rate of the virus by innate immune cells, and **(B)** as a function of s_{ZR}, the source of innate immune cells. We use $t = 300$ h as the endpoint for the innate sensitivity analysis. The PRCC between k_{VZ} and the susceptible tumor population is $\mathcal{P}(k_{VZ}, 300) = 0.6591$, and between s_{ZR} and the susceptible tumor population is $\mathcal{P}(s_{ZR}, 300) = 0.3241$.

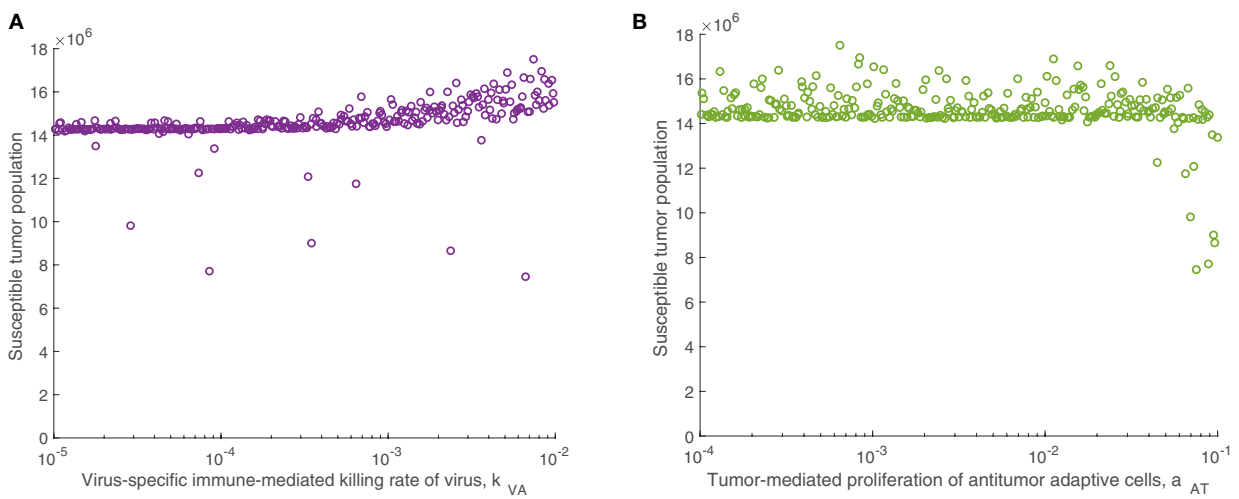

FIGURE 5 | Tumor size in adaptive immune sensitivity analysis. The tumor size is shown **(A)** as a function of k_{VA}, the killing rate of the virus by virus-specific adaptive immune cells, and **(B)** as a function of a_{AT}, the tumor-mediated proliferation of tumor-specific adaptive immune cells. We use $t = 300$ h as the endpoint for the adaptive sensitivity analysis. The partial rank correlation coefficient (PRCC) between k_{VA} and the susceptible tumor population is $\mathcal{P}(k_{VA}, 300) = 0.7961$, and between a_{AT} and the susceptible tumor population is $\mathcal{P}(a_{AT}, 300) = -0.1228$.

related to the innate immune response and fixing all other parameters, and subsequently varying only the parameters directly related to the adaptive immune response. In the innate immune case, the results were very similar to those with OVT alone, and we summarize these in the **Appendix**. Similarly to the global parameter sensitivity analysis, when we vary only adaptive immune-related parameters, the parameter a_{AT} is much more significant with anti-PD-1 than without this treatment. With anti-PD-1, the PRCC is $\hat{\mathcal{P}}(a_{AT}, 300) = -0.3213$, as compared to $\mathcal{P}(a_{AT}, 300) = -0.1228$ with OVT alone. This is the second most significant parameter in this analysis, surpassing the innate immune-mediated activation rate of virus-specific adaptive immune cells, a_{VZ}, with $\hat{\mathcal{P}}(a_{VZ}, 300) = 0.2685$. The most significant parameter is again the adaptive immune-mediated viral killing rate, k_{VA}, with $\hat{\mathcal{P}}(k_{VA}, 300) = 0.6995$, reduced from the PRCC value without anti-PD-1 of $\mathcal{P}(k_{VA}, 300) = 0.7961$. Although $|\hat{\mathcal{P}}(a_{AT}, 300)|$ is smaller than $|\hat{\mathcal{P}}(k_{VA}, 300)|$ in this adaptive immune parameter sensitivity analysis, large values of a_{AT} seem to contribute to tumor clearance, as shown in **Figure 9**. In contrast, the parameter k_{VA} does not seem to contribute to a reduction in tumor size, but rather, high values of k_{VA} lead to larger tumors. Hence, the tumor antigenicity level, a_{AT}, seems to be the most important adaptive immune-related parameter, with respect to tumor size reduction and clearance when treated with

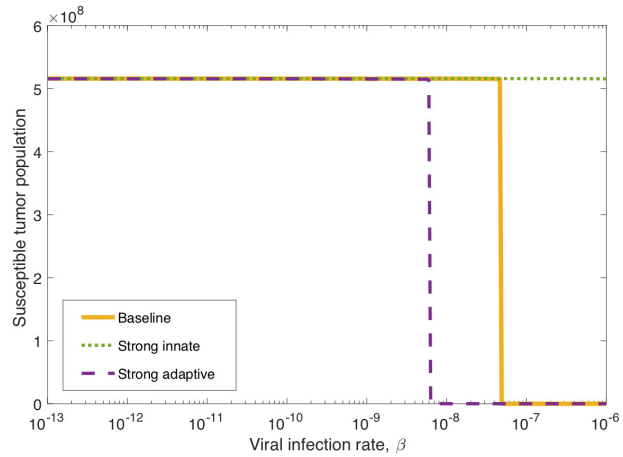

FIGURE 6 | Long-term tumor size, after 3,000 h, as the viral infection rate, β varies. The tumors were treated with a single oncolytic viral dose, and we compare between a baseline tumor and a tumor with strong innate or a strong adaptive immune system. In the strong innate case, $k_{VZ} = 2$ and $s_{ZR} = 0.2$. In the strong adaptive case, $a_{AT} = 0.05$.

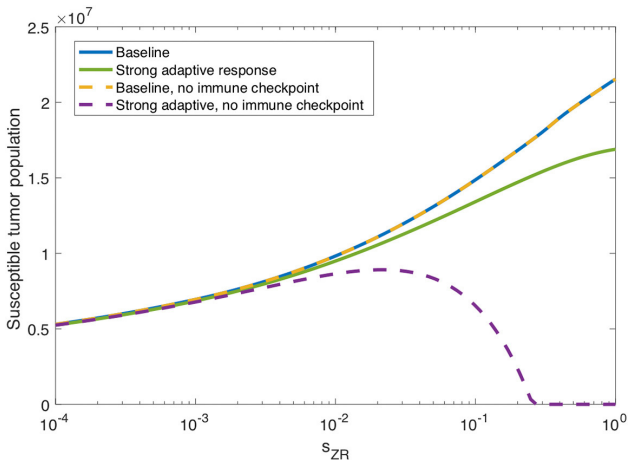

FIGURE 7 | Innate immunity tradeoff. We display the susceptible tumor population after 300 h without anti-PD-1, as the source of innate immune cells, s_{ZR} varies. We compare the baseline parameter regime with a stronger adaptive immune response, given by $a_{AT} = 0.05$, in the presence of suppression of T-cell activity via the PD-1/PD-L1 pathway and also in the absence of T cell suppression via the PD-1/PD-L1 pathway.

both anti-PD-1 and OVT. Compare **Figure 9** with **Figure 5** to see that the tumor reduction for large a_{AT} is much more striking with anti-PD-1 than we observed without anti-PD-1.

3.2.2. Treatment Dependence on Viral Infection Rate, With Anti-PD-1

We determined in the global sensitivity analysis that the effectiveness of OVT and anti-PD-1 immunotherapy to treat GBM is dependent on the viral infection rate, but this dependence is less severe with anti-PD-1 than without. We investigate this further by varying only the viral infection rate, β, while fixing all other parameters, and comparing the tumor size after 3,000 h. In **Figure 10**, the blue curve displays the susceptible tumor population after treatment with anti-PD-1, with all parameters outside of β at their baseline levels. In this case, there is a larger viral infection range that will lead to tumor clearance, as compared to the yellow curve showing the tumor size after treatment with OVT alone. The threshold for tumor clearance without anti-PD-1 is $\beta = 4.9 \times 10^{-8}$, whereas with anti-PD-1, tumor clearance occurs for all $\beta \geq 2.5 \times 10^{-8}$, and $5 \times 10^{-10} < \beta < 3.2 \times 10^{-9}$ will likely also lead to tumor clearance. In this range of β values, treatment success is highly sensitive to the timing of the viral infection and to the timing of the immune response. Hence, for $5 \times 10^{-10} \leq \beta \leq 3.2 \times 10^{-9}$, treatment success is likely, but the treatment results are less predictable.

We also observe the sensitivity to infection and immune response timing when varying the dosing of the virus. **Figure 11** shows the difference between the cell and viral populations when one viral dose is administered at $t = 0$, in 11(a) and 11(b), and when one initial dose is followed 7 days later by a second viral dose, in 11(c) and 11(d). In both cases, anti-PD-1 is administered intravenously for 1 h, every 2 weeks. We observe that the combination therapy results in tumor clearance when a single viral dose is administered. Interestingly, when an additional viral dose is administered 1 week after the first dose, the treatment actually becomes ineffective, with the tumor rebounding to its carrying capacity level. One possible explanation for this phenomenon is that the administration of an additional viral dose after stimulating an immune response can counteract treatment progress by diverting the attention of the immune response from the tumor alone to additional viral particles. The absence of a viral oscillation in the simulation with two doses, in comparison with the rapid viral oscillation just before 800 h in the single dose case, suggests more active immune-mediated killing of the virus when two doses are administered. It is also possible that this effect may be the result of an increased innate immune cell population, stemming from the second viral dose, which in turn produces a larger concentration of PD-L1 within the tumor microenvironment. This observation warrants follow-up work, to experimentally study the effect of multiple viral doses in combination with immune checkpoint inhibitors. The discrepancy in tumor response in these two cases emphasizes the sensitivity of the tumor response to viral and immune response timing. Additionally it suggests that the primary role of the oncolytic virus is its stimulation of the immune system, rather than its cytotoxic effect on tumor cells.

We also consider treatment dependence on viral infection rate as the immune landscape changes. In the case of a strong innate immune response, simulated using $k_{VZ} = 2$ and $s_{ZR} = 0.2$, there is a range of large β values that lead to treatment success with both anti-PD-1 and OVT, shown in red in **Figure 10**, in contrast to no tumor size reduction for any β with OVT alone, shown in green. However, the β range for tumor clearance with a strong innate immune response is quite high, suggesting the rapid innate immune-mediated clearance the virus prevents treatment success unless the virus is infectious enough to persist until a sufficient adaptive immune response has been initiated. Note that the results for a strong adaptive immune response, treated with both

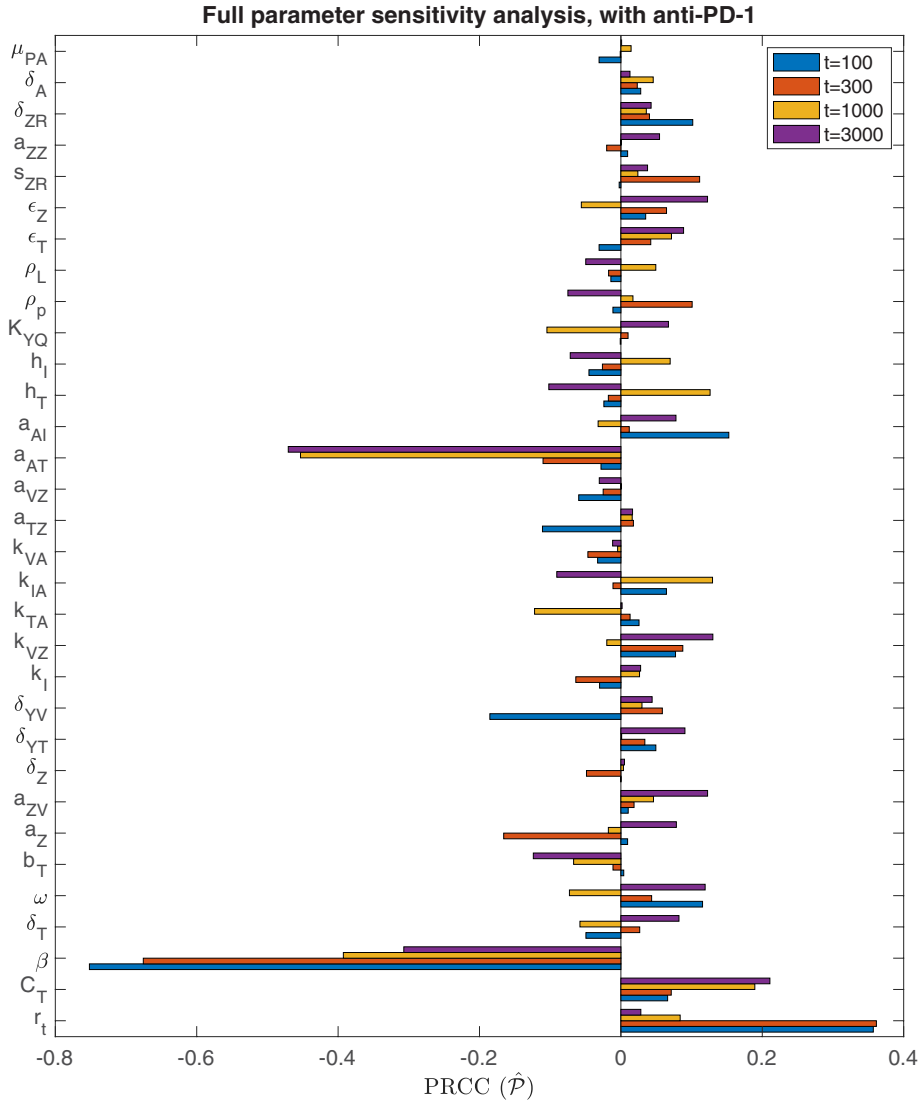

FIGURE 8 | Parameter sensitivity analysis with anti-PD-1. Bar plot of PRCC values for each model parameter, shown for simulation end points at $t = 100$, $t = 300$, $t = 1,000$, and $t = 3,000$ h.

OVT and anti-PD-1 immunotherapy, are not shown in the figure because this case leads to eventual tumor clearance for all viral infection rates. Hence, for any oncolytic virus, without the PD-1/PD-L1 checkpoint suppression of adaptive immune activity, a high level of tumor-mediated adaptive immune cell proliferation is sufficient to successfully clear the tumor.

We find that in all cases, combining OVT with anti-PD-1 decreases the viral infection rate threshold for effective treatment, increasing the likelihood of developing an oncolytic virus that is sufficiently infectious to successfully treat murine GBM. However, a strong innate immune response on its own makes the therapy less effective, so we next investigate the dynamics that occur in a microenvironment equipped with both strong innate and strong adaptive immune responses.

3.2.3. Innate Immunity Tradeoff, With Anti-PD-1

We find in the previous sections that the source of innate immune cells, s_{ZR}, is positively correlated with post-treatment tumor size, and that increasing the innate immune cell presence in the tumor microenvironment leads to an increase in the viral infection rate threshold required for effective treatment. Hence, in a typical tumor environment, the net contribution of the innate immune cells to the combination therapy success is negative, due to their role in viral clearance. In section 3.1.3, we determined that this was the case with OVT alone, even as the strength of the adaptive immune response increased. When the tumor is treated with both OVT and anti-PD-1, **Figure 12A** shows the tumor size after 300 h as the source of innate immune cells, s_{ZR} varies, in the baseline case and when paired with a strong adaptive immune response, represented by an increased a_{AT}. This figure is analogous for

combination therapy to **Figure 7** for OVT alone, and we observe that for sufficiently large s_{ZR}, the tumor size actually reaches a maximum and then declines as s_{ZR} increases. This behavior confirms our hypothesis from the previous model that combining OVT with anti-PD-1 treatment allows the antitumor immune response to reach its full potential; the strong innate response, combined with a strong adaptive immune response is sufficient to clear the tumor relatively quickly. Without anti-PD-1, such a parameter regime would yield a larger tumor than in the baseline case.

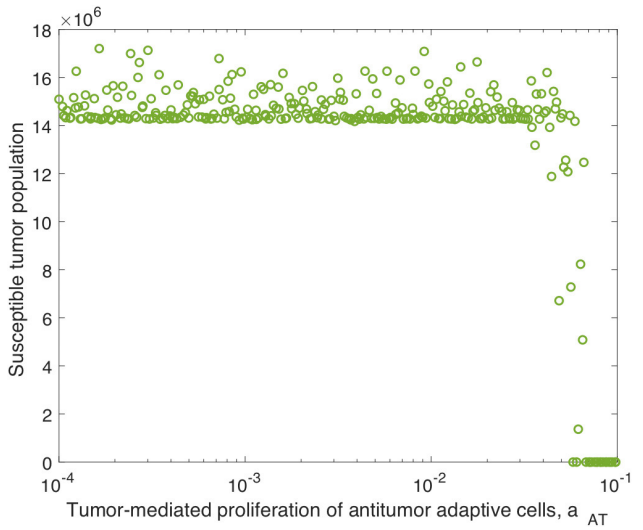

FIGURE 9 | Tumor size in adaptive immune sensitivity analysis with anti-PD-1, as a function of a_{AT}, the tumor-mediated proliferation of tumor-specific adaptive immune cells. We use $t = 300$ h as the endpoint for the adaptive sensitivity analysis. The PRCC between a_{AT} and the susceptible tumor population is $\hat{\mathcal{P}}(a_{AT}, 300) = -0.3213$.

In **Figure 12B**, we consider the dynamics within the tumor microenvironment on a longer time scale, until $t = 1,000$ h, as s_{ZR} varies. In this figure, all other parameters are set to their baseline level, and we observe that there is a large range of s_{ZR} that leads to eventual tumor clearance. There is one small blip occurring around $s_{ZR} = 4 \times 10^{-4}$, in which the tumor returns to carrying capacity, due to sensitivity to the timing of the immune response; For small values of s_{ZR}, there are a few discontinuities in the long-term tumor size, due to sensitivity to the timing of the immune response. The tumor rebounds to its carrying capacity when the innate immune population decline is precisely timed to prevent an oscillation of the viral population, driven by the bursting of infected cells. This viral oscillation is required to stimulate a surge in adaptive immune activity that ultimately clears the tumor. Outside of this small range, tumor clearance occurs, except in the highest ranges of s_{ZR}, in which we hypothesize the influx of innate immune cells clears the virus too quickly.

Next we vary the strength of the innate immune response and the adaptive immune response simultaneously. In **Figure 13**, we display the parameter values in the $s_{ZR} - a_{AT}$ space that yield post-treatment tumor clearance or recurrence to tumor carrying capacity by $t = 4,000$ h. **Figure 13A** shows the long-term results when the tumor is treated with OVT alone, while **Figure 13B** shows the results with both OVT and anti-PD-1 immunotherapy. With OVT alone, tumor clearance only occurs for a very small range of large a_{AT} values, i.e., when the tumor is highly antigenic. After combining OVT with anti-PD-1, tumor clearance occurs for a larger upper range of a_{AT}, but for weak and

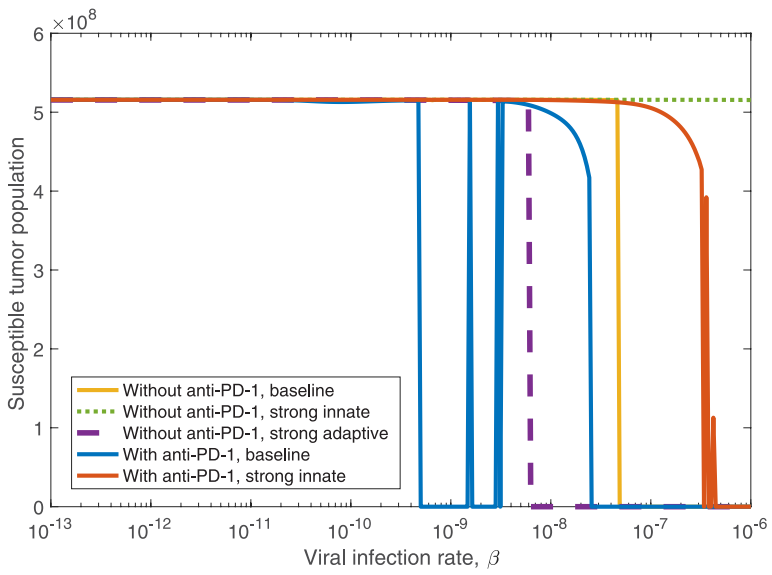

FIGURE 10 | The susceptible tumor population after 3,000 h as the viral infection rate, β varies. The susceptible population is shown both with and without anti-PD-1, and we compare between a baseline tumor and a tumor with strong innate or a strong adaptive immune system. In the strong innate case, $k_{VZ} = 2$ and $s_{ZR} = 0.2$. In the strong adaptive case, $a_{AT} = 0.05$.

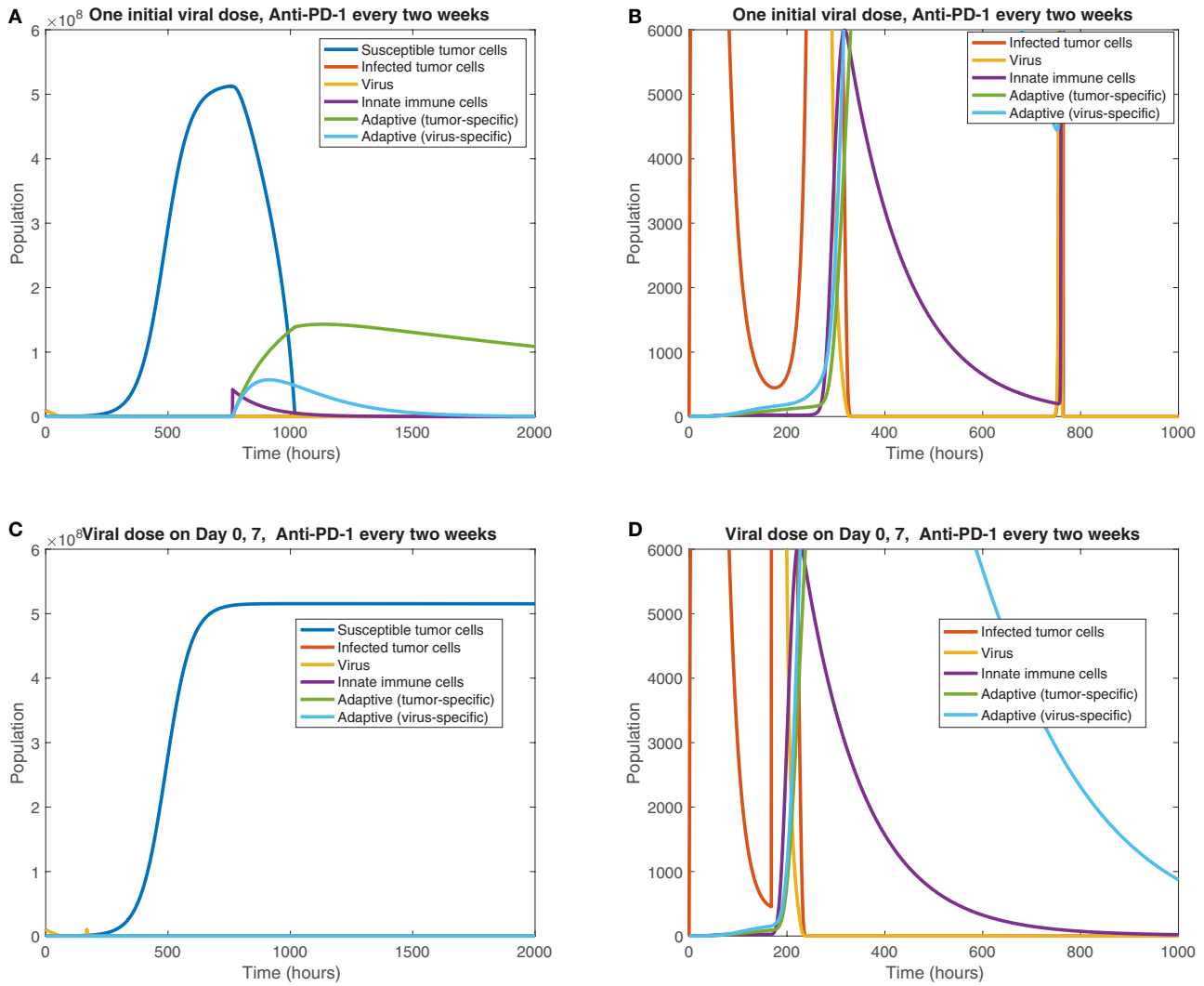

FIGURE 11 | Treatment interference from a second viral dose. We display the cell populations over time in the baseline case, with anti-PD-1 dosing every 2 weeks and one initial viral dose, shown in **(A,B)**, and with one initial viral dose and a second viral dose at 1 week, shown in **(C,D)**. The second column, consisting of figures **(B,D)**, shows a zoomed in version of each of the plots to the left.

intermediate values of a_{AT}, there are ranges of innate immune levels leading to tumor clearance, interspersed with ranges leading to tumor growth. This suggests a much more complex relationship between the two facets of the immune system, when exposed to both therapies. For large and small s_{ZR} values, the tumor rebounds to its carrying capacity for all low-intermediate values of a_{AT}, confirming that the long-term behavior described above for the baseline a_{AT} is representative of the behavior in the extreme s_{ZR} ranges as a_{AT} decreases. Similarly to the discontinuities seen in **Figure 12B**, for intermediate parameter values there are a few irregular instances interspersed within the blue clearance region, in which the innate immune response timing precludes an essential viral oscillation, thus leading to tumor rebound, rather than immune-mediated clearance of the tumor. Note that in the uncolored regions, namely for intermediate values of s_{ZR} and low values of a_{AT}, the tumor starts to shrink early on, but then slowly rebounds when the adaptive immune populations begin to decline. At $t = 4,000$, the susceptible population falls between 4×10^8 and 5×10^8 for all simulations in this range, illustrated by a representative simulation in Figure S3 in the Supplementary Material, and eventually by about 10^5 h, the susceptible population falls within 0.1% of the carrying capacity.

Overall, we see that there is a significantly larger range of a_{AT} that makes the combination therapy effective, as compared to OVT alone. Additionally, there are ranges of innate immune strength that can be beneficial to the combination therapy, yielding eventual tumor clearance, which we did not see in the absence of anti-PD-1. The precise relationship between the innate and adaptive immune response to OVT and anti-PD-1 immunotherapy is still not well-understood, but our work suggests there are parameters regimes in which these operate

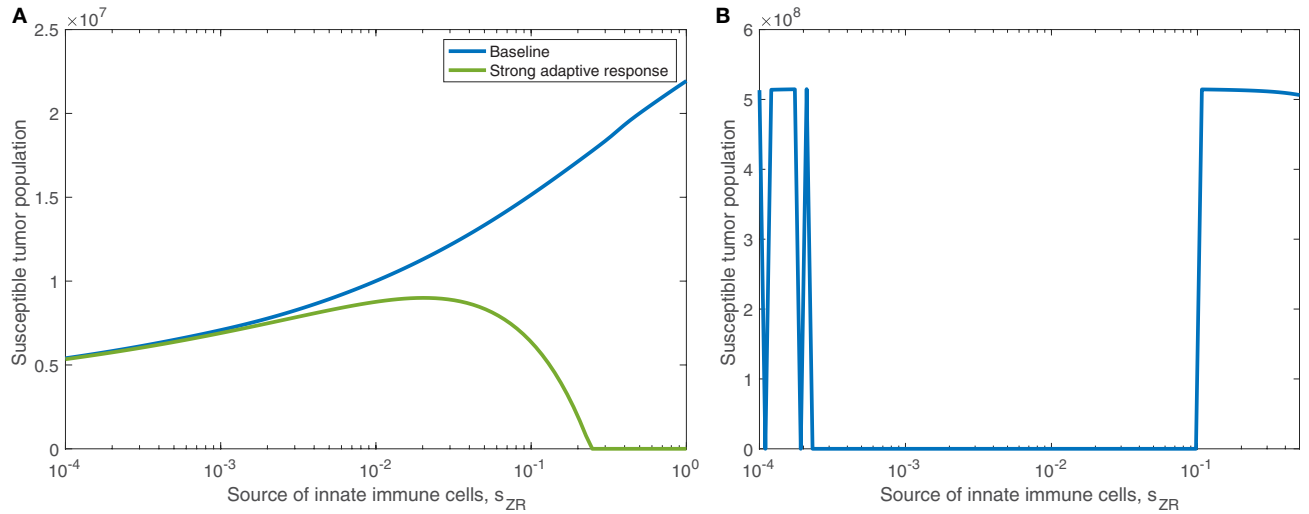

FIGURE 12 | (A) Displays the susceptible tumor population after 300 h with anti-PD-1, as the source of innate immune cells, s_{ZR} varies. We compare the baseline parameter regime with a stronger adaptive immune response, given by $a_{AT} = 0.05$. **(B)** Displays the number of susceptible tumor cells, after $t = 1,000$ h, as a function of the source of innate immune cells, s_{ZR}.

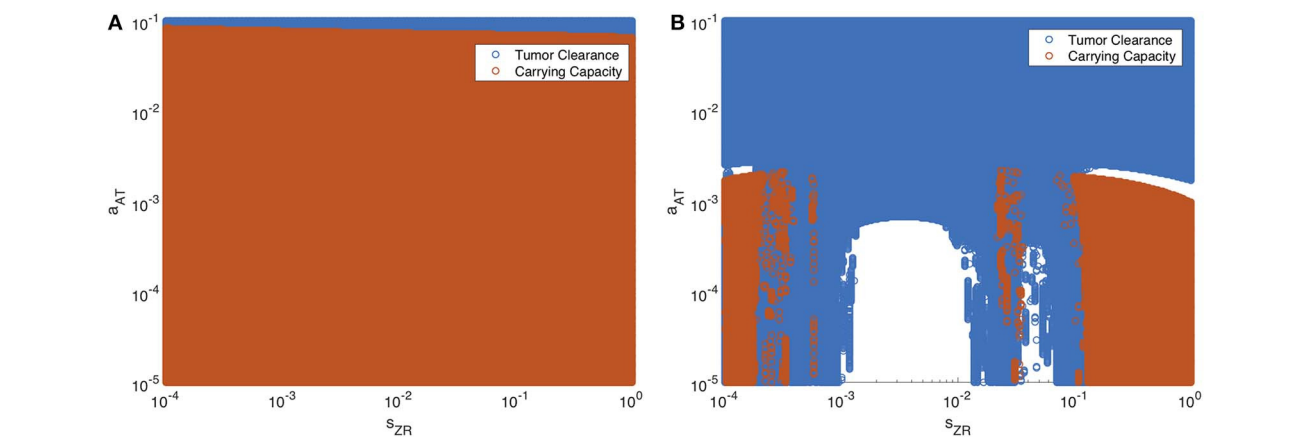

FIGURE 13 | Long-term behavior as innate and adaptive immune response vary, without and with anti-PD-1. We display parameter values in the $s_{ZR} - a_{AT}$ space that yield post-treatment tumor clearance in blue, and recurrence to tumor carrying capacity in red. The figure on the left shows the results after a single oncolytic viral dose is administered, while on the right, treatment included the initial viral dose, followed by a dose of anti-PD-1 every 2 weeks. All other parameters are set at their baseline values. **(A)** Oncolytic virus only **(B)** Oncolytic virus and anti-PD-1.

in synergy, when the anti-PD-1 allows the antitumor adaptive immune cells to be sufficiently active.

4. DISCUSSION

In this work, we first developed a model of GBM response to OVT and the resulting response from innate and adaptive immune cells. We parameterized the model using *in vivo* data from murine GBM models and performed sensitivity analyses to determine which parameters most significantly impact the tumor response to treatment. In Friedman et al. (2006), they concluded that a tumor cannot be eradicated by OVT unless the burst size is large. We found a similar limiting threshold, but in our model, this is a viral infection rate threshold, rather than burst size, below which tumor eradication is not possible. The infection rate is a modifiable viral feature, but effective oncolytic viral treatment requires an infectivity level that may not be biologically achievable. With a viral infection rate on the order of 10^{-9} $\text{pfu}^{-1}\text{h}^{-1}$, varying the strength of the adaptive immune response does not significantly improve tumor response to OVT alone, but it does increase the viral infection range under which the tumor can be eliminated. We found that a stronger innate immune response, driven primarily by an increase in the localization of the innate immune cells and the innate immune-mediated viral killing rate, leads to a less effective treatment, due to more

rapid viral clearance by macrophages and natural killer cells. Even when combined with a strong adaptive immune response, the innate immune response has an antagonistic effect on OVT efficacy. Our results suggest this is due to the limitations on T cell productivity, imposed by the PD-1/PD-L1 immune checkpoints.

Thus, we chose to incorporate a second cancer treatment within the model, via an immune checkpoint inhibitor, in order to investigate the effect of this immunotherapy in combination with OVT. In this case, the viral infection rate is still the most significant parameter on the short-to-intermediate time frame. However, the tumor antigenicity level is much more significant when the tumor is treated with the combination therapy than with OVT alone. This is indicative of the fact that the adaptive immune system plays a much more significant role in response to the combination therapy than to OVT alone. Under the combination therapy, there is a larger viral infectivity range under which the tumor can be eliminated, increasing the possibility of developing a sufficiently infectious virus to combine with anti-PD-1 to eliminate murine GBM. However, there is a high degree of sensitivity to the timing of viral infection and immune response, suggesting that subsequent doses following an initial viral dose may interfere with the stimulated immune response.

In addition, there is a much more complex relationship between innate and adaptive immune cells in the presence of both OVT and anti-PD-1; under some circumstances, when treating a highly antigenic tumor, increasing the strength of the innate immune response can improve treatment efficacy. Hence, on its own, OVT is unlikely to effectively treat GBM, but combining with anti-PD-1 can lead to successful treatment, particularly when treating highly antigenic tumors. In such cases, a more rapid innate immune response enhances, rather than counteracts, the treatment. Our work builds upon Wodarz' investigations into oncolytic viral and adaptive immune interactions in Wodarz (2001), by determining innate immune conditions required for effective viral treatment.

We supplement the study in Eftimie and Eftimie (2018), which focused on the role of macrophages in response to OVT, by combining this focus with the interactions between innate and adaptive immune cells. With the inclusion of immune checkpoints within our model, our results suggest that outside of very extreme cases, tumor elimination is not possible with OVT alone. However, when combining OVT with anti-PD-1, in tumors below a certain antigenicity threshold, we confirm Eftimie's conclusion that tumor elimination strongly depends on the total number of innate immune cells. For sufficiently high levels of antigenicity, the influence of innate immune activity diminishes. In the future, we would like to include both M1 and M2 macrophages within our model framework to determine whether this distinction affects our model results. Our model suggests that it may be beneficial to perform testing of immune cell levels within the tumor microenvironment and of tumor antigenicity, in order to improve predictions of treatment efficacy. Additionally, vaccinating the host with tumor-specific antigen could help to enhance the antitumor adaptive immune response, thereby improving treatment outcomes.

A limitation of this model is that it is not spatially explicit, so it does not account for the spatial distribution of various cell types and the diffusion of the virus and anti-PD-1 drug. We plan to extend this work by incorporating spatial heterogeneity within the tumor, in order to investigate the degree to which this heterogeneity impacts treatment efficacy. Additionally we calibrate our model parameters using data from mouse models, which prevents direct translation to human patients. However, our work provides information that can be used to inform a clinical trial. We would like to follow up this work by first validating our computational predictions using experimental mouse models, administering a combination of HSV and the immunotherapy nivolumab to GBM in a range of immune landscapes. Our work also suggests that investigating the maximum level of tolerable infectivity for oncolytic viruses would benefit GBM treatment development. Subsequently, if the experiments confirm the necessary conditions and dosing protocol that yield tumor control or elimination, then this could provide the impetus for a clinical trial for GBM patients.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

KS, SL, and TJ conceived the model and computational experiments. KS performed the computational experiments. KS and TJ analyzed the simulation model outputs. KS, SL, and TJ prepared and edited the manuscript.

FUNDING

This work was supported by Simon's Foundation Collaboration Grant 312622 (TJ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.00151/full#supplementary-material

REFERENCES

Alexander, B. M., and Cloughesy, T. F. (2017). Adult glioblastoma. *J. Clin. Oncol.* 35, 2402–2409. doi: 10.1200/JCO.2017.73.0119

Bagheri, N., Shiina, M., Lauffenburger, D., and Korn, W. (2011). A dynamical systems model for combinatorial cancer therapy enhances oncolytic adenovirus efficacy by MEK-inhibition. *PLoS Comput. Biol.* 7:e1001085. doi: 10.1371/journal.pcbi.1001085

Banerjee, S., Khajanchi, S., and Chaudhuri, S. (2015). A mathematical model to elucidate brain tumor abrogation by immunotherapy with T11 target structure. *PLoS ONE* 10:e0123611. doi: 10.1371/journal.pone.0123611

Barber, D., Wherry, E., Masopust, D., Zhu, B., Allison, J. P., Sharpe, A. H., et al. (2006). Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature* 439, 682–687. doi: 10.1038/nature04444

Brahmer, J., Drake, C., Wollner, I., Powderly, J., Picus, J., Sharfman, W., et al. (2010). Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates. *J. Clin. Oncol.* 28, 3167–3175. doi: 10.1200/JCO.2009.26.7609

Branwood, A., Noble, K., and Schindhelm, K. (1992). Phagocytosis of carbon particles by macrophages *in vitro*. *Biomaterials* 19, 646–648. doi: 10.1016/0142-9612(92)90035-M

Cheng, X., Veverka, V., Radhakrishnan, A., Waters, L. C., Muskett, F. W., Morgan, S. H., et al. (2013). Structure and interactions of the human programmed cell death 1 receptor. *J. Biol. Chem.* 288, 11771–11785. doi: 10.1074/jbc.M112.448126

de Pillis, L., Gu, W., and Radunskaya, A. (2006). Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations. *J. Theor. Biol.* 238, 841–862. doi: 10.1016/j.jtbi.2005.06.037

Eftimie, R., Dushoff, J., Bridle, B., Bramson, J., and Earn, D. (2011). Multi-stability and multi-instability phenomena in a mathematical model of tumor-immune-virus interactions. *Bull. Math. Biol.* 73, 2932–2961. doi: 10.1007/s11538-011-9653-5

Eftimie, R., and Eftimie, G. (2018). Tumour-associated macrophages and oncolytic virotherapies: a mathematical investigation into a complex dynamics. *Lett. Biomath.* 5, S6–S35. doi: 10.1080/23737867.2018.1430518

Friedman, A., Tian, J., Fulci, G., Chiocca, E., and Wang, J. (2006). Glioma virotherapy: effects of innate immune suppression and increased viral replication capacity. *Cancer Res.* 66, 2314–2319. doi: 10.1158/0008-5472.CAN-05-2661

He, J., Hu, Y., Hu, M., and Li, B. (2015). Development of PD-1/PD-L1 pathway in tumor immune microenvironment and treatment for non-small cell lung cancer. *Sci. Rep.* 5:13110. doi: 10.1038/srep13110

Hegi, M., Diserens, A., Gorlia, T., Hamou, M. F., de Tribolet, N., Weller, M., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003. doi: 10.1056/NEJMoa043331

Kaufman, H., Kohlhapp, F., and Zloza, A. (2015). Oncolytic viruses: a new class of immunotherapy drugs. *Nat. Rev. Drug Discov.* 14, 642–662. doi: 10.1038/nrd4663

Kim, Y., Lee, J., Lee, D., and Othmer, H. (2019). Synergistic effects of bortezomib-ov therapy and anti-invasive strategies in glioblastoma: a mathematical model. *Cancers (Basel)* 11:215. doi: 10.3390/cancers11020215

Kim, Y., Yoo, J., Lee, T., Yu, J., Caligiuri, M. A., Kaur, B., et al. (2018). Complex role of NK cells in regulation of oncolytic virus–Bortezomib therapy. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4927–4932. doi: 10.1073/pnas.1715295115

Kirschner, D., and Panetta, J. (1998). Modeling immunotherapy of the tumor-immune interaction. *J. Math. Biol.* 37, 235–252. doi: 10.1007/s002850050127

Komarova, N., and Wodarz, D. (2010). ODE models for oncolytic virus dynamics. *J. Theor. Biol.* 263, 530–543. doi: 10.1016/j.jtbi.2010.01.009

Lai, X., and Friedman, A. (2017). Combination therapy of cancer with cancer vaccine and immune checkpoint inhibitors: a mathematical model. *PLoS ONE* 12:e0178479. doi: 10.1371/journal.pone.0178479

Lee, K., Lee, D., Kang, J., Park, J. O., Kim, S. H., Hong, Y. S., et al. (2018). Phase I pharmacokinetic study of nivolumab in korean patients with advanced solid tumors. *Oncologist* 23, 155-e17. doi: 10.1634/theoncologist.2017-0528

Linsenmann, T., Jawork, A., Westermaier, T., Homola, G., Monoranu, C. M., Vince, G. H., et al. (2019). Tumor growth under rhGM? CSF application in an orthotopic rodent glioma model. *Oncol. Lett.* 17, 4843–4850. doi: 10.3892/ol.2019.10179

Mahasa, K., Eladdadi, A., de Pillis, L., and Ouifki, R. (2017). Oncolytic potency and reduced virus tumor-specificity in oncolytic virotherapy. A mathematical modelling approach. *PLoS ONE* 12:e0184347. doi: 10.1371/journal.pone.0184347

Mautea, R., Gordona, S., Mayere, A., McCrackena, M., Natarajane, A., Ring, N., et al. (2015). Engineering high-affinity PD-1 variants for optimized immunotherapy and immuno-PET imaging. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6506–E6514. doi: 10.1073/pnas.1519623112

McDonald, D., and Levy, O. (2019). "3–innate immunity," in *Clinical Immunology. 5th Edn.*, eds R. Rich, T. Fleisher, W. Shearer, H. Schroeder, A. Frew, and C. Weyand (London: Elsevier), 39–53.e1.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21, 239–245. doi: 10.1080/00401706.1979.10489755

Nikolopoulou, E., Johnson, L., Harris, D., Nagy, J., Stites, E., and Kuan, Y. (2018). Tumour-immune dynamics with an immune checkpoint inhibitor. *Lett. Biomath.* 5, S137–S159. doi: 10.1080/23737867.2018.1440978

Okamoto, K., Amarasekare, P., and Petty, I. (2014). Modeling oncolytic virotherapy: is complete tumor-tropism too much of a good thing? *J. Theor. Biol.* 358, 166–178. doi: 10.1016/j.jtbi.2014.04.030

Reynolds, A., Rubina, J., Clermont, G., Day, J., Vodovotz, Y., and Bard Ermentrout, G. (2006). A reduced mathematical model of the acute inflammatory response: I. Derivation of model and analysis of anti-inflammation. *J. Theor. Biol.* 242, 220–236. doi: 10.1016/j.jtbi.2006.02.016

Shi, L., Chen, S., Yang, L., and Li, Y. (2013). The role of PD-1 and PD-L1 in T-cell immune suppression in patients with hematological malignancies. *J. Hematol. Oncol.* 6:74. doi: 10.1186/1756-8722-6-74

Speranza, M., Passaro, C., Ricklefs, F., Kasai, K., Klein, S. R., Nakashima, H., et al. (2018). Preclinical investigation of combined gene-mediated cytotoxic immunotherapy and immune checkpoint blockade in glioblastoma. *Neuro Oncol.* 20, 225–235. doi: 10.1093/neuonc/nox139

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* 352, 987–996. doi: 10.1056/NEJMoa043330

Wishart, D., Feunang, Y., and Guo, A. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wodarz, D. (2001). Viruses as antitumor weapons. *Cancer Res.* 61, 3501–3507. Available online at: https://cancerres.aacrjournals.org/content/61/8/3501.

Wodarz, D., and Komarova, N. (2009). Towards predictive computational models of oncolytic virus therapy: basis for experimental validation and model selection. *PLoS ONE* 4:e4271. doi: 10.1371/journal.pone.0004271

Zhang, J., Stevens, M., and Bradshaw, T. (2012). Temozolomide: mechanisms of action, repair and resistance. *Curr. Mol. Pharmacol.* 5, 102–114. doi: 10.2174/1874467211205010102

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 April 2020
doi: 10.3389/fgene.2020.00314

Modeling Basins of Attraction for Breast Cancer Using Hopfield Networks

Alessandra Jordano Conforte[1,2], *Leon Alves*[3], *Flávio Codeço Coelho*[4], *Nicolas Carels*[1] *and Fabrício Alves Barbosa da Silva*[2*]

[1] *Laboratory of Biological Systems Modeling, Center for Technological Development in Health, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil,* [2] *Laboratory of Computational Modeling of Biological Systems, Scientific Computing Program, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil,* [3] *Applied Math School, Getúlio Vargas Foundation, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil,* [4] *Applied Math School, Getúlio Vargas Foundation, Rio de Janeiro, Brazil*

Cancer is a genetic disease for which traditional treatments cause harmful side effects. After two decades of genomics technological breakthroughs, personalized medicine is being used to improve treatment outcomes and mitigate side effects. In mathematical modeling, it has been proposed that cancer matches an attractor in Waddington's epigenetic landscape. The use of Hopfield networks is an attractive modeling approach because it requires neither previous biological knowledge about protein-protein interactions nor kinetic parameters. In this report, Hopfield network modeling was used to analyze bulk RNA-Seq data of paired breast tumor and control samples from 70 patients. We characterized the control and tumor attractors with respect to their size and potential energy and correlated the Euclidean distances between the tumor samples and the control attractor with their corresponding clinical data. In addition, we developed a protocol that outlines the key genes involved in tumor state stability. We found that the tumor basin of attraction is larger than that of the control and that tumor samples are associated with a more substantial negative energy than control samples, which is in agreement with previous reports. Moreover, we found a negative correlation between the Euclidean distances from tumor samples to the control attractor and patient overall survival. The ascending order of each node's density in the weight matrix and the descending order of the number of patients that have the target active only in the tumor sample were the parameters that withdrew more tumor samples from the tumor basin of attraction with fewer gene inhibitions. The combinations of therapeutic targets were specific to each patient. We performed an initial validation through simulation of trastuzumab treatment effects in HER2+ breast cancer samples. For that, we built an energy landscape composed of single-cell and bulk RNA-Seq data from trastuzumab-treated and non-treated HER2+ samples. The trajectory from the non-treated bulk sample toward the treated bulk sample was inferred through the perturbation of differentially expressed genes between these samples. Among them, we characterized key genes involved in the trastuzumab response according to the literature.

Keywords: breast cancer, Hopfield network, basin region of attraction of a minimizer, systems biology, dynamic system

OPEN ACCESS

Edited by:
Russell C. Rockne, City of Hope National Medical Center, United States

Reviewed by:
Ping Ao, Shanghai University, China
Carlos Espinosa-Soto, Universidad Autónoma de San Luis Potosí, Mexico

***Correspondence:**
Fabrício Alves Barbosa da Silva
fabricio.silva@fiocruz.br

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *19 December 2019*
Accepted: *16 March 2020*
Published: *07 April 2020*

Citation:
Conforte AJ, Alves L, Coelho FC, Carels N and Silva FAB (2020) Modeling Basins of Attraction for Breast Cancer Using Hopfield Networks. Front. Genet. 11:314. doi: 10.3389/fgene.2020.00314

1. INTRODUCTION

Cancer may be caused by genetic and epigenetic factors that deregulate cellular homeostasis. Hanahan and Weinberg (2011) classified cancer deregulated processes in terms of ten hallmarks, which include unlimited proliferative potential, cell death evasion, and angiogenesis, among others.

This disease was described as a pre-existing attractor in Waddington's epigenetic landscape in 2009 by Huang et al. (2009). An attractor is defined as a stable cell state of minimum energy and is associated with a cell phenotype. Due to the stochastic behavior of gene regulation, the attractor is surrounded by a basin of attraction resulting from other gene expression profiles (states) that sustain the same phenotype. The authors proposed that a cancer attractor would be enclosed within epigenetic barriers that should prevent its access. As a cell accumulates mutations and gene deregulation, these epigenetic barriers are lost, and the cancer attractor becomes accessible. This indicates that the epigenetic landscape is not rigid and may change over time (Ao et al., 2008; Huang et al., 2009). In this context, Ao et al. (2008) proposed that cancer can be classified as preventable, curable or incurable, according to its respective functional landscape.

Cancer was also described as an intrinsic robust state of the endogenous network (Ao et al., 2008; Su et al., 2017; Yuan et al., 2017b,c). The endogenous network theory (ENT) is a realistic network dynamic approach, in which the molecular-cellular network is composed of oncogenes, tumor suppressors, and other related agents, and covers most molecular functions. It represents a non-linear stochastic dynamical system able to generate multiple stable states and paths between them (Ao et al., 2008). In this context, coarse-grained modeling was applied using the non-linear Hill functions. The attractors found by this strategy matched gene expression profiles of cell phenotypes from colorectal, prostate, hepatocellular, and gastric cancer. This approach was also applied to acute promyelocytic leukemia and myelopoiesis (Su et al., 2017; Yuan et al., 2017b,c).

The most used methods in gene regulatory networks (GRNs) modeling are stochastic differential equations (SDE), ordinary differential equations (ODEs), and Boolean networks. SDEs have been used to identify the appropriate therapeutic approach against cancer, following the concept of ENT (Su et al., 2017; Yuan et al., 2017b,c). For instance, Yuan et al. (2017b) proposed perturbations that would lead colorectal cancer phenotypes toward the normal intestine phenotype. Either ODEs or SDEs have been used to model the regulation of p53 by MDM2 and MDMX (Leenders and Tuszynski, 2013), the tamoxifen-induced apoptosis in breast cancer (Rouhimoghadam et al., 2018), to predict the impact of combined therapies on myeloma growth (Ji et al., 2016), and to quantify the landscape for cell differentiation and cancer development (Li and Wang, 2013). On the other hand, Cornelius et al. (2013) used differential equations derived from a Boolean network to understand how a leukemia GRN could be switched from an active cell proliferation to an active cell death state. These methods required previous biological knowledge about protein-protein interactions and/or kinetic parameter rates, which may limit the network size and requires an extensive literature search.

The Hopfield network modeling is an alternative method that does not require kinetic parameter rates or protein-protein interactions knowledge. It uses the gene expression profile as input, and the GRN size is only limited by the available computational capacity. This method is a form of a recurrent artificial neural network and was popularized in 1982 by Hopfield (1982). It considers symmetric and asymmetric connections and ensures that sample states converge toward stored attractor patterns during computational modeling.

This method has been used to elucidate cell and cancer development. For instance, Fard et al. (2016) and Guo and Zheng (2017) analyzed single-cell data and identified attractors in the Waddington's epigenetic landscape related to developmental trajectories. In cancer-related reports, Hopfield networks have been used to identify attractors associated with cancer subtypes (Maetschke and Ragan, 2014) and stages (Taherian Fard and Ragan, 2017). Moreover, Szedlak et al. (2014) have used asymmetric Hopfield networks to test densely connected nodes as therapeutic targets and inferred the minimum number of genes necessary for treatment. Meanwhile, Cantini and Caselle (2019) developed a methodology to identify molecular similarities to stratify cancer patients and improve their therapies.

Stratification of patients may improve treatment outcomes through the identification of molecular targets common to a group of patients (He et al., 2019). For instance, trastuzumab is a monoclonal antibody, used as adjuvant treatment against breast and stomach cancers that overexpress the HER2 protein (Wang et al., 2019). Also, triple-negative breast cancer patients may present resistance to neoadjuvant chemotherapy due to pre-existing resistant cell phenotypes (Kim et al., 2018). Both studies were performed considering single-cell sequencing of tumor samples aiming to identify gene expression signatures.

Most one-size-fits-all medicine approach may cause harmful side effects due to low selectivity that might affect both tumor and healthy cells (Siegel et al., 2012). In contrast, personalized medicine considers the tumor of a patient as unique, and identifies genes differentially expressed in tumors in comparison to the surrounding tissue (stroma), which is used as a control (Carels et al., 2015; Conforte et al., 2019). For this reason, personalized medicine is expected to mitigate the side effects and improve treatment efficacy. *In vitro* validation of this approach showed that simultaneous inhibition of target combinations exhibited a more substantial disruptive effect on malignant cells than the sum of single inhibitions (Tilli et al., 2016).

In this report, we identified differentially expressed genes between tumors and their control paired samples from breast cancer patients and used them in Hopfield network modeling. After the characterization of tumor and control attractors, we developed a protocol to identify the best target combination, for each patient, that would minimize potential side effects and withdraw tumor samples from their basin of attraction. For this purpose, we prioritized gene selection according to four criteria: density, node degree, association with cancer-related biological

processes, and rate of gene activation in tumor samples. We also performed a further validation of our approach by simulating trastuzumab treatment effects. For that, we used single-cell and bulk RNA-Seq data from three HER2+ breast cancer samples, one treated and two untreated with trastuzumab.

To our knowledge, this is the first report that combines single-cell and bulk RNA-Seq data, personalized treatment concepts, and Hopfield network modeling with the aim of disrupting tumor sample stability.

2. MATERIALS AND METHODS

2.1. Identification and Characterization of Differentially Expressed Genes

Bulk RNA-Seq data was obtained from The Cancer Genome Atlas (TCGA) project housed by the Genomic Data Commons (GDC, portal.gdc.cancer.gov), accessed in June 2019. This data set comprises paired tumor and control (stroma) samples from 70 breast cancer patients. We used the FPKM version (Trapnell et al., 2010) normalized by the upper quartile method (Hyndman and Fan, 1996).

scRNA-Seq data, accession number GSE 75688, was obtained in the NCBI Gene Expression Omnibus database, accessed in February 2020. This data set comprises single-cell and pooled samples (bulk RNA-Seq) of primary breast tumor tissue from three HER2+ patients. Among them, one received adjuvant trastuzumab treatment and had 75 RNA-Seq samples available, while two patients did not receive any treatment and had 48 and 18 RNA-Seq samples, respectively (Chung et al., 2017).

We analyzed both RNA-Seq data sets by fold change (FC), aiming to identify differentially expressed genes (DEGs). This method quantifies the change between an initial and final value as the ratio of the final value over the initial one. For the RNA-Seq obtained from TCGA, we considered the tumor expression data as the final value and the control expression data as the initial value. Consequently, positive logFC values indicated higher expression values in tumor samples, while negative logFC values indicated higher expression values in control samples. The logFC values of all genes were calculated individually considering the paired samples of each patient, and then we calculated the average of each gene among all patients. On the other hand, for the RNA-Seq data obtained from NCBI Gene Expression Omnibus, we considered the treated expression data as the final value and the non-treated expression data as the initial value. Consequently, positive logFC values indicated higher expression values in treated samples, while negative logFC values indicated higher expression values in non-treated samples. The logFC values of all genes were calculated considering the bulk RNA-Seq samples because it represents a weighted average of the heterogeneous cells present in the sample.

For both RNA-Seq data sets, we used a p-value ≤ 0.01 and a false discovery rate (FDR) ≤ 0.01 as a threshold to select the DEGs. This threshold was associated with an average logFC >3 or < -3.

The DEGs found were characterized using the Gene List tool from the Panther Classification System (pantherdb.org) with respect to their biological process categories, following the Gene Ontology (GO) classification (Thomas et al., 2003; Mi et al., 2010) (**Supplementary Table 1**).

2.2. Clinical Data

We obtained the clinical data of each patient from the TCGA data set in TCGA-GDC, accessed in June 2019 (**Supplementary Table 2**), and appended data of molecular subtype, entropy, and overall survival. The molecular subtypes were defined according to the classification of The Cancer Genome Atlas Network (2012); the entropy values were obtained from Supplementary File 5 of Conforte et al. (2019); the overall survival (OS) of each patient were determined based on the OS data available in Liu et al. (2018). The OS data was analyzed with the Kaplan-Meier curve using GraphPad Prism software, where 1 is indicative of death, and 0 is indicative of censored data. The resulting curve indicates the percentage of patients alive after the OS time. We considered the overall survival of each patient as the percentage of patients alive on his/her OS time.

2.3. Hopfield Network

We applied the discrete neural Hopfield network to perform our analysis. This method implements an auto-associative network that can recover a pattern from partial discrete information (Hopfield, 1982). As input vectors, we used the binarized gene expression profile of each sample. For this, we considered the normal distributions of the logarithm of expression values from DEGs identified for each RNA-Seq data set and each condition (tumor/control or treated/non-treated) separately. We used the geometric mean as the threshold to binarize the gene state in each sample expression profile (Limpert et al., 2001) (**Supplementary Table 3**). This method allows the identification of genes with different states between samples, which is expected from DEGs. More importantly, it also allows for the identification of genes with the same state between samples, which must be considered since we selected DEGs based on the average logFC value of each gene among all samples from each data set.

The attractors were characterized as the centroids of their respective samples. Each centroid was composed of the average of states for each gene among its samples. The gene state was assigned a value of 1 for an average value >0.5, and 0 otherwise. Since we used the samples to define each attractor, our method for energy surface construction is parametric. Contrary to the non-parametric method used in Taherian Fard and Ragan (2017), the one applied in this work ensures the existence of basins of attraction related to each attractor.

The Hopfield network analysis was performed with Neupy, a library for neural networks in Python (www.neupy.com). Each attractor from the analyzed data set was used in the training phase to define its weight matrix. The weight matrix (W_a) is defined in Equation (1), where P is the attractor's gene expression profile, P^{T} is its transpose, and I is the identity matrix necessary to impose symmetric behavior with diagonal equal to zero. Since W may be composed of more than one stored pattern, its value is equal to the sum of all weight matrices (Equation 2).

$$W_a = (PP^{\mathrm{T}}) - I \tag{1}$$

$$W = W_{a1} + W_{a2} \tag{2}$$

The dynamic trajectory of each sample ($P_{(t+1)}$, defined in Equation 3), was predicted by following the synchronous approach as shown by Equation (3), where W is the final weight matrix, and $P_{(t)}$ is the sample gene expression profile at time t. The $sgn(x)$ function (Equation 4) determines the binarized output pattern.

$$P_{(t+1)} = sgn(P_{(t)}W) \tag{3}$$

$$sgn(x) = \begin{cases} 1 & : x \geq 0 \\ 0 & : x < 0 \end{cases} \tag{4}$$

By analogy to a physical system, the discrete Hopfield network energy (E) is calculated using the Lyapunov function, which guarantees convergence to a low-energy attractor state (see Equation 5) (Taherian Fard and Ragan, 2017).

$$E[P(s)] = \frac{-1}{2} PWP^{\mathrm{T}} \tag{5}$$

where $E[P(s)]$ is the energy of network state s for sample vector P and time t.

2.4. Samples Characterization

For the RNA-Seq data set from TCGA, the Euclidean distances (EDs) were calculated based on all network dimensions and implemented following three strategies: (i) calculation of the EDs between each sample and all other samples that converged to the same attractor, whose respective average was used to infer the sizes of the basins of attraction for the tumor and control sample attractors; (ii) calculation of the EDs between tumor samples and the control attractor (centroid); and (iii) calculation of the EDs between tumor samples and the tumor attractor (centroid). These values were correlated with the patients' clinical data.

We set two conditions to ensure the statistical significance of data correlations despite the data heterogeneity. First, the data of each clinical variable (tumor stage, molecular subtype, entropy, and overall survival) should group into at least three classes. Second, each class should include at least three patients to infer the average of the respective class.

The non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952) and the pairwise Wilcoxon signed-rank test (Wilcoxon, 1945) were performed to evaluate if all classes of the same clinical variable were significantly different. The null hypothesis of the non-parametric Kruskal-Wallis test is that all classes have the same average ED. When the null hypothesis was rejected, a pairwise Wilcoxon signed-rank test was performed to identify which class significantly deviated from the average. This test was performed for tumor stage, molecular subtype, and entropy clinical variables. For overall survival, we performed the Pearson correlation test. These statistical analyses were performed in R.

For the RNA-Seq data set from NCBI Gene Expression Omnibus database, we characterized the samples using principal component analysis (PCA) and a t-distributed stochastic neighbor embedding analysis (t-SNE) (Maaten and Hinton, 2008). Both tests were performed in Python.

2.5. Target Identification

Each DEG found for paired tumor and control samples was classified according to four parameters. For each parameter, the gene priority was screened in ascending and descending order. Parameter 1: density of each gene in the Hopfield network; Parameter 2: number of GOs related to cancer development associated with each of the DEGs; Parameter 3: number of patients with the gene under consideration active (1) in their tumor samples (biomarker); and Parameter 4: node degree of each gene.

Parameter 1 was determined following Equation (6), where the density (D) of node i is the sum of all weights in W for node i divided by the number of network nodes (n). Negative values for w_{ij} indicated different states for nodes i and j in the stored patterns, while the opposite is true for positive values.

$$D_i = \frac{1}{n} \sum_{j=1}^{j=n} w_{ij} \tag{6}$$

The second parameter is determined by the number of GOs, identified in the Panther Classification System as related to cancer development, associated with each DEG. On the other hand, parameter 3 identified the number of patients with an active DEG in the tumor sample and inactive in it respective control sample. In this case, we may hypothesize that genes active in many tumor samples, and inactive in their respective control samples, may be considered as breast cancer biomarkers.

The node degree of each gene, parameter 4, was determined according to the human interactome, obtained from the intactmicluster.txt file (version updated December 2017) accessed on January 11, 2018, at ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.txt. This file presents 151,631 interactions among 15,526 human proteins with UniProtKB accessions. The node degree of each protein was calculated through automated counting of their edges (**Supplementary Table 4**). We analyzed the node degrees of DEGs for which we found equivalence between the Ensemble and UniProtKB accessions (215/324), used for the RNA-Seq data and the interactome, respectively.

As stated in algorithm 1, we tested the effect of switching off 1–20 genes, according to priority lists, and analyzed the resulting energy values. This experiment was performed with the aim of identifying the number of genes that needed to be inhibited to move tumor samples away from their tumor basin of attraction. Our strategy followed the personalized medicine concept, considering the paired tumor and control samples of each patient. Moreover, to avoid potential side effects, we only switched off genes that were active (1) in the patient's tumor sample and inactive (0) in the paired control sample (stroma).

The implementation of algorithm 1 is available upon request. The algorithm considers two functions: **length**, which is the vector extent, and **energy**, which calculates the sample-related energy as described above. We analyzed each patient's tumor (*patientTumorSample*) and control (*patientControlSample*) sample gene expression profiles and searched for genes (*gene*) in the gene priority lists (*listOfGenes*). If a gene was active in

Algorithm 1

1: *Input*: Priority list of genes for inhibition (*listOfGenes*) / Patient control sample expression profiles (*patientControlSample*) / Patient tumor sample expression profiles (*patientTumorSample*).
2: *Output*: Combination and number of gene inhibitions recommended for each patient (*inhibitedGenes*, length of *inhibitedGenes*).
3: **procedure** GENE INHIBITION(*listOfGenes*, *patientControlSample*, *patientTumorSample*)
4: *Energy* = 0
5: *attemps* = 0
6: *patientTreated* = *patientTumorSample*
7: *inhibitedGenes* = {}
8: **while** (length (*inhibitedGenes*) < 20) and (*attemps* < 100) **do**
9: *attemps* += 1
10: **for all** *gene* in *listOfGenes* **do**
11: **if** *gene* = 0 in *patientControlSample* **then**
12: **if** *gene* = 1 in *patientTumorSample* **then**
13: *gene* = 0 in *patientTreated*
14: **add** *gene* to *InhibitedGenes*
15: **if** energy(*patientTreated*) ≥ −35000 **then**
16: **break**
17: **return** *inhibitedGenes*, length(*inhibitedGenes*)

the patient tumor sample and inactive in its paired control, it would be inhibited, and a new patient gene expression profile (*patientTreated*) was retrieved. If the change was not sufficient to reach an energy level equivalent to the barrier height between the two attractors (−35,000), the algorithm would continue to the second gene (*gene*) in the gene priority list (*listOfGenes*). When the barrier height energy was reached, the algorithm would leave the "for" loop, and the next patient gene expression profile would be analyzed. This algorithm returns the number of inhibitions indicated to move each tumor sample away from the tumor basin of attraction.

A similar algorithm was used to test the gene target combinations able to move tumor samples toward the control attractor. In this case, we tested switching off 1–50 genes, according to the priority lists. Instead of calculating the energy of the new patient gene expression profile (*patientTreated*) in line 15 of algorithm 1, we predicted its convergence toward the tumor or control attractor. If the changes were not sufficient to induce tumor sample convergence toward the control attractor, then the algorithm would continue to the next gene (*gene*) in the gene priority list (*listOfGenes*); otherwise, it would leave the "for" loop, and the next patient gene expression profile would be analyzed. In this case, the algorithm returns the number of inhibitions indicated to move each tumor sample toward the control attractor.

For trastuzumab treated and non-treated patients, we simulated the effect of trastuzumab treatment in the non-treated RNA-Seq sample, aiming to further validate our personalized approach. To do this, we built an energy landscape with single-cell and bulk RNA-Seq samples from non-treated patient 1 and the treated patient. We could not perform this experiment for non-treated patient 2 because it did not have enough samples to build its basin of attraction.

The state transition from the non-treated bulk RNA-Seq sample toward the treated bulk RNA-Seq sample was inferred through the perturbation of genes differentially expressed between those samples. Each DEG would receive the same state than the one in the treated bulk RNA-Seq sample, creating a new transitory state. The connection between all transitory states defines the trajectory from the non-treated basin of attraction toward the treated one. Besides, we characterized each DEG according to their role in signaling pathways associated with trastuzumab response (**Supplementary Table 7**).

We used bulk RNA-Seq samples as reference in the state trajectory because they comprise single-cell heterogeneity and abundance, as a weighted average of all single-cell samples available.

3. RESULTS

3.1. Characterization of Differentially Expressed Genes

We identified 324 DEGs among the paired tumor and control samples. We binarized the gene expression profiles from each patient following the normal distribution found for the logarithms of expression values (RNA-Seq data) of all tumor and control samples, separately, then used the geometric mean as a threshold (Limpert et al., 2001) (**Figure 1**). It is important to highlight that the results are sensitive to the chosen threshold and the geometric mean is the best fit for our samples (**Supplementary Table 3**).

From the 324 DEGs, 295 were recognized by the Panther Classification System. The Gene List tool found 1,918 GO codes related to biological processes, among which 111 were related to cancer development or response. Of all DEGs, 65.4% were classified with at least one cancer-related GO. All of these results can be analyzed in more detail using **Supplementary Table 1**.

We classified the cancer-related processes into ten onco- or tumor suppressor processes (**Figure 2**), from which seven corresponded to well-described hallmarks, and three were cancer-related pathways. The identified hallmarks were proliferation, cell death, cell migration, metabolic process, inflammatory response, cell growth, and angiogenesis, while the cancer-related pathways were the MAPK, WNT, and Rho GTPase pathways (Hanahan and Weinberg, 2011).

Note that inflammatory-related GOs could not be differentiated into acute or chronic response, which hampers elucidation of association with cancer formation, since it can only be triggered by chronic inflammation (Gonzalez et al., 2018). Furthermore, we included the regulation of mTOR signaling in both the proliferation and cell death categories due to its essential role in these hallmarks (Tian et al., 2019). Lastly, the hypoxia response characterized the metabolic cancer hallmark and is triggered by the hypoxia-inducible factor (HIF),

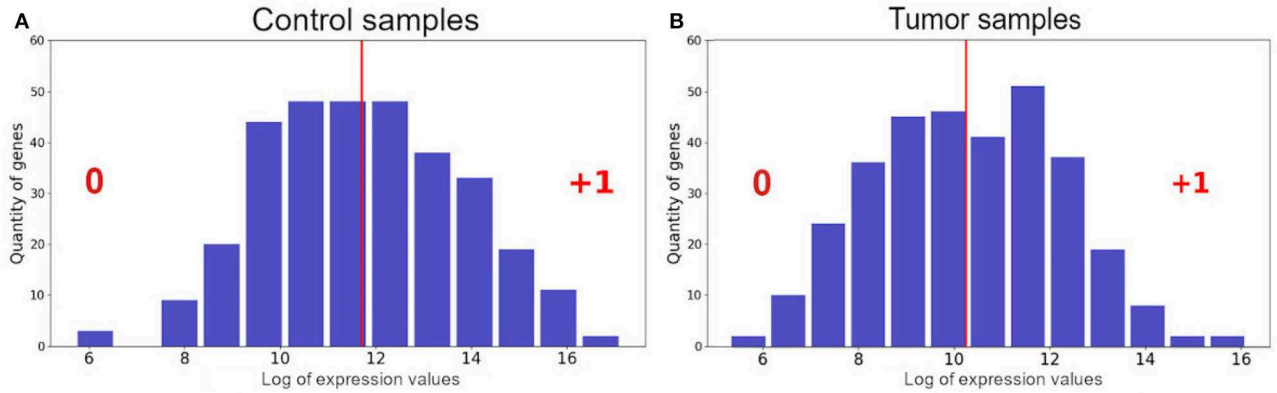

FIGURE 1 | Distribution of the logarithm of expression values from control samples **(A)** and tumor samples **(B)**. The red line represents the geometric mean. The numbers "0" and "+1" are the binarized states that each gene received according to each sample expression profile.

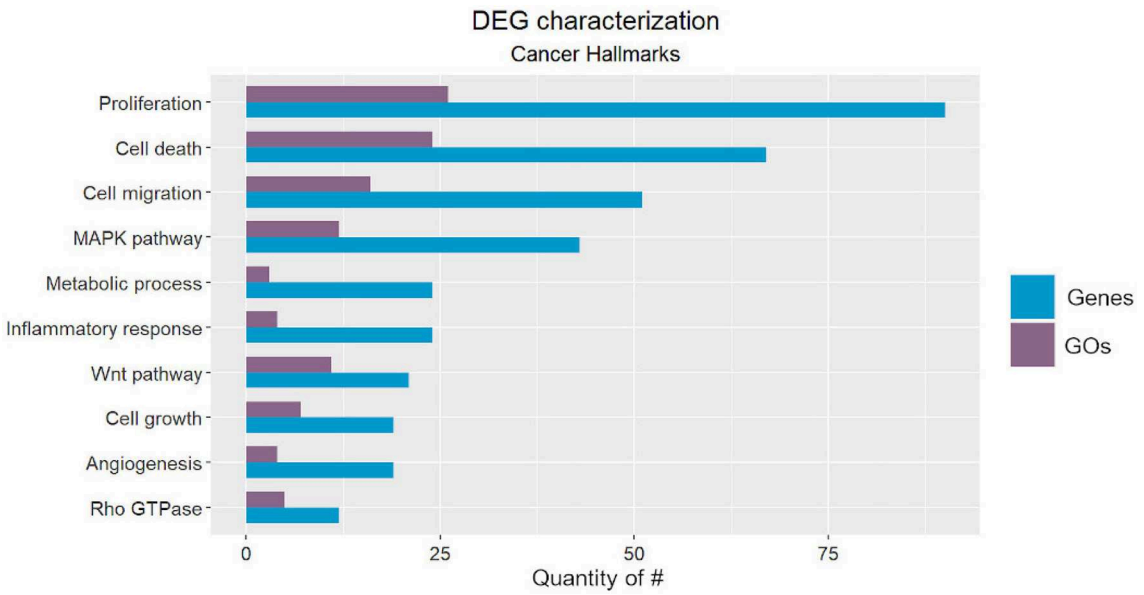

FIGURE 2 | Number of Gene Ontology (GO) codes found for each cancer-related hallmark/pathway (purple) and number of Differentially Expressed Genes (DEGs) classified according to each GO (blue).

which is directly related to the establishment of the "Warburg Effect." This metabolic rewiring occurs when tumor cells activate ATP generation via glycolysis (Simon, 2006; Liberti and Locasale, 2016). This process is essential for cancer cell survival under hypoxic stress. HIF transcriptionally regulates hundreds of genes that are also related to invasion, metastasis, genetic instability, and immune response (Yan et al., 2019).

Among the pathways that we identified, the MAPK pathway has been characterized as a key regulator of cancer development and is associated with several cellular processes, such as proliferation, growth, apoptosis, and migration. It involves other essential kinases (ERK and JNK) and proteins (RAS, Raf, and MEK), which can be reviewed in more detail in Dhillon et al. (2007).

The WNT pathway is divided into two main types: canonical and non-canonical (via JNK cascade) signaling pathways. Its function was first described in the developmental processes of *Drosophila melanogaster* and has been recently associated with cancer cell proliferation, stemness, metastasis, and immune evasion. This pathway can be reviewed in Zhan et al. (2017).

Finally, the Rho GTPase pathway has been associated with remodeling of the actin cytoskeleton, which is related to cell division and phenotype transition. It participates in cancer cell migration, proliferation, survival, and death. Its role in all of these signaling pathways can be reviewed in more detail in Haga and Ridley (2016).

TABLE 1 | Tumor and control basin of attraction sizes according to the Euclidean distances of all samples and samples that converged to each attractor.

	ED for all samples	ED for samples that converged to each attractor
Tumor	8.34	8.19
Control	7.92	8.17

3.2. Attractors Analysis

The Euclidean distances (EDs) were calculated considering all data dimensions. The EDs between all tumor and control paired samples revealed that the tumor basin of attraction was, indeed, larger than that of the control, when considering all samples that converged to the expected attractor. However, the size difference between both basins of attraction, considering samples that converged to each attractor, depended on the second decimal value and may not be considered as meaningful (**Table 1**). The difference among ED considering all samples and ED considering samples that converged to each attractor can be explained by the fact that five tumor samples converged to the control attractor, and because of that, were included in the ED measurement of the control basin of attraction. Interestingly, these samples were classified by (i) molecular subtypes with good prognosis (LumA or LumB); (ii) being in the initial stages of tumor development (stages i and iib); and (iii) most of them (A0BM, A0C3, A1EU, and A2FF) presenting the smallest entropies among patients (**Supplementary Table 2**), which is associated with a low level of aggressiveness (Breitkreutz et al., 2012; Conforte et al., 2019).

When comparing the clinical data with the EDs between the control attractor and the tumor samples that converged toward the tumor attractor, we found that the averages among the groups of tumor stages, molecular subtypes, and entropies were not significantly different when considering the non-parametric Kruskal-Wallis test. The result is the same when considering the ED between the tumor attractor and the tumor samples that converged toward the tumor attractor. Nevertheless, the Pearson correlation test, performed for the overall survival groups, showed a significant negative correlation with the EDs between the control attractor and the tumor samples that converged toward the tumor attractor ($r = -0.85$). A p-value = 0.001 and a slope = -0.07, with a 95% confidence interval between -0.11 and -0.03, which indicates that the regression line slope is different than zero. As expected, smaller distances are related to higher overall survival. No correlation was found among the overall survival groups considering the ED between the tumor attractor and the tumor samples that converged toward the tumor attractor. The Kruskal-Wallis test and the Pearson correlation test results are shown in **Figure 3**.

The same energy value was found for both tumor and control attractors ($-55{,}000$), corresponding to the local minimum of the energy function. The potential energy analysis revealed that tumor samples are more associated with a lower energy level and are closer to their attractor minimum energy than the control samples (**Table 2**) (for more details, see **Supplementary Table 5**). This result follows a common biological trend of tumors presenting alternative pathways that ensure tumor stability and promote tumor resistance to chemotherapy. The energy landscape based on those samples is presented in **Figure 4**.

3.3. Attractor Transitions

Hypothesizing that the genes involved in tumor state stability are essential for tumor maintenance, it may be reasonable to argue that the disruption of their products (mRNA, protein) might lead to tumors moving toward an attractor associated with active cell death. Accordingly, we searched for the best key gene combinations for each patient by considering their gene expression profiles that, when switched off, would minimize side effects and move tumor samples away from the tumor basin of attraction. This exercise provides a measure of how many key genes should be inhibited in each tumor sample to withdraw it from the tumor basin of attraction.

Supplementary Table 6 presents the values attributed to each gene when considering the four prioritization parameters: density of each gene in the Hopfield network; number of GOs related to cancer development associated with each gene; number of patients with the gene under consideration active (1) only in their tumor samples (biomarker); and node degree. Each parameter was analyzed in ascending and descending order.

Figure 5 shows that the inhibition of two targets would be sufficient to move ∼55 tumor samples (78.6%) away from their basin of attraction. Moreover, the parameters regarding the number of GOs and node degree were not effective for identifying key genes with the potential to change gene expression patterns in the Hopfield network since their ascending and descending orders exhibit similar behavior. On the other hand, the biomarker and density parameters showed different behaviors when considering their ascending and descending orders. The descending biomarker curve moved more tumor samples away from their basin of attraction than its respective ascending curve, while the opposite trend was observed for the node density parameter. These curves are similar due to their node selection strategy.

The descending biomarker curve prioritizes genes that are active in several tumor samples and inactive in their respective paired control samples. In other words, these genes present different states between tumor and control samples for most patients. Similarly, the ascending density curve prioritizes nodes with negative connection weights in the weight matrix. As revealed in the methodology section, interacting nodes with different states present a negative connection weight. Additionally, we impose the restriction that only nodes active in tumor samples and inactive in control samples are suitable for inhibition. For these reasons, the descending biomarker order and the ascending density order of prioritization matched.

According to the biomarker classification, CNTFR-alpha presented the highest value (70), which means that it was active in all tumor samples analyzed and inactive in all normal samples. CNTFR-alpha has been associated with proliferation and poor prognosis and been proposed as a biomarker of low-grade gliomas (Lu et al., 2012; Fan et al., 2017). SGK2 and PLP1 appeared in the second position of biomarker classification, being active in the tumor samples of 69 patients and inactive in their respective controls. SGK2 has been associated

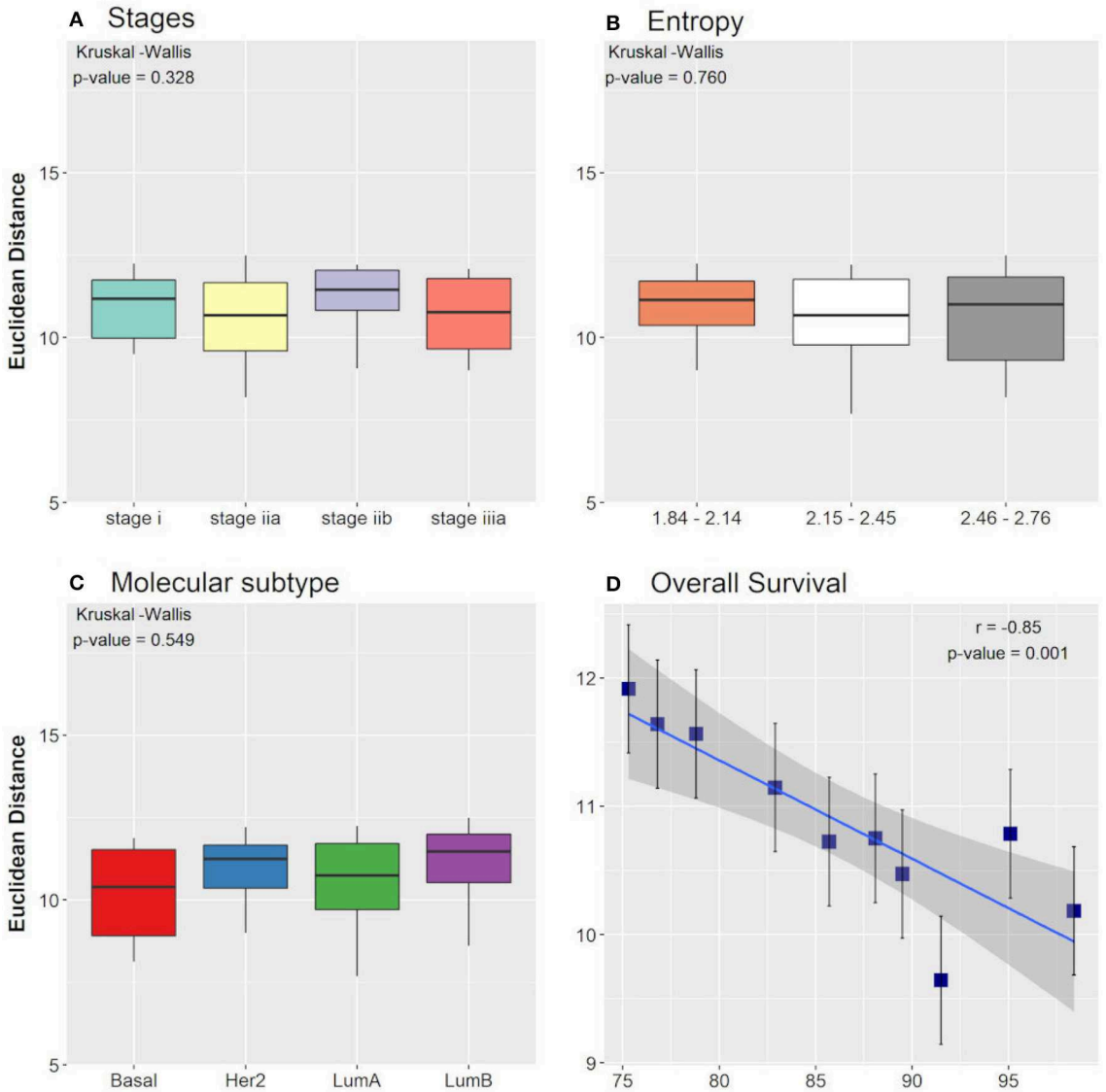

FIGURE 3 | Boxplot of Euclidean distances found between the control attractor and tumor samples, considering cancer stage **(A)**, entropy **(B)**, and molecular subtype **(C)** clinical data classifications. The *p*-values found in Kruskal-Wallis test are indicated in each figure. **(D)** Pearson correlation between the overall survival clinical data classification and the Euclidean distances found between the control attractor and tumor samples. The correlation coefficient (r) and the *p*-value found are indicated in the figure.

with hepatocarcinoma progression and bladder cancer cell proliferation, migration, and invasion (Liu et al., 2017; Chen et al., 2018). PLP1, although active in most tumor samples analyzed in this work, has been recently described as consistently downregulated in several cancer types, including breast cancer (Li et al., 2017). We did not find any lines of evidence in the literature that enable its association with cancer development. Nevertheless, it is described as a cancer gene in GeneCards, and its antibody is an effective inhibitor of cell growth in breast cancer (www.mybiosource.com—#7005540).

We also tested the number of targets necessary to bring tumor samples toward the control attractor. **Figure 6** shows that the inhibition of 50 targets is necessary to bring 18–26 tumor samples (25.7–37.2%) to the control attractor. As stated above, the descending biomarker curve and ascending density curve were the most promising gene selection parameters since they enabled movement of the largest number of tumor samples toward the control attractor, through the inhibition of few genes. This result indicates that it is not feasible to bring a tumor sample back to the normal phenotype.

3.4. Simulation of Trastuzumab Treatment Effect

The two-dimensionality reduction methods, PCA and t-SNE, were able to separate gene expression profiles from treated and non-treated single-cell and bulk samples according to their

similarities. The PCA analysis is plotted in **Figure 7A**. This figure shows that the PCA was able to explain the variance among the samples with two principal components (PCs), and separated treated and non-treated samples into two clusters. The first PC explained 22.03% of the total variance, while the second PC explained 8.14% of it.

The t-SNE analysis (**Figure 7B**) also separated the samples into two main clusters: trastuzumab treated and non-treated samples. Moreover, it distinguished non-treated samples of each HER2+ patient. These results indicated that the DEG selected for this data set not only represents the trastuzumab treatment but also, the differences between the gene expression profile from HER2+ patients. In addition, these results are in agreement with previously published results (Wang et al., 2019).

TABLE 2 | Average energy, average energy distance between samples and their respective attractor energy minimum and attractor energy minimum for tumor and control samples.

	Average energy of samples	Average energy distance between samples and attractors	Minimum (attractor) energy
Tumor	−31,338	24,188	−55,000
Control	−29,426	26,100	−55,000

The energy landscape built for HER2+ treated and non-treated samples is shown in **Figure 8A**. It is interesting to note that samples from both non-treated patients shared the same basin of attraction, even though we performed the Hopfield training phase considering three attractors, one for each patient. Besides, the trastuzumab treated samples composed a different basin of attraction with a minimum energy higher than the one found in the non-treated basin of attraction. This result indicated that non-treated samples have a higher tumor phenotype stability than treated samples, and agrees with the fact that trastuzumab treatment is, indeed, an adjuvant therapy approach rather than a healing one.

We can also observe in **Figure 8** that the bulk RNA-Seq samples (squares) may be away from their respective single-cells (spheres). This may occur because bulk RNA-Seq is a weighted average of all single-cells in the tumor tissue, and because of that, influenced by single-cell relative amounts. Besides, the bulk RNA-Seq may comprise single-cell phenotypes that were not obtained during sequencing.

Paired samples from the same patient, before and after trastuzumab treatment, were not available. We used samples from three different patients, two non-treated with trastuzumab (patients 1 and 2) and one treated (patient 3). Since samples from patient 1 and 2 belonged to the same region of the epigenetic landscape (see **Figure 8**), we hypothesized that their corresponding treated samples would also be in the same basin of attraction composed of samples from patient 3.

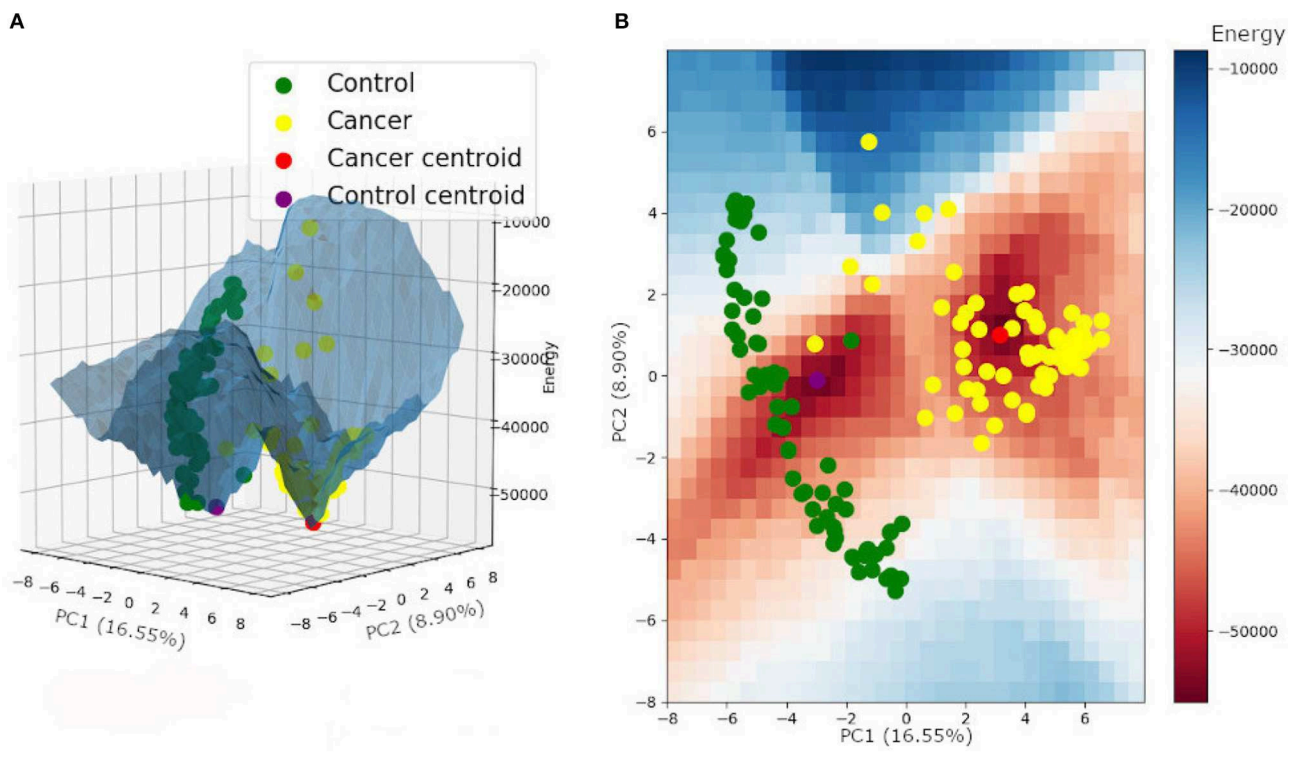

FIGURE 4 | Energy landscape built for control and tumor attractors, and samples, plotted on a three-dimensional **(A)** and a two-dimensional grid **(B)**.

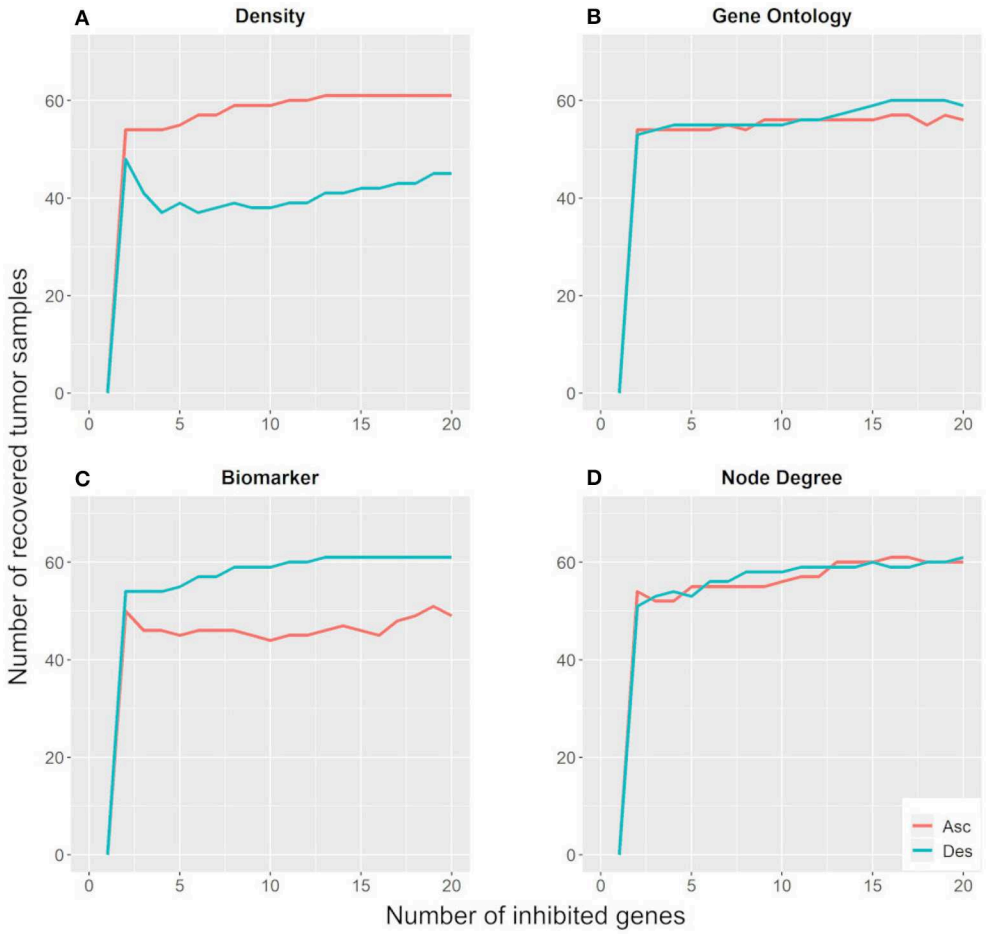

FIGURE 5 | Number of tumor samples that moved away from the tumor basin of attraction, according to the number of genes inhibited for each parameter: density **(A)**, gene ontology **(B)**, biomarker **(C)**, and node degree **(D)**. "Asc" represents the ascending order, while "Des" represents the descending order.

In this context, we tested the effects of trastuzumab treatment in the non-treated bulk RNA-Seq from patient 1. To do this, we built an energy landscape based on non-treated patient 1 samples and the treated patient samples. We identified 172 differentially expressed genes between the non-treated and treated bulk RNA-Seq samples. All DEGs were perturbed according to the treated gene expression profile. Those changes created new transitory states that, when connected, formed the trajectory between both basins of attractions (**Figure 8B**).

The target profile was reached after the perturbation of all 172 DEGs, and each triangle in **Figure 8B** represents the perturbation of a subset with ∼30 DEGs. The trajectory found was mapped in the energy landscape defined by the Hopfield network, and is one among many other possibilities. This trajectory was not optimized because we did not consider the associated signaling pathways, which would reduce the number of DEGs to be perturbed.

This experiment was not performed for the non-treated samples from patient 2 because it did not have enough samples to build its basin of attraction.

Changing 172 genes expression values is not feasible in the medical context, if we consider each intervention individually. However, key genes may initiate a cascade response that involves many others. For this reason, we characterized the DEGs concerning their respective biological processes in the Panther Classification System. Among the 172 DEGs, 92 were characterized by the Panther Classification System with at least one biological process (**Supplementary Table 7**). We reduced this analysis considering the biological processes involved in the trastuzumab treatment response. Among them, we can highlight cell proliferation, transcription, apoptosis, motility, and immune response (Herbst, 2004; Shi et al., 2014). Through a literature search, we saw that the genes involved in those biological processes have their role, in trastuzumab treatment, characterized.

The STAT1 gene is involved in proliferation and transcription biological processes. This gene plays an important role in HER2 inhibition and is activated after the trastuzumab treatment, through interferon-gamma production by the mobilized natural killer cells (Shi et al., 2014). This interferon-gamma production

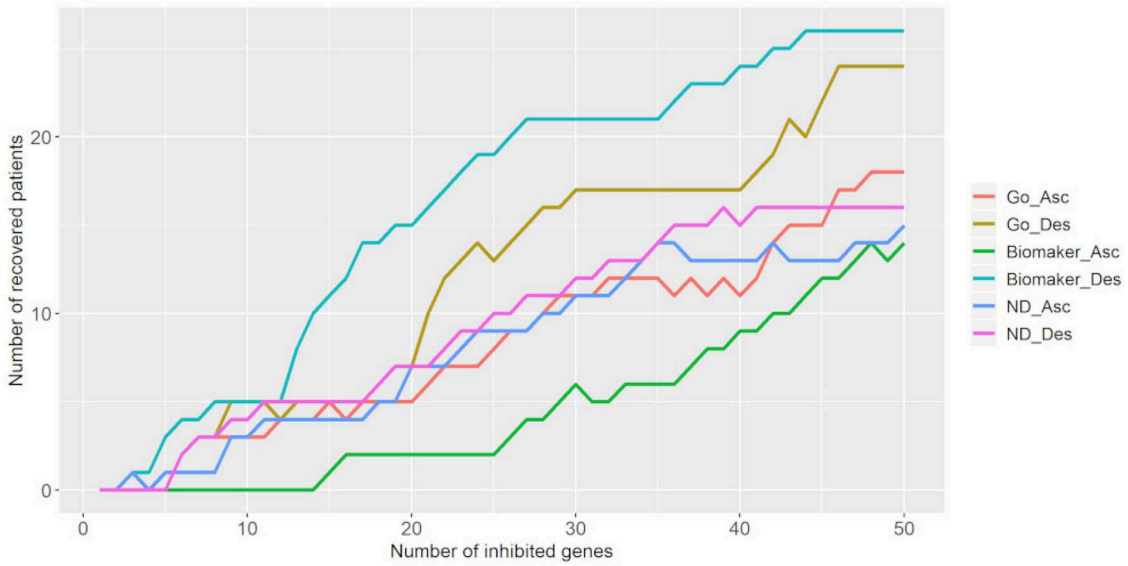

FIGURE 6 | Number of tumor samples that converged toward the control attractor, according to the number of genes inhibited for each parameter: gene ontology (GO), biomarker, and node degree (ND). "Asc" represents the ascending order, while "Des" represents the descending order. The biomarker ascending curve matches the density descending curve, and the biomarker descending curve matches the density ascending curve.

also activates the HLA-A antigen, involved in the immune response (Chaganty et al., 2015). The MEL-18 gene (or PCGF2) was classified in the transcription biological process. This gene is described as essential for trastuzumab treatment since its inhibition may result in a trastuzumab-resistant phenotype (Lee et al., 2019).

In this context, CCR7, PIP, and GBP1 genes were classified in the immune response biological process. CCR7 determines a cancer stem cell phenotype through the Notch signaling pathway, and PIP belongs to the PI3K signaling pathway. Both signaling pathways are related to trastuzumab treatment resistance (Pohlmann et al., 2009; Baker et al., 2014; Boyle et al., 2017). Besides, GBP1 and IFI27 were associated with the apoptotic biological process, and were related to breast cancer phenotype resistant to trastuzumab treatment (von der Heyde et al., 2015).

New therapeutic targets have been explored to overcome trastuzumab resistance. The ITGB6 gene, associated with motility biological process, has been proposed as a therapeutic target through inhibition by 264RAD antibody. Its combination with trastuzumab treatment was able to stop tumor growth even in trastuzumab-resistant cells (Moore et al., 2014).

The results obtained in this section propose a further validation of our personalized approach.

4. DISCUSSION

The Hopfield network was efficient in revealing cancer attractors related to molecular subtypes and developmental stages in previous works (Maetschke and Ragan, 2014; Taherian Fard and Ragan, 2017). In this report, we considered the gene expression profile of paired tumor and control samples from breast cancer patients to analyze both normal and tumor attractors, infer the best target combinations able to withdraw the tumor sample from its basin of attraction and simulate the trastuzumab treatment effect in non-treated bulk RNA-Seq sample.

Among our data, only five tumor samples converged to the control attractor. These samples presented molecular subtypes with a good prognosis, were in initial stages of cancer development, and had low entropy values. The Shannon entropy has been widely explored as a cancer development measure and aggressiveness indicator. Higher entropy values are associated with aggressive tumor phenotypes. This correlation was found when comparing cancer and control cells, advanced and initial stages of tumor development, aggressive cancer types and good prognosis cancer types (Breitkreutz et al., 2012; Winterbach et al., 2013; Banerji et al., 2015; Conforte et al., 2019).

The energy values of tumor samples were closer to the tumor attractor minimum compared to control samples. Although this result differs from the one published by Taherian Fard and Ragan (2017), this behavior is expected because tumor samples have alternative pathways that ensure their phenotypic stability (Fumi and Martins, 2013; Taherian Fard and Ragan, 2017).

Also, the correlation between the Euclidean distances from tumor samples to the control attractor and the patient's overall survival indicates that there is a higher chance of treatment success when the gene expression profile is not yet fully reprogrammed for cancer development. However, those results did not hold considering the Euclidean distance from tumor samples to the tumor attractor. This analysis also revealed that the tumor basin of attraction is larger than that of the control and should comprise more heterogeneous data, as indicated by Taherian Fard and Ragan (2017).

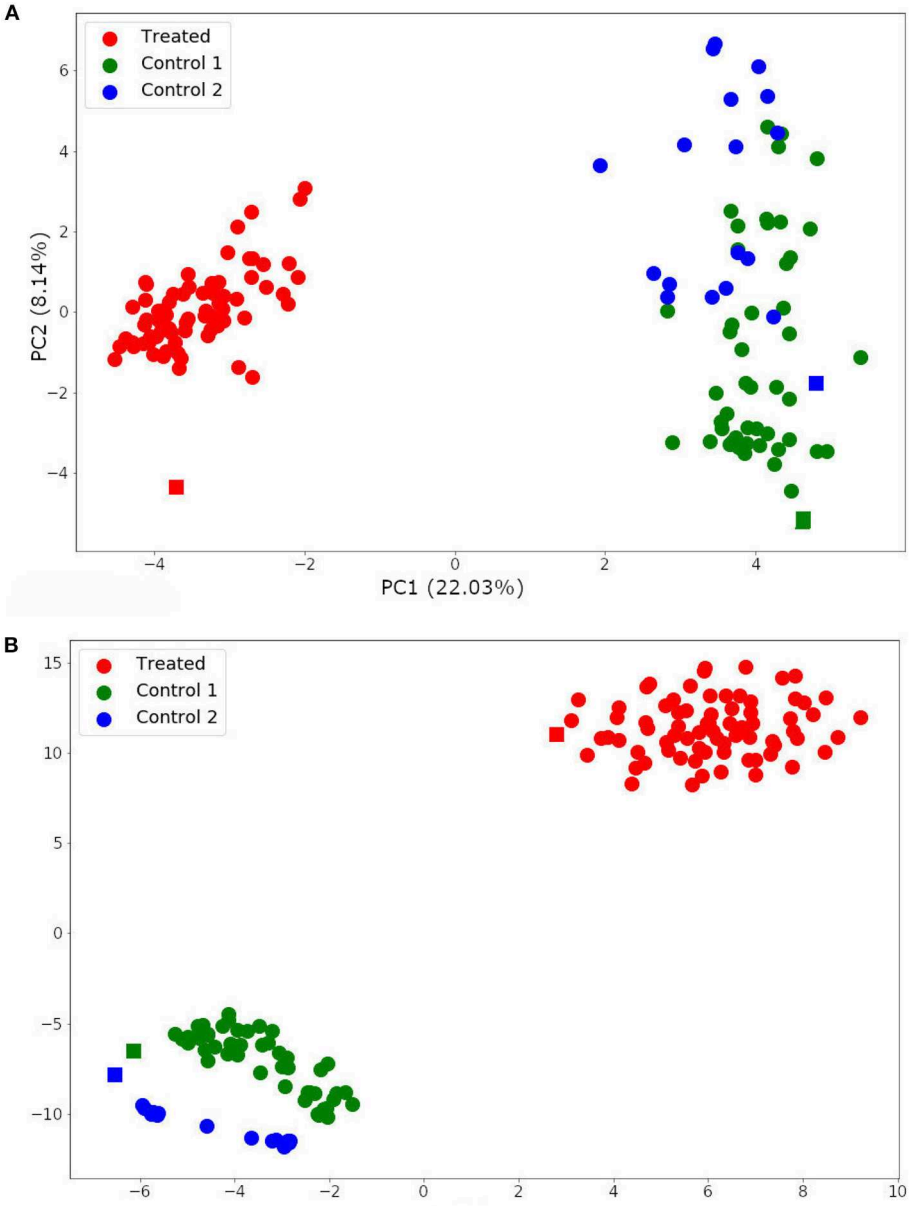

FIGURE 7 | Single-cell and bulk RNA-Seq data analysis by Principal Component Analysis (PCA) **(A)** and t-distributed stochastic neighbor embedding (t-SNE) **(B)**. Spheres represent single-cell samples and squares represent their respective bulk RNA-Seq samples.

The analysis of single-cell data allowed us to see the tumor basin of attraction in more detail and, along with the results discussed above, indicates that there may be multiple basins of attraction related to cancer development, rather than one large basin of attraction that comprises all cancer samples. Cancer basins of attraction could be composed of similar gene expression profiles. For instance, samples from the same molecular subtype.

In this context, our results showed that non-treated patient 2 did not have enough samples to build its own basin of attraction, but its samples were distributed in the basin of attraction built for the non-treated patient 1. Both were characterized as HER2+ breast cancer molecular subtype. Moreover, the trastuzumab-treated samples composed a new basin of attraction with higher minimum energy than the non-treated one. This result agrees with the trastuzumab adjuvant role in cancer therapy.

The protocol developed for the identification of potential therapeutic targets matches the concept of personalized medicine. Specific target combinations were derived from the gene expression profile of each patient, with the potential to mitigate side effects and enhance the treatment outcome.

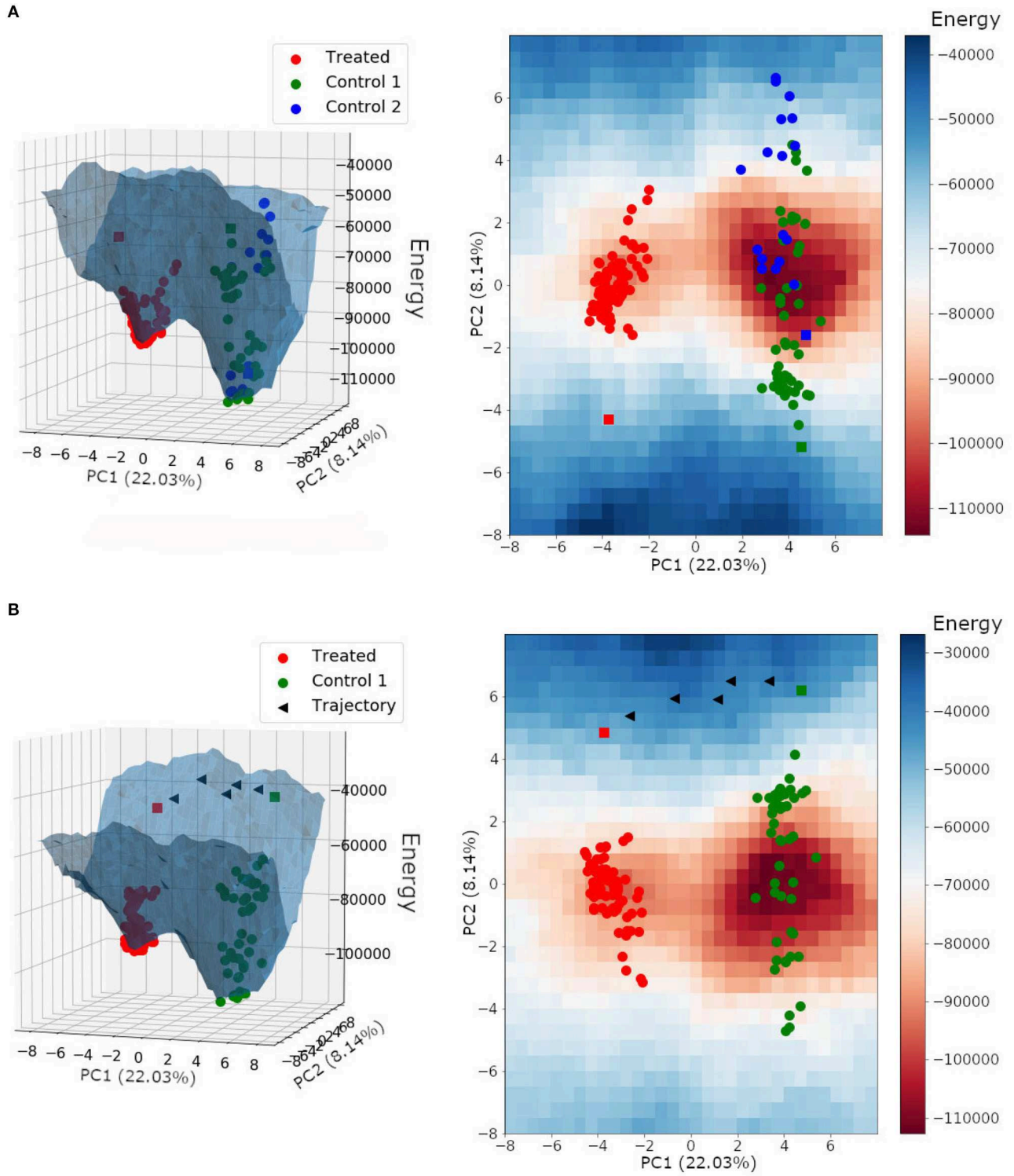

FIGURE 8 | (A) Energy landscape for HER2+ breast cancer samples plotted on a three-dimensional grid and a two-dimensional grid. **(B)** Energy landscape plotted on a three-dimensional grid and a two-dimensional grid, with the trajectory between the non-treated bulk RNA-Seq sample from patient 1 (control 1) and the treated bulk RNA-Seq sample (treated). Each triangle represents the perturbation of a subset with ∼30 DEGs. Spheres represent single-cell samples and squares represent their respective bulk RNA-Seq sample.

Among the parameters tested to indicate gene priority, the node degree has been indicated as a key factor (Carels et al., 2015; Tilli et al., 2016; Conforte et al., 2019). Other topological measures of gene regulatory networks (GRNs) have also been widely used in the identification of new therapeutic targets (Peng and Schork, 2014; Azevedo and Moreira-Filho, 2015). However, they could not be inferred in this work because the Hopfield network approach does not consider protein-protein interactions.

The best parameters, according to our results, were genes with high number of patients that present the node active only in tumor samples and low-density values. Both had the largest potential for tumor sample destabilization with fewer gene inhibitions. The difference between their effect and the effect of its opposite priority order was small. This small effect difference was also observed for other parameters. This may occur because the Hopfield network can be viewed as a highly connected network, which hampers the characterization of each node's impact on the network. Yet, the observations in both cases are biologically coherent.

The Hopfield network is highly connected, but each interaction has its own weight. Consequently, the Hopfield network is heterogeneous, and the weighted interactions allows the differentiation between important and non-important connections. The biological coherence is related to higher effects expected after inhibition of nodes that are active in most tumor samples and inactive in their respective control samples. This indicates that those genes have an essential role in tumor development. For instance, PLP1 was active in most tumor samples and inhibited in the respective control samples. Its antibody is an effective inhibitor of cell growth in breast cancer.

The identification of therapeutic targets was further validated through simulation of the trastuzumab treatment effect in the non-treated bulk RNA-Seq data from patient 1. We determined the trajectory from non-treated to treated basins of attraction for patient 1 and identified key genes involved in the trastuzumab treatment response.

The trastuzumab adjuvant treatment is indicated for HER2+ breast cancer patients, but there are cases of trastuzumab treatment resistance (Han et al., 2019). In this context, the energy landscape obtained for HER2+ samples could determine the state space of gene expression profiles that could be indicated for effective trastuzumab treatment. Also, the trajectory between the treated and the non-treated basins of attraction may indicate new potential therapeutic targets. These could be used in combination with trastuzumab, such as the ITGB6 specific antibody 264RAD, to increase the state space of gene expression profiles with available treatment. This approach could be explored in the context of personalized medicine in future studies.

The Hopfield network succeeded in modeling the basins of attraction for both bulk and single-cell RNA-Seq. This method is entirely based on the gene expression data and considers the differentially expressed genes among our samples, which is essential due to tumor heterogeneity. As an advantage, it does not require a fully-featured network or literature search about protein-protein interactions.

Other modeling methods have also been proposed in order to identify appropriate therapeutic approaches against cancer. For instance, Su et al. (2017) and Yuan et al. (2017b,c) applied the Endogenous Network Theory (ENT) with a coarse-grained modeling, using the non-linear Hill function. They found attractors that matched gene expression profiles of cell phenotypes related to colorectal, prostate, hepatocellular, and gastric cancer (Su et al., 2017; Yuan et al., 2017b,c). In this context, Yuan et al. (2017b) proposed that colorectal cancer could be treated, and reach a normal intestine phenotype, through suppression or promotion of the inflammation program, suppression of retinoic acid signaling, and suppression of anti-inflammation process, according to the cancer cell phenotype. Moreover, the ENT indicates that cancer can be classified as preventable, curable or incurable according to its respective functional landscape.

As proposed by ENT, our results indicated the existence of different functional landscapes for tumor samples. However, the ENT considers the transition from cancer to the normal state as feasible. In this research, we observed that more than 20 gene inhibitions are required to move a tumor sample from the tumor toward the control basin of attraction. This result indicates that recovering a tumor sample back to the control state is not feasible, which may be explained by accumulation of genetic mutations, alteration in genes copy-number, and other processes that may not be regulated by drug administration (Van Bockstal et al., 2020). Rather, we identified the key genes responsible for attractor stability. By extension, one could argue that key genes are essential to tumor biology and that their inhibition would lead to cell death (Tilli et al., 2016).

Biological networks are typically asymmetrical, and several modeling paradigms consider this asymmetry explicitly. For instance, Kwon et al. proposed a stochastic dynamic decomposition method to analyze the dynamics near stable or unstable states. This modeling approach can generate the landscape and the associated energy function, considering the inherent asymmetry of biological networks (Kwon et al., 2005; Yuan et al., 2017a). However, this approach is based on stochastic differential equations, and requires parameters related to each interaction. Those parameters are normally defined by extensive literature search or wet-lab experiments. Furthermore, Li and Wang showed that the potential landscape based on differential equations is susceptible to parameter changes. Nevertheless, stochastic differential equations may offer several advantages, such as a global landscape, reduced sampling space of paths between two states, and relative stability between stable states in the presence of the system's noise (Tang et al., 2017). Besides, Yuan et al. (2017c) presented a non-linear and coupled SDE system that models stable states with relatively large basins of attraction, and showed that this model is insensitive to interaction details at the core network level, by performing random parameter tests. Furthermore, Toulouse et al. (2005) suggested that the attractor robustness to small amounts of noise on SDE models is related to the presence of network motifs.

Asymmetric Hopfield networks do not have a general method to obtain the energy function. Previously published works used

symmetric Hopfield network with a non-parametric training approach (Maetschke and Ragan, 2014; Taherian Fard and Ragan, 2017). These authors discussed that the resulting landscape is a gross approximation of the biological reality. We improved this aspect in our work using a parametric training approach, which receives more biological information and limits the number of possible basins of attraction. Consequently, our approach better characterizes the landscape and approximates it to the biological reality. The energy landscape obtained is data-driven and may represent the biological reality without the noise of estimated parameters.

Our approach combined Hopfield networks with the application of personalized medicine, by considering a large subset of data from real patients effectively involved in oncogenesis. Hopfield network modeling was successful in (i) identifying and characterizing both tumor and normal attractors; (ii) associating tumor sample locations, in the epigenetic landscape, with clinical data; (iii) identifying target combinations whose inhibition would be more efficient in moving tumor samples away from their basin of attraction; (iv) simulating the effects of trastuzumab treatment in non-treated bulk RNA-Seq data; and (v) inferring the trajectory between trastuzumab treated and non-treated basins of attraction.

5. CONCLUSION

We used Hopfield network modeling to analyze cancer and control attractors based on real patient data and associated their locations, in the epigenetic landscape, with clinical data. Our results indicate that the larger the Euclidean distance between the tumor sample and the control attractor, the lower the patient overall survival is. Besides, tumor samples' energies imply a stable phenotype that requires a combination of changes, specific to tumor sample, to move them away from its basin of attraction. We developed and applied a protocol to identify the key genes in tumor phenotype stability. Since these key genes are essential for sustaining tumor biology, we suggest that their combined inhibition would be helpful in patient treatment. This protocol followed the personalized medicine concept in its three main aspects: considering each tumor as unique, mitigating harmful side effects, and enhancing the treatment outcome. We further validated our approach by simulating the trastuzumab effect in non-treated RNA-seq data and identifying the trajectory from the non-treated to the treated basin of attraction. The key genes involved in the state transition were characterized according to their biological processes and participation in trastuzumab-related pathways.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

AC and LA participated in the conception, design, analysis, and interpretation of the work. FC participated in the conception, design, and drafting of the work. NC participated in the design of the work and substantially revised it. FS participated in the conception, design, analysis, and interpretation of the work and substantially revised it. All authors participated in the report writing and approved the final version.

FUNDING

This study was supported by a fellowship from Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) to AC.

ACKNOWLEDGMENTS

We are grateful for support from PrInt Fiocruz-CAPES Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00314/full#supplementary-material

Supplementary Table 1 | Characterization of each DEG identified in the RNA-Seq data set from TCGA concerning their respective biological process, following the Gene Ontology (GO) classification.

Supplementary Table 2 | Clinical data, molecular subtype, entropy, and overall survival for each patient from TCGA data set.

Supplementary Table 3 | Analysis of sample convergence after threshold variation. The geometric mean best fitted our samples and is indicated as the threshold to binarize the gene state in each sample expression profile.

Supplementary Table 4 | Node degree for differentially expressed genes.

Supplementary Table 5 | Samples energy and Euclidean distances between samples and attractors.

Supplementary Table 6 | Values attributed to each gene considering the four prioritization parameters: density in the Hopfield network; number of GOs related to cancer development; number of patients with the gene active only in the tumor sample (biomarker); and node degree.

Supplementary Table 7 | Characterization of each DEG identified in the RNA-Seq data set from NCBI Gene Expression Omnibus with respect to biological processes involved in the trastuzumab treatment response.

REFERENCES

Ao, P., Galas, D., Hood, L., and Zhu, X. (2008). Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. *Med. Hypotheses* 70, 678–684. doi: 10.1016/j.mehy.2007.03.043

Azevedo, H., and Moreira-Filho, C. A. (2015). Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. *Sci. Rep*. 5:16830. doi: 10.1038/srep16830

Baker, A. T., Zlobin, A., and Osipo, C. (2014). Notch-EGFR/HER2 bidirectional crosstalk in breast cancer. *Front. Oncol.* 4:360. doi: 10.3389/fonc.2014.00360

Banerji, C. R. S., Severini, S., Caldas, C., and Teschendorff, A. E. (2015). Intra-tumour signalling entropy determines clinical outcome in breast and lung cancer. *PLoS Comput. Biol.* 11:e1004115. doi: 10.1371/journal.pcbi.1004115

Boyle, S. T., Gieniec, K. A., Gregor, C. E., Faulkner, J. W., McColl, S. R., and Kochetkova, M. (2017). Interplay between CCR7 and Notch1 axes

promotes stemness in MMTV-PyMT mammary cancer cells. *Mol. Cancer* 16:19. doi: 10.1186/s12943-017-0592-0

Breitkreutz, D., Hlatky, L., Rietman, E., and Tuszynski, J. A. (2012). Molecular signaling network complexity is correlated with cancer patient survivability. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9209–9212. doi: 10.1073/pnas.1201416109

Cantini, L., and Caselle, M. (2019). Hope4genes: a Hopfield-like class prediction algorithm for transcriptomic data. *Sci. Rep.* 9:337. doi: 10.1038/s41598-018-36744-y

Carels, N., Tilli, T., and Tuszynski, J. A. (2015). A computational strategy to select optimized protein targets for drug development toward the control of cancer diseases. *PLoS ONE* 10:e0115054. doi: 10.1371/journal.pone.0115054

Chaganty, B. K. R., Lu, Y., Qiu, S., Somanchi, S. S., Lee, D. A., and Fan, Z. (2015). Trastuzumab upregulates expression of HLA-ABC and T cell costimulatory molecules through engagement of natural killer cells and stimulation of IFN secretion. *Oncoimmunology* 5:e1100790. doi: 10.1080/2162402X.2015.1100790

Chen, J.-B., Zhang, M., Zhang, X.-L., Cui, Y., Liu, P.-H., Hu, J., et al. (2018). Glucocorticoid-inducible kinase 2 promotes bladder cancer cell proliferation, migration and invasion by enhancing beta-catenin/c-Myc signaling pathway. *J. Cancer* 9, 4774–4782. doi: 10.7150/jca.25811

Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8:15081. doi: 10.1038/ncomms15081

Conforte, A. J., Tuszynski, J. A., Silva, F. A. B., and Carels, N. (2019). Signaling complexity measured by shannon entropy and its application in personalized medicine. *Front. Genet.* 10:930. doi: 10.3389/fgene.2019.00930

Cornelius, S. P., Kath, W. L., and Motter, A. E. (2013). Realistic control of network dynamics. *Nat. Commun.* 4:1942. doi: 10.1038/ncomms2939

Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* 26, 3279–3290. doi: 10.1038/sj.onc.1210421

Fan, K., Wang, X., Zhang, J., Ramos, R. I., Zhang, H., Li, C., et al. (2017). Hypomethylation of CNTFR-alpha is associated with proliferation and poor prognosis in lower grade gliomas. *Sci. Rep.* 7:7079. doi: 10.1038/s41598-017-07124-9

Fard, A. T., Srihari, S., Mar, J. C., and Ragan, M. A. (2016). Not just a colourful metaphor: modelling the landscape of cellular development using Hopfield networks. *NPJ Syst. Biol. Appl.* 2, 1–9. doi: 10.1038/npjsba.2016.1

Fumi, H. F., and Martins, M. L. (2013). Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PLoS ONE* 8:e69008. doi: 10.1371/journal.pone.0069008

Gonzalez, H., Hagerling, C., and Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev.* 32, 1267–1284. doi: 10.1101/gad.314617.118

Guo, J., and Zheng, J. (2017). HopLand: single-cell pseudotime recovery using continuous Hopfield network-based modeling of Waddington's epigenetic landscape. *Bioinformatics* 33, i102–9. doi: 10.1093/bioinformatics/btx232

Haga, R. B., and Ridley, A. J. (2016). Rho GTPases: Regulation and roles in cancer cell biology. *Small GTPases* 7, 207–221. doi: 10.1080/21541248.2016.1232583

Han, Y., Wang, J., Liu, W., Yuan, P., Li, Q., Zhang, P., et al. (2019). Trastuzumab treatment after progression in HER2-positive metastatic breast cancer following relapse of trastuzumab-based regimens: a meta-analysis. *Cancer Manag. Res.* 11, 4699–4706. doi: 10.2147/CMAR.S198962

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

He, Z., Zhang, J., Yuan, X., Xi, J., Liu, Z., and Zhang, Y. (2019). Stratification of breast cancer by integrating gene expression data and clinical variables. *Molecules* 24:E631. doi: 10.3390/molecules24030631

Herbst, R. S. (2004). Review of epidermal growth factor receptor biology. *Int. J. Radiat. Oncol. Biol. Phys.* 59, 21–26. doi: 10.1016/j.ijrobp.2003.11.041

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* 20, 869–876. doi: 10.1016/j.semcdb.2009.07.003

Hyndman, R. J., and Fan, Y. (1996). Sample quantiles in statistical packages. *Am. Stat.* 50, 361–365. doi: 10.1080/00031305.1996.10473566

Ji, Z., Su, J., Wu, D., Peng, H., Zhao, W., Zhao, B. N., et al. (2016). Predicting the impact of combined therapies on myeloma cell growth using a hybrid multi-scale agent-based model. *Oncotarget* 8, 7647–7665. doi: 10.18632/oncotarget.13831

Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., et al. (2018). Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* 173, 879–893.e13. doi: 10.1016/j.cell.2018.03.041

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583–621. doi: 10.1080/01621459.1952.10483441

Kwon, C., Ao, P., and Thouless, D. J. (2005). Structure of stochastic dynamics near fixed points. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13029–13033. doi: 10.1073/pnas.0506347102

Lee, J.-Y., Joo, H.-S., Choi, H.-J., Jin, S., Kim, H.-Y., Jeong, G.-Y., et al. (2019). Role of MEL-18 amplification in anti-HER2 therapy of breast cancer. *J. Natl. Cancer Inst.* 111, 609–619. doi: 10.1093/jnci/djy151

Leenders, G. B., and Tuszynski, J. A. (2013). Stochastic and deterministic models of cellular p53 regulation. *Front. Oncol.* 3:64. doi: 10.3389/fonc.2013.00064

Li, C., and Wang, J. (2013). Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLoS Comput. Biol.* 9:e1003165. doi: 10.1371/journal.pcbi.1003165

Li, C., and Wang, J. (2015). Quantifying the landscape for development and cancer from a core cancer stem cell circuit. *Cancer Res.* 75, 2607–2618. doi: 10.1158/0008-5472.CAN-15-0079

Li, M., Sun, Q., and Wang, X. (2017). Transcriptional landscape of human cancers. *Oncotarget* 8, 34534–34551. doi: 10.18632/oncotarget.15837

Liberti, M. V., and Locasale, J. W. (2016). The Warburg effect: how does it benefit cancer cells? *Trends Biochem. Sci.* 41, 211–218. doi: 10.1016/j.tibs.2015.12.001

Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal distributions across the sciences: keys and clues on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability-normal or log-normal: that is the question. *Bioscience* 51, 341–352. doi: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.e11. doi: 10.1016/j.cell.2018.02.052

Liu, J., Zhang, G., Lv, Y., Zhang, X., Ying, C., Yang, S., et al. (2017). SGK2 promotes hepatocellular carcinoma progression and mediates GSK-3beta/beta-catenin signaling in HCC cells. *Tumor Biol.* 39:1010428317700408. doi: 10.1177/1010428317700408

Lu, J., Ksendzovsky, A., Yang, C., Mehta, G. U., Yong, R. L., Weil, R. J., et al. (2012). CNTF receptor subunit alpha as a marker for glioma tumor-initiating cells and tumor grade. *J. Neurosurg.* 117, 1022–1031. doi: 10.3171/2012.9.JNS1212

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Maetschke, S. R., and Ragan, M. A. (2014). Characterizing cancer subtypes as attractors of Hopfield networks. *Bioinformatics* 30, 1273–1279. doi: 10.1093/bioinformatics/btt773

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P. D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* 38, D204–10. doi: 10.1093/nar/gkp1019

Moore, K. M., Thomas, G. J., Duffy, S. W., Warwick, J., Gabe, R., Chou, P., et al. (2014). Therapeutic targeting of integrin v6 in breast cancer. *J. Natl. Cancer Inst.* 106:dju169. doi: 10.1093/jnci/dju169

Peng, Q., and Schork, N. (2014). Utility of network integrity methods in therapeutic target identification. *Front. Genet.* 5:12. doi: 10.3389/fgene.2014.00012

Pohlmann, P. R., Mayer, I. A., and Mernaugh, R. (2009). Resistance to trastuzumab in breast cancer. *Clin. Cancer Res.* 15, 7479–7491. doi: 10.1158/1078-0432.CCR-09-0636

Rouhimoghadam, M., Safarian, S., Carroll, J. S., Sheibani, N., and Bidkhori, G. (2018). Tamoxifen-induced apoptosis of MCF-7 cells via GPR30/PI3k/MAPKs interactions: verification by ODE modeling and RNA sequencing. *Front. Physiol.* 9:907. doi: 10.3389/fphys.2018.00907

Shi, Y., Fan, X., Meng, W., Deng, H., Zhang, N., and An, Z. (2014). Engagement of immune effector cells by trastuzumab induces HER2/ERBB2 downregulation

in cancer cells through STAT1 activation. *Breast Cancer Res.* 16:R33. doi: 10.1186/bcr3637

Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., et al. (2012). Cancer treatment and survivorship statistics, 2012. *CA Cancer J. Clin.* 62, 220–241. doi: 10.3322/caac.21149

Simon, M. C. (2006). Coming up for air: HIF-1 and mitochondrial oxygen consumption. *Cell Metab.* 3, 150–151. doi: 10.1016/j.cmet.2006.02.007

Su, H., Wang, G., Yuan, R., Wang, J., Tang, Y., Ao, P., et al. (2017). Decoding early myelopoiesis from dynamics of core endogenous network. *Sci. China Life Sci.* 60, 627–646. doi: 10.1007/s11427-017-9059-y

Szedlak, A., Paternostro, G., and Piermarocchi, C. (2014). Control of asymmetric hopfield networks and application to cancer attractors. *PLoS ONE* 9:e105842. doi: 10.1371/journal.pone.0105842

Taherian Fard, A., and Ragan, M. A. (2017). Modeling the attractor landscape of disease progression: a network-based approach. *Front. Genet.* 8:48. doi: 10.3389/fgene.2017.00048

Tang, Y., Yuan, R., Wang, G., Zhu, X., and Ao, P. (2017). Potential landscape of high dimensional nonlinear stochastic dynamics with large noise. *Sci. Rep.* 7:15762. doi: 10.1038/s41598-017-15889-2

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi: 10.1101/gr.772403

Tian, T., Li, X., and Zhang, J. (2019). mTOR signaling in cancer and mTOR inhibitors in solid tumor targeting therapy. *Int. J. Mol. Sci.* 20:755. doi: 10.3390/ijms20030755

Tilli, T. M., Carels, N., Tuszynski, J. A., and Pasdar, M. (2016). Validation of a network-based strategy for the optimization of combinatorial target selection in breast cancer therapy: siRNA knockdown of network targets in MDA-MB-231 cells as an *in vitro* model for inhibition of tumor development. *Oncotarget* 7, 63189–63203. doi: 10.18632/oncotarget.11055

Toulouse, T., Ao, P., Shmulevich, I., and Kauffman, S. (2005). Noise in a small genetic circuit that undergoes bifurcation. *Complexity* 11, 45–51. doi: 10.1002/cplx.20099

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Van Bockstal, M. R., Agahozo, M. C., van Marion, R., Atmodimedjo, P. N., Sleddens, H. F. B. M., Dinjens, W. N. M., et al. (2020). Somatic mutations and copy number variations in breast cancers with heterogeneous HER2 amplification. *Mol. Oncol.* doi: 10.1002/1878-0261.12650. [Epub ahead of print].

von der Heyde, S., Wagner, S., Czerny, A., Nietert, M., Ludewig, F., Salinas-Riester, G., et al. (2015). mRNA profiling reveals determinants of trastuzumab efficiency in HER2-positive breast cancer. *PLoS ONE* 10:e0117818. doi: 10.1371/journal.pone.0117818

Wang, J., Xu, R., Yuan, H., Zhang, Y., and Cheng, S. (2019). Single-cell RNA sequencing reveals novel gene expression signatures of trastuzumab treatment in HER2+ breast cancer: a pilot study. *Medicine* 98:e15872. doi: 10.1097/MD.0000000000015872

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometr. Bull.* 1, 80–83. doi: 10.2307/3001968

Winterbach, W., Mieghem, P. V., Reinders, M., Wang, H., and Ridder, D., d. (2013). Topology of molecular interaction networks. *BMC Syst. Biol.* 7:90. doi: 10.1186/1752-0509-7-90

Yan, R., Sun, L., and Zhang, H. (2019). Metabolic reprogramming and tumor immunity under hypoxic microenvironment. *Curr. Opin. Physiol.* 7, 53–59. doi: 10.1016/j.cophys.2019.01.001

Yuan, R., Tang, Y., and Ao, P. (2017a). SDE decomposition and A-type stochastic interpretation in nonequilibrium processes. *Front. Phys.* 12:120201. doi: 10.1007/s11467-017-0718-2

Yuan, R., Zhang, S., Yu, J., Huang, Y., Lu, D., Cheng, R., et al. (2017b). Beyond cancer genes: colorectal cancer as robust intrinsic states formed by molecular interactions. *Open Biol.* 7:170169. doi: 10.1098/rsob.170169

Yuan, R., Zhu, X., Wang, G., Li, S., and Ao, P. (2017c). Cancer as robust intrinsic state shaped by evolution: a key issues review. *Rep. Prog. Phys.* 80:042701. doi: 10.1088/1361-6633/aa538e

Zhan, T., Rindtorff, N., and Boutros, M. (2017). Wnt signaling in cancer. *Oncogene* 36, 1461–1473. doi: 10.1038/onc.2016.304

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 17 April 2020
doi: 10.3389/fphys.2020.00319

Drug-Induced Resistance in Micrometastases: Analysis of Spatio-Temporal Cell Lineages

*Judith Pérez-Velázquez[1] and Katarzyna A. Rejniak[2,3]**

[1] *Mathematical Modeling of Biological Systems, Centre for Mathematical Science, Technical University of Munich, Garching, Germany,* [2] *Integrated Mathematical Oncology Department, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, United States,* [3] *Department of Oncologic Sciences, Morsani College of Medicine, University of South Florida, Tampa, Tampa, FL, United States*

Resistance to anti-cancer drugs is a major cause of treatment failure. While several intracellular mechanisms of resistance have been postulated, the role of extrinsic factors in the development of resistance in individual tumor cells is still not fully understood. Here we used a hybrid agent-based model to investigate how sensitive tumor cells develop drug resistance in the heterogeneous tumor microenvironment. We characterized the spatio-temporal evolution of lineages of the resistant cells and examined how resistance at the single-cell level contributes to the overall tumor resistance. We also developed new methods to track tumor cell adaptation, to trace cell viability trajectories and to examine the three-dimensional spatio-temporal lineage trees. Our findings indicate that drug-induced resistance can result from cells adaptation to the changes in drug distribution. Two modes of cell adaptation were identified that coincide with microenvironmental niches—areas sheltered by cell micro-communities (protectorates) or regions with limited drug penetration (refuga or sanctuaries). We also recognized that certain cells gave rise to lineages of resistant cells (precursors of resistance) and pinpointed three temporal periods and spatial locations at which such cells emerged. This supports the hypothesis that tumor micrometastases do not need to harbor cell populations with pre-existing resistance, but that individual tumor cells can adapt and develop resistance induced by the drug during the treatment.

Keywords: cell viability trajectory, cell spatio-temporal evolution, lineage tree of survivors, precursor of resistance, agent-based models

OPEN ACCESS

Edited by:
Russell C. Rockne, City of Hope National Medical Center, United States

Reviewed by:
James Greene, Clarkson University, United States
Hermann Frieboes, University of Louisville, United States

***Correspondence:**
Katarzyna A. Rejniak
kasia.rejniak@moffitt.org

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *13 January 2020*
Accepted: *20 March 2020*
Published: *17 April 2020*

Citation:
Pérez-Velázquez J and Rejniak KA (2020) Drug-Induced Resistance in Micrometastases: Analysis of Spatio-Temporal Cell Lineages. Front. Physiol. 11:319. doi: 10.3389/fphys.2020.00319

INTRODUCTION

Drug resistance is one of the main impediments in effective anti-cancer therapy. While tumors may first respond well to chemotherapeutic agents, they often start growing back during or after the treatment period and become tolerant to the treatment. Several different intrinsic mechanisms of drug resistance have been postulated (Holohan et al., 2013; Housman et al., 2014; Cree and Charlton, 2017), including alteration of drug targets, changes in the expression of efflux pumps, increased ability to repair DNA damage, reduced apoptosis, elevated cell death inhibition, and altered proliferation. Some extrinsic factors contributing to drug resistance have also been postulated. A pivotal role can be played by the tumor microenvironment

(Correia and Bissell, 2012; Sun, 2016) due to reciprocal communication between tumor cells and the surrounding stromal components. This includes interactions with fibroblasts and emergence of cancer associated fibroblasts, cross-talk with immune cells, and sensing cues from the extracellular matrix (ECM), as well as ECM remodeling. Tumor cells can modify their microenvironment by creating specific niches, including pre-cancerous, pre-metastatic or stem cell niches (Barcellos-Hoff et al., 2013; Hambardzumyan and Bergers, 2015; Huch and Rawlins, 2017). Additionally, the changes in tumor vasculature and interstitial fluid pressure may lead to creation of regions that are poorly penetrated by therapeutics, forming drug-limited pharmacologic sanctuaries or refugia (Cory et al., 2013; Puhalla et al., 2015) that influence tumor response to therapeutics. However these extrinsic factors are still not well understood.

Of particular interest is the heterogeneous and dynamically changing tumor microenvironment. As a result, spatially and temporarily variable gradients of drugs can be formed in the stroma, and tumor cells can be, therefore, exposed to different drug levels during the treatment period. It has been shown experimentally by Wu et al. (2013) that aggressive breast tumor cells can respond to drug gradients by migrating toward the regions of high concentration of doxorubicin and low cell population, and are able to adapt to high drug levels. Subsequently, they become tolerant to the drug and repopulate the region despite the high drug concentration. Fu et al. (2015) used mathematical modeling to investigate how the heterogeneity in drug penetration through the microenvironment effects tumor response to treatment. They showed that resistance arises first in cells located in regions with poor drug penetration, named pharmacological sanctuaries, and then populate areas with higher drug levels. Our own research showed that the non-homogeneous drug distribution within the tumor tissue that results in emergence of tissue regions with poor drug penetration but with normal oxygenation levels may lead to the emergence of acquired resistance (Gevertz et al., 2015; Perez-Velazquez et al., 2016). Similar results were previously generated using different mathematical models. Chisholm et al. (2015) investigated transient emergence of a drug tolerant population of cells using models of reversible phenotypic evolution, and concluded that a combination of non-genetic instability, stress-induced adaptation and selection are responsible for the emergence of weakly-proliferative and drug-tolerant tumor cells. Cho and Levy (2017) used a continuous model to show that cancer cells of different resistance levels can coexist in spatially-different areas in tumor tissue. Feizabadi (2017) used mathematical modeling to show that certain chemotherapy strategies are highly unsuccessful, and even damaging to the patient, under the assumption that the drug can induce resistance during the treatment period. Greene et al. (2018, 2019) developed mathematical approach to differentiate between spontaneous and induced resistance to drugs and proposed *in vitro* experiments that can determine whether treatment can induce resistance. The authors also designed optimized treatment protocols that can prolong the time before resistance develops.

Several experimental studies considered scenarios in which resistance is acquired by the tumor cells as a result of their exposure to the drug, either through epigenetic alteration, drug-induced genetic changes or non-genetic phenotype switching. Pisco et al. (2013) and Pisco and Huang (2015) used a combination of laboratory experiments and mathematical modeling to show that the emergence of multi-drug resistance in leukemic cells can be induced by the lasting stress response to the drug. In this case, the tumor cells exploited their phenotypic plasticity by modifying efflux capacity in a non-genetic but inheritable way. Goldman et al. (2015) and Goldman (2016) showed that exposure of breast tumor cells to high concentration of taxanes can induce phenotypic transitions toward chemotherapy-tolerant stem-like state that can confer drug resistance. Moreover, the authors demonstrated that this adaptive resistance process can be halted by carefully designed order of administered drug combinations. Other examples of drug-induced resistance pointed to modifications in chromatin configuration in lung cancer cells (Dannenberg and Berns, 2010; Sharma et al., 2010), changes in expression of stress adaptation-related proteins in prostate cancer cells (Ferrari et al., 2017), or switching to mesenchymal phenotype in melanoma cells (Su et al., 2017) as the mechanisms of increased cell tolerance to the drug. In all these studies, the exposure of tumor cells to chemotherapy caused non-genetic changes that allowed the tumor cells to tolerate drug treatment and evade drug-induced death.

In this paper, we used mathematical modeling to examine how individual tumor cells can adapt to alterations in drug distribution within the tumor microenvironment in order to acquire resistance to the drug. By tracking cells individually and reconstructing their behavioral history, we were able to provide insights into the complex spatio-temporal changes that occur in cell microcommunities and to explain how they avoid drug-induced death leading to therapy failure. In particular, we developed a concept of 3D spatio-temporal lineage trees that trace both genealogy and spatial locations of cells that survived the simulated treatment. This is an extension of classical lineage trees used to depict tumor clonal expansion in a form of a flat graph with an initiating cell connected to its children cells, that are connected to their descendants until the terminal nodes are reached (Navin and Hicks, 2010; Davis et al., 2017). The 3D spatio-temporal lineage trees allow us to identify the cells that drive a resistant phenotype in the sense that all their successors are resistant to the drug. The existence of such "special" cells has been reported previously under various names: drivers (Hutchinson, 2016; Nikbakht et al., 2016), superstars (Cheeseman et al., 2014a,b), or starter cells (Perez-Velazquez et al., 2015). We refer to these cells as precursors of drug resistance. The current study focuses on analyzing the behavior of individual resistant cells which is an extension of our previous work at the population level. This approach allowed us to develop novel evaluation methods, such as the 3D lineage trees, and also to identify the third microenvironmental niche prone to the emergence of resistant cells. Overall, this paper contributes to a better understanding of drug-induced resistance.

MATERIALS AND METHODS

We used a hybrid multi-cell lattice-free model (*MultiCell-LF*) that combines the off-lattice individual tumor cells with the continuous description of oxygen and a cytotoxic drug. The cells can physically interact with one another, and respond to levels of oxygen and cytotoxic drug absorbed from cell's vicinity. Low levels of oxygen (hypoxia) result in cell quiescence (Qiu et al., 2017). Exposure to the drug leads to cell damage – while we model this as a generic process, one can consider a more specific processes, such as the DNA damage (genotoxicity; Swift and Golsteyn, 2014) or cell membrane damage (lysis; Collins and Kao, 1989). Moreover, cells can become more tolerant to the drug they are exposed to, as shown in Sharma et al. (2010) and Pisco and Huang (2015). This cell's response to different levels of oxygen and drug is a mechanism of cell adaptation to the microenvironment.

Drug and Oxygen Kinetics

The model is defined on a small patch of the tumor tissue with four irregularly positioned stationary vessels (**Figures 1A,B**). Both drug γ and oxygen ξ are intravenously supplied, diffuse through the tissue, are absorbed by the cells, and the drug is subject to decay. We model a small drug molecule of diffusivity comparable to oxygen diffusion (Schmidt and Wittrup, 2009) and with the same supply rate from the vessels. However, the drug is absorbed by the cells twice faster than oxygen (Schmidt and Wittrup, 2009). Drug γ and oxygen ξ kinetics are given by the following equations:

$$\frac{\partial \gamma(\boldsymbol{x},t)}{\partial t} = \underbrace{D_\gamma \cdot \Delta\gamma(\boldsymbol{x},t)}_{diffusion} - \underbrace{d_\gamma \cdot \gamma(\boldsymbol{x},t)}_{decay} - \underbrace{\rho_\gamma \sum\nolimits_i \chi_R(\boldsymbol{X}_i,\boldsymbol{x})}_{uptake\ by\ the\ cells} + \underbrace{S_\gamma \sum\nolimits_j \chi_{Rv}(V_j,\boldsymbol{x})}_{supply} \quad (1)$$

$$\frac{\partial \xi(\boldsymbol{x},t)}{\partial t} = \underbrace{D_\xi \cdot \Delta\xi(\boldsymbol{x},t)}_{diffusion} - \underbrace{\rho_\xi \sum\nolimits_i \chi_R(\boldsymbol{X}_i,\boldsymbol{x})}_{uptake\ by\ the\ cells} + \underbrace{S_\xi \sum\nolimits_j \chi_{Rv}(V_j,\boldsymbol{x})}_{supply}. \quad (2)$$

where, D_γ and D_ξ are the drug and oxygen diffusion coefficients, ρ_γ and ρ_ξ are the drug and oxygen uptake rates, S_γ and S_ξ are the drug and oxygen supply rates, and d_γ is the drug decay rate. In numerical implementation, we take the smaller of the cellular demand ρ_γ /ρ_ξ and the current drug/oxygen level to assure that both concentrations do not fall below zero. Here, x represents the Cartesian coordinate system, X_i are the coordinates of discrete cells, V_j are the coordinates of discrete vessels, and χ is the indicator function of the local neighborhood of radius R around the cells X_i or of radius Rv around the vessels V_j, respectively:

$$\chi_R(Y,x) = \begin{cases} 1 & if\ \ ||x-Y|| < R \\ 0 & otherwise \end{cases} \quad (3)$$

The initial condition consists of a stable oxygen gradient and no drug. The sink-like boundary conditions are imposed to implicitly represent the lymphatic system.

Individual Cell Dynamics

Each cell $C_i(t)$ is defined by its position $X_i(t)$, a fixed radius R, cell current age $A_i(t)$ and cell maturation age A_i^{mat}. Cell division takes place upon reaching maturation age (30 h with 5% fluctuations between the cells to avoid synchronized cell division (Mehrara et al., 2007; Hafner et al., 2016), provided that the host cell is not overcrowded by other cells (14 cells within 2 cell diameters), and it is not located in the hypoxic areas (Qiu et al., 2017). If the level of oxygen in a cell's neighborhood falls below the hypoxia level (5% of vascular supply), the cell becomes quiescent and will not proliferate (flowchart in **Figure 1C**). Upon division of cell $C_i(t)$, two daughter cells $C_{i1}(t)$ and $C_{i2}(t)$ are created instantaneously. One daughter cell takes the coordinates of the mother cell, whereas the second daughter cell is placed near the mother cell at a random angle θ:

$$C_{i2}(t) = C_i(t) + R\ (\cos\theta,\ \sin\theta). \quad (4)$$

The current ages of both cells are initialized to zero, and their cell maturation ages are inherited from their mother cell with a small noise term. To preserve cell volume, the repulsive forces are introduced between overlapping cells (Gevertz et al., 2015; Perez-Velazquez et al., 2016). Since both daughter cells are placed in a distance equal to one cell radius, the repulsive forces are exerted to push the cells apart until they reach the distance equal to cell diameter. The repulsive forces are defined as overdamped springs:

$$\frac{d\boldsymbol{X}_i}{dt} = \frac{1}{\nu}\sum_{j=1\ldots M} \boldsymbol{f}_{i,j}^{rep},\ where$$

$$\boldsymbol{f}_{i,j}^{rep} = \begin{cases} F^{rep}\left(2R - ||\boldsymbol{X}_i - \boldsymbol{X}_j||\right)\frac{\boldsymbol{X}_i - \boldsymbol{X}_j}{||\boldsymbol{X}_i - \boldsymbol{X}_j||}, & if\ ||\boldsymbol{X}_i - \boldsymbol{X}_j|| < 2R \\ 0, & otherwise, \end{cases} \quad (5)$$

here, ν is the damping coefficient, F^{rep} is the repulsive spring stiffness, and $2R$ is the spring resting length; M denotes the number of cells that overlap with X_i.

Upon division, both daughter cells inherit mothers' damage level and tolerance level, while the drug absorbed by the mother cell is split into half between both daughter cells (Schmidt and Wittrup, 2009; Greene et al., 2019).

Cell Resistance Mechanism

Cell's resistance mechanism is modeled as a competition between the level of drug-induced damage accumulated by the cell and the level of damage that the cell can withstand (tolerance) without committing to death. However, the cell can also adapt by increasing its tolerance level if it is exposed to the drug for a certain time (flowchart in **Figure 1C**; Gevertz et al., 2015; Perez-Velazquez et al., 2016).

Cell damage $C_i^{dam}(t)$ is increased proportionally to the newly absorbed amount of drug (we assume that the drug absorbed in the past has already exerted its damage effect). The rate of internal drug decay is taken to be the same as in the extracellular microenvironment. This drug-induced damage can be counterbalanced by cell natural ability for damage repair at

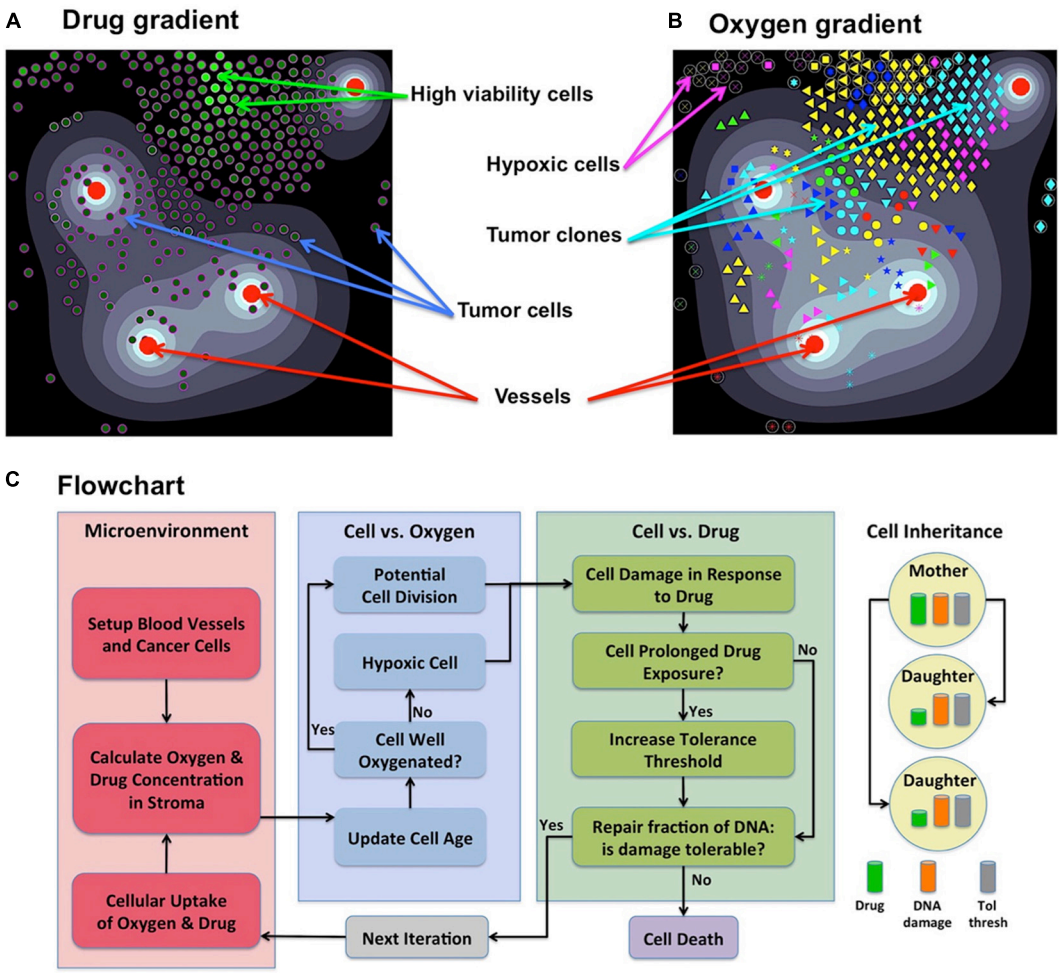

FIGURE 1 | Components of the *MultiCell-LF* model. **(A)** A snapshot showing an irregular drug gradient (high level-white, low level-black) and individual tumor cells color-coded according to their viability (low viability-dark green, high viability-light green). **(B)** The same time snapshot showing oxygen gradient (high level-white, low level-black) and tumor clones marked by a unique symbol assigned to their initial ancestor cell (65 different symbols). Red circles in both panels represent four non-symmetrically located vessels. **(C)** A flowchart showing the relationship between cell behavior and (from left to right) the tumor microenvironment; oxygen levels that regulate cell proliferation or quiescence; drug levels that regulate cell survival, adaptation or death; upon cell division daughter cells inherit from the mother cell: the damage, tolerance level and half of the accumulated drug. Panel **(C)** adopted from Shah et al. (2016).

rate p_γ (three times faster that the damage rate (Gevertz et al., 2015; Perez-Velazquez et al., 2016):

$$\frac{dC_i^{dam}}{dt} = \underbrace{\rho_\gamma \sum_x \chi_R(\boldsymbol{X_i}, \boldsymbol{x}) \underbrace{(1-d_\gamma)}_{drug\ decay} - \underbrace{p_\gamma C_i^{dam}(t)}_{repair}}_{newly\ absorbed\ drug} \quad (6)$$

Cell exposure to high drug concentrations γ_{exp} (at least 1% of the vascular supply) for a long enough time t_{exp} (at least 2% of the cell cycle) results in cell adaptation and in increased cell tolerance to the drug $C_i^{tol}(t)$ (at a slow rate of $\Delta_{tol} = 0.01\%$ of the baseline tolerance value; Gevertz et al., 2015; Perez-Velazquez et al., 2016):

$$\frac{dC_i^{tol}}{dt} = \begin{cases} \Delta_{tol} & if\ C_i^\gamma(t) \geq \gamma_{exp}\ \ for\ a\ time \geq t_{exp} \\ 0 & otherwise \end{cases} \quad (7)$$

where the amount of accumulated drug $C_i^\gamma(t)$ depends on its continuous absorption (at a constant rate ρ_γ) and decay (at a rate d_γ):

$$\frac{dC_i^\gamma}{dt} = \underbrace{\rho_\gamma \sum_x \chi_R(\boldsymbol{X_i}, \boldsymbol{x})}_{uptake} - \underbrace{d_\gamma C_i^\gamma(t)}_{decay} \quad (8)$$

Similarly as for Equation (1), in numerical implementation we take into account that cell demand for the drug may exceed the amount available in cell microenvironment, thus we take the smaller of the cellular demand ρ_γ and the current drug level to assure that drug concentration is non-negative.

Cell death depends on whether cell damage $C_i^{dam}(t)$ exceeds the tolerance level $C_i^{tol}(t)$. The dead cells are removed from the system. Thus, cell resistance to the drug depends on competition

between the level of its damage and the level of damage the cell can withstand without committing to death.

Initially, there is a small micrometastasis consisting of 65 cells with the same baseline tolerance level, no damage, and identical proliferative properties. Each cell responds to the environmental cues, such as the level of sensed oxygen (that regulates cell quiescence or proliferation) and the amount of absorbed drug (that induces cell damage and modulates cell adaptability). The levels of drug and oxygen that the cell is exposed to during its lifetime can vary because the cell can move from one part of the tissue to another, and because drug gradient can change if the overall number of tumor cells changes. The full flowchart of cell behavior is shown in **Figure 1C**. During the simulation, we trace location and viability (the difference between tolerance and actual damage) of each individual cell. When the values of cell damage and tolerance steadily diverge over time, the cell is considered resistant to the drug.

Viability Trajectories of Individual Cells

Cell viability is defined as a difference between the level of cell tolerance to drug-induced damage and actually accumulated damage. The larger the viability value, the more non-responsive to the absorbed drug the cell is. The viability trajectory shows how the viability value is changing in time for a given cell and all its predecessors. The viability trajectory is generated backward starting from the last iteration at which the cell was alive, and going back the cells' lifespan, the lifespan of that cell's mother, the mother's mother, and so on until the one of the initial 65 cells is reached (compare **Figures 2B**, **3A–D**, **4B**, **5**).

Classification of Viability Trajectories and Cell Adaptation Process

To classify how a given cell adapts to the drug exposure, we took into account both its viability trajectory and its recorded drug uptake over the last 20 cell cycles. Visually, there were two significantly distinguishable patterns: cells with constant drug uptake, and cells with rapidly increasing viability trajectories (concave shape). Therefore, the following classification criteria were chosen: (i) the amount of absorbed drug is constant; (ii) the viability curve is monotonically increasing over at least 95% of the considered time interval (numerical second derivative of the viability curve is negative); (iii) the remaining cases. As a result, we identified: (i) linear adaptation pattern (constant drug uptake and almost linear viability trajectories); (ii) a superlinear adaptation pattern (concave viability trajectories with diminished drug uptake); (iii) intermediate pattern where cellular uptake was diminishing but the viability trajectory was fluctuating for the majority of time (compare **Figures 3B–D** and the figure insets).

Cells 3D Spatio-Temporal Routes/3D Lineage Trees

A 3D cell route shows a spatio-temporal evolutionary history of a given tumor cell; that is, it shows all recorded locations of that cell and all cell's predecessors within the tissue patch. The 3D route is created backwards in time by linking positions (in the x-z plane, locations within a tissue at a given time) of a given cell taken at consecutive time points (y-axis) until the cell's birth time is reached, and then repeating this procedure for all cell's predecessors until the beginning of the simulation. The 3D spatio-temporal routes can be traced for multiple cells of the same predecessor forming a 3D spatio-temporal lineage tree (compare **Figures 4A**, **5A,Bii**). These 3D lineage trees are an extension of classical lineage trees used to depict tumor clonal expansion. These figures synthesize information regarding cell locations, cell lineage and time in one single image.

Lineage Trees of Survivors

For each of the initial 65 cells, the whole classical binary lineage tree can be constructed that contains all successors of this cell. A lineage tree of survivors is a subtree of the whole lineage tree and contains only these tree branches that lead to cells that survived the whole treatment (compare **Figures 5A,Bi** and **Supplementary Figures S4–S18**).

Precursors of Resistance

The precursors of resistance are these cells for which all successors survived the treatment at the end of simulation. The precursor of resistance is identified by inspecting the lineage tree generated by that cell; if the lineage tree does not contain any dead cells, its initiating cell is considered to be a

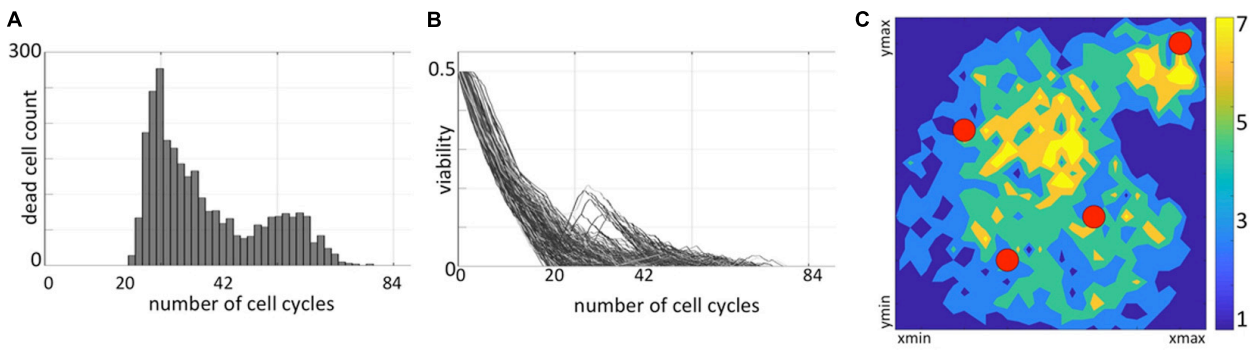

FIGURE 2 | Distribution of non-resistant cells. **(A)** A temporal histogram of dying cells. **(B)** Individual viability trajectories for each dying cell. **(C)** A density map of final locations of cells before they were annihilated by the drug-induced damage. Colors correspond to the number of cells killed at the given location during the whole simulation.

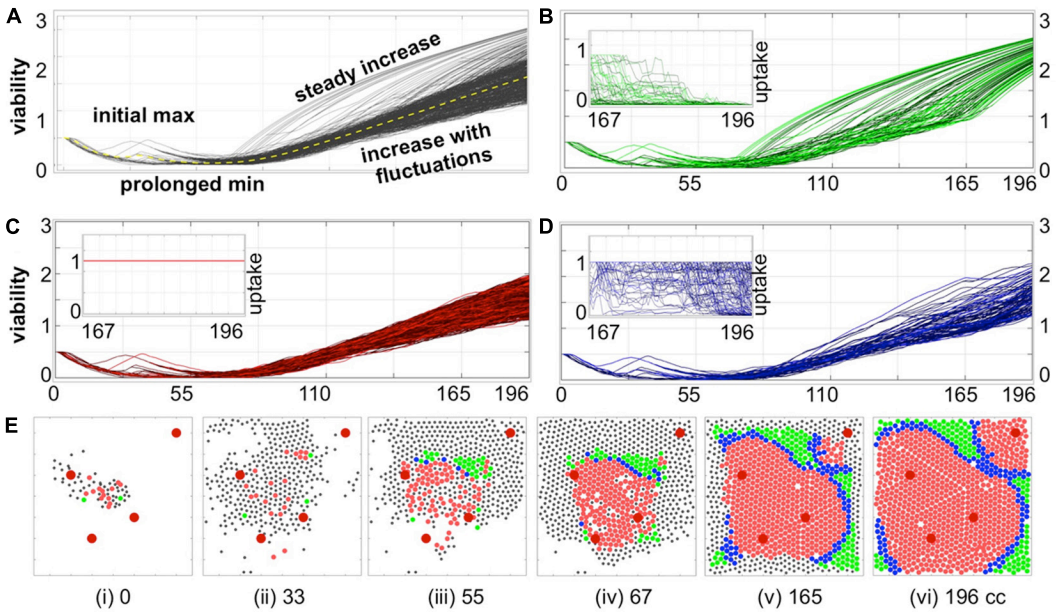

FIGURE 3 | Individual cell adaptation to drug exposure. **(A)** Temporal evolution of viability trajectories of all tumor cells that survived the treatment (gray lines) and the average viability value (yellow dashed line). **(B–D)** Viability trajectories for individual cells from three distinct populations of cells with **(B)** superlinear, **(C)** linear, and **(D)** intermediate viability patterns; the insets show drug uptake by each cell from a given subpopulation during the final period of treatment corresponding to 29 cell cycles of the simulated time: **(B)** significant decrease in drug uptake, **(C)** constant drug uptake, and **(D)** small or short-time decrease in drug uptake. **(E)** Spatial configurations of tumor cells at six different time points; colors correspond to cell subpopulations from **(B–D)**; black dots indicate cells that did not survive to the end of simulation. The larger red circles represent four vessels. All time points are shown in terms of the number of cell cycles (cc).

precursor. We treat the cells that left the computational domain as alive, thus allow them to be successors of the precursor cells (compare **Figures 5A,Bi**).

RESULTS

We previously analyzed a parameter space of this model and identified regimes for which the whole tumor developed resistance (Gevertz et al., 2015; Perez-Velazquez et al., 2016). Here, we summarize these results briefly. A small colony of 65 sensitive tumor cells was exposed to a drug diffusing from four irregularly placed vessels for the period of about 200 cell cycles. Initially, the tumor increased in size until some cells started responding to drug-induced damage and dying; but the remaining cells finally adapted their tolerance. After about 84 cell cycles, the tumor reached a stable population. The average cell viability showed also a steady increase that confirmed the emergence of a resistant tumor. The evolution of tumor resistance on the population level is presented in **Supplementary Figure S1**. This showed that a small homogeneous cell colony exposed to a drug gradient can acquire resistance. The final tumor contained the offsprings of 15 initial cells only; the successors of the remaining 50 initial cells went extinct. The rest of the paper is devoted to analysis of resistance at the individual cell level, whether spatial structure of the tumor and tumor microenvironment play a role in the emergence of resistance, and which cell lineages drive resistance of the whole tumor.

Temporal Distributions of Dying Cells Confirm the Drug-Induced Resistance

The fate of each cell depends on both the accumulated damage and the level of damage that the cell can withstand without committing to death. To determine how the resistance is acquired in individual cells, we need to understand the conditions leading to cell death. Initially, each cell has some baseline tolerance level and no accumulated damage. With time, the absorbed drug induces damage to the cell, while the cell can also adapt to the surrounding extracellular conditions that leads to increase in its tolerance. The cell dies when the level of cell damage exceeds the level of cell tolerance to damage. During the simulated treatment, the initially sensitive cells either develop resistance or respond to the treatment and die. In fact, about 75% of the initial 65 cells did not produce offsprings that were able to survive to the end of the treatment period. Since some cells located near the domain boundaries might have been pushed outside of the observed tissue region by the pressure from their growing neighbors, these cells are assumed to move to the other tissue areas and are removed from our system. Here, we only consider cells that remained inside the tissue domain until they were annihilated by the drug. The summary of spatial and temporal analysis of dying cells is shown in **Figure 2**.

During the initial period of treatment, no cells were dying (no cell counts in histogram in **Figure 2A**) since they must accumulate the drug-induced damage to overcome the baseline tolerance level. However, the viability trajectories for numerous cells decreased during this time (**Figure 2B**) confirming that

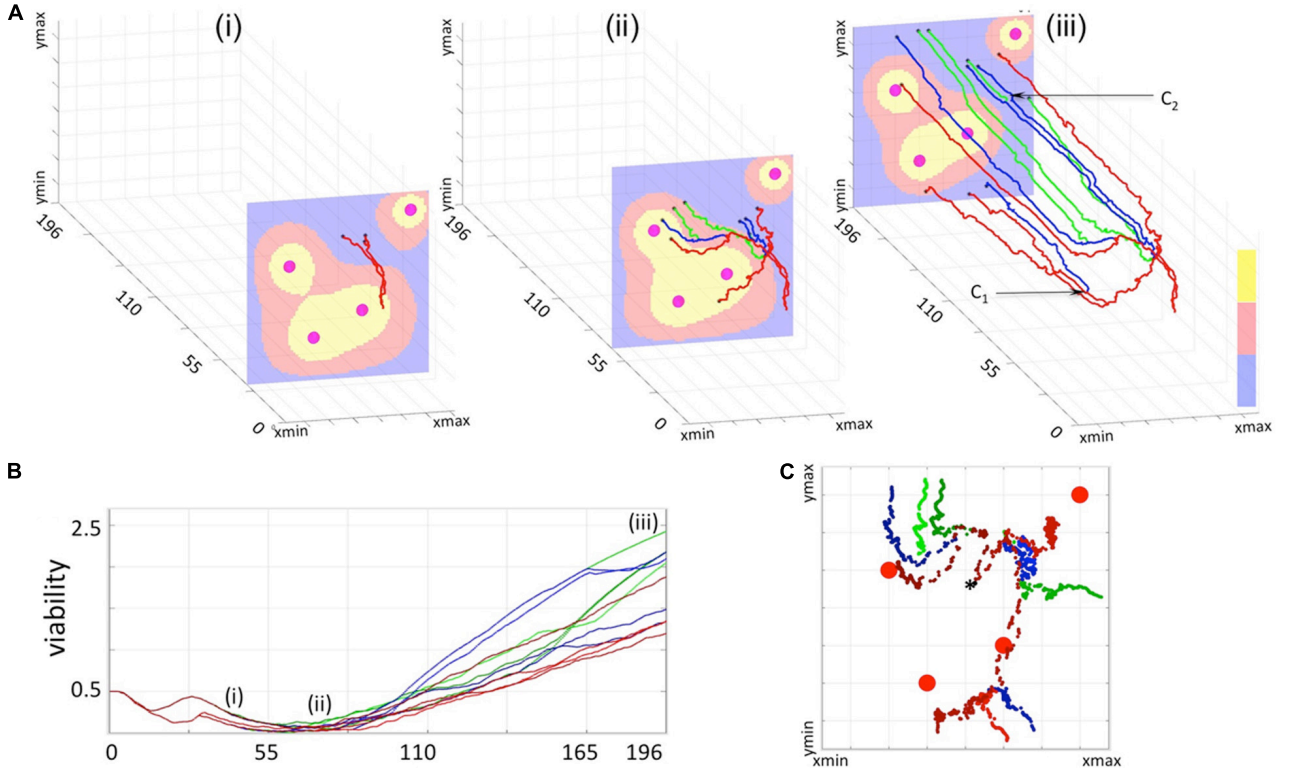

FIGURE 4 | 3D Spatio-temporal routes of the representative resistant cells. Twelve resistant cells with a common predecessor were traced in space and time. All line colors correspond to cell viability categories: concave (green), linear (red), and intermediate (blue). **(A)** Three snapshots of routes traversed by the selected cells within a dynamically changing drug gradient in the background (high to low: yellow-red-blue; four large circles indicate the vessels). **(Ai)** initial overlap in cell routes (time corresponding to 33 cell cycles, 3 cells); **(Aii)** separation of individual cell routes and significant distances traveled by the cells (about 67 cell cycles, 9 cells); **(Aiii)** final map of cell routes showing very small changes in cell locations over an extended time (about 196 cell cycles, 12 cells). **(B)** Viability trajectories for the selected cells (i–iii correspond to the time points from **A**). **(C)** Projections on the xz plane of the individual routes for all 12 cells. A black star indicates the initial position of the common predecessor cell.

these cells were accumulating damage. Each curve in this graph corresponds to one cell and traces in time the viability values of this cell and all its predecessors, back to one of the initial 65 cells. This period corresponds to a steady tumor growth shown in **Supplementary Figures S1i,ii**. The first peak in cell death histogram and a time interval when cell viability trajectories reached zero match the significant reduction in the overall tumor size (**Supplementary Figure S1iii**). The second peak in the death histogram is much smaller since a large number of cells have already developed resistance and only a small subpopulation of cells remained still sensitive to the drug (compare to steady tumor growth in **Supplementary Figures S1iv,v**). After the time corresponding to about 84 simulated cell cycles, no more cells have died. Similarly, all viability trajectories for these dying cells reached the zero value at or before this time (**Figure 2B**). This confirms that all remaining tumor cells in the observable tissue patch have developed a drug-induced resistance. Spatially, the tissue regions that are most prone to cell death are situated either near the single vessel in the top-right corner or in the region near the tissue center between the remaining three vessels (**Figure 2C**). It is worth noting, that in our previous work (Gevertz et al., 2015), we identified the model parameter regimes for which the tumors got extinct, thus the development of drug-induced resistance is not an intrinsic property of our model.

Cell Adaptation to Drug Exposure Can Progress in Three Distinct Ways

To determine how individual cells contributed to the overall tumor resistance, we analyzed the viability trajectories of each cell that survived the treatment (**Figure 3A**). These graphs confirm our previous observations of several phases in the evolution of resistance in the individual tumor cells: from initial identical viability values, to viability decrease due to the damage being accumulated, to a transient increase in viability when the mechanism of tolerance became activated (initial max), to prolonged reduction in viability values due to accumulated damage approaching the individual cell tolerance level leading to cells adaptation to the drug (prolonged min), to a continuous increase in cell viability when the tolerance mechanism gains a lead. Despite the fact that all surviving cells originated from identical predecessor cells and that they shared very similar viability trajectories for the first 55 cell cycles, we identified

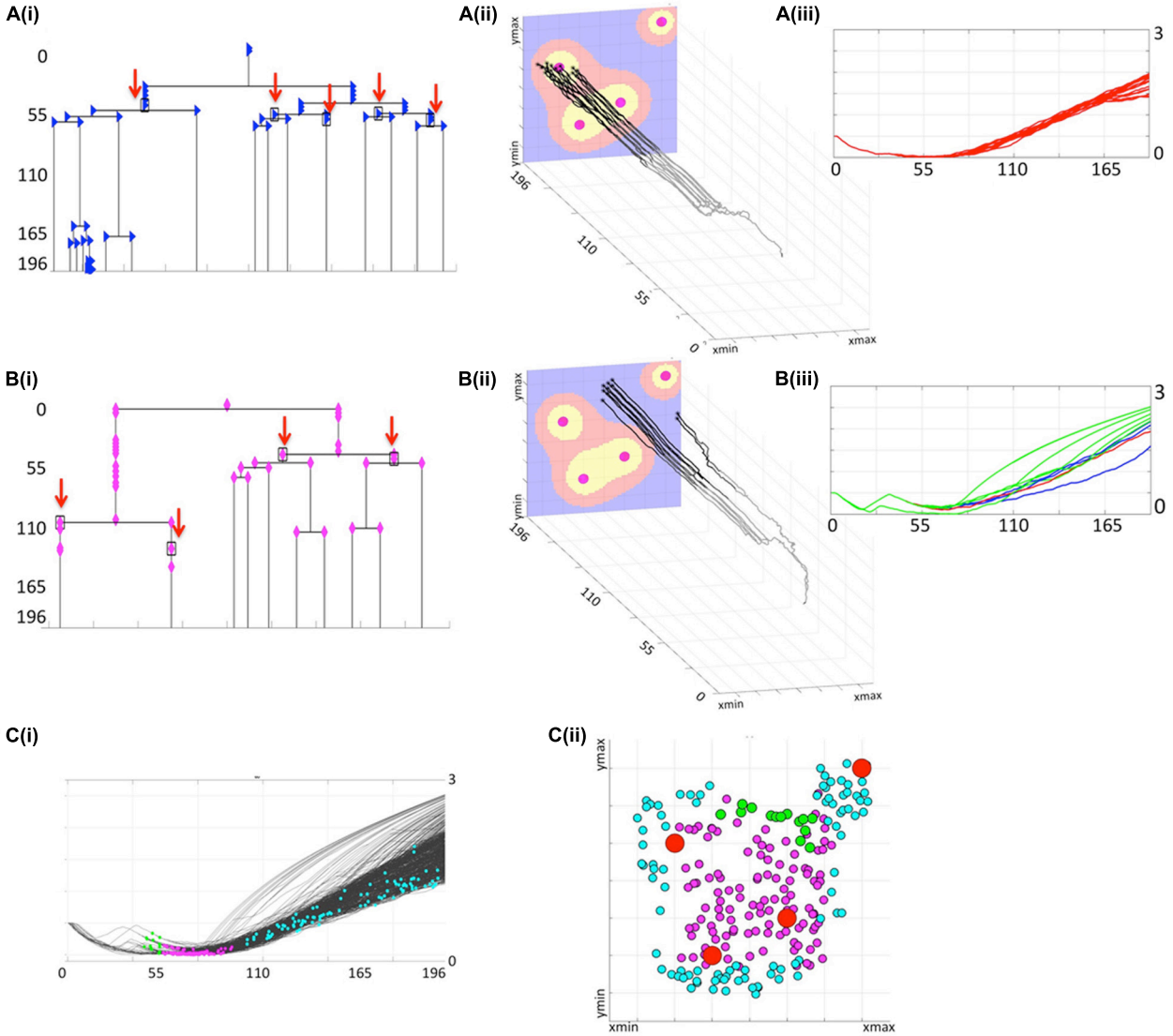

FIGURE 5 | Lineage trees of survivors and cellular precursors of drug resistance. **(Ai)** shows all successors of cell #6 that survived the treatment (lineage tree of survivors). The cells which have no dead successors (the precursor cells of resistance) are shown by red arrows. The corresponding spatio-temporal routes of all surviving cells that arose from cell #6 are shown in **(Aii)** and indicate changes in cells position during the treatment, and final cell locations within the tissue with a drug gradient. The corresponding viability trajectories are shown in **(Aiii)**; line colors correspond to those in **Figure 4**. **(Bi–iii)** shows the lineage tree of survivors, the precursor of resistance and, the 3D spatio-temporal routs, and the viability curves for all daughter cells generated by cell #26. **(Ci)** shows distributions of all precursor cells along the full set of viability trajectories, with spatial localizations of all precursor cells within the tumor tissue shown in **(Cii)**; colors represent time of cell birth: early (green), middle (magenta), and late (cyan); red circles show vessel locations. All time points indicate time corresponding to the number of cell cycles.

three patterns of cell adaptation that resulted in drug-induced resistance (**Figures 3B,D**).

The first cell subpopulation is characterized by rapid increase in viability values that form concave curves of distinct durations (**Figure 3B**). In all these cases, there is also a reduced absorption of the drug for a significant length of time (at least 29 cell cycles, inset in **Figure 3B**). The diminished drug uptake is a result of drug concentration being below the cell's demand. A closer analysis of cell spatial distributions over time shows that this subpopulation occupied tissue regions distant from the vessels and, more importantly, was surrounded by other cells (green circles in **Figures 3Ei–vi**). This was a combined effect of cells' proliferation and their passive relocation due to physical pressure from other growing cells. Since the cells remained in the areas poorly penetrated by the drug for a prolonged time, it resulted in rapidly increasing cell viability that is manifested by the concave shape of the viability curves. The second subpopulation consists of cells with nearly linear increase in viability values (**Figure 3C**) and with constantly high drug absorption (inset in **Figure 3C**). The early predecessors were located in between the three central vessels (red circles in **Figures 3Eii,iii**), and thus were exposed to moderate drug concentrations. This resulted in faster increase

of drug-induced tolerance than drug-induced damage and in the steady increase in cell viability values. The fluctuations in almost linear viability patterns arose from competition between gained tolerance, acquired damage and damage repair. This led to repopulation of the space between the blood vessels (**Figure 3Eiv**). Subsequently, the cells were able to survive in the areas well penetrated by the drug, even in the vicinity of the blood vessels **Figures 3Eiv–vi**). The remaining resistant cells manifest an intermediate behavior with regards to drug absorption, as it decreases over a very short time near the end of the treatment period (**Figure 3D**) but not as pronounced as in the subpopulations with concave viability curves. This subpopulation also acquired a quite distinct spatial pattern on a border between two other subpopulations (**Figures 3Eiii–vi**). These cells are transient in the sense that their characteristics may change during the treatment. For example, a cell with linear viability values may become transient if it gets surrounded by other cells, and becomes protected from drug exposure (cell C_1 in **Figure 4Aiii**). Similarly, a transient cell can give birth to a cell that falls into the category of concave viability if it moves to the poorly penetrated area (cell C_2 in **Figure 4Aiii**).

The 3D Cellular Routes Delineate Spatio-Temporal Dynamics of Cell Adaptation

To more closely examine how cells from all three categories can adapt to the treatment, we selected 12 cells (four from each category) with one common predecessor (**Figure 4**) and traced their locations within the tissue during the whole simulation. The 3D spatio-temporal routes traversed by each cell are shown in **Figure 4A** at three different time points together with the drug profile at that time. Here, the xz-plane represents cell positions within the tissue, and y-axis represents the time. Note, that the drug distribution profile at each time point is different despite the continuous drug influx from the vessels because drug absorption depends on the total number of cells in the tissue, and this cell number varies in time. The presented exemplary cells were chosen intentionally to show a variety of spatial and temporal dynamics that may lead to cell survival, adaptation and acquired resistance. The corresponding 12 viability trajectories are presented in **Figure 4B** to confirm characteristics of each cell. Since these cells have a common predecessor, there is a period of time when both the viability trajectories and the 3D routes overlap and thus the number of observable curves is limited (**Figures 4A,Bi**). However, these curves eventually split up in both figures into eight separate lines (**Figures 4A,Bii**). Furthermore, individual cells were able to move at significant distances from the position of their common predecessor (**Figures 4Aii,C**). This was due to the pressure imposed by other growing cells. From this point on, the viability trajectories steadily increased (**Figure 4Biii**), but cells' routes deviated only insignificantly forming almost horizontal lines (**Figure 4Aiii**). This was due to cell overcrowding by numerous neighbors that resulted in cells' prolonged dormancy, without division. This ultimately contributed toward cell survival and steady increase in cell viability. We intentionally selected a case in which the initial predecessor cell was able to give rise to successors from each of the three categories. However, out of 15 initial cells which successors survived the whole treatment, four generated cells in all three categories, three produced cells in two categories and eight gave rise to cells in a single category.

Lineage Tree Analysis Identifies the Cells That Drive Resistance

Less than a quarter of cells that formed the initial micrometastasis (15 out of 65) produced successors that survived the whole chemotherapeutic treatment. Here we examined the lineages of each subpopulation in order to identify how drug resistance developed for each of them. We inspected the full lineage trees for each of the survived subpopulation and identified subtrees containing only those branches that led from the initial cell to cells that survived the whole treatment. The branches leading to dead cells were omitted. If one of the daughter cells left the domain, but the other survived, its symbol was indicated along the vertical line connecting that cell with its mother cell. These structures represent the lineage trees of survivors. Two representative examples generated by the initial cells with indices #6 and #26 are shown in **Figures 5A,Bi**. The corresponding spatio-temporal routes traversed by these cells are shown in **Figures 5A,Bii**. Additional snapshots of spatio-temporal routes at different time points are shown in **Supplementary Figures S2, S3**. The cell viability trajectories are shown in **Figures 5A,Biii**. These examples illustrate different cases of cellular adaptation observable among all survived subpopulations. The cells for which viability increases linearly are located in well-penetrated areas. These cells were able to survive the drug insult for a prolonged time since they were surrounded by other cells that absorbed the drug creating a protective niche (**Figures 5Aii,iii**). Cells with concave viability trajectories are located in poorly penetrated areas, often equidistant from the vessels, where damage induced by the drug is lower that the ability of the cell to repair damage (**Figures 5Bii,iii**). For some lineage trees of survivors, their spatio-temporal routes may have multiple spatially separated branches due to the proliferation and pressure from neighboring cells. In other cases, the routes do not deviate significantly in space and form horizontal lines. This is due to overcrowding that limits cell proliferation and migration (other 3D routs are discussed in **Supplementary Figures S4–S18**).

For each lineage tree of survivors, we identified the subtrees that do not contain any dead cells; that is, all branches of these subtrees point either to cells that survived the whole treatment or to cells that left the domain (these cells have positive viability values, so they are alive). The roots of such subtrees are considered to be the precursors of drug resistance, since none of their successors underwent drug-induced death. The precursor cells are indicated by black rectangles and red arrows in the trees shown in **Figures 5A,Bi** (for clarity, the branches leading to the cells that left the domain are omitted from the graphs). In total, there were 224 precursor cells emerging from all 15

lineage trees of survivors. They all are pictured in **Figure 5Ci** along the viability trajectories to indicate the time at which they emerge. In **Figure 5Cii**, these cells are cumulatively projected on the tissue space to show the initial locations of the precursor cells. The cell colors correspond to a time period at which they first appeared. The very first precursor cells have arisen in the area poorly penetrated by the drug between the single vessel in the top-right corner, and the three other vessels (cells shown in green in **Figure 5C**). Such areas are known as drug sanctuaries or refugia. The next cohort of precursor cells emerged in the center of the tissue between the three blood vessels (indicated by magenta dots in **Figure 5C**). While, in principle, these areas can be better penetrated by the drug, they actually form protective niches (protectorates) in which the precursor cells may be shielded from the exposure to the drug by the surrounding cells. The final cohort of precursor cells (indicated by cyan dots in **Figure 5C**) was emerging over a longer period of time and mostly in the areas located closer to the tissue boundaries in the hypoxic or nearly-hypoxic niches. Interestingly, none of the precursor cells were located directly at the concave viability trajectories. This indicates that all precursor cells emerged as a result of a direct competition between drug-induced cell damage and acquired tolerance, and that the increase in cell viability was amplified (in fast superlinear fashion) in cells that have already developed resistance.

DISCUSSION

We presented here a study analyzing how resistant cell lineages arise in micrometastases exposed to a systemic chemotherapeutic treatment. This research is an extension of our previous work (Gevertz et al., 2015; Perez-Velazquez et al., 2016) that focused on the emergence of drug-induced resistance on a cell population level. While we followed the previous mathematical model setup and considered a small tumor growing in a heterogeneous microenvironment, the individual-cell perspective and novel evaluation methods allowed us to identify a new microenvironmental niche prone to the emergence of resistant cells. In addition to previously reported *refugia* characterized by low drug penetration due to their distance from the vasculature and the *hypoxic or near-hypoxic niches* in which cells were able to thrive and repair the drug-induced damage, we also located areas in which cells were not exposed to lethal drug concentrations because they were shielded by other cells absorbing the excess of the drug—the *protectorates*. We also recognized that certain cells gave rise to lineages of resistant cells (precursors of resistance) and correlated three temporal periods with three different spatial locations at which such cells emerged. This supports the hypothesis that tumor micrometastases do not need to harbor cell populations with pre-existing resistance, but that individual tumor cells can adapt and develop resistance induced by the drug during the treatment.

The novel analysis and visualization methods developed here, such as the lineage trees of survivors, the method to identify the precursors of resistance and the 3D sptatio-temporal routes and 3D lineage trees can enhance the library of tools used with other hybrid mathematical models (Kim et al., 2013; Karolak and Rejniak, 2019; Chamseddine and Rejniak, 2020) to analyze tumor evolution and clonality.

Moreover, we showed that once the cells have developed resistance, they were able to elevate their viability either in a fast superlinear manner or in a slower, linear fashion, depending whether they moved toward the refugia areas or not; a small population of transient cells that could transfer from the linear to superlinear populations was also observable. This is in line with the theory of mixed models of tumor evolution (Davis et al., 2017), in which different evolution forms can occur in parallel or can shift from one form to another as a result of changes in tumor size or due to microenvironmental selection forces.

Our results can be also placed within a context of tumor ecology (Kenny et al., 2006; Korolev et al., 2014), such as the ecological concepts of microenvironmental niche partitioning and niche construction (Scott and Marusyk, 2017). In the former case, different cell subpopulations are driven into distinct tissue compartments by the microenvironmental selection forces – we observed that certain cell subpopulations were harbored within the refugia areas or within the hypoxic niches. In the latter, the cells are able to modify their own surroundings to create a favorable microenvironment – we observed the formation of cellular protectorates characterized by microenvironmental conditions distinct from the surrounding areas. This spatial heterogeneity in tumor microenvironments is often referred as ecological habitats (Chang et al., 2017; Sala et al., 2017) that can lead to unique fitness landscapes and selection for different cell phenotypes and genotypes, even under the same extrinsic pressure such as anti-cancer therapy. Our simulations showed that individual cell viability was changing over time that encourages revisiting the idea of a static fitness landscape, and supports the view that cell fitness is not a constant value, but a function of the environmental context (Rozhok and DeGregori, 2015; Scott and Marusyk, 2017). While we did not explicitly model any genetic mutations, the observable changes in tumor cell viability could be potentially linked to changes in cell gene expression.

Ultimately, the link between ecological changes within the tumor microenvironment and tumor evolutionary changes will reflect on patients' clinical outcome. While the systemic chemotherapy is often used in the clinical protocol in order to minimize the tumor metastatic spread, it should be taken into account that such therapy may stimulate progression of the nearly-killed cells toward resistance. Therefore, the approaches targeting the resistance-inducing strategies may prove more effective than targeting the tumor cells directly. This is similar in concept to eco-evo drugs from the field of microbial antibiotic resistance (Baquero et al., 2011). Some such preconditioning mechanisms have been tested in cancer cells and already showed promise (McDunn and Cobb, 2005; Pisco et al., 2013; Huang, 2014); however, more research in this area is needed.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

The computational model and analysis methods were conceived by KR and JP-V. All computer simulations were performed by KR and JP-V. KR and JP-V contributed to the manuscript writing.

FUNDING

This work was supported in part by the U01-CA202229 Physical Sciences Oncology Project (PS-OP) grant from the US National Institutes of Health (to KR). JP-V thanks the "Global Challenges for Women in Math Sciences" program of the Mathematics Faculty of the Technical University of Munich for the Entrepreneurial Award to further this collaborative project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.00319/full#supplementary-material

REFERENCES

Baquero, F., Coque, T. M., and de la Cruz, F. (2011). Ecology and evolution as targets: the need for novel eco-evo drugs and strategies to fight antibiotic resistance. *Antimicrob. Agents Chemother.* 55, 3649–3660. doi: 10.1128/AAC.00013-11

Barcellos-Hoff, M. H., Lyden, D., and Wang, T. C. (2013). The evolution of the cancer niche during multistage carcinogenesis. *Nat. Rev. Cancer* 13, 511–518. doi: 10.1038/nrc3536

Chamseddine, I. M., and Rejniak, K. A. (2020). Hybrid modeling frameworks of tumor development and treatment. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 12:e1461. doi: 10.1002/wsbm.1461

Chang, Y. C., Ackerstaff, E., Tschudi, Y., Jimenez, B., Foltz, W., Fisher, C., et al. (2017). Delineation of Tumor Habitats based on Dynamic Contrast Enhanced MRI. *Sci. Rep.* 7:9746. doi: 10.1038/s41598-017-09932-5

Cheeseman, B. L., Newgreen, D. F., and Landman, K. A. (2014a). Spatial and temporal dynamics of cell generations within an invasion wave: a link to cell lineage tracing. *J. Theor. Biol.* 363, 344–356. doi: 10.1016/j.jtbi.2014.08.016

Cheeseman, B. L., Zhang, D., Binder, B. J., Newgreen, D. F., and Landman, K. A. (2014b). Cell lineage tracing in the developing enteric nervous system: superstars revealed by experiment and simulation. *J. R. Soc. Interf.* 11:20130815. doi: 10.1098/rsif.2013.0815

Chisholm, R. H., Lorenzi, T., Lorz, A., Larsen, A. K., de Almeida, L. N., Escargueil, A., et al. (2015). Emergence of drug tolerance in cancer cell populations: an evolutionary outcome of selection, nongenetic instability, and stress-induced adaptation. *Cancer Res.* 75, 930–939. doi: 10.1158/0008-5472.CAN-14-2103

Cho, H., and Levy, D. (2017). Modeling the dynamics of heterogeneity of solid tumors in response to chemotherapy. *Bull. Math. Biol.* 79, 2986–3012. doi: 10.1007/s11538-017-0359-1

Collins, J. L., and Kao, M. S. (1989). The anticancer drug, cisplatin, increases the naturally occurring cell-mediated lysis of tumor cells. *Cancer Immunol. Immunother.* 29, 17–22.

Correia, A. L., and Bissell, M. J. (2012). The tumor microenvironment is a dominant force in multidrug resistance. *Drug Resist. Update* 15, 39–49. doi: 10.1016/j.drup.2012.01.006

Cory, T. J., Schacker, T. W., Stevenson, M., and Fletcher, C. V. (2013). Overcoming pharmacologic sanctuaries. *Curr. Opin. HIV AIDS* 8, 190–195. doi: 10.1097/COH.0b013e32835fc68a

Cree, I. A., and Charlton, P. (2017). Molecular chess? Hallmarks of anti-cancer drug resistance. *BMC Cancer* 17:10. doi: 10.1186/s12885-016-2999-1

Dannenberg, J. H., and Berns, A. (2010). Drugging drug resistance. *Cell* 141, 18–20. doi: 10.1016/j.cell.2010.03.020

Davis, A., Gao, R., and Navin, N. (2017). Tumor evolution: linear, branching, neutral or punctuated? *Biochim. Biophys. Acta* 1867, 151–161. doi: 10.1016/j.bbcan.2017.01.003

Feizabadi, M. S. (2017). Modeling multi-mutation and drug resistance: analysis of some case studies. *Theor. Biol. Med. Model.* 14:6. doi: 10.1186/s12976-017-0052-y

Ferrari, N., Granata, I., Capaia, M., Piccirillo, M., Guarracino, M. R., Vene, R., et al. (2017). Adaptive phenotype drives resistance to androgen deprivation therapy in prostate cancer. *Cell Commun. Signal.* 15:51. doi: 10.1186/s12964-017-0206-x

Fu, F., Nowak, M. A., and Bonhoeffer, S. (2015). Spatial heterogeneity in drug concentrations can facilitate the emergence of resistance to cancer therapy. *PLoS Comput. Biol.* 11:e1004142. doi: 10.1371/journal.pcbi.1004142

Gevertz, J. L., Aminzare, Z., Norton, K. A., Perez-Velazquez, J., Volkening, A., and Rejniak, K. A. (2015). "Emergence of anti-cancer drug resistance exploring the importance of the microenvironmental niche via a spatial model," in *Applications of Dynamial Systems in Biology and Medicine*, Vol. 158, eds A. Radunskaya and T. Jackson (Berlin: Springer), 1–34. doi: 10.1007/978-1-4939-2782-1_1

Goldman, A. (2016). Tailoring combinatorial cancer therapies to target the origins of adaptive resistance. *Mol. Cell Oncol.* 3:e1030534. doi: 10.1080/23723556.2015.1030534

Goldman, A., Majumder, B., Dhawan, A., Ravi, S., Goldman, D., Kohandel, M., et al. (2015). Temporally sequenced anticancer drugs overcome adaptive resistance by targeting a vulnerable chemotherapy-induced phenotypic transition. *Nat. Commun.* 6:6139. doi: 10.1038/ncomms7139

Greene, J. M., Gevertz, J. L., and Sontag, E. D. (2019). Mathematical approach to differentiate spontaneous and induced evolution to drug resistance during cancer treatment. *JCO Clin. Cancer Inform.* 3, 1–20. doi: 10.1200/CCI.18.00087

Greene, J. M., Sanchez-Tapia, C., and Sontag, E. D. (2018). "Control structures of drug resistance in cancer chemotherapy," in *Proceedings of the IEEE Conference on Decision and Control*, Tampa, FL.

Hafner, M., Niepel, M., Chung, M., and Sorger, P. K. (2016). Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods* 13, 521–527. doi: 10.1038/nmeth.3853

Hambardzumyan, D., and Bergers, G. (2015). Glioblastoma: defining tumor niches. *Trends Cancer* 1, 252–265. doi: 10.1016/j.trecan.2015.10.009

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* 13, 714–726. doi: 10.1038/nrc3599

Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., et al. (2014). Drug resistance in cancer: an overview. *Cancers* 6, 1769–1792. doi: 10.3390/cancers6031769

Huang, S. (2014). The war on cancer: lessons from the war on terror. *Front. Oncol.* 4:293. doi: 10.3389/fonc.2014.00293

Huch, M., and Rawlins, E. L. (2017). Cancer: tumours build their niche. *Nature* 545, 292–293. doi: 10.1038/nature22494

Hutchinson, L. (2016). Genetics: defining driver mutations in the genomic landscape of breast cancer. *Nat. Rev. Clin. Oncol.* 13:327. doi: 10.1038/nrclinonc.2016.75

Karolak, A., and Rejniak, K. A. (2019). Micropharmacology: an in silico approach for assessing drug efficacy within a tumor tissue. *Bull. Math. Biol.* 81, 3623–3641. doi: 10.1007/s11538-018-0402-x

Kenny, P. A., Nelson, C. M., and Bissell, M. J. (2006). The Ecology of Tumors: by perturbing the microenvironment, wounds and infection may be key to tumor development. *Scientist* 20:30.

Kim, M., Gillies, R. J., and Rejniak, K. A. (2013). Current advances in mathematical modeling of anti-cancer drug penetration into tumor tissues. *Front. Oncol.* 3:278. doi: 10.3389/fonc.2013.00278

Korolev, K., Xavier, J., and Gore, J. (2014). Turning ecology and evolution against cancer. *Nat. Rev. Cancer* 14, 371–380. doi: 10.1038/nrc3712

McDunn, J. E., and Cobb, J. P. (2005). That which does not kill you makes you stronger: a molecular mechanism for preconditioning. *Sci STKE.* 2005:e34.

Mehrara, E., Forssell-Aronsson, E., Ahlman, H., and Bernhardt, P. (2007). Specific growth rate versus doubling time for quantitative characterization of tumor growth rate. *Cancer Res.* 67, 3970–3975. doi: 10.1158/0008-5472.can-06-3822

Navin, N. E., and Hicks, J. (2010). Tracing the tumor lineage. *Mol. Oncol.* 4, 267–283. doi: 10.1016/j.molonc.2010.04.010

Nikbakht, H., Panditharatna, E., Mikael, L. G., Li, R., Gayden, T., Osmond, M., et al. (2016). Spatial and temporal homogeneity of driver mutations in diffuse intrinsic pontine glioma. *Nat. Commun.* 7:11185. doi: 10.1038/ncomms11185

Perez-Velazquez, J., Gevertz, J. L., Karolak, A., and Rejniak, K. A. (2016). Microenvironmental niches and sanctuaries: a route to acquired resistance. *Adv. Exp. Med. Biol.* 936, 149–164. doi: 10.1007/978-3-319-42023-3_8

Perez-Velazquez, J., Quinones, B., Hense, B. A., and Kuttler, C. (2015). A mathematical model to investigate quorum sensing regulation and its heterogeneity in *pseudomonas* syringae on leaves. *Ecol. Compl.* 21, 128–141. doi: 10.1016/j.ecocom.2014.12.003

Pisco, A. O., Brock, A., Zhou, J., Moor, A., Mojtahedi, M., Jackson, D., et al. (2013). Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nat. Commun.* 4:2467. doi: 10.1038/ncomms3467

Pisco, A. O., and Huang, S. (2015). Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br. J. Cancer* 112, 1725–1732. doi: 10.1038/bjc.2015.146

Puhalla, S., Elmquist, W., Freyer, D., Kleinberg, L., Adkins, C., Lockman, P., et al. (2015). Unsanctifying the sanctuary: challenges and opportunities with brain metastases. *Neuro Oncol.* 17, 639–651. doi: 10.1093/neuonc/nov023

Qiu, G. Z., Jin, M. Z., Dai, J. X., Sun, W., Feng, J. H., and Jin, W. L. (2017). Reprogramming of the tumor in the hypoxic niche: the emerging concept and associated therapeutic strategies. *Trends Pharmacol. Sci.* 38, 669–686. doi: 10.1016/j.tips.2017.05.002

Rozhok, A. I., and DeGregori, J. (2015). Toward an evolutionary model of cancer: considering the mechanisms that govern the fate of somatic mutations. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8914–8921. doi: 10.1073/pnas.1501713112

Sala, E., Mema, E., Himoto, Y., Veeraraghavan, H., Brenton, J. D., Snyder, A., et al. (2017). Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clin. Radiol.* 72, 3–10. doi: 10.1016/j.crad.2016.09.013

Schmidt, M. M., and Wittrup, K. D. (2009). A modeling analysis of the effects of molecular size and binding affinity on tumor targeting. *Mol. Cancer Ther.* 8, 2861–2871. doi: 10.1158/1535-7163.MCT-09-0195

Scott, J., and Marusyk, A. (2017). Somatic clonal evolution: a selection-centric perspective. *Biochim. Biophys. Acta Rev. Cancer* 1867, 139–150. doi: 10.1016/j.bbcan.2017.01.006

Shah, A. B., Rejniak, K. A., and Gevertz, J. L. (2016). Limiting the development of anti-cancer drug resistance in a spatial model of micrometastases. *Math. Biosci. Eng.* 13, 1185–1206. doi: 10.3934/mbe.2016038

Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., et al. (2010). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* 141, 69–80.

Su, Y., Wei, W., Robert, L., Xue, M., Tsoi, J., Garcia-Diaz, A., et al. (2017). Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. *Proc. Natl. Acad. Sci. U.S.A.* 114, 13679–13684. doi: 10.1073/pnas.1712064115

Sun, Y. (2016). Tumor microenvironment and cancer therapy resistance. *Cancer Lett.* 380, 205–215. doi: 10.1016/j.canlet.2015.07.044

Swift, L. H., and Golsteyn, R. M. (2014). Genotoxic anti-cancer agents and their relationship to DNA damage, mitosis, and checkpoint adaptation in proliferating cancer cells. *Int. J. Mol. Sci.* 15, 3403–3431. doi: 10.3390/ijms15033403

Wu, A., Loutherback, K., Lambert, G., Estevez-Salmeron, L., Tlsty, T. D., Austin, R. H., et al. (2013). Cell motility and drug gradients in the emergence of resistance to chemotherapy. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16103–16108. doi: 10.1073/pnas.1314385110

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 June 2020
doi: 10.3389/fgene.2020.00478

Bioinformatics Analysis of Prognostic miRNA Signature and Potential Critical Genes in Colon Cancer

Weigang Chen[1,2†], Chang Gao[1,2†], Yong Liu[1,2], Ying Wen[1,2], Xiaoling Hong[1,2] and Zunnan Huang[1,3,4]*

[1] Key Laboratory of Big Data Mining and Precision Drug Design of Guangdong Medical University, Research Platform Service Management Center, Guangdong Medical University, Dongguan, China, [2] Key Laboratory for Research and Development of Natural Drugs of Guangdong Province, School of Pharmacy, Guangdong Medical University, Dongguan, China, [3] The Second School of Clinical Medicine, Guangdong Medical University, Dongguan, China, [4] Institute of Marine Biomedical Research, Guangdong Medical University, Zhanjiang, China

OPEN ACCESS

Edited by:
George Bebis, University of Nevada, Reno, United States

Reviewed by:
Bhanwar Lal Puniya, University of Nebraska–Lincoln, United States
Leda Torres, National Institute of Pediatrics (Mexico), Mexico

***Correspondence:**
Zunnan Huang
zn_huang@yahoo.com;
zn_huang@gdmu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *06 January 2020*
Accepted: *17 April 2020*
Published: *09 June 2020*

Citation:
Chen W, Gao C, Liu Y, Wen Y, Hong X and Huang Z (2020) Bioinformatics Analysis of Prognostic miRNA Signature and Potential Critical Genes in Colon Cancer. Front. Genet. 11:478. doi: 10.3389/fgene.2020.00478

This study aims to lay a foundation for studying the regulation of microRNAs (miRNAs) in colon cancer by applying bioinformatics methods to identify miRNAs and their potential critical target genes associated with colon cancer and prognosis. Data of differentially expressed miRNAs (DEMs) and genes (DEGs) downloaded from two independent databases (TCGA and GEO) and analyzed by R software resulted in 472 DEMs and 565 DEGs in colon cancers, respectively. Next, we developed an 8-miRNA (hsa-mir-6854, hsa-mir-4437, hsa-mir-216a, hsa-mir-3677, hsa-mir-887, hsa-mir-4999, hsa-mir-34b, and hsa-mir-3189) prognostic signature for patients with colon cancer by Cox proportional hazards regression analysis. To predict the target genes of these miRNAs, we used TargetScan and miRDB. The intersection of DEGs with the target genes predicted for these eight miRNAs retrieved 112 consensus genes. GO and KEGG pathway enrichment analyses showed these 112 genes were mainly involved in protein binding, one-carbon metabolic process, nitrogen metabolism, proteoglycans in cancer, and chemokine signaling pathways. The protein–protein interaction network of the consensus genes, constructed using the STRING database and imported into Cytoscape, identified 14 critical genes in the pathogenesis of colon cancer (*CEP55*, *DTL*, *FANCI*, *HMMR*, *KIF15*, *MCM6*, *MKI67*, *NCAPG2*, *NEK2*, *RACGAP1*, *RRM2*, *TOP2A*, *UBE2C*, and *ZWILCH*). Finally, we verified the critical genes by weighted gene co-expression network analysis (WGCNA) of the GEO data, and further mined the core genes involved in colon cancer. In summary, this study identified an 8-miRNA model that can effectively predict the prognosis of colon cancer patients and 14 critical genes with vital roles in colon cancer carcinogenesis. Our findings contribute new ideas for elucidating the molecular mechanisms of colon cancer carcinogenesis and provide new therapeutic targets and biomarkers for future treatment and prognosis.

Keywords: colon cancer, microRNA, bioinformatics, prognosis, biomarker, TCGA, GEO

INTRODUCTION

Colon cancer is one of the common malignant tumors of the digestive tract and occurs in the colon. With the development of the economy and the improvement of people's living standards, the incidence of colon cancer in recent years has increased, and the age of onset lowered, posing a serious threat to people's life and health (Arnold et al., 2017). Patients with colon cancer have no specific clinical symptoms in the early stage (Cappell, 2008). Most patients are in the middle and late stages when they seek medical treatment, and the treatment and prognosis are poor (Cappell, 2008). Most of the deaths of colon cancer patients are a result of tumor metastasis (Siegel et al., 2017). The 5-year survival rate of patients with metastatic colon cancer is much lower than that of non-metastatic colon cancer patients (Zhang et al., 2015). Therefore, it is necessary to identify new biomarkers and find potential therapeutic targets for early detection and treatment of colon cancer through effective strategies.

MicroRNAs (miRNAs) are short non-coding RNAs of approximately 18–25 nucleotides in length. Since their discovery, there has been a plethora of research indicating the aberrant expression of miRNAs in various types of cancers, including those of the colon, liver, and lung (Wang et al., 2015; Yang et al., 2015; Ding et al., 2018). MiRNAs can act as tumor suppressor genes or oncogenes in tumor tissues. Studies show that down-regulation of miR-708 expression could inhibit the progress of colon cancer cells by targeting the tumor promoter zinc finger E-box binding homeobox 1 (ZEB1), and overexpressed miR-155 could promote the proliferation of cancer cells by targeting the tumor suppressor cbl proto-oncogene (*CBL*) (Yu et al., 2017; Sun et al., 2019). Multiple high-throughput studies have shown high correlations between miRNA expression levels and the treatment and diagnosis of cancer patients (Bolmeson et al., 2011; Toiyama et al., 2014; Tan et al., 2018). In colon cancer, miRNAs are associated with the transmission and inhibition of numerous signaling pathways, and have great potential in diagnosis, prognosis, and personalized targeted therapy (Cekaite et al., 2016). It follows that in-depth studies of miRNAs will contribute to understanding the mechanism of colon cancer development and its biological functions, providing a theoretical basis for its prevention, diagnosis, and treatment.

Bioinformatics uses computational tools to store, search, and analyze biological information. A wide array of computational techniques related to database design and construction, protein structure and function prediction, gene discovery, and expression data clustering, are provided as bioinformatics methods for researching cancer and several other diseases (Luscombe et al., 2001). Access to The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), the Gene Expression Omnibus (GEO) (Barrett et al., 2007), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), the Gene Ontology (GO) database (Ashburner et al., 2000), and other databases are pertinent to cancer research. These resources enable relevant tumor data to be searched, processed, and analyzed by using differential expression analysis, survival analysis, functional enrichment analysis, pathway enrichment analysis, and the other functional tools available. Early biomarkers and potential therapeutic targets of tumors identified by these methods have assisted in exploring the molecular mechanisms of tumor pathogenesis and provide clues for further understanding of related tumors. For example, functional enrichment and survival analysis showed that miR-19b-3p might affect the apoptosis and proliferation of human colon cancer cells through SMAD family member 4 (SMAD4) and serve as a prognostic marker for colon cancer (Jiang et al., 2017). In another study, differentially expressed genes (DEGs) identified in colon cancer by differential expression analysis were further analyzed using function and survival analysis approaches (Yong et al., 2018). The results implicated protein phosphatase 2 catalytic subunit alpha (PPP2CA) in the occurrence and development of colon cancer, and its potential to serve as a therapeutic target in colon cancer (Yong et al., 2018). With the gradual development of molecular biology technology, bioinformatics has become increasingly important in cancer research, performing a major role in elucidating cancer mechanisms and finding novel targets for cancer treatment and patient prognosis.

Colon cancer is a multifactorial disease caused by assorted factors, such as genetic, environmental, and lifestyle influences, but its pathogenesis is not fully clarified (Aran et al., 2016). Exploring and studying the molecular mechanism and critical genes of colon cancer is key in improving the prevention and treatment of colon cancer. In this paper, we performed differential expression analysis to screen out miRNA (DEMs) and DEGs from colon cancer data downloaded from two independent databases (TCGA and GEO). To identify prognostic miRNAs, we constructed a Cox proportional hazards regression model. Then, we identified the overlapping genes between the predicted DEM targets and the DEGs and performed a functional enrichment analysis to understand the potential biological functions of these

Abbreviations: CBL, Cbl proto-oncogene; CCNB1, cyclin B1; CEP55, centrosomal protein 55; CIN, chromosomal instability; COAD, colon adenocarcinoma; CXCR2, C-X-C motif chemokine receptor 2; DAVID, Database for Annotation, Visualization and Integrated Discovery; DEGs, differentially expressed genes; DEMs, differentially expressed miRNAs; DTL, denticleless E3 ubiquitin protein ligase homolog; ENTPD5, ectonucleoside triphosphate diphosphohydrolase 5; FANCI, FA complementation group I; GEO, Gene Expression Omnibus Database; GO, Gene Ontology Database; GUCA2A, guanylate cyclase activator 2A; HMMR, hyaluronan mediated motility receptor; HSPB8, heat shock protein family B (small) member 8; KCNMB1, potassium calcium-activated channel subfamily M regulatory beta subunit 1; KEGG, Kyoto Encyclopedia of Genes and Genomes Database; KIF15, kinesin family member 15; KOBAS, KEGG Orthology Based Annotation System; LMOD1, leiomodin 1; MCM6, minichromosome maintenance complex component 6; MCODE, Molecular Complex Detection; MiRNAs, microRNAs; MKI67, marker of proliferation Ki-67; MS4A12, membrane spanning 4-domains A12; NCAPG2, non-SMC condensin II complex subunit G2; NEK2, NIMA related kinase 2; NSCLC, non-small cell lung cancer; PADI2, peptidyl arginine deiminase 2; PLK1, polo-like kinase 1; PPI, protein–protein interaction; PPP2CA, protein phosphatase 2 catalytic subunit alpha; PRC1, protein regulator of cytokinesis 1; PTTG1, PTTG1 regulator of sister chromatid separation, securing; RACGAP1, Rac GTPase activating protein 1; RAD51, RAD51 recombinase; ROC, receiver operating characteristic; RRM2, ribonucleotide reductase regulatory subunit M2; SCNN1B, Sodium channel epithelial 1 subunit beta; SMAD4, SMAD family member 4; SNRPA1, small nuclear ribonucleoprotein polypeptide A; STRING, Search Tool for the Retrieval of Interacting Genes; TCGA, The Cancer Genome Atlas Database; TNBC, triple-negative breast cancer; TNS1, Tensin 1; TOP2A, DNA topoisomerase II alpha; UBE2C, ubiquitin conjugating enzyme E2 C; UPS, ubiquitin–proteasome system; WGCNA, weighted gene co-expression network analysis; ZEB1, zinc finger E-box binding homeobox 1. ZWILCH, zwilch kinetochore protein.

consensus genes. Finally, we constructed the protein–protein interaction (PPI) network of the consensus genes to illuminate the critical genes. These results might provide new ideas for future research and treatment of colon cancer by exploring prognostic miRNAs and therapeutic targets in colon cancer.

MATERIALS AND METHODS

Tumor Data and Differential Expression Analysis

We downloaded 467 miRNA transcriptomes of 459 colon cancer and 8 normal tissue samples from the TCGA database on June 3, 2019, as well as the GSE24514 microarray data of 34 tumor tissues and 15 normal tissues from the GEO database. Both datasets were analyzed using R software (version 3.4.4) packages edgeR and limma, to identify DEMs and DEGs, respectively. The cutoff criteria were $P_{adj} < 0.05$ and $|\log_2 FC| > 1.0$, where FC denotes fold change (Robinson et al., 2010; Ritchie et al., 2015).

Cox Proportional Hazards Regression Model Based on DEMs

To evaluate the effect of single independent miRNAs on the survival time of colon cancer patients, we performed univariate Cox proportional hazard regression analysis (Ahmed et al., 2007) on DEMs using the survival package of R software and screened miRNAs related to patient survival according to the cutoff criterion of $P < 0.01$. Multivariate Cox proportional hazards regression analysis (Ahmed et al., 2007) with stepwise regression methods and a mathematical model allowed identifying prognostic miRNAs and evaluating the impact of these miRNAs on the survival distribution of patients. From the constructed Cox proportional hazards regression model, we used the following formula to compute the risk scores for each patient: miRNA risk score = $\beta_{miRNA1} \times \exp(\text{miRNA}_1) + \beta_{miRNA2} \times \exp(\text{miRNA}_2) + \ldots + \beta_{miRNAn} \times \exp(\text{miRNA}_n)$, where β is the regression coefficient derived from the multivariate Cox proportional hazards regression model, and exp() is the expression level of prognostic miRNAs (Sui et al., 2017). This study divided the patients into a high-risk group and a low-risk group based on the median value of the risk score. The Kaplan–Meier survival curves of both groups were estimated. Then, we calculated the 5-year survival rates of the high-risk and low-risk groups and plotted the receiver operating characteristic (ROC) (Heagerty and Zheng, 2005) curve to test whether the predictive ability of the model was reliable.

Target Genes Prediction for Prognostic miRNAs

To predict the target genes of the prognostic miRNAs, we used the online analysis tools TargetScan (Agarwal et al., 2015) and miRDB (Wong and Wang, 2015) on June 14, 2019. To further improve the reliability of these results, we identified the overlapping target genes by using the VennDiagram package of R software. Then, these overlapping target genes were crossed with DEGs by using the VennDiagram package of R software to obtain the consensus genes.

Functional Enrichment Analysis of Consensus Genes

For GO and KEGG pathway enrichment analyses, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang da et al., 2009) and the KEGG Orthology-Based Annotation System (KOBAS) (Xie et al., 2011), respectively. $P < 0.05$ was set as the cutoff criterion.

Construction and Analysis of PPI Networks With Consensus Genes

The Search Tool for the Retrieval of Interacting Genes (STRING) can aid in understanding the PPI by integrating a large number of known and predicted correlation data between proteins (Szklarczyk et al., 2017). To study the interactions between the consensus genes and to obtain potential critical genes, we constructed their PPI network using the STRING database on July 8, 2019. Genes with significant interactions were screened out based on a confidence score ≥0.4 (Sun et al., 2017), and the filtered results were imported into Cytoscape software (version 3.7.0) for network visualization (Shannon et al., 2003). We used the CentiScaPe plugin (Scardoni et al., 2014) for topology analysis of the entire network to calculate the central parameters, such as the degree value of each node in the PPI network (Williams and Del Genio, 2014). In consideration of the degree value of each node differing significantly, we calculated the average value of the degree of all nodes. Simultaneously, to obtain more meaningful target genes, we selected nodes with scores larger than twice the average as candidate hub nodes. Then, we used the Molecular Complex Detection (MCODE) plugin (Bader and Hogue, 2003) in Cytoscape to screen out the important functional modules in the PPI network of the consensus genes. The MCODE plugin parameters were degree cutoff ≥10, node score cutoff ≥0.2, *k*-core ≥2, and max depth = 100 (Zhao et al., 2018).

Weighted Gene Co-expression Network Analysis

Weighted gene co-expression network analysis (WGCNA) allows analyzing the gene expression patterns of multiple samples for mining the core genes in the pathogenesis of patients with colon cancer (Langfelder and Horvath, 2008). This study analyzed 13,640 genes from the transcriptome data (GSE24514) using the WGCNA algorithm, and 49 samples were clustered through the systematic cluster tree to determine any outliers. Then, we set an appropriate soft threshold of 15 to make the co-expression network meet the scale-free distribution, and genes with similar expression patterns were merged into the same module using a dynamic tree-cutting algorithm (module size = 30) (Ning and Sun, 2020). Subsequently, three different-colored modules containing the most DEGs were further selected to mine the core genes. The edges with topological overlap measures greater than 0.30 were selected and input into Cytoscape for network visualization (Deng et al., 2018). Using the CentiScaPe plugin in Cytoscape, we

calculated the degree value of each gene. Genes with degrees more than twice the average value were considered the core genes of the network.

Running Scripts

All running scripts used above can be found in **Supplementary Material**.

RESULTS

Differential Expression Analysis of Colon Cancer

From the analysis of the TCGA data, we identified 472 DEMs with statistical significance, composed of 201 up-regulated miRNAs and 271 down-regulated miRNAs (**Figure 1A**). In addition, the analysis of the GSE24514 dataset identified 565 DEGs with statistical significance, which included 266 up-regulated genes and 299 down-regulated genes (**Figure 1B**).

Cox Proportional Hazards Regression Model of DEMs

Univariate and multivariate Cox proportional hazards regression analyses identified 12 miRNAs associated with survival in colon cancer patients ($P < 0.01$; **Table 1**) and a further 8 prognostic miRNAs (hsa-mir-6854, hsa-mir-4437, hsa-mir-216a, hsa-mir-3677, hsa-mir-887, hsa-mir-4999, hsa-mir-34b, and hsa-mir-3189), respectively (**Table 2**). Among prognostic miRNAs, hsa-mir-3677, hsa-mir-216a, hsa-mir-4437, and hsa-mir-6854 were also independent prognostic miRNAs ($P < 0.05$). The risk score was calculated as follows: miRNA risk score = (−0.4034 × hsa-mir-6854) + (1.6106 × hsa-mir-4437) + (0.2508 × hsa-mir-216a) + (−0.2327 × hsa-mir-3677) + (0.2306 × hsa-mir-887) + (0.2045 × hsa-mir-4999) + (0.161 × hsa-mir-34b) + (−0.2008 × hsa-mir-3189). **Figure 2A** presents the detailed information of the risk score. Kaplan–Meier survival analysis showed that the 5-year survival rate was 50.5% in the high-risk group and 76.3% in the low-risk group (**Figure 2B**). The area under the ROC curve was 0.729, demonstrating that the model could effectively predict the prognosis of colon cancer patients (**Figure 2C**).

Target Genes Prediction for Prognostic miRNAs

To predict the target genes of the eight prognostic miRNAs, we used two independent online analytical tools (TargetScan and miRDB). **Figures 3A–H** shows that the intersections between the predicted results from the two servers provided 460, 553, 855, 214, 618, 552, 992, and 697 overlapping target genes of hsa-mir-6854, hsa-mir-4437, hsa-mir-216a, hsa-mir-3677, hsa-mir-887, hsa-mir-4999, hsa-mir-34b, and hsa-mir-3189, separately. Overlapping target genes of eight prognostic miRNAs and 565 DEGs from differential expression analysis of colon cancer intersected to obtain 9, 11, 14, 17, 18, 11, 30, and 19 consensus genes, respectively, for these miRNAs, with a total of 112 consensus genes (**Table 3**).

Functional Enrichment Analysis of Consensus Genes

Gene Ontology enrichment analysis, performed using the DAVID database, showed 35 GO terms noticeably enriched with these 112 consensus genes included protein binding, one-carbon metabolic

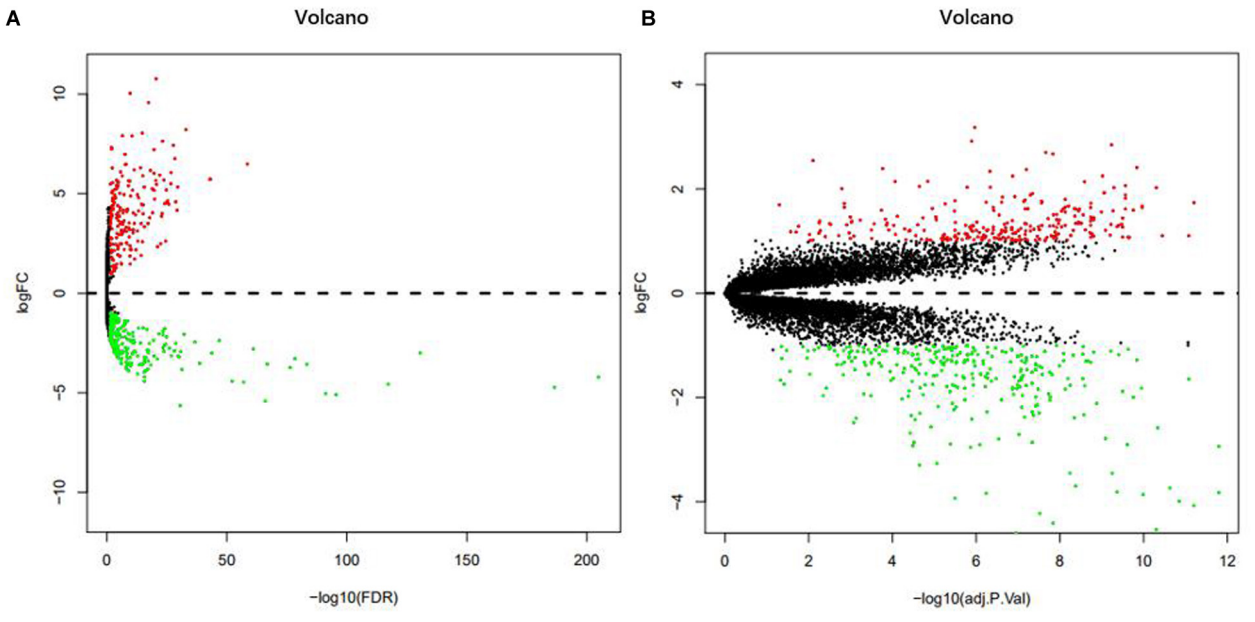

FIGURE 1 | Volcano plot of DEMs in TCGA **(A)**. Volcano plot of DEGs in GSE24514 **(B)**. Red dots represent up-regulation and green dots represent down-regulation.

TABLE 1 | Univariate Cox regression analysis of the 12 miRNAs associated with survival in colon cancer patients.

miRNA	HR	*z*	*P*-value
hsa-mir-887	1.488449	3.418183	0.000630
hsa-mir-3677	0.729468	−3.29453	0.000986
hsa-mir-216a	1.349487	3.274952	0.001057
hsa-mir-149	1.333374	3.184117	0.001452
hsa-mir-4437	4.482079	3.068887	0.002149
hsa-mir-4999	1.390901	3.047926	0.002304
hsa-mir-1271	1.351069	2.990206	0.002788
hsa-mir-3189	0.685402	−2.91866	0.003515
hsa-mir-187	1.201949	2.841883	0.004485
hsa-mir-6854	0.726455	−2.81219	0.004921
hsa-mir-34b	1.297501	2.781959	0.005403
hsa-mir-130a	1.380213	2.744909	0.006053

HR, hazard ratio.

TABLE 2 | Multivariate Cox regression analysis of the 8-miRNA signature associated with survival in colon cancer patients.

miRNA	Coefficient	HR	SE	*P*-value
hsa-mir-887	0.2306	1.2594	0.1194	0.05338
hsa-mir-3677	−0.2327	0.7924	0.1047	0.02619
hsa-mir-216a	0.2508	1.2851	0.0938	0.00750
hsa-mir-4437	1.6106	5.0059	0.4972	0.00120
hsa-mir-4999	0.2045	1.2269	0.1149	0.07519
hsa-mir-3189	−0.2008	0.8181	0.1406	0.15327
hsa-mir-6854	−0.4034	0.6681	0.1183	0.00065
hsa-mir-34b	0.1610	1.1747	0.1044	0.12306

HR, hazard ratio; SE, standard error of coefficient.

process, bicarbonate transport, cytoplasm, and membrane, among others (**Figure 4A**). The GO term "protein binding function" had the smallest *P*-value (P = 5.52e-04) and was enriched with the largest number of consensus genes, with a total of 72, indicating the strongest correlation between them. The KEGG pathway enrichment analysis of these consensus genes, performed using the KOBAS database, revealed 56 pathways were noticeably enriched, including nitrogen metabolism, the thyroid hormone signaling pathway, proteoglycans in cancer, chemokine signaling pathways, and focal adhesion, among others (**Figure 4B**). Of these pathways, nitrogen metabolism had the smallest *P*-value (P = 2.38e-05) and was associated with three consensus genes. Proteoglycans in cancer had the largest number of genes involved, and a *P*-value of 3.07e-05.

Construction and Analysis of PPI Networks for Consensus Genes

To study their PPIs, we entered all the 112 consensus genes into the STRING database to construct the PPI network. Next, for visualization, we imported the genes with confidence scores above 0.4 into Cytoscape. The constructed network was an undirected graph. Each node in the network represented a gene, and the connections between the nodes symbolized the interactions between the proteins encoded by the corresponding genes (Kohler et al., 2008). The network contained 75 nodes and 198 interactions (**Figure 5A**). According to a criterion larger than twice the average (average = 5.28), we identified 16 candidate hub genes: *CCND1*, *CEP55*, *DTL*, *FANCI*, *HMMR*, *KIF15*, *MCM6*, *MKI67*, *MYC*, *NCAPG2*, *NEK2*, *RACGAP1*, *RRM2*, *TOP2A*, *UBE2C*, and *ZWILCH* (**Figure 5B**). The module analysis of the PPI network, performed using the MCODE plugin, revealed two functional modules (**Figures 5C,D**). Except for *CCND1* and *MYC*, the remaining 14 of the 16 candidate hub genes appeared in module 1, indicating that these 14 genes may play important biological functions in the PPI network, and thus, were defined as the critical genes of the network.

Weighted Gene Co-expression Network Analysis

The cluster analysis in WGCNA showed no abnormal value in the 49 GSE24514 samples. According to the independence and average connectivity of networks with different power values (power values ranging from 1 to 20), the soft threshold was determined to be 15 (**Figure 6A**). Ultimately, there were 17 modules of different colors generated. The co-expression degree of genes in the same module was high, and the co-expression degree of genes from different modules was low (**Figure 6B**). Among them, the midnight blue, red, and yellow-green modules contained the most DEGs, which were 200, 103, and 126, respectively. We constructed three weighted gene co-expression networks using edges with topological overlap measures greater than 0.30 in these modules. Ultimately, 19 core genes, which were all DEGs, were identified, according to a degree value criterion of greater than twice the average of the degree. These DEGs were *CCNB1*, *DTL*, *ENTPD5*, *FANCI*, *GUCA2A*, *HSPB8*, *KCNMB1*, *LMOD1*, *MKI67*, *MS4A12*, *NEK2*, *PADI2, PRC1*, *PTTG1*, *RRM2*, *SCNN1B*, *TNS1*, *TOP2A,* and *UBE2C* (**Figures 7A–C**).

DISCUSSION

MicroRNAs play important roles in cell differentiation, biological development, and the occurrence and progression of diseases, attracting increasing attention from researchers. Despite progress in understanding the role of miRNAs in the regulation of tumor growth and evolution, miRNAs are easily affected by a variety of factors during their activity in cancer, and they have the specificity of spatiotemporal expression in different types of tumors or different stages of the same tumors. Therefore, the specific relationships between miRNAs and tumors remain unclear and need to be further clarified.

In this study, we identified DEMs and DEGs of colon cancer from the TCGA and GEO databases, respectively. Then, we derived a prognostic model using Cox proportional hazards regression analysis based on eight miRNAs, namely hsa-miR-6854, hsa-mir-4437, hsa-mir-216a, hsa-mir-3677, hsa-mir-887, hsa-mir-4999, hsa-mir-34b, and hsa-mir-3189. We further obtained a total of 112 consensus genes from the intersection of DEGs with the target genes predicted for these eight miRNAs using TargetScan and miRDB tools. Subsequent GO and KEGG pathway enrichment analyses revealed that these consensus genes

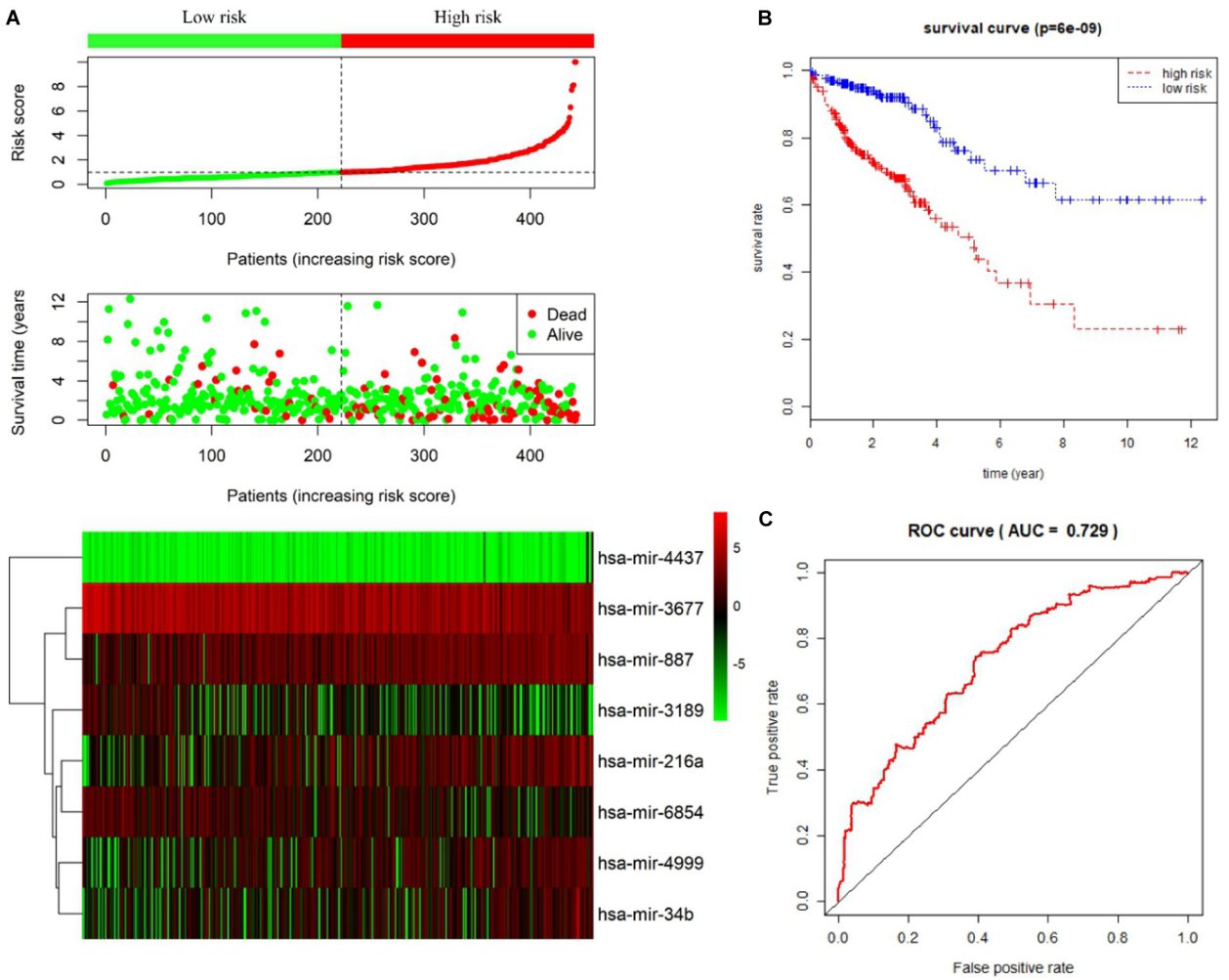

FIGURE 2 | Prognostic risk score model analysis of eight prognostic miRNAs in colon cancer patients. **(A)** From top to bottom are the risk score distribution, patients' survival status distribution, and the heatmap of eight miRNA expression profiles ranked by risk score. **(B)** Kaplan–Meier curves for high-risk and low-risk groups. **(C)** The ROC curves for predicting survival in colon cancer patients by the risk score.

TABLE 3 | One hundred and twelve consensus genes shared by the target genes of 8 prognostic miRNAs and DEGs from differential expression analysis of colon cancer.

miRNA	Consensus genes
hsa-mir-887	*SLC36A1*, *C7*, *HNRNPL*, ***MCM6***, *HSPB6*, *SVIL*, *PPP1R12B*, ***TOP2A***, *SLC17A4*, *MARCKSL1*, *ATP1A2*, *CXCL9*, *METTL7A*, *SLC25A32*, *VSNL1*, *VCL*, *FOXF1*, *CLIP3*
hsa-mir-3677	*SORBS1*, *ARNTL2*, ***KIF15***, *RAB15*, *AKAP12*, *SSR3*, *PJA2*, *CXCL12*, *CA12*
hsa-mir-216a	*CA7*, *HSPD1*, ***NEK2***, *HMGB3*, *NPTX1*, *TXNIP*, *ZCCHC24*, *CA12*, *HOXC6*, *MAN2A1*, *FOSB*
hsa-mir-4437	*CCNB1IP1*, *SPIB*, ***UBE2C***, *SULT1A1*, *CDHR5*, *HHLA2*, *BACE2*, *TRANK1*, *PTPRH*, *CD79A*, *NDC1*, *HSD11B2*, *LGR5*, *MLEC*
hsa-mir-4999	*IFITM1*, *SORBS1*, *SLC17A4*, *LMO3*, ***ZWILCH***, *CCND1*, *RCN1*, *XPOT*, *PDZRN4*, *LRRC19*, *CAV1*
hsa-mir-3189	*FAM57A*, *TMEM97*, ***FANCI***, *CDH3*, *SETBP1*, *ITIH5*, *MAB21L2*, *VIPR1*, *RETSAT*, *GOLT1B*, *MEIS1*, *NPTX1*, *JAM3*, *TXNIP*, *ZCCHC24*, *SLC4A4*, *A1CF*, *NR3C2*, *DUSP1*
hsa-mir-6854	*STMN2*, *SYNM*, *SORBS1*, ***DTL***, ***RACGAP1***, *PLN*, *C1orf115*, ***MKI67***, *GPD1L*, ***HMMR***, *SLC25A32*, *FNBP1*, *PRKACB*, *CAV1*, *TDP2*, *CXCL14*, *DCN*
hsa-mir-34b	*STMN2*, ***NCAPG2***, ***RRM2***, ***CEP55***, *CA1*, *ENC1*, *CXCL1*, *SLC17A4*, *BCAS1*, *PBX1*, *FAM47E-STBD1*, *CCDC59*, *MEST*, *MYC*, *PUS1*, *CCND1*, *SPP1*, *SATB1*, *NDC1*, *AHCYL2*, *KRT20*, *PALLD*, *MLEC*, *SSR3*, *PJA2*, *PAPSS2*, *TGFBI*, *CAV1*, *PDZRN3*, *CLDN8*

Bold represents the critical genes.

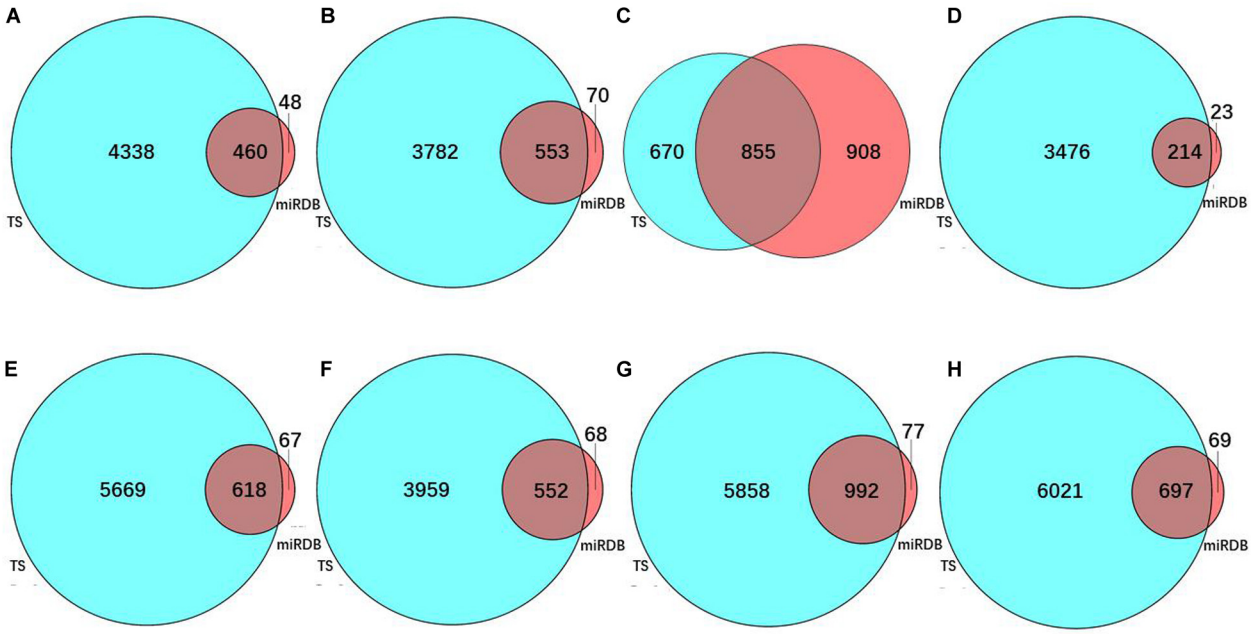

FIGURE 3 | The number of predicted target genes of eight prognostic miRNAs. Target gene number predicted for **(A)** hsa-mir-6854, **(B)** hsa-mir-4437, **(C)** hsa-mir-216a, **(D)** hsa-mir-3677, **(E)** hsa-mir-887, **(F)** hsa-mir-4999, **(G)** hsa-mir-34b, and **(H)** hsa-mir-3189. In these sub-figures, blue represents the predicted results of TargetScan, and red represents the predicted results of miRDB.

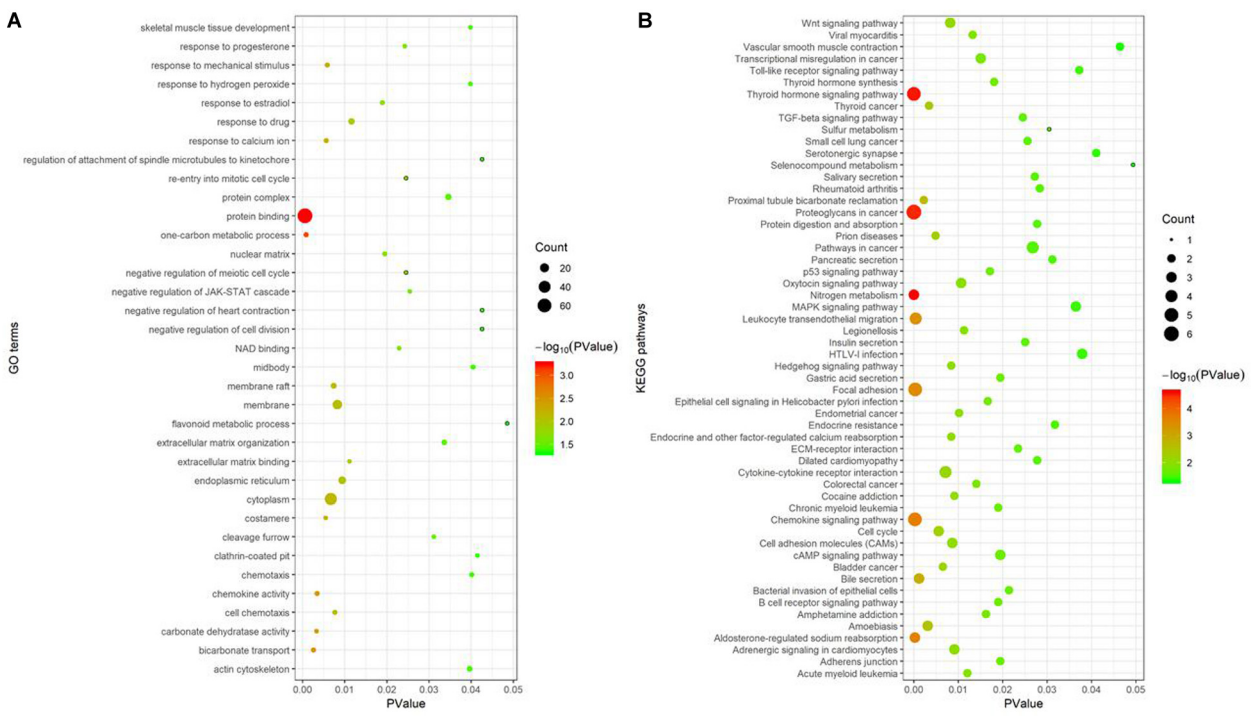

FIGURE 4 | Functional enrichment analysis of 112 consensus genes. **(A)** GO enrichment analysis; **(B)** KEGG pathway enrichment analysis. In these two sub-figures, the *x*-axis represents the *P*-value, and the *y*-axis represents the different GO terms and the KEGG pathways, respectively. The size of the bubbles grows as the number of involved genes increases.

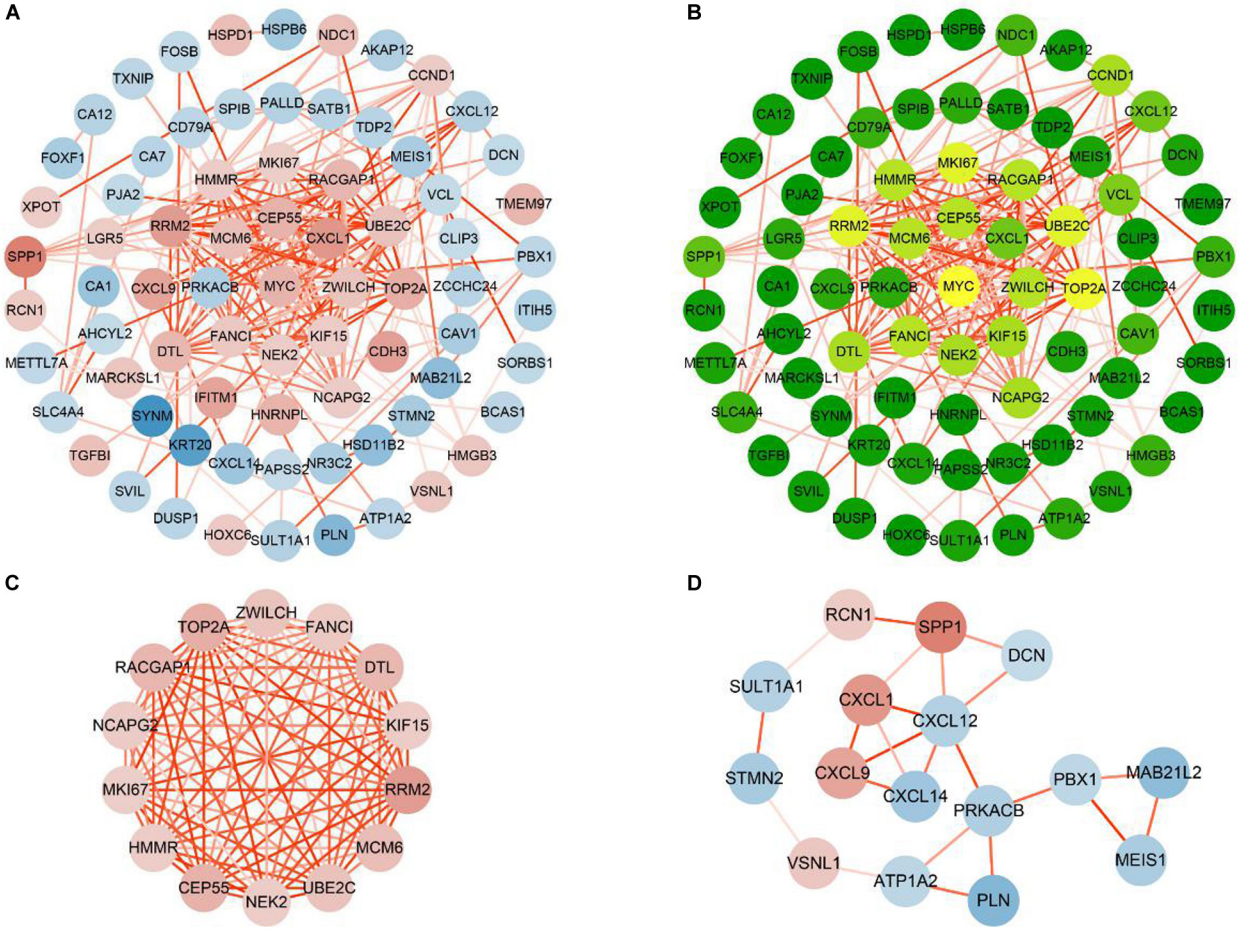

FIGURE 5 | Construction and analysis PPI networks of consensus genes. **(A)** PPI network of 75 consensus genes. Red nodes represent up-regulated genes, and blue nodes represent down-regulated genes. The color of the node deepens as the value of |log2FC| increases. The color of the line connecting the circles deepens as the confidence scores increase. **(B)** Degree values of 75 consensus genes were obtained by CentiScaPe. As the degree values increase, the color of the node changes from green to yellow. **(C)** Module 1 (MCODE score = 13.8466). **(D)** Module 2 (MCODE score = 3.067).

were mainly involved in protein binding, one-carbon metabolic process, nitrogen metabolism, proteoglycans in cancer, and chemokine signaling pathways. Finally, we used the STRING database to construct the PPI network of the 112 consensus genes. With the two Cytoscape plugins CentiScaPe and MCODE, 14 critical genes were recognized (*CEP55*, *DTL*, *FANCI*, *HMMR*, *KIF15*, *MCM6*, *MKI67*, *NCAPG2*, *NEK2*, *RACGAP1*, *RRM2*, *TOP2A*, *UBE2C*, and *ZWILCH*).

Among eight prognostic miRNAs in colon cancer, the expression of hsa-mir-6854, hsa-mir-216a, hsa-mir-3677, hsa-mir-4999, hsa-mir-34b, and hsa-mir-3189 was up-regulated, and that of hsa-mir-4437 and hsa-mir-887 was down-regulated. Among these eight miRNAs, hsa-mir-216a and hsa-mir-34b have been validated in experiments previously, proving they have crucial roles in colon cancer. Wang et al. (2018) showed that the up-regulation of miR-216a-3p inhibited the expression of its target genes *ALOX5* and *COX-2* in colon cancer cells, consequently enhancing the proliferation of colon cancer cells. Hiyoshi et al. (2015) used quantitative RT-PCR to detect overexpression of miR-34b in colon cancer tissues and confirmed that it was associated with poor prognosis in patients.

For the other six prognostic miRNAs, including hsa-mir-6854, hsa-mir-4437, hsa-mir-3677, hsa-mir-887, hsa-mir-4999, and hsa-mir-3189, although their roles have not yet been shown in colon cancer, some experimental studies demonstrated that the expression change of hsa-mir-887 and hsa-mir-3189 played crucial roles in other cancer cells. Jiang et al. (2016) illustrated that miR-887-5p was overexpressed in the serum of patients with endometrial cancer and might be a potential biomarker for endometrial cancer. Jones et al. (2015) demonstrated that overexpression of miR-3189-3p up-regulated p53 and many p53 target genes, which could effectively induce apoptosis and inhibit cell proliferation in colorectal cancer (CRC). In glioblastoma and gastric cancer, overexpressed miR-3189 could markedly inhibit cell proliferation and migration (Jeansonne et al., 2015; Bian et al., 2018). These studies showed miR-3189 as a tumor suppressor. The above results illustrated that the expression of miR-887 and miR-3189 in these cancers was contrary to ours. MiRNA

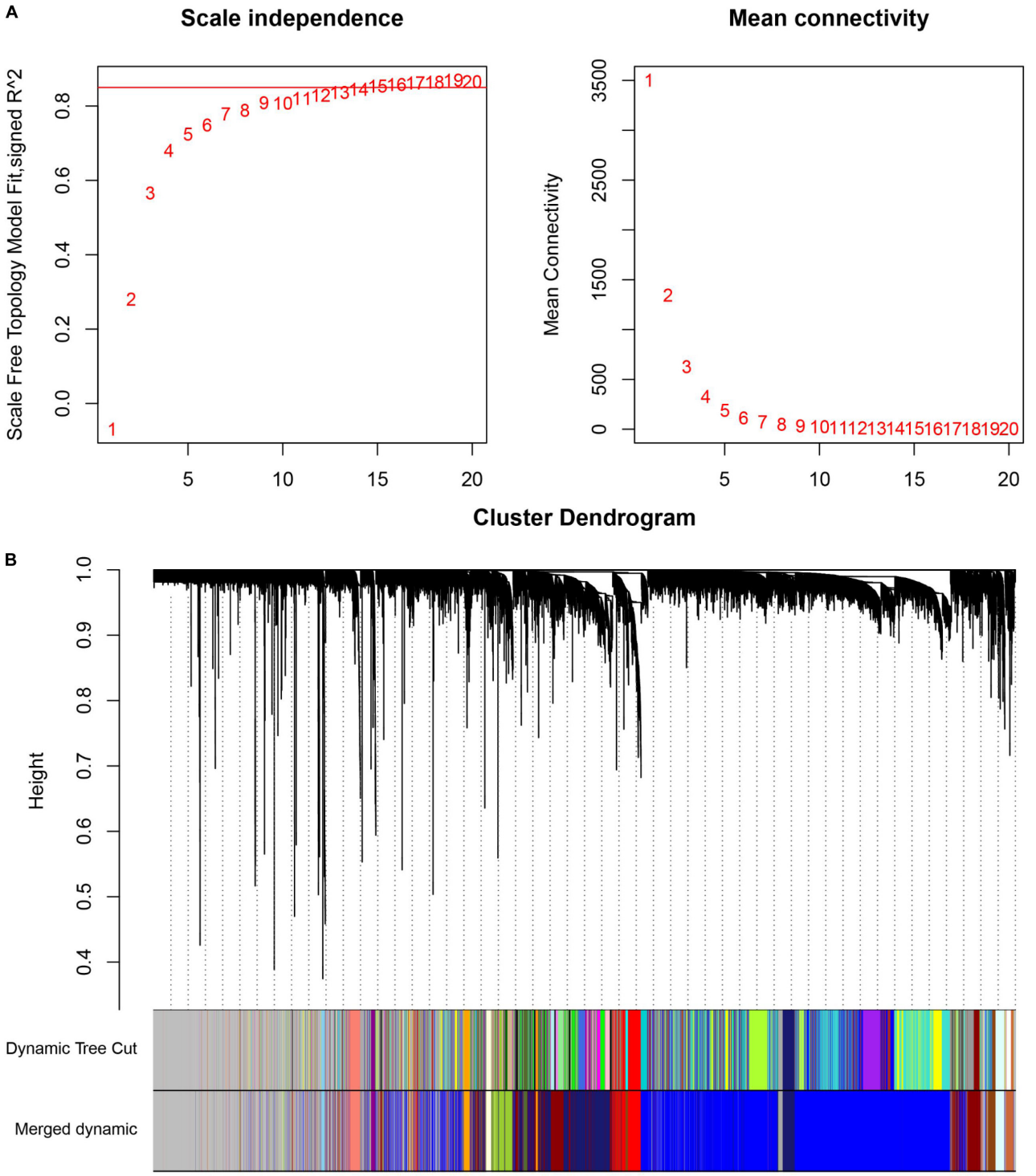

FIGURE 6 | Weighted co-expression gene network analysis. **(A)** Determination of the soft threshold in the WGCNA algorithm. When the soft thresholding power was 15, the gene distribution conformed to the scale-free network. **(B)** The cluster dendrogram of all the genes in GSE24514. Each leaf represents a separate gene, and each branch represents a co-expression gene module.

expression may differ among cancer types, so the expression and specific mechanism of miR-887 and miR-3189 in colon cancer need to be further clarified experimentally.

He et al. (2018) showed that hsa-mir-4437 could directly act on C-X-C motif chemokine receptor 2 (CXCR2), which can increase tumor inflammation and angiogenesis. Saintigny et al. (2013)

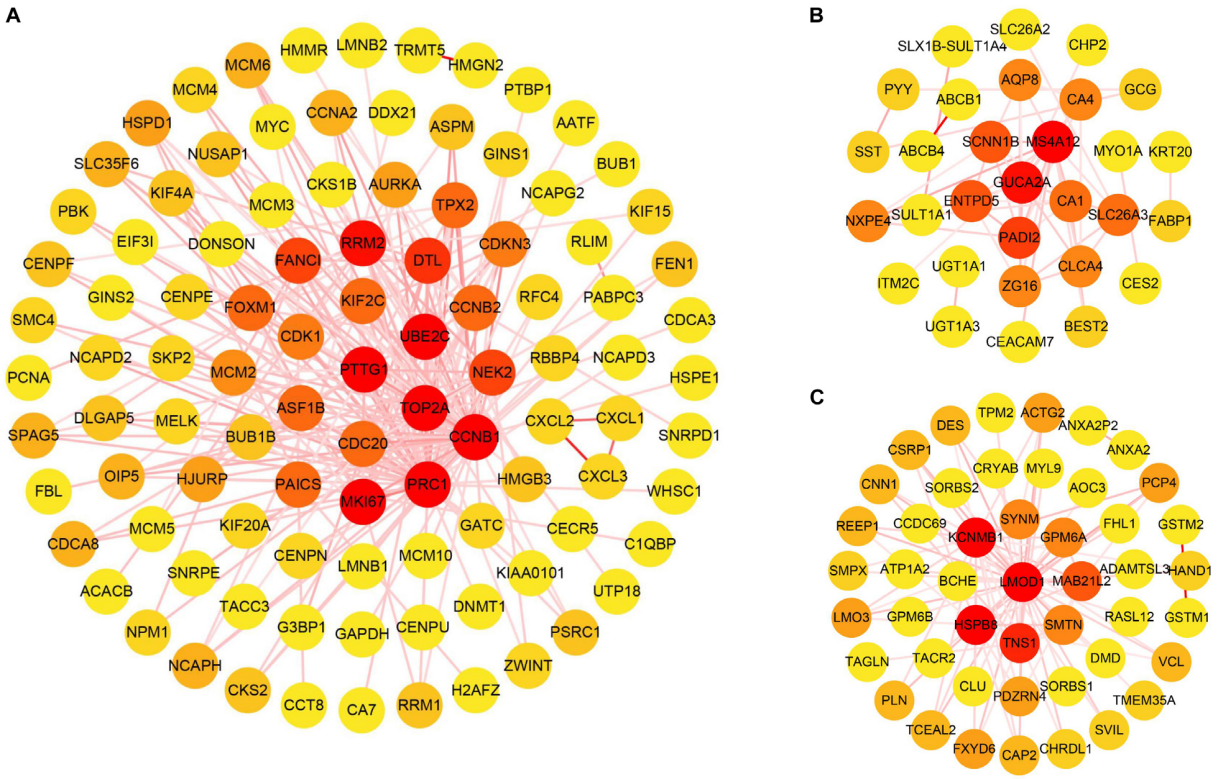

FIGURE 7 | The visualization of co-expression gene modules. **(A)** Midnight blue module. **(B)** Yellow-green module. **(C)** Red module. The color of the line connecting the circles deepens as the topological overlap measures increases. The color of the node changes from yellow to red as the degree values increases.

indicated that in lung adenocarcinoma, the overexpression of CXCR2 caused invasion, metastasis, and poor prognosis of tumor patients. Wu et al. (2015) also showed enhanced expression of CXCR2 in colon cancer tissues, particularly in advanced-stage tumor cells or tumor cells with lymph node metastasis, indicating the potential to use the expression level of CXCR2 for evaluating tumor growth and invasion in CRC. Our results showed that hsa-mir-4437 was an independent prognostic factor for colon cancer, and CXCR2 was found from the target prediction by both TargetScan and miRDB. Therefore, hsa-mir-4437 may affect the proliferation and apoptosis of colon cancer cells by targeting CXCR2.

According to our prediction results, all 14 critical genes of colon cancer we identified from the PPI network were up-regulated in colon cancer cells. Among these 14 genes, abnormal overexpression of *CEP55*, *DTL*, *HMMR*, *MCM6*, *MKI67*, *NEK2*, *RACGAP1*, *RRM2*, *TOP2A*, and *UBE2C* have previously been reported in colon cancer (**Table 4**).

Denticleless E3 ubiquitin–protein ligase homolog (DTL) complex is a nuclear protein that targets centrosomes in mitosis, with an important role in DNA synthesis, cell cycle regulation, cytokinesis, proliferation, and differentiation (Pan et al., 2006). Baraniskin et al. (2012) demonstrated that miR-30a-5p could produce a tumor suppressor effect by repressing the overexpression of DTL in colon cancer. Karaayvaz et al. (2011) showed miR-215 achieved a similar outcome. The Rac GTPase activating protein 1 (RACGAP1) is a member of the GTPase-active protein family, with a regulatory role in cell division, cell growth differentiation, and tumor metastasis and proliferation (Milde-Langosch et al., 2013; Yeh et al., 2016). According to Yeh et al. (2016), patients with high expression of cytoplasmic RACGAP1 in CRC had a favorable prognosis, whereas those with high expression of nuclear RACAGAP1 had a poor prognosis. Imaoka et al. (2015) demonstrated that RACGAP1 expression was dramatically high in CRC with advanced tumor stage, vessel invasion, and lymph node and distant metastasis, causing poor overall survival. The *marker of proliferation Ki-67* (*MKI67*) is a nucleoprotein gene involved in cell proliferation and expressed at all stages of the cell cycle (Yang et al., 2017). Lin et al. (2008) detected high expression of MKI67 in CRC based on immunohistochemistry. Zeng et al. (2019) showed that in CRC, the knockdown of oncogenic gene *small nuclear ribonucleoprotein polypeptide A* (*SNRPA1*) caused the down-regulation of its other downstream genes, including *MK167*, inhibiting the proliferation of CRC cells. The hyaluronan-mediated motility receptor (HMMR), also known as RHAMM, plays a key role in the occurrence and development of tumors by mediating the migration of hyaluronan to tumor cells and is closely related to cell proliferation, migration, signal transduction, adhesion, and metastasis (Hatano et al., 2011). Koelzer et al. (2015) indicated that HMMR was overexpressed in tumor-budding cells of CRC and associated with advanced tumor

TABLE 4 | Fourteen critical genes reported in cancer from previous studies.

Gene	*CEP55*	***DTL***	***FANCI***	*HMMR*	*KIF5*	*MCM6*	***MKI67***
Feature	●	●	○	●	⊜	●	●
Gene	*NCAPG2*	***NEK2***	*RACGAP1*	***RRM2***	***TOP2A***	***UBE2C***	*ZWILCH*
Feature	⊜	●	●	●	●	●	○

● *Gene experimentally up-regulated in colon cancer, and the expression was consistent with our calculation result of colon cancer.* ⊜ *Gene experimentally up-regulated in other cancers, and the expression was consistent with our calculation result of colon cancer.* ○ *Gene not yet experimentally verified in colon cancer or other cancers. Bold represents the genes that are both critical genes and WGCNA core genes.*

grade, invasion, metastasis, and poor prognosis. HMMR is also a biomarker for poor prognosis in several cancers, including those of the colon, stomach, lung, and breast (Chen et al., 2018).

The ubiquitin-conjugating enzyme E2C (UBE2C) is the central component of the ubiquitin–proteasome system (UPS), an ATP-dependent protein degradation pathway in the cytoplasm and nucleus (Rousseau and Bertolotti, 2018). By immunohistochemical analysis, Fujita et al. (2009) confirmed that the UBE2C content was higher in colon cancer tissues than in normal colon epithelium, and overexpressed UBE2C could change the cell cycle and promote tumor proliferation. Okamoto et al. (2003) noted that UBE2C was highly expressed in a variety of tumors, including CRC, causing cell growth promotion and malignant transformation. *NIMA-related kinase 2* (*NEK2*) encodes a serine/threonine protein kinase involved in the centrosome cell cycle and mitosis regulation. The expression of *NEK2* is closely associated with the prognosis and pathological features of cancer, including colon cancer (Ren et al., 2018). Takahashi et al. (2014) found that the high expression of *NEK2* in CRC was associated with advanced tumor stage, invasion, dissemination, and poor prognosis, but that mir-128 could repress *NEK2* expression, and inhibited cell proliferation. As a member of the MCM family, *mini-chromosome maintenance complex component 6* (*MCM6*) is highly expressed in human malignant cells. The encoded product of *MCM6* is a key protein for DNA replication and is involved in the regulation of the cell cycle (Lei, 2005). *MCM6* is highly expressed in colon cancer tissues (Hendricks et al., 2019). Huang et al. (2018) showed that the suppression of *MCM6* in colon cancer cells could inhibit the foci-forming and chromatin localization of RAD51 recombinase (RAD51), a protein essential for DNA damage recovery. DNA topoisomerase II alpha (TOP2A) is a key enzyme that controls the topological state of DNA and is involved in processes, such as chromosome condensation, chromatid separation, and gene expression (Tsavaris et al., 2009). Zhang R. et al. (2018) detected up-regulated TOP2A in colon cancer tissues compared with adjacent non-cancerous tissues and found that down-regulated TOP2A in colon cancer cells could dramatically inhibit proliferation and invasion of colon cancer cells. The ribonucleotide reductase regulatory subunit M2 (RRM2) plays a vital role in DNA synthesis and repair, as well as many key cellular processes, such as cell proliferation, invasion, migration, senescence, and tumorigenesis (Nordlund and Reichard, 2006). In colon cancer, the increased expression of RRM2 can noticeably enhance the invasive ability of cancer cells (Liu et al., 2007). Liu et al. (2013) proved that increasing the expression of RRM2 in colon cancer cells substantially enhanced cell migration and invasion ability, which indicated that RRM2 was an independent prognostic biomarker for colon cancer and could predict the low survival rate of colon cancer patients. The centrosomal protein 55 (CEP55) is involved not only in the process of cytokinesis but also in the invasion, metastasis, and prognosis of many malignancies (Jeffery et al., 2016). Bioinformatics analysis performed by Hauptman et al. (2019) indicated that CEP55 was overexpressed in CRC and could be used as a potential biomarker in colon cancer tissues, as validated in clinical samples. Similarly, Sakai et al. (2006) reported that inhibiting the expression of CEP55 caused a marked reduction in the growth rate of colon cancer cells. These experiment results on the expression of the above 10 genes in colon cancer are consistent with our calculation results, which further verified the reliability of our computational analysis.

As mentioned above, 4 of the 14 key genes recognized in this study as playing major roles in colon cancer (*FANCI*, *KIF15*, *NCAPG2*, and *ZWILCH*) have not been experimentally shown to be up-regulated in colon cancer. However, abnormal overexpression of *KIF15* and *NCAPG2* has been detected in many other types of cancer. Kinesin family member 15 (KIF15) is involved in important biological processes, including mitosis, cell signaling pathways, gene translation, and protein trafficking (Penna et al., 2017). According to Sheng et al. (2019), the up-regulation of *KIF15* in breast cancer led to poor overall survival, indicating that *KIF15* could serve as a potential therapeutic target for triple-negative breast cancer. In a study by Wang et al. (2017), overexpression of *KIF15* in pancreatic cancer promoted the expression of p-MEK and p-ERK, inducing activation of the MEK–ERK signaling pathway and causing G_1/S phase transition and cancer growth. The non-SMC condensin II complex subunit G2 (NCAPG2) is a component of the condensin II complex that interacts with Polo-like kinase 1 (PLK1) during the anterior-to-metaphase transition of mitosis, thereby regulating correct chromosome segregation (Kim et al., 2014). The up-regulation of *NCAPG2* in non-small cell lung cancer (NSCLC) cells caused a short survival time, whereas suppressing *NCAPG2* expression led to proliferation inhibition and G_2/S cycle arrest (Zhan et al., 2017). Zhan et al. (2017) concluded that *NCAPG2* expression was closely related to the progression of NSCLC and could act as a prognostic factor. In liver cancer, Meng et al. (2019) determined that highly expressed *NCAPG2* promoted tumor cell proliferation, migration, and invasion, mediated by activation of the STAT3 and NF-κB pathways. Such findings confirmed *NCAPG2* as both an oncogene of liver cancer and a biomarker

predicting poor patient prognosis. In summary, *KIF15* and *NCAPG2* might be involved in the development and progression of colon cancer, and they serve as prognostic markers or therapeutic targets for colon cancer.

Zwilch kinetochore protein (ZWILCH) is an important component of the Rod–Zw10–Zwilch complex and is crucial for maintaining the normal function of mitotic checkpoints (Kops et al., 2005). The abnormal function of mitotic checkpoints is associated with the appearance of chromosomal instability, a consensus sign of many human malignancies. According to Shih et al. (2001), chromosomal instability occurs in the early stages of colon cancer, resulting in genomic instability that might promote tumor development. *Fanconi anemia* (*FA*) *complementation group I* (*FANCI*) is a gene belonging to the FA–breast cancer pathway, and the mono-ubiquitination of the FANCI–FANCD2 protein complex is the key to the normal function of the FA pathway (Smogorzewska et al., 2007). A dysfunctional FA pathway reduces the ability of DNA repair, causing genomic instability, which increases the incidence of tumor development (Deans and West, 2011). *FANCI* is one of the most pathogenic mutated genes in CRC (Zhunussova et al., 2019). This gene is a negative regulator of the consensus oncogene *Akt* (Zhang et al., 2016). Although no current study concerns the direct correlation between these two genes and cancer, both genes can act as components of cancer progression pathways and play certain roles in the formation of cancer, yet the specific mechanism remains unclear. Our results identified *FANCI* and *ZWILCH* as critical target genes of colon cancer, suggesting that they might provide a potential pathway for the treatment and intervention of colon cancer.

In this study, we also performed WGCNA to mine the core genes via the analysis of gene expression patterns of multiple samples in GSE24514 and constructed 17 co-expression modules from 13,640 genes of these transcriptome data. Among them, three modules (midnight blue, red, and yellow-green) containing the most DEGs were the key functional ones significantly related to colon cancer. These modules comprised 19 core genes (*CCNB1*, *DTL*, *ENTPD5*, *FANCI*, *GUCA2A*, *HSPB8*, *KCNMB1*, *LMOD1*, *MKI67*, *MS4A12*, *NEK2*, *PADI2*, *PRC1*, *PTTG1*, *RRM2*, *SCNN1B*, *TNS1*, *TOP2A*, and *UBE2C*). 7 genes, including *DTL*, *FANCI*, *MKI67*, *NEK2*, *RRM2*, *TOP2A*, and *UBE2C*, appeared in the above-discussed 14 critical genes from the consensus genes (**Table 4**), which, at least partially verified the reliability of the main results of this work. In addition, although the remaining 12 core genes from WGCNA were not target genes for our derived prognostic miRNAs, they were DEGs of colon cancer, and their relationship with colon cancer deserves further study.

Bioinformatics is indispensable for mining data related to colon cancer. In 2018, Wei et al. (2018) constructed a 10-miRNA prognostic model composed of hsa-mir-891a, **hsa-mir-6854**, **hsa-mir-216a**, hsa-mir-378d-1, hsa-mir-92a-1, hsa-mir-4709, hsa-mir-92a-2, hsa-mir-210, hsa-mir-940, and **hsa-mir-887**, by analyzing the genome-wide miRNA sequencing dataset and corresponding clinical information of 425 colon adenocarcinoma patients from TCGA. In the same year, Zhang H. et al. (2018) identified DEMs between 457 colon cancer tissues and 8 normal tissues from TCGA. Subsequent Cox proportional hazards regression analysis provided a prognostic model of six miRNAs, including hsa-mir-149, **hsa-mir-3189**, **hsa-mir-3677**, hsa-mir-3917, **hsa-mir-4999**, and **hsa-mir-6854**. Thus, six of eight prognostic factors (**hsa-mir-6854**, hsa-mir-4437, **hsa-mir-216a**, **hsa-mir-3677**, **hsa-mir-887**, **hsa-mir-4999**, hsa-mir-34b, and **hsa-mir-3189**) we calculated were consistent with their results. It is worth noting that although the expressions of miR-887 and miR-3189 in other cancers were experimentally different from our predicted ones, the above two studies provided the same analysis results as ours showed in colon cancer, namely down-regulated miR-887 and up-regulated miR-3189. However, different from these two previous works, our study further explored the potential critical target genes of the prognostic miRNAs by using combination methods, including target prediction calculation, differential expression screening, intersection analysis, and PPI network construction and visualization.

As shown in **Table 3**, 14 targeting relationships are available between 8 prognostic miRNAs and 14 critical genes. Specifically, *DTL*, *HMMR*, *MKI67*, and *RACGAP1* were predicted as the target genes of hsa-mir-6854. *FANCI*, *KIF15*, *NEK2*, *UBE2C*, and *ZWILCH* were the target genes of hsa-mir-3189, hsa-mir-3677, hsa-mir-216a, hsa-mir-4437, and hsa-mir-4999, respectively. *MCM6* and *TOP2A* were the target genes of hsa-mir-887, and *CEP55*, *NCAPG2*, and *RRM2* were the target genes of hsa-mir-34b (**Table 3**). To date, there have been no experiments to confirm any of the above targeting relationships. However, the targeting relationships of *KIG15* with has-mir-3677 (Zorniak et al., 2018), and *CEP55* with has-mir-34b (Liang, 2008) were also predicted in the functional analysis of miRNA in patients with gastric antral vascular ectasia and expression meta-analysis of lung cancer miRNA targets, respectively. As discussed above, experiments have shown that these miRNAs and genes are mostly, directly or indirectly, related to colon cancer. Therefore, we speculate the existence of these targeting relationships for further study, which might clarify the mechanisms of colon cancer and provide novel methods for future exploration of prevention and treatment.

CONCLUSION

In summary, we used bioinformatics methods to construct a prognostic model of colon cancer patients with eight prognostic miRNAs, including hsa-mir-6854, hsa-mir-4437, hsa-mir-216a, hsa-mir-3677, hsa-mir-887, hsa-mir-4999, hsa-mir-34b, and hsa-mir-3189. Fourteen potential critical target genes of these independent prognostic biomarkers were identified in the PPI network. These genes were *CEP55*, *DTL*, *FANCI*, *HMMR*, *KIF15*, *MCM6*, *MKI67*, *NCAPG2*, *NEK2*, *RACGAP1*, *RRM2*, *TOP2A*, *UBE2C*, and *ZWILCH*. One miRNA (hsa-mir-4437) and four genes (*FANCI*, *KIF15*, *NCAPG2*, and *ZWILCH*) have not yet been confirmed to be associated with colon cancer in previous experiments and calculations. In addition, the targeting relationship between the 8 prognostic miRNAs and the 14 critical genes deserves further study. Furthermore, 12 core genes obtained from WGCNA are also worthy of future research. Our

results indicate that these prognostic miRNAs and their target genes could have valuable potential for prognosis and targeted therapy of colon cancer, and thereby could provide new guidance for the diagnosis and treatment of colon cancer in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://portal.gdc.cancer.gov/ and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24514.

AUTHOR CONTRIBUTIONS

WC, CG, and ZH contributed to the design and conception of the study. WC, CG, YL, YW, and XH did information retrieval and analysis. WC and CG wrote the manuscript. WC, YL, YW, and XH created tables and figures. ZH guided manuscript writing, revised the manuscript and provided financial support. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (31770774), the Provincial Major Project of Basic or Applied Research in Natural Science, Guangdong Provincial Education Department (2016KZDXM038), and the higher education reform project of Guangdong Province (2019268).

ACKNOWLEDGMENTS

We thank Wordvice for their help in revising the English grammar.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00478/full#supplementary-material

REFERENCES

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4:e05005. doi: 10.7554/eLife.05005

Ahmed, F. E., Vos, P. W., and Holbert, D. (2007). Modeling survival in colon cancer: a methodological review. *Mol. Cancer* 6:15. doi: 10.1186/1476-4598-6-15

Aran, V., Victorino, A. P., Thuler, L. C., and Ferreira, C. G. (2016). Colorectal cancer: epidemiology, disease mechanisms and interventions to reduce onset and mortality. *Clin. Colorectal. Cancer* 15, 195–203. doi: 10.1016/j.clcc.2016.02.008

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683–691. doi: 10.1136/gutjnl-2015-310912

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* 4:2. doi: 10.1186/1471-2105-4-2

Baraniskin, A., Birkenkamp-Demtroder, K., Maghnouj, A., Zollner, H., Munding, J., Klein-Scory, S., et al. (2012). MiR-30a-5p suppresses tumor growth in colon carcinoma by targeting DTL. *Carcinogenesis* 33, 732–739. doi: 10.1093/carcin/bgs020

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* 35, D760–D765. doi: 10.1093/nar/gkl887

Bian, Y., Guo, J., Qiao, L., and Sun, X. (2018). miR-3189-3p mimics enhance the effects of S100A4 siRNA on the inhibition of proliferation and migration of gastric cancer cells by targeting CFL2. *Int. J. Mol. Sci.* 19:E236. doi: 10.3390/ijms19010236

Bolmeson, C., Esguerra, J. L., Salehi, A., Speidel, D., Eliasson, L., and Cilio, C. M. (2011). Differences in islet-enriched miRNAs in healthy and glucose intolerant human subjects. *Biochem. Biophys. Res. Commun.* 404, 16–22. doi: 10.1016/j.bbrc.2010.11.024

Cappell, M. S. (2008). Pathophysiology, clinical presentation, and management of colon cancer. *Gastroenterol. Clin. North Am.* 37, 1–24. doi: 10.1016/j.gtc.2007.12.002

Cekaite, L., Eide, P. W., Lind, G. E., Skotheim, R. I., and Lothe, R. A. (2016). MicroRNAs as growth regulators, their function and biomarker status in colorectal cancer. *Oncotarget* 7, 6476–6505. doi: 10.18632/oncotarget.6390

Chen, Y. T., Chen, Z., and Du, Y. N. (2018). Immunohistochemical analysis of RHAMM expression in normal and neoplastic human tissues: a cell cycle protein with distinctive expression in mitotic cells and testicular germ cells. *Oncotarget* 9, 20941–20952. doi: 10.18632/oncotarget.24939

Deans, A. J., and West, S. C. (2011). DNA interstrand crosslink repair and cancer. *Nat. Rev. Cancer* 11, 467–480. doi: 10.1038/nrc3088

Deng, J., Kong, W., Mou, X., Wang, S., and Zeng, W. (2018). Identifying novel candidate biomarkers of RCC based on WGCNA analysis. *Per. Med.* 15, 381–394. doi: 10.2217/pme-2017-0091

Ding, L., Lan, Z., Xiong, X., Ao, H., Feng, Y., Gu, H., et al. (2018). The dual role of microRNAs in colorectal cancer progression. *Int. J. Mol. Sci.* 19:E2791. doi: 10.3390/ijms19092791

Fujita, T., Ikeda, H., Taira, N., Hatoh, S., Naito, M., and Doihara, H. (2009). Overexpression of UbcH10 alternates the cell cycle profile and accelerate the tumor proliferation in colon cancer. *BMC Cancer* 9:87. doi: 10.1186/1471-2407-9-87

Hatano, H., Shigeishi, H., Kudo, Y., Higashikawa, K., Tobiume, K., Takata, T., et al. (2011). RHAMM/ERK interaction induces proliferative activities of cementifying fibroma cells through a mechanism based on the CD44-EGFR. *Lab. Invest.* 91, 379–391. doi: 10.1038/labinvest.2010.176

Hauptman, N., Jevsinek Skok, D., Spasovska, E., Bostjancic, E., and Glavac, D. (2019). Genes CEP55, FOXD3, FOXF2, GNAO1, GRIA4, and KCNA5 as potential diagnostic biomarkers in colorectal cancer. *BMC Med. Genomics* 12:54. doi: 10.1186/s12920-019-0501-z

He, Q., Shi, X., Zhou, B., Teng, J., Zhang, C., Liu, S., et al. (2018). Interleukin 8 (CXCL8)-CXC chemokine receptor 2 (CXCR2) axis contributes to MiR-4437-associated recruitment of granulocytes and natural killer cells in ischemic stroke. *Mol. Immunol.* 101, 440–449. doi: 10.1016/j.molimm.2018.08.002

Heagerty, P. J., and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* 61, 92–105. doi: 10.1111/j.0006-341X.2005.030814.x

Hendricks, A., Gieseler, F., Nazzal, S., Brasen, J. H., Lucius, R., Sipos, B., et al. (2019). Prognostic relevance of topoisomerase II alpha and minichromosome maintenance protein 6 expression in colorectal cancer. *BMC Cancer* 19:429. doi: 10.1186/s12885-019-5631-3

Hiyoshi, Y., Schetter, A. J., Okayama, H., Inamura, K., Anami, K., Nguyen, G. H., et al. (2015). Increased microRNA-34b and -34c predominantly expressed in

stromal tissues is associated with poor prognosis in human colon cancer. *PLoS One* 10:e0124899. doi: 10.1371/journal.pone.0124899

Huang, J., Luo, H. L., Pan, H., Qiu, C., Hao, T. F., and Zhu, Z. M. (2018). Interaction between RAD51 and MCM complex is essential for RAD51 foci forming in colon cancer HCT116 cells. *Biochemistry (Mosc.)* 83, 69–75. doi: 10.1134/s0006297918010091

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Imaoka, H., Toiyama, Y., Saigusa, S., Kawamura, M., Kawamoto, A., Okugawa, Y., et al. (2015). RacGAP1 expression, increasing tumor malignant potential, as a predictive biomarker for lymph node metastasis and poor prognosis in colorectal cancer. *Carcinogenesis* 36, 346–354. doi: 10.1093/carcin/bgu327

Jeansonne, D., DeLuca, M., Marrero, L., Lassak, A., Pacifici, M., Wyczechowska, D., et al. (2015). Anti-tumoral effects of miR-3189-3p in glioblastoma. *J. Biol. Chem.* 290, 8067–8080. doi: 10.1074/jbc.M114.633081

Jeffery, J., Sinha, D., Srihari, S., Kalimutho, M., and Khanna, K. K. (2016). Beyond cytokinesis: the emerging roles of CEP55 in tumorigenesis. *Oncogene* 35, 683–690. doi: 10.1038/onc.2015.128

Jiang, T., Ye, L., Han, Z., Liu, Y., Yang, Y., Peng, Z., et al. (2017). miR-19b-3p promotes colon cancer proliferation and oxaliplatin-based chemoresistance by targeting SMAD4: validation by bioinformatics and experimental analyses. *J. Exp. Clin. Cancer Res.* 36:131. doi: 10.1186/s13046-017-0602-5

Jiang, Y., Wang, N., Yin, D., Li, Y. K., Guo, L., Shi, L. P., et al. (2016). Changes in the expression of serum MiR-887-5p in patients with endometrial cancer. *Int. J. Gynecol. Cancer* 26, 1143–1147. doi: 10.1097/igc.0000000000000730

Jones, M. F., Li, X. L., Subramanian, M., Shabalina, S. A., Hara, T., Zhu, Y., et al. (2015). Growth differentiation factor-15 encodes a novel microRNA 3189 that functions as a potent regulator of cell death. *Cell Death Differ.* 22, 1641–1653. doi: 10.1038/cdd.2015.9

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karaayvaz, M., Pal, T., Song, B., Zhang, C., Georgakopoulos, P., Mehmood, S., et al. (2011). Prognostic significance of miR-215 in colon cancer. *Clin. Colorectal. Cancer* 10, 340–347. doi: 10.1016/j.clcc.2011.06.002

Kim, J. H., Shim, J., Ji, M. J., Jung, Y., Bong, S. M., Jang, Y. J., et al. (2014). The condensin component NCAPG2 regulates microtubule-kinetochore attachment through recruitment of Polo-like kinase 1 to kinetochores. *Nat. Commun.* 5:4588. doi: 10.1038/ncomms5588

Koelzer, V. H., Huber, B., Mele, V., Iezzi, G., Trippel, M., Karamitopoulou, E., et al. (2015). Expression of the hyaluronan-mediated motility receptor RHAMM in tumor budding cells identifies aggressive colorectal cancers. *Hum. Pathol.* 46, 1573–1581. doi: 10.1016/j.humpath.2015.07.010

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Kops, G. J., Weaver, B. A., and Cleveland, D. W. (2005). On the road to cancer: aneuploidy and the mitotic checkpoint. *Nat. Rev. Cancer* 5, 773–785. doi: 10.1038/nrc1714

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Lei, M. (2005). The MCM complex: its role in DNA replication and implications for cancer therapy. *Curr. Cancer Drug Targets* 5, 365–380. doi: 10.2174/1568009054629654

Liang, Y. (2008). An expression meta-analysis of predicted microRNA targets identifies a diagnostic signature for lung cancer. *BMC Med. Genomics* 1:61. doi: 10.1186/1755-8794-1-61

Lin, M. X., Wen, Z. F., Feng, Z. Y., and He, D. (2008). Expression and significance of Bmi-1 and Ki67 in colorectal carcinoma tissues. *Ai Zheng* 27, 1321–1326.

Liu, X., Zhang, H., Lai, L., Wang, X., Loera, S., Xue, L., et al. (2013). Ribonucleotide reductase small subunit M2 serves as a prognostic biomarker and predicts poor survival of colorectal cancers. *Clin. Sci. (Lond.)* 124, 567–578. doi: 10.1042/cs20120240

Liu, X., Zhou, B., Xue, L., Yen, F., Chu, P., Un, F., et al. (2007). Ribonucleotide reductase subunits M2 and p53R2 are potential biomarkers for metastasis of colon cancer. *Clin. Colorectal. Cancer* 6, 374–381. doi: 10.3816/CCC.2007.n.007

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* 40, 346–358.

Meng, F., Zhang, S., Song, R., Liu, Y., Wang, J., Liang, Y., et al. (2019). NCAPG2 overexpression promotes hepatocellular carcinoma proliferation and metastasis through activating the STAT3 and NF-kappaB/miR-188-3p pathways. *EBioMedicine* 44, 237–249. doi: 10.1016/j.ebiom.2019.05.053

Milde-Langosch, K., Karn, T., Muller, V., Witzel, I., Rody, A., Schmidt, M., et al. (2013). Validity of the proliferation markers Ki67, TOP2A, and RacGAP1 in molecular subgroups of breast cancer. *Breast Cancer Res. Treat.* 137, 57–67. doi: 10.1007/s10549-012-2296-x

Ning, X., and Sun, L. (2020). Gene network analysis reveals a core set of genes involved in the immune response of Japanese flounder (*Paralichthys olivaceus*) against Vibrio anguillarum infection. *Fish Shellfish Immunol.* 98, 800–809. doi: 10.1016/j.fsi.2019.11.033

Nordlund, P., and Reichard, P. (2006). Ribonucleotide reductases. *Annu. Rev. Biochem.* 75, 681–706. doi: 10.1146/annurev.biochem.75.103004.142443

Okamoto, Y., Ozaki, T., Miyazaki, K., Aoyama, M., Miyazaki, M., and Nakagawara, A. (2003). UbcH10 is the cancer-related E2 ubiquitin-conjugating enzyme. *Cancer Res.* 63, 4167–4173.

Pan, H. W., Chou, H. Y., Liu, S. H., Peng, S. Y., Liu, C. L., and Hsu, H. C. (2006). Role of L2DTL, cell cycle-regulated nuclear and centrosome protein, in aggressive hepatocellular carcinoma. *Cell Cycle* 5, 2676–2687. doi: 10.4161/cc.5.22.3500

Penna, L. S., Henriques, J. A. P., and Bonatto, D. (2017). Anti-mitotic agents: are they emerging molecules for cancer treatment? *Pharmacol. Ther.* 173, 67–82. doi: 10.1016/j.pharmthera.2017.02.007

Ren, Q., Li, B., Liu, M., Hu, Z., and Wang, Y. (2018). Prognostic value of NEK2 overexpression in digestive system cancers: a meta-analysis and systematic review. *Oncol. Targets Ther.* 11, 7169–7178. doi: 10.2147/ott.s169911

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rousseau, A., and Bertolotti, A. (2018). Regulation of proteasome assembly and activity in health and disease. *Nat. Rev. Mol. Cell Biol.* 19, 697–712. doi: 10.1038/s41580-018-0040-z

Saintigny, P., Massarelli, E., Lin, S., Ahn, Y. H., Chen, Y., Goswami, S., et al. (2013). CXCR2 expression in tumor cells is a poor prognostic factor and promotes invasion and metastasis in lung adenocarcinoma. *Cancer Res.* 73, 571–582. doi: 10.1158/0008-5472.can-12-0263

Sakai, M., Shimokawa, T., Kobayashi, T., Matsushima, S., Yamada, Y., Nakamura, Y., et al. (2006). Elevated expression of C10orf3 (chromosome 10 open reading frame 3) is involved in the growth of human colon tumor. *Oncogene* 25, 480–486. doi: 10.1038/sj.onc.1209051

Scardoni, G., Tosadori, G., Faizan, M., Spoto, F., Fabbri, F., and Laudanna, C. (2014). Biological network analysis with CentiScaPe: centralities and experimental dataset integration. *F1000Res* 3:139. doi: 10.12688/f1000research.4477.2

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sheng, J., Xue, X., and Jiang, K. (2019). Knockdown of kinase family 15 inhibits cancer cell proliferation in vitro and its clinical relevance in triple-negative breast cancer. *Curr. Mol. Med.* 19, 147–155. doi: 10.2174/1566524019666190308122108

Shih, I. M., Zhou, W., Goodman, S. N., Lengauer, C., Kinzler, K. W., and Vogelstein, B. (2001). Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. *Cancer Res.* 61, 818–822.

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., et al. (2017). Colorectal cancer statistics, 2017. *CA Cancer J. Clin.* 67, 177–193. doi: 10.3322/caac.21395

Smogorzewska, A., Matsuoka, S., Vinciguerra, P., McDonald, E. R. III, Hurov, K. E., Luo, J., et al. (2007). Identification of the FANCI protein, a monoubiquitinated

FANCD2 paralog required for DNA repair. *Cell* 129, 289–301. doi: 10.1016/j.cell.2007.03.009

Sui, J., Xu, S. Y., Han, J., Yang, S. R., Li, C. Y., Yin, L. H., et al. (2017). Integrated analysis of competing endogenous RNA network revealing lncRNAs as potential prognostic biomarkers in human lung squamous cell carcinoma. *Oncotarget* 8, 65997–66018. doi: 10.18632/oncotarget.19627

Sun, C., Yuan, Q., Wu, D., Meng, X., and Wang, B. (2017). Identification of core genes and outcome in gastric cancer using bioinformatics analysis. *Oncotarget* 8, 70271–70280. doi: 10.18632/oncotarget.20082

Sun, S., Hang, T., Zhang, B., Zhu, L., Wu, Y., Lv, X., et al. (2019). miRNA-708 functions as a tumor suppressor in colorectal cancer by targeting ZEB1 through Akt/mTOR signaling pathway. *Am. J. Transl. Res.* 11, 5338–5356. eCollection 2019.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Takahashi, Y., Iwaya, T., Sawada, G., Kurashige, J., Matsumura, T., Uchi, R., et al. (2014). Up-regulation of NEK2 by microRNA-128 methylation is associated with poor prognosis in colorectal cancer. *Ann. Surg. Oncol.* 21, 205–212. doi: 10.1245/s10434-013-3264-3

Tan, W., Liu, B., Qu, S., Liang, G., Luo, W., and Gong, C. (2018). MicroRNAs and cancer: key paradigms in molecular therapy. *Oncol. Lett.* 15, 2735–2742. doi: 10.3892/ol.2017.7638

Toiyama, Y., Hur, K., Tanaka, K., Inoue, Y., Kusunoki, M., Boland, C. R., et al. (2014). Serum miR-200c is a novel prognostic and metastasis-predictive biomarker in patients with colorectal cancer. *Ann. Surg.* 259, 735–743. doi: 10.1097/SLA.0b013e3182a6909d

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* 19, A68–A77. doi: 10.5114/wo.2014.47136

Tsavaris, N., Lazaris, A., Kosmas, C., Gouveris, P., Kavantzas, N., Kopterides, P., et al. (2009). Topoisomerase I and IIalpha protein expression in primary colorectal cancer and recurrences following 5-fluorouracil-based adjuvant chemotherapy. *Cancer Chemother. Pharmacol.* 64, 391–398. doi: 10.1007/s00280-008-0886-4

Wang, D., Li, Y., Zhang, C., Li, X., and Yu, J. (2018). MiR-216a-3p inhibits colorectal cancer cell proliferation through direct targeting COX-2 and ALOX5. *J. Cell Biochem.* 119, 1755–1766. doi: 10.1002/jcb.26336

Wang, J., Guo, X., Xie, C., and Jiang, J. (2017). KIF15 promotes pancreatic cancer proliferation via the MEK-ERK signalling pathway. *Br. J. Cancer* 117, 245–255. doi: 10.1038/bjc.2017.165

Wang, L. Y., Li, B., Jiang, H. H., Zhuang, L. W., and Liu, Y. (2015). Inhibition effect of miR-577 on hepatocellular carcinoma cell growth via targeting beta-catenin. *Asian Pac. J. Trop. Med.* 8, 923–929. doi: 10.1016/j.apjtm.2015.10.001

Wei, H. T., Guo, E. N., Liao, X. W., Chen, L. S., Wang, J. L., Ni, M., et al. (2018). Genomescale analysis to identify potential prognostic microRNA biomarkers for predicting overall survival in patients with colon adenocarcinoma. *Oncol. Rep.* 40, 1947–1958. doi: 10.3892/or.2018.6607

Williams, O., and Del Genio, C. I. (2014). Degree correlations in directed scale-free networks. *PLoS One* 9:e110121. doi: 10.1371/journal.pone.0110121

Wong, N., and Wang, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.* 43, D146–D152. doi: 10.1093/nar/gku1104

Wu, W., Sun, C., Xu, D., Zhang, X., Shen, W., Lv, Y., et al. (2015). Expression of CXCR2 and its clinical significance in human colorectal cancer. *Int. J. Clin. Exp. Med.* 8, 5883–5889. eCollection 2015.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322. doi: 10.1093/nar/gkr483

Yang, C. K., Yu, T. D., Han, C. Y., Qin, W., Liao, X. W., Yu, L., et al. (2017). Genome-wide association study of MKI67 expression and its clinical implications in HBV-related hepatocellular carcinoma in Southern China. *Cell Physiol. Biochem.* 42, 1342–1357. doi: 10.1159/000478963

Yang, L., Yang, J., Li, J., Shen, X., Le, Y., Zhou, C., et al. (2015). MircoRNA-33a inhibits epithelial-to-mesenchymal transition and metastasis and could be a prognostic marker in non-small cell lung cancer. *Sci. Rep.* 5:13677. doi: 10.1038/srep13677

Yeh, C. M., Sung, W. W., Lai, H. W., Hsieh, M. J., Yen, H. H., Su, T. C., et al. (2016). Opposing prognostic roles of nuclear and cytoplasmic RACGAP1 expression in colorectal cancer patients. *Hum. Pathol.* 47, 45–51. doi: 10.1016/j.humpath.2015.09.002

Yong, L., YuFeng, Z., and Guang, B. (2018). Association between PPP2CA expression and colorectal cancer prognosis tumor marker prognostic study. *Int. J. Surg.* 59, 80–89. doi: 10.1016/j.ijsu.2018.09.020

Yu, H., Xu, W., Gong, F., Chi, B., Chen, J., and Zhou, L. (2017). MicroRNA-155 regulates the proliferation, cell cycle, apoptosis and migration of colon cancer cells and targets CBL. *Exp. Ther. Med.* 14, 4053–4060. doi: 10.3892/etm.2017.5085

Zeng, Q., Lei, F., Chang, Y., Gao, Z., Wang, Y., Gao, Q., et al. (2019). An oncogenic gene, SNRPA1, regulates PIK3R1, VEGFC, MKI67, CDK1 and other genes in colorectal cancer. *Biomed. Pharmacother.* 117:109076. doi: 10.1016/j.biopha.2019.109076

Zhan, P., Xi, G. M., Zhang, B., Wu, Y., Liu, H. B., Liu, Y. F., et al. (2017). NCAPG2 promotes tumour proliferation by regulating G2/M phase and associates with poor prognosis in lung adenocarcinoma. *J. Cell Mol. Med.* 21, 665–676. doi: 10.1111/jcmm.13010

Zhang, H., Wang, Z., Ma, R., Wu, J., and Feng, J. (2018). MicroRNAs as biomarkers for the progression and prognosis of colon carcinoma. *Int. J. Mol. Med.* 42, 2080–2088. doi: 10.3892/ijmm.2018.3792

Zhang, R., Xu, J., Zhao, J., and Bai, J. H. (2018). Proliferation and invasion of colon cancer cells are suppressed by knockdown of TOP2A. *J. Cell Biochem.* 119, 7256–7263. doi: 10.1002/jcb.26916

Zhang, X., Lu, X., Akhter, S., Georgescu, M. M., and Legerski, R. J. (2016). FANCI is a negative regulator of Akt activation. *Cell Cycle* 15, 1134–1143. doi: 10.1080/15384101.2016.1158375

Zhang, Y., Lin, C., Liao, G., Liu, S., Ding, J., Tang, F., et al. (2015). MicroRNA-506 suppresses tumor proliferation and metastasis in colon cancer by directly targeting the oncogene EZH2. *Oncotarget* 6, 32586–32601. doi: 10.18632/oncotarget.5309

Zhao, X., Liang, M., Li, X., Qiu, X., and Cui, L. (2018). Identification of key genes and pathways associated with osteogenic differentiation of adipose stem cells. *J. Cell Physiol.* 233, 9777–9785. doi: 10.1002/jcp.26943

Zhunussova, G., Afonin, G., Abdikerim, S., Jumanov, A., Perfilyeva, A., Kaidarova, D., et al. (2019). Mutation spectrum of cancer-associated genes in patients with early onset of colorectal cancer. *Front. Oncol.* 9:673. doi: 10.3389/fonc.2019.00673

Zorniak, M., Garczorz, W., Wosiewicz, P., Marek, T., Blaszczynska, M., Waluga, M., et al. (2018). Mucosal miR-3677 is over-expressed in cirrhotic patients with gastric antral vascular ectasia (GAVE). *Scand. J. Gastroenterol.* 53, 1503–1508. doi: 10.1080/00365521.2018.1547922

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

METHODS
published: 24 June 2020
doi: 10.3389/fonc.2020.01052

A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis

*Duc Tran[1], Hung Nguyen[1], Uyen Le[2], George Bebis[1], Hung N. Luu[3,4] and Tin Nguyen[1]**

[1] Department of Computer Science and Engineering, University of Nevada, Reno, NV, United States, [2] NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, [3] Division of Cancer Control and Population Sciences, Hillman Cancer Canter, University of Pittsburgh Medical Center, Pittsburgh, PA, United States, [4] Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, United States

Cancer is an umbrella term that includes a range of disorders, from those that are fast-growing and lethal to indolent lesions with low or delayed potential for progression to death. One critical unmet challenge is that molecular disease subtypes characterized by relevant clinical differences, such as survival, are difficult to differentiate. With the advancement of multi-omics technologies, subtyping methods have shifted toward data integration in order to differentiate among subtypes from a holistic perspective that takes into consideration phenomena at multiple levels. However, these integrative methods are still limited by their statistical assumption and their sensitivity to noise. In addition, they are unable to predict the risk scores of patients using multi-omics data. Here, we present a novel approach named Subtyping via Consensus Factor Analysis (SCFA) that can efficiently remove noisy signals from consistent molecular patterns in order to reliably identify cancer subtypes and accurately predict risk scores of patients. In an extensive analysis of 7,973 samples related to 30 cancers that are available at The Cancer Genome Atlas (TCGA), we demonstrate that SCFA outperforms state-of-the-art approaches in discovering novel subtypes with significantly different survival profiles. We also demonstrate that SCFA is able to predict risk scores that are highly correlated with true patient survival and vital status. More importantly, the accuracy of subtype discovery and risk prediction improves when more data types are integrated into the analysis. The SCFA software and TCGA data packages will be available on Bioconductor.

Keywords: multi-omics integration, risk score prediction, cancer subtyping, survival analysis, factor analysis

OPEN ACCESS

Edited by:
Timothy I. Shaw, St. Jude Children's Research Hospital, United States

Reviewed by:
Shengli Li, University of Texas Health Science Center at Houston, United States
Prashanth N. Suravajhala, Birla Institute of Scientific Research, India

****Correspondence:***
Tin Nguyen
tinn@unr.edu

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *31 January 2020*
Accepted: *27 May 2020*
Published: *24 June 2020*

Citation:
Tran D, Nguyen H, Le U, Bebis G, Luu HN and Nguyen T (2020) A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis. Front. Oncol. 10:1052. doi: 10.3389/fonc.2020.01052

1. INTRODUCTION

After 20 years of cancer screening, the chance of a person being diagnosed with prostate or breast cancer has nearly doubled (1–4). However, this has only marginally reduced the number of patients with advanced disease, suggesting that screening has resulted in the substantial harm of excess detection and over-diagnosis. At the same time, 30–50% of patients with non-small cell lung cancer (NSCLC) develop recurrence and die after curative resection (5), suggesting that a subset of patients would have benefited from more aggressive treatments at early stages. Although not routinely recommended as the initial course of treatment, adjuvant and neoadjuvant chemotherapy have been shown to significantly improve the survival of patients with advanced early-stage disease (6–8). The ability to prognosticate outcomes would allow us to manage these diseases better: patients

whose cancer is likely to advance quickly or recur would receive the necessary treatment. The important challenge is to discover the molecular subtypes of disease and subgroups of patients (9–12).

Cluster analysis has been a basic tool for subtype discovery using gene expression data. These include hierarchical clustering (HC), neural networks (13–17), mixture model (18–20), matrix factorization (21, 22), and graph-theoretical approaches (23–25). Arguably, the state-of-the-art approach in this area is Consensus Clustering (CC) (26, 27), which is a resampling-based methodology of class discovery and cluster validation (28–30). However, these approaches are not able to combine multiple data types. Although analyses on a single data type could reveal some distinct characteristics for different subtypes, it is not sufficient to explain the mechanism that happens across multiple biological levels.

With the advancement of multi-omics technologies, recent subtyping methods have shifted toward multi-omics data integration. The goal is to differentiate among subtypes from a holistic perspective, that can take into consideration phenomena at various levels (e.g., transcriptomics, proteomics, epigenetics). These methods can be grouped into three categories: simultaneous data decomposition methods, joint statistical models, and similarity-based approaches. Methods in the first category (data decomposition) include md-modules (31), intNMF (32), and LRAcluster (33). These methods assume that there exist molecular patterns that are shared across multiple types of data. Therefore, these methods aim at finding a low dimensional representation of the high-dimensional multi-omics data that retains those patterns. For example, both md-modules and intNMF utilize a joint non-negative matrix factorization to simultaneously factorize the data matrices of multiple data types. In their design, the basis vectors are shared across all data types while the coefficient matrices vary from data type to data type. These two methods, md-modules and intNMF, only differ in the way they iteratively estimate the coefficient matrices. Another method is LRAcluster, which applies the low-rank approximation and singular vector decomposition to generate low dimensional representations of the data and then performs k-means clustering to identify the subtypes. These methods strongly rely on the assumption that all molecular signals can be linearly and simultaneously reconstructed.

Methods in the second category (statistical modeling) include BCC (34), MDI (35), iClusterBayes (36), iClusterPlus (37), and iCluster (38, 39). These methods assume that each data type follows a mixture of distributions and then integrate multiple types of data using a joint statistical model. The parameters of the mixture models are estimated by maximizing the likelihood of observed data. These methods strongly depend on the correctness of their statistical assumptions. Also, due to a large number of parameters and iterations involved, the computation complexity of statistical methods is usually extensive. Therefore, these methods often rely on pre-processing and gene filtering to ease the computational burden.

Methods in the third category (similarity-based) typically construct the pair-wise connectivity between patients (that represents how often the patients are grouped together) for each data type and then integrate multiple data types by fusing the individual connectivity matrices. As these methods perform data integration in the sample space, their computational complexity depends mostly on the number of patients, not the dimensions of features/genes. Therefore, these methods are capable of performing subtyping on a genomic scale. Methods in this category include SNF (40), rMKL-DR (41), NEMO (42), CIMLR (43), and PINS (44, 45). SNF creates a patient-to-patient network by fusing connectivity matrices and then partitions the network using spectral clustering (46). rMKL-DR projects samples into a lower-dimensional subspace and then partitions the patients using k-means. NEMO follows a similar strategy with the difference is that it incorporates only partial data into the integrative analysis. Though powerful, these methods do not account for the noise and unstable nature of quantitative assays. PINS and CIMLR follow two different strategies to address noise and instability. PINS introduces Gaussian noise to the data in order to obtain subtypes that are robust against data perturbation. CIMLR combines multiple gaussian kernels per data type to measure the similarity between each pair of samples. The resulted similarity matrix is then subjected to dimension reduction and k-means to determine the subtypes. Though powerful, the similarity metrics used in these methods (i.e., Gaussian kernel, Euclidean distance) make them susceptible to noise and the "curse of dimensionality" (47) from the high-dimensional multi-omics data.

Here we propose a novel approach, named Subtyping via Consensus Factor Analysis (SCFA), that follows a three-stage hierarchical process to ensure the robustness of the discovered subtypes. First, the method uses an autoencoder to filter out genes with an insignificant contribution in characterizing each patient. Second, it applies a modified factor analysis to generate a collection of factor representations of the high-dimensional multi-omics data. Finally, it utilizes a consensus ensemble to find subtypes that are shared across all factor representations. The software package also includes a model based on Cox regression and Elastic net that is able to predict the risk scores of new patients. In an extensive analysis using 7,973 samples related to 30 different cancer diseases, we demonstrate that our method outperforms other state-of-the-art methods in discovering subtypes with significantly different survival profiles. We also demonstrate that data integration indeed improves the subtyping procedure as subtypes obtained from multi-omics data have more significant Cox p-values than subtypes obtained from individual data types. Finally, we demonstrate that the method is able to predict the risk factor of new patients with high accuracy.

2. METHODS

The high-level workflow of SCFA is shown in **Figure 1**. The framework consists of two main modules: disease subtyping (**Figure 1A**) and risk assessment (**Figure 1B**). The input of the subtyping module is a list of data matrices (e.g., mRNA, methylation, miRNA) in which rows represent patients while columns represent genes/features. For each matrix, the method

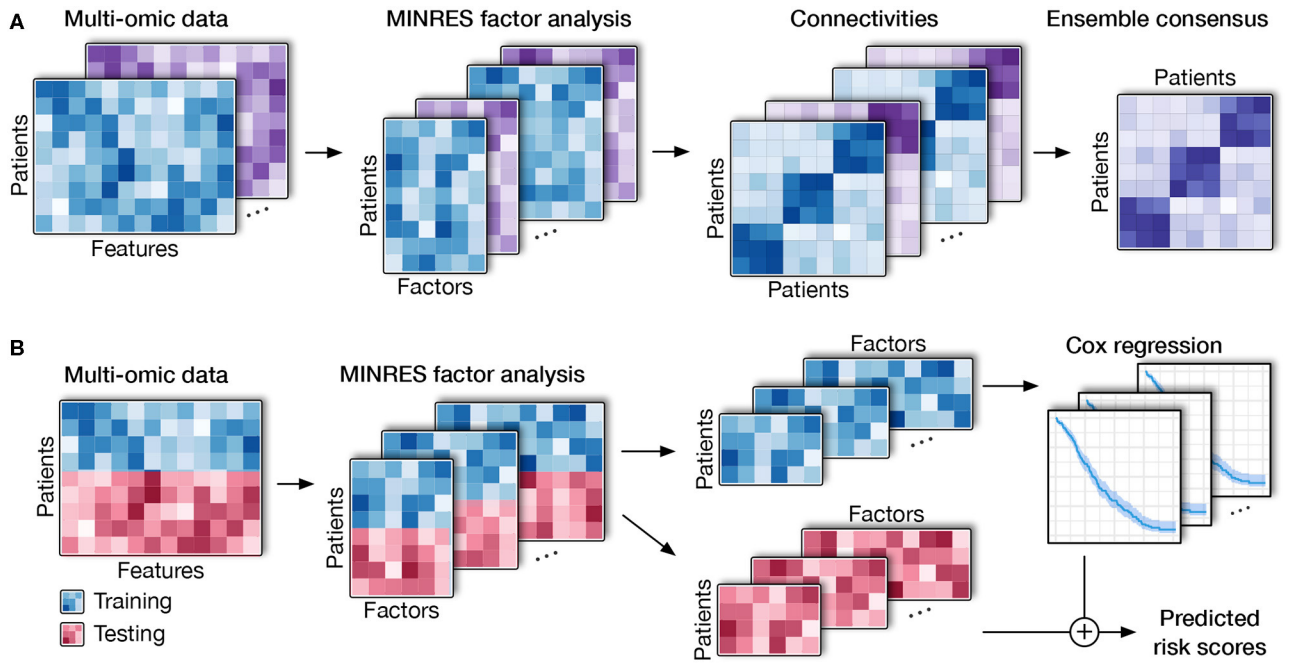

FIGURE 1 | Overall SCFA pipeline. **(A)** Cancer subtyping using multi-omics data. For each of the data matrix, SCFA repeatedly performs factor analysis to generate multiple data representations with different numbers of factors. For each representation, SCFA clusters the data to construct a connectivity matrix. The method next merges all connectivity matrices using an ensemble strategy to obtain the final clustering. **(B)** Risk prediction. SCFA is able to learn from training data (patients with survival information) in order to predict risk scores of patients in testing data (patients without survival information). SCFA first merges training and testing sets together and then performs factor analysis. Using the factor representations of the training set, the method trains a Cox regression model, which will be utilized to predict risk factor of patients in the testing set.

first performs a filtering step using an autoencoder and then repeatedly performs factor analysis (48) to represent the data with different numbers of factors. By representing data with different numbers of factors, we can improve on situations where the projected data do not accurately represent the original data due to noise. Using an ensemble strategy, SCFA combines all of the factor representations to determine the final subtypes.

In the second module, SCFA focuses on predicting the risk scores of patients with unknown survival information. In this module, SCFA combines factor analysis with Cox regression (49, 50) and elastic net (51) to build a prediction model. The method first performs factor analysis on both training (patients with survival information) and testing data (patients without survival information) and then builds a Cox regression model, which can be used to predict the risk scores of patients from the testing data. By default, our software package includes data obtained from The Cancer Genome Atlas (TCGA) that can be used as the training data by default. However, users are free to provide new training data. Using the training data, users can train the model and then predict the risk score of new patients using molecular data.

In the following sections, we will describe in detail the techniques used in the SCFA framework: (i) dimension reduction and factor analysis, (ii) the ensemble strategy for subtyping, and (iii) Cox model and elastic net for risk assessment.

2.1. Dimension Reduction and Factor Analysis

Both modules start with dimension reduction and factor analysis. The purpose of dimension reduction is to remove features/genes that play no role in differentiating between patients. This technique was originally introduced in our scDHA method for single-cell analysis (52). Briefly, we utilize a non-negative kernel autoencoder which consists of two components: encoder and decoder. The encoder aims at representing the data in a low dimensional space whereas the decoder tries to reconstruct the original input from the compressed data. By forcing the weights of the network to be non-negative, we capture the positive correlation between the original features and the representative features. Selecting features with high variability in weights would result in a set of features that are informative, non-redundant, and capable of representing the original data.

After the filtering step using the non-negative autoencoder, we perform another dimension reduction step using Factor Analysis (FA) (48). In general, factor analysis aims at minimizing the difference of feature-feature correlation matrix between the latent space and original data. Correlation is a standardized metric, where it takes into account the number of observations and variance of the features during the calculation process. This makes factor analysis robust against scaling and high number of dimensions compared to traditional decomposition such as principle component analysis (PCA), which uses

Euclidean distance as the distance metric. To further improve the performance of factor analysis, we adjust the objective of FA to maintain the patient-patient correlation.

Starting with the original correlation matrix, FA finds k (number of factors) largest principle components and tries to reproduce the original matrix using those principal components (model matrix). FA iteratively fits the model matrix to the original matrix using optimization algorithms. In our model, we employ the Minimum Residual (MINRES) optimization because it copes better with the small and medium sample size of the input data (53). Also, instead of preserving the relationship between variables, we aim to maintain the overall patient-patient relationships by preserving their Pearson correlations in the representations. By changing the objective, the computational power required is significantly lower as the number of patients (in the scale of hundreds) is much lower than the number of features (in the scale of tens of thousands). Moreover, maintaining the distance between patients in the low dimensional representation would be more beneficial for our desired applications. To avoid overfitting, we repeatedly perform factor analysis with different numbers of factors, resulting in multiple representations of each input matrix. In the clustering module (**Figure 1A**), all factor representations of all data types (data matrices) are combined using an ensemble strategy to determine the subtypes. In the risk prediction module (**Figure 1B**), the factor representations of the training data are combined to build the prediction model.

2.2. Subtyping Using Consensus Ensemble

Given a collection of factor representations from all data types, we aim at finding patient subgroups that are consistently observed together in all representations (**Figure 1A**). For each representation, we first determine the optimal number of clusters using two indices: (i) the ratio of *between sum of squares* over the *total sum of squares*, and (ii) the increase of *within sum of squares* when the number of cluster increases (52). After the optimal number of clusters is determined, we use k-means to cluster the underlying factor representation to build a connectivity matrix. To avoid the convergence to a local minimum, we perform k-means clustering using multiple starting points and choose the results with the smallest sum of square error. This process is repeated for all of the representations to obtain a collection of connectivity matrices for all data types.

Finally, we use the Weighted-based meta-clustering algorithm (54) to combine all clustering results from each data representation to determine the final subtyping. In short, the meta-clustering first calculates the weight for each pair of patients regarding their chance to be grouped together. Next, it assigns a weight for each patient by accumulating the weights of all pairs containing this patient. It then computes the weighted cluster-to-cluster similarity from all connectivity matrices. Finally, it partitions the cluster-to-cluster similarity matrix using hierarchical clustering to determine the final subtypes.

2.3. Risk Score Prediction

The goal of this module is to calculate the risk score of new patients using their molecular data. This is a supervised learning method that learns from a training set in order to predict the risk scores each patient in the testing set. More specifically, the training set consists of a set of patients with molecular data (e.g., mRNA, methylation, miRNA) and known survival information while the testing set consists of patients with only molecular data. By default, we provide TCGA datasets in our package as training data, but users are free to provide training data if necessary. Using the training data, this module will train the Cox regression model that can be used to predict the risk scores of new patients. Below is the description of the method for one data type and for multi-omics data.

Given a single data type as input, we merge the testing data with training data and then perform dimension reduction and factor analysis to generate multiple representations of this data. For each representation, we use the training data to train the Cox regression model. This model aims at estimating a coefficient β_i for each corresponding predictor x_i of the input data. After the model is trained, the risk scores for new patients can be calculated as $\exp(\sum_{i=1}^{n} \beta_i x_i)$, where n is the number of features in the factor representation. In the Cox model, the risk score is defined as $\frac{h(t)}{h_0(t)}$, where $h(t)$ is the expected hazard at time t, and $h_0(t)$ is the baseline hazard when all the predictors are equal zero. Patients with a higher risk score are likely to suffer the event of interest (e.g., vital status or disease recurrence) earlier than the one with a lower risk score. Here we use elastic net (51) implemented in the R-package "glmnet" (55) to fit the model to better cope with the dynamic number of predictors. Elastic net linearly combines Lasso and Ridge penalty during the training process to select only the most relevant predictors that have important effects on the response (the risk scores in this case). We use five-fold cross-validation to select the parameters for the model. The final risk score for each patient is the geometric average of the risk scores resulted from all representations.

In the case of multi-omics data, we repeat the same process (described above) for each data type. We perform factor analysis to produce multiple representations, resulting in a collection of representations from all data types. For a new patient, each representation will produce an estimated risk score. The final risk score for the patient is calculated as the geometric average of all predictions from all representations.

3. RESULT

Here we assess the performance of SCFA using data obtained from 7,973 patients related to 30 different cancer diseases downloaded from The Cancer Genome Atlas (TCGA). For each of the 30 cancer datasets, we downloaded mRNA, miRNA, and methylation data. We also downloaded the clinical data for these patients, which includes vital status and survival information. Using clinical information, we assess the ability of SCFA in both unsupervised subtyping and supervised risk prediction.

3.1. Subtypting on 30 TCGA Datasets

Here we compare the performance of SCFA with four state-of-the-art methods: Consensus Clustering (CC) (26, 27), Similarity Network Fusion (SNF) (40), Cancer Integration via Multikernel

TABLE 1 | Cox *p*-values of subtypes identified by SCFA, CC, SNF, iClusterBayes (iCB), and CIMLR for 30 TCGA datasets.

	SCFA	CC	SNF	iCB	CIMLR
ACC	3.4e-03	5.4e-04	4.3e-05	9.2e-04	3.4e-01
BLCA	7.2e-03	1.1e-01	1.1e-01	5.1e-01	4.7e-01
BRCA	3.2e-04	2.9e-02	1.2e-01	2.7e-02	4.9e-03
CESC	9.4e-03	5.8e-02	5.1e-01	2e-02	1.9e-01
DLBC	4.3e-06	5.1e-01	7.5e-01	2.9e-01	7.4e-01
ESCA	7.3e-05	7.7e-01	3.9e-01	7.9e-01	5.6e-01
GBM	2.3e-03	3.2e-01	2.1e-02	1.1e-01	8.1e-02
GBMLGG	5.8e-14	1.6e-04	4.8e-14	8e-02	6.4e-10
HNSC	4e-02	5e-01	3.7e-01	3.7e-01	4e-01
KICH	2.3e-13	8.7e-01	7e-01	6.9e-01	4.6e-01
KIPAN	1.4e-19	9.3e-08	2.1e-07	1.6e-09	9.8e-05
KIRP	1.7e-03	4.5e-01	5.3e-03	3e-03	1.9e-02
LAML	5.8e-04	3.9e-02	1.7e-03	9e-01	1.4e-04
LGG	6.5e-15	6.6e-07	1.6e-14	1.1e-01	8.3e-15
MESO	1.6e-04	3.1e-01	4.2e-04	3.7e-02	1.1e-02
PAAD	6.9e-04	1.1e-02	7.4e-04	2.3e-03	2e-03
SARC	3.3e-03	2.4e-01	4.4e-02	4.3e-02	5.6e-02
SKCM	1.6e-03	6.3e-01	4.8e-01	8.4e-03	7.4e-05
STES	3.9e-02	2e-01	1.6e-01	4.1e-03	3.4e-02
THCA	7.8e-03	7.9e-01	6.2e-01	7.8e-01	8.6e-03
THYM	8.1e-04	1.5e-01	9.7e-02	9e-03	1.2e-01
UCEC	6.5e-03	8.9e-02	1.8e-02	5.9e-02	4.6e-02
UCS	3.4e-02	1.6e-01	8.6e-01	9.6e-01	3.6e-01
UVM	1.3e-06	6.1e-04	1.7e-04	6.6e-02	5.8e-04
CHOL	3.1e-01	7.9e-02	5.7e-01	9.1e-01	3.4e-01
COAD	4.7e-01	5.8e-01	1.3e-01	2.2e-01	5.6e-01
KIRC	1e-01	8.3e-01	6.9e-01	8.3e-01	9.1e-02
LIHC	3.8e-01	8.8e-01	3.3e-01	9.3e-02	1.9e-01
OV	4.2e-01	6.1e-01	4.4e-01	4.6e-01	5.4e-01
TGCT	3.9e-01	7.4e-01	8.4e-01	7.1e-01	8.4e-01
#Significant	24	8	12	11	13

The cells highlighted in yellow have Cox p*-values smaller than 5%. In each row, cells highlighted in green have the most significant* p*-value. SCFA outperforms other methods by having significant* p*-values in most datasets (24 out of 30 datasets).*

LeaRning (CIMLR) (43), and iClusterBayes (iCB) (36). CC is a resampling-based approach, while SNF and CIMLR are graph-theoretical approaches. The fourth method, iClusterBayes is a model-based approach and is the enhanced version iClusterPlus. These methods were selected to represent three distinctively different subtyping strategies. Among these methods, CC is the only method that cannot integrate multiple data types. For CC, we concatenate the three data types for the integrative analysis. We demonstrate that SCFA outperforms these methods in identifying subtypes with significantly different survival profiles.

Note that here we focus on unsupervised learning, in which each dataset is partitioned independently without using any external information. For example, when analyzing the glioblastoma multiforme (GBM) dataset, we use only the molecular data (mRNA, miRNA, and methylation) of this dataset to determine the subtypes. For each cancer dataset, we first use each of the five methods (SCFA, CC, SNF, CIMLR, and iClusterBayes) to integrate the molecular data (mRNA, miRNA, and methylation) in order to determine patient subgroups. For each method, we calculate the Cox *p*-value that measures the statistical significance in survival differences between the discovered subtypes. The Cox *p*-values of subtypes discovered by the five methods for the 30 datasets are shown in **Table 1**. Among the 30 datasets, there are 6 datasets (CHOL, COAD, KIRC, LIHC, OV, and TGCT) for which no method is able to identify subtypes with significant survival differences. In the remaining 24 datasets, SCFA is able to obtain significant Cox *p*-values in all of them while CC, SNF, iClusterBayes, and CIMLR have significant *p*-values in only 8, 12, 11, and 13 datasets, respectively. Also, SCFA has the most significant *p*-values in 19 out of 24 datasets. Regarding time complexity, SCFA, CC, SNF, and CIMLR are able to analyze each dataset in minutes, whereas iClusterBayes can take up to hours to analyze a dataset.

To better understand the usefulness of data integration, we also calculated the Cox *p*-values obtained from individual data types and compared them to Cox *p*-values obtained from data integration (when mRNA, miRNA, and methylation are analyzed together). For each dataset, we perform subtyping using SCFA for each data type and report the Cox *p*-value of the discovered subtypes. The distributions of Cox *p*-values for data integration and for individual data types using SCFA are shown in **Figure 2**. Among 30 cancer datasets, the Cox *p*-values obtained from data integration has the median $-log10(p)$ of 2.6, compared to 1.7, 1.1, and 1.1 from gene expression, methylation and miRNA data. Interestingly, subtypes discovered using gene expression data have significantly different survival in 18 over 30 datasets, compared to 10 and 14 of methylation and miRNA data, respectively. The figure also shows that the Cox *p*-values obtained from gene expression data are more significant than those obtained from methylation and miRNA data ($p = 0.046$ using one-sided Wilcoxon test). However, we note that miRNA and methylation also provide valuable information in data integration, when all data types are analyzed together. As shown in **Figure 2**, the Cox *p*-values obtained from data integration are more significant than those of any individual data type (including mRNA) with a one-sided Wilcoxon test *p*-value of 0.004. This means that each of the three data types provides meaningful contributions to the data integration. To understand how other methods perform with respect to each data type, we also plot the distributions of Cox *p*-values obtained from each data type using CC, SNF, iClusterBayes, and CIMLR (**Figure S1**). CC is the only method that produces comparable Cox *p*-values across the three data types. SNF and CIMLR perform better using miRNA, while iClusterBayes favors mRNA and miRNA data.

There are four important clinical variables that are available in more than 10 TCGA datasets: age (21 datasets), gender (25 datasets), cancer stages (24 datasets), and tumor grades (12 datasets). To understand the association between these variables and the discovered subtypes, we perform the following analyses: (1) Fisher's exact test to assess the association between gender (male and female) and the discovered subtypes; (2) ANOVA test to assess the age difference between the discovered subtypes;

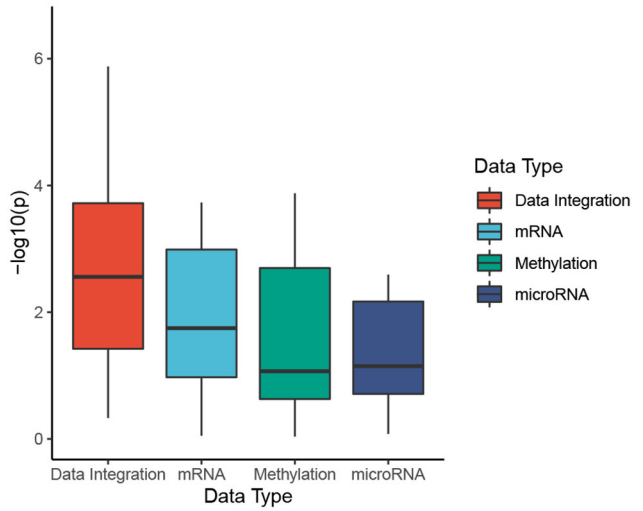

FIGURE 2 | Cox *p*-values of subtypes identified by SCFA. To better understand the usefulness of data integration, we calculate the Cox *p*-values obtained from individual data types and compared them to Cox *p*-values obtained from data integration (when mRNA, miRNA, and methylation are analyzed together). The horizontal axis shows the data types while the vertical axis shows the minus *log*10 *p*-values. Overall the Cox *p*-values obtained from data integration are significantly smaller than those obtained from individual data types ($p = 0.004$ using one-sided Wilcoxon test).

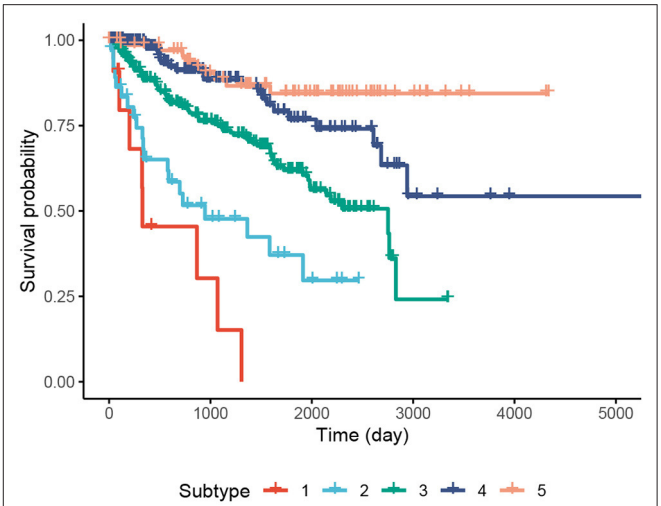

FIGURE 3 | Kaplan–Meier survival analysis of the Pan-kidney (KIPAN) dataset. The horizontal axis represents the time (day) while the vertical axis represents the estimated survival probability.

and finally (3) calculate the agreement between the discovered subtypes and known cancer stages/tumor grades using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). **Figure S2** and **Tables S1, S2** show the *p*-values obtained for gender and age. Overall, the four methods, SCFA, CC, SNF, and CIMLR, are not biased toward gender with only some significant *p*-values (**Table S1**). In contrast, iClusterBayes is subject to gender bias with significant *p*-values in 12 out of 25 datasets (**Table S1**). The *p*-values of iClusterBayes are significantly smaller than those of other methods ($p = 0.0007$ using one-sided Wilcoxon test). Regarding age, all methods have comparable *p*-values (**Table S2**). **Figure S3** and **Table S3** show the ARI values that represent the agreement between the discovered subtypes and known cancer stages and tumor grades. The median ARI of SCFA and SNF are comparable and they are higher than those of CC, iClusterBayes, and CIMLR. Regarding tumor grade, the ARI values of SCFA are higher than the rest. **Figure S4** and **Table S4** show the NMI values. SCFA has higher NMI values in both comparisons. However, the low NMI and ARI values show that there is a low agreement between the discovered subtypes and known stages/grades.

We perform an in-depth analysis for the Pan-kidney (KIPAN) dataset. For this dataset, SCFA discovers five subtypes, each with a very different survival probability (**Figure 3**). Subtype 1 has the lowest survival probability while Subtype 5 has the highest survival probability. All patients of Subtype 1 die within 3 years whereas 85% of patients in Subtype 5 survive at the end of the study (after 15 years). We also perform variant analysis to look for mutations that are highly abundant in the short-term survival groups (Subtypes 1, 2, and 3) but not in the long-term survival groups (Subtypes 4 and 5), and vice versa. In **Figure 4**, each point represents a gene and its coordinates represent the number of patients having at least a variant in that gene in each group. In principle, we would look for mutated genes in the top left and the bottom right corners. From this figure, we can identify four notable markers: VHL, PBRM1, MUC4, and FRG1B. Among these, MUC4 has been reported to be associated with exophytic growth of clear cell renal cell carcinoma (56) while VHL linked to a primary oncogenic driver in kidney cancers (57). PBRM1 is also a major clear cell renal cell carcinoma (ccRCC) gene (58). See **Supplementary Section 2** and **Figures S5–S8** for details.

3.2. Risk Score Prediction Using Multi-Omics Data

We also use the same set of data to demonstrate the ability of SCFA in predicting risk score of each patient. For each of the TCGA datasets, we randomly split the data into two equal sets of patients: a training set and a testing set. We use the training set to train the model and then predict the risk for patients in the testing set. The predicted risk scores are then compared with the true vital status and survival information using Cox *p*-value and concordance index (C-index) (59). Concordance index represents the probability that, for a pair of randomly chosen patients, the patient with higher predicted risk will experience death event before the other patient. On the other hand, Cox *p*-value measures how significant the difference in survival when correlating with predicted risk scores. This process is repeated 20 times for each dataset, and the average C-index and $-log10(p)$ for each dataset are calculated using results from these 20 runs. We note that some datasets do not have enough patients with either event (survive or death), which leads to errors for Cox regression. For that reason, we removed five datasets (DLBC, KIRP, TGCT,

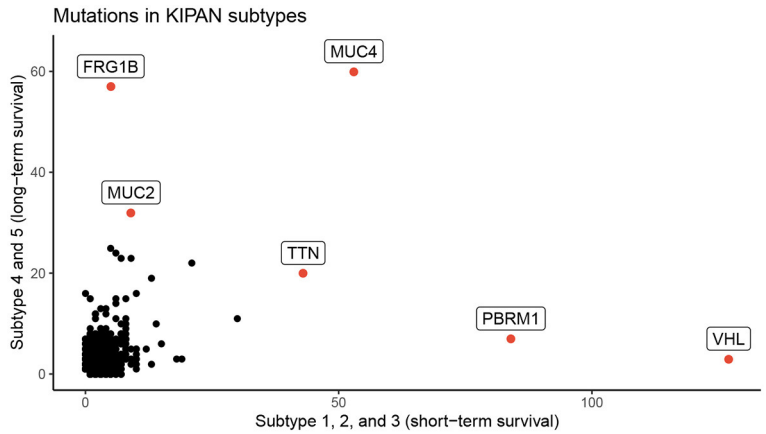

FIGURE 4 | Number of patients in each group for each mutated gene for KIPAN. The horizontal axis represents the count in subtypes with low survival rate (subtype 1, 2, and 3), while the vertical axis shows the count for subtypes with high survival (subtype 4 and 5) rate.

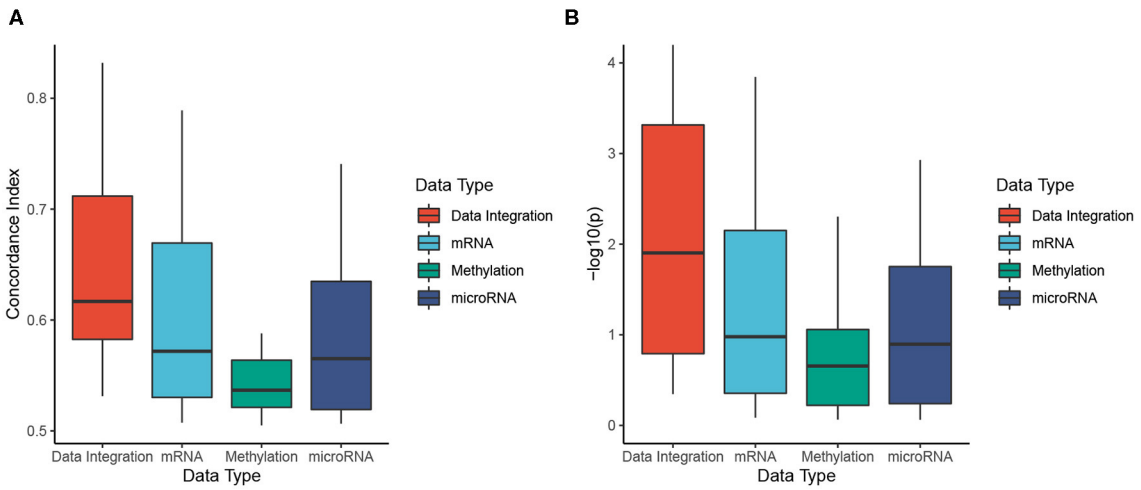

FIGURE 5 | Evaluation of risk prediction using concordance index (C-index) and Cox *p*-values. For each dataset, we calculate the C-index and Cox *p*-values between predicted risk scores and known survival of patients. To better understand the usefulness of data integration, we calculate the C-index and Cox *p*-value obtained from individual data types and compared them to those obtained from data integration. **(A)** Distributions of C-indices for data integration and individual data types. **(B)** Distributions of Cox *p*-values for data integration and individual data types. SCFA is able to predict risk scores that are highly correlated to true survival with a median C-index of 0.62 and Cox *p*-value of 0.01. In addition, the prediction is more accurate when all data types are analyzed together. The C-indices are significantly higher and the *p*-values are significantly smaller when all data types are combined ($p = 0.0007$ and $p = 0.002$ using one-sided Wilcoxon test).

THYM, UCEC) from the analysis, and report survival prediction for only 25 datasets without errors.

Figure 5 shows the distributions of C-indices and Cox *p*-values (in minus log10 scale), while **Table 2** shows the exact values calculated for each dataset. We calculate the C-index and Cox *p*-value obtained from individual data types and compared them to those obtained from data integration (when mRNA, miRNA, and methylation are analyzed together). As shown in **Figure 5A**, the accuracy of the prediction using data integration is generally higher than the accuracy obtained from individual data types. Predictions using data integration have a median C-index of 0.62, compared to 0.57, 0.54, and 0.57 when using mRNA, methylation, and miRNA, respectively. Similar results are also observed in the evaluation using Cox *p*-values (**Figure 5B**). The Cox *p*-values obtained from data integration has the median $-log10(p)$ of 1.9, compared to 1.0, 0.7, and 0.9 for mRNA, methylation, and miRNA. The results demonstrate that we can potentially predict the risk score of each patient using only molecular data. More importantly, the prediction using multi-omics data is generally more accurate than using individual data types.

4. CONCLUSION

In this article, we presented a novel method (SCFA) for disease subtyping and risk assessment using multi-omics data. The

TABLE 2 | Risk score prediction evaluated by concordance index (C-index) and Cox *p*-values.

Dataset	C-index				−*log*10(*p*)			
	Integration	mRNA	Methylation	microRNA	Integration	mRNA	Methylation	microRNA
ACC	0.78	0.79	0.59	0.72	3.32	3.84	0.66	2.73
BLCA	0.59	0.55	0.55	0.54	2.44	1.1	0.9	0.73
BRCA	0.62	0.55	0.52	0.51	1.38	0.77	0.28	0.14
CESC	0.68	0.63	0.54	0.64	3.42	2.15	1.4	2.02
CHOL	0.56	0.56	0.51	0.55	0.38	0.36	0.2	0.24
COAD	0.56	0.52	0.51	0.57	0.52	0.09	0.09	0.48
ESCA	0.53	0.52	0.5	0.51	0.35	0.09	0.18	0.06
GBM	0.55	0.51	0.53	0.53	2.44	0.3	1.04	1.12
GBMLGG	0.77	0.79	0.72	0.73	14.1	11.56	4.83	5.14
HNSC	0.59	0.59	0.51	0.55	1.41	1.81	0.22	0.48
KICH	0.68	0.6	0.63	0.57	1.35	0.62	2.3	1.31
KIPAN	0.79	0.77	0.73	0.74	24.42	14.53	11.65	20.54
KIRC	0.58	0.59	0.54	0.6	0.79	1.24	0.5	0.94
LAML	0.63	0.61	0.56	0.59	2.45	1.94	1.06	1.16
LGG	0.77	0.78	0.73	0.73	14.02	11.44	5.21	7.53
LIHC	0.62	0.53	0.55	0.57	1.9	0.36	0.86	0.9
MESO	0.72	0.69	0.53	0.63	4.46	3.72	0.22	2.93
OV	0.54	0.51	0.53	0.51	0.41	0.12	0.72	0.14
PAAD	0.71	0.67	0.56	0.59	3.35	2.58	0.79	1.75
SARC	0.62	0.57	0.53	0.53	1.19	0.98	0.19	0.26
SKCM	0.61	0.53	0.53	0.52	2.32	0.55	0.32	0.24
STES	0.54	0.51	0.52	0.51	0.4	0.11	0.29	0.16
THCA	0.66	0.53	0.54	0.51	1.26	0.44	0.33	0.57
UCS	0.58	0.53	0.51	0.51	0.68	0.15	0.06	0.08
UVM	0.83	0.67	0.69	0.72	2.62	1.14	2.87	1.33

contribution of SCFA is two-fold. First, it utilizes a robust dimension reduction procedure using autoencoder and factor analysis to retain only essential signals. Second, it allows researchers to predict risk scores of patients using multi-omics data—the attribute that is missing in current state-of-the-art subtyping methods.

To evaluate the developed method, we examined data obtained from 7,973 patients related to 30 cancer diseases downloaded from The Cancer Genome Atlas (TCGA). SCFA was compared against four state-of-the-art subtyping methods, CC, SNF, iClusterBayes, and CIMLR. We demonstrate that SCFA outperforms existing approaches in discovering novel subtypes with significantly different survival profiles. We also demonstrate that the method is capable of exploiting complementary signals available in different types of data in order to improve the subtypes. Indeed, the Cox *p*-values obtained from data integration are more significant than those obtained from individual data types.

To further demonstrate the usefulness of the developed method, we also performed a risk assessment using molecular data. We demonstrate that SCFA is able to predict risk scores that are highly correlated with vital status and survival probability. The correlation between predicted risk scores and survival information has a median of 0.62 and can be as high as 0.83. More importantly, we demonstrate that the risk prediction becomes more accurate when more data types are involved.

DATA AVAILABILITY STATEMENT

TCGA datasets were downloaded from http://firebrowse.org/. The docker contains the environment and scripts used in this article are available at: http://scfa.tinnguyen-lab.com. The current version of SCFA can be found at: https://github.com/duct317/SCFA. The SCFA software and TCGA data package will be available in the next release of Bioconductor.

AUTHOR CONTRIBUTIONS

DT and TN conceived of and designed the approach. DT implemented the method in R, performed the data analysis and computational experiments. HN and UL helped with data preparation and some data analysis. DT, HL, GB, and TN wrote the manuscript. All authors reviewed and approved the manuscript.

FUNDING

This work was partially supported by NASA under grant number 80NSSC19M0170, and by Research Enhancement Grant at UNR to TN. This work was also partially support by the UPMC Start-up Fund to HL. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.01052/full#supplementary-material

REFERENCES

1. Esserman LJ, Thompson IM, Reid B, Nelson P, Ransohoff DF, Welch HG, et al. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol*. (2014) 15:e234–42. doi: 10.1016/S1470-2045(13)70598-9
2. Esserman L, Shieh Y, Thompson I. Rethinking screening for breast cancer and prostate cancer. *J Am Med Assoc*. (2009) 302:1685–92. doi: 10.1001/jama.2009.1498
3. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clinic. (2009) 59:225–49. doi: 10.3322/caac.20006
4. Seidman H, Mushinski MH, Gelb SK, Silverberg E. Probabilities of eventually developing or dying of cancer–United States, 1985. *CA Cancer J Clinic*. (1985) 35:36–56. doi: 10.3322/canjclin.35.1.36
5. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res*. (2014) 3:242–9. doi: 10.3978/j.issn.2218-6751.2013.12.05
6. Booth CM, Shepherd FA. Adjuvant chemotherapy for resected non-small cell lung cancer. *J Thoracic Oncol*. (2006) 1:180–7. doi: 10.1097/01243894-200602000-00016
7. NSCLC Meta-analysis Collaborative Group. Preoperative chemotherapy for non-small-cell lung cancer: a systematic review and meta-analysis of individual participant data. Lancet. (2014) 383:1561–71. doi: 10.1016/S0140-6736(13)62159-5
8. Felip E, Rosell R, Maestre JA, Rodriguez-Paniagua JM, Moran T, Astudillo J, et al. Preoperative chemotherapy plus surgery versus surgery plus adjuvant chemotherapy versus surgery alone in early-stage non-small-cell lung cancer. *J Clin Oncol*. (2010) 28:3138–45. doi: 10.1200/JCO.2009.27.6204
9. Collisson EA, Bailey P, Chang DK, Biankin AV. Molecular subtypes of pancreatic cancer. *Nat Rev Gastroenterol Hepatol*. (2019) 16:207–20. doi: 10.1038/s41575-019-0109-y
10. Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat Rev Cancer*. (2017) 17:79–92. doi: 10.1038/nrc.2016.126
11. Yam C, Mani SA, Moulder SL. Targeting the molecular subtypes of triple negative breast cancer: understanding the diversity to progress the field. *Oncologist*. (2017) 22:1086–93. doi: 10.1634/theoncologist.2017-0095
12. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Investig*. (2011) 121:2750–67. doi: 10.1172/JCI45014
13. Kohonen T. The self-organizing map. *Proc IEEE*. (1990) 78:1464–80. doi: 10.1109/5.58325
14. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA*. (1999) 96:2907–12. doi: 10.1073/pnas.96.6.2907
15. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. (1999) 286:531–7. doi: 10.1126/science.286.5439.531
16. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*. (2001) 17:126–36. doi: 10.1093/bioinformatics/17.2.126
17. Luo F, Khan L, Bastani F, Yen IL, Zhou J. A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. *Bioinformatics*. (2004) 20:2605–17. doi: 10.1093/bioinformatics/bth292
18. McLachlan GJ, Bean R, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*. (2002) 18:413–22. doi: 10.1093/bioinformatics/18.3.413
19. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*. (2002) 18:275–86. doi: 10.1093/bioinformatics/18.2.275
20. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Trans Knowledge Data Eng*. (2004) 16:1370–86. doi: 10.1109/TKDE.2004.68
21. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA*. (2004) 101:4164–9. doi: 10.1073/pnas.0308531101
22. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*. (2005) 21:3970–5. doi: 10.1093/bioinformatics/bti653
23. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. In: *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. Vol. 8. La Jolla, CA (2000) p. 16.
24. Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. *Inform Process Lett*. (2000) 76:175–81. doi: 10.1016/S0020-0190(00)00142-3
25. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol*. (1999) 6:281–97. doi: 10.1089/106652799318274
26. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. (2010) 26:1572–3. doi: 10.1093/bioinformatics/btq170
27. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*. (2003) 52:91–118. doi: 10.1023/A:1023949509487
28. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol*. (2002) 3:1–21. doi: 10.1186/gb-2002-3-7-research0036
29. Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. In: *Pacific Symposium on Biocomputing*. Vol. 7 (2001). p. 6–17. doi: 10.1142/9789812799623_0002
30. Tseng GC, Wong WH. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*. (2005) 61:10–6. doi: 10.1111/j.0006-341X.2005.031032.x
31. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. (2012) 40:9379–91. doi: 10.1093/nar/gks725
32. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS ONE*. (2017) 12:e0176278. doi: 10.1371/journal.pone.0176278
33. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*. (2015) 16:1022. doi: 10.1186/s12864-015-2223-8
34. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics*. (2013) 29:2610–6. doi: 10.1093/bioinformatics/btt425

35. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. (2012) 28:3290–7. doi: 10.1093/bioinformatics/bts595
36. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*. (2017) 19:71–86. doi: 10.1093/biostatistics/kxx017
37. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA*. (2013) 110:4245–50. doi: 10.1073/pnas.1208949110
38. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. (2009) 25:2906–12. doi: 10.1093/bioinformatics/btp543
39. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE*. (2012) 7:e35236. doi: 10.1371/journal.pone.0035236
40. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. (2014) 11:333–37. doi: 10.1038/nmeth.2810
41. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple Kernel learning with application to cancer subtype discovery. *Bioinformatics*. (2015) 31:i268–75. doi: 10.1093/bioinformatics/btv244
42. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*. (2019) 35:3348–56. doi: 10.1093/bioinformatics/btz058
43. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat Commun*. (2018) 9:4453. doi: 10.1038/s41467-018-06921-8
44. Nguyen T, Tagett R, Diaz D, Draghici S. A novel approach for data integration and disease subtyping. *Genome Res*. (2017) 27:2025–39. doi: 10.1101/gr.215129.116
45. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*. (2019) 35:2843–6. doi: 10.1093/bioinformatics/bty1049
46. Von Luxburg U. A tutorial on spectral clustering. *Stat Comput*. (2007) 17:395–416. doi: 10.1007/s11222-007-9033-z
47. Bellman R. Dynamic Programming. Princeton, NJ: Princeton University Press (1957).
48. Harman HH, Jones WH. Factor analysis by minimizing residuals (minres). *Psychometrika*. (1966) 31:351–68. doi: 10.1007/BF02289468
49. Breslow NE. Analysis of survival data under the proportional hazards model. *Int Stat Rev*. (1975) 43:45–57. doi: 10.2307/1402659
50. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B*. (1972) 34:187–220. doi: 10.1111/j.2517-6161.1972.tb00899.x
51. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x
52. Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T. Fast and precise single-cell data analysis using hierarchical autoencoder. *bioRxiv*. (2019) 799817:1–10. doi: 10.1101/799817
53. Li CH. The Performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychol Methods*. (2016) 21:369–87. doi: 10.1037/met0000093
54. Wan S, Kim J, Won KJ. SHARP: Single-cell RNA-seq hyper-fast and accurate processing via ensemble random projection. *bioRxiv*. (2018) 461640:1–25. doi: 10.1101/461640
55. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. (2011) 39:1–3. doi: 10.18637/jss.v039.i05
56. Shinagare AB, Vikram R, Jaffe C, Akin O, Kirby J, Huang E, et al. Radiogenomics of clear cell renal cell carcinoma: preliminary findings of the cancer genome atlas-renal cell carcinoma (TCGA-RCC) imaging research group. *Abdominal Imaging*. (2015) 40:1684–92. doi: 10.1007/s00261-015-0386-z
57. Thomas GV, Tran C, Mellinghoff IK, Welsbie DS, Chan E, Fueger B, et al. Hypoxia-inducible factor determines sensitivity to inhibitors of mTOR in kidney cancer. *Nat Med*. (2006) 12:122–7. doi: 10.1038/nm1337
58. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*. (2011) 469:539–42. doi: 10.1038/nature09639
59. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *J Am Med Assoc*. (1982) 247:2543–6. doi: 10.1001/jama.1982.03320430047030

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 15 July 2020
doi: 10.3389/fams.2020.00025

A New Bayesian Methodology for Nonlinear Model Calibration in Computational Systems Biology

Fortunato Bianconi[1]*, *Lorenzo Tomassoni*[2†], *Chiara Antonini*[2†] and *Paolo Valigi*[2]

[1] Independent Researcher, Perugia, Italy, [2] Department of Engineering, University of Perugia, Perugia, Italy

Computational modeling is a common tool to quantitatively describe biological processes. However, most model parameters are usually unknown because they cannot be directly measured. Therefore, a key issue in Systems Biology is model calibration, i.e., estimate parameters from experimental data. Existing methodologies for parameter estimation are divided in two classes: frequentist and Bayesian methods. The first ones optimize a cost function while the second ones estimate the parameter posterior distribution through different sampling techniques. Here, we present an innovative Bayesian method, called Conditional Robust Calibration (CRC), for nonlinear model calibration and robustness analysis using omics data. CRC is an iterative algorithm based on the sampling of a proposal distribution and on the definition of multiple objective functions, one for each observable. CRC estimates the probability density function of parameters conditioned to the experimental measures and it performs a robustness analysis, quantifying how much each parameter influences the observables behavior. We apply CRC to three Ordinary Differential Equations (ODE) models to test its performances compared to the other state of the art approaches, namely Profile Likelihood (PL), Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC), and Delayed Rejection Adaptive Metropolis (DRAM). Compared with these methods, CRC finds a robust solution with a reduced computational cost. CRC is developed as a set of Matlab functions (version R2018), whose fundamental source code is freely available at https://github.com/fortunatobianconi/CRC.

Keywords: parameter estimation, ODE models, Bayesian algorithms, robustness analysis, model calibration, computational systems biology

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, United States

Reviewed by:
Takao Shimayoshi,
Kyushu University, Japan
Luca Martino,
Rey Juan Carlos University, Spain

***Correspondence:**
Fortunato Bianconi
fortunato.bianconi@gmail.com

These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *16 December 2019*
Accepted: *08 June 2020*
Published: *15 July 2020*

Citation:
Bianconi F, Tomassoni L, Antonini C and Valigi P (2020) A New Bayesian Methodology for Nonlinear Model Calibration in Computational Systems Biology. Front. Appl. Math. Stat. 6:25. doi: 10.3389/fams.2020.00025

1. INTRODUCTION

In recent years *omics* technologies have tremendously advanced allowing the identification and quantification of molecules at the DNA, RNA and protein level [1, 2]. These high-throughput experiments produce huge amounts of data which need to be managed and analyzed in order to extract useful information [3]. In this context, mathematical models play an important role since they process these data and simulate complex biological phenomena. The main purpose of mathematical modeling is to study cellular and extracellular biological processes from a quantitative point of view and highlight the dynamics of cellular components interactions [4]. Moreover, models represent an excellent tool to predict the value of biological parameters that may not be directly accessible through biological experiments because they would be time consuming, expensive or

not feasible [5]. One of the most common modeling techniques consists in representing a biological event, such as a signaling pathway, through a system of Ordinary Differential Equations (ODEs), which describes the dynamic behavior of state variables, i.e., the variation of species concentration in the system as a function of time [6]. Currently, the most used kinetic laws in ODE models can be divided into three types: the law of mass action, the Michaelis-Menten kinetic and the Hill function [7, 8]. However, these equations contain unknown parameters which have to be estimated in order to properly simulate the model and represent the problem under study. Typically, the calibration process of a model consists in the inference of parameters in order to make output variables as close as possible to the experimental dataset [9]. Hence, a calibrated model can be used to predict the time evolution of substances for which enough information or measures are not available. The most common methodologies for parameter estimation can be divided in two classes: the frequentist and the Bayesian approach [5, 10]. Frequentist methods aim at maximizing the likelihood function $f(\mathbf{y}|\mathbf{p})$, which is the probability density of observing the dataset $\mathbf{y}$ given parameter values $\mathbf{p}$ [11]. Under the hypothesis of independent additive Gaussian noise with constant and known variance for each measurement, the Maximum Likelihood Estimation (MLE) problem is equivalent to the minimization of an objective function, which compares simulated and experimental data [12, 13]. Common objective functions are the sum of squared residuals or the negative log-likelihood [14]. When the variances of measurement noise are not known, they are included in the likelihood as additional terms to estimate [13]. Different optimization algorithms are then employed to estimate the best parameter values. They implement global and/or local techniques and return in output the best fit between simulated and real data.

Since these optimization methods return only one solution for the parameter vector, i.e., the best fit, then parameter estimation is usually combined with identifiability analysis, in order to assess how much uncertainty there is in the parameter estimate [15]. Identifiability analysis is typically performed through the computation of confidence intervals for all estimated parameters. A confidence interval is the range where the true parameter value is located with a certain frequency. In this context, Profile Likelihood (PL) is a widely used data-based algorithm for structural and practical identifiability analysis [16].

On the other hand, the Bayesian approach considers parameters as random variables, whose joint posterior distribution $f_{\mathbf{P}|\mathbf{y}}(\mathbf{p})$ is estimated through the Bayes theorem: $f_{\mathbf{P}|\mathbf{y}}(\mathbf{p}) = \frac{f(\mathbf{y}|\mathbf{p})f_{\mathbf{P}}(\mathbf{p})}{f(\mathbf{y})}$, where $f(\mathbf{y})$ is the marginal density for $\mathbf{y}$ and $f_{\mathbf{P}}(\mathbf{p})$ is the prior probability density of parameters [17]. Through $f_{\mathbf{P}}(\mathbf{p})$ it is possible to include *a priori* beliefs about parameter values [18]. The joint posterior density automatically provides an indication of the uncertainty of the parameter inference and gives major insights about the robustness of the solution [19]. Since computing the posterior distribution analytically is usually not feasible, sampling based techniques are used to estimate it [12, 17]. Two classes of sampling methods widely used are the Markov chain Monte Carlo (MCMC) and the Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC) [12]. MCMC algorithms approximate the posterior distribution with a Markov chain, whose states are samples from the parameter space. Their major advantage is the ability to infer the posterior distribution which is known only up to a normalizing constant [20]. The ABC-SMC algorithms evaluate an approximation of the posterior distribution through a series of intermediate distributions, obtained by iteratively perturbing the parameter space. Each iteration selects only those parameters that give rise to a distance function under a predefined threshold [21].

In this paper, our main purpose is to introduce a new version of the standard ABC-SMC approach, called Conditional Robust Calibration (CRC), for parameter estimation of mathematical models.

As in all ABC-SMC methods, CRC is an iterative procedure based on the parameter space sampling and on the estimation of the probability density function (pdf) for each model parameter. However, it presents different aspects that differentiate it from the other existent methodologies. The distinctive features of CRC are: (i) a major control of the computational costs of the procedure, (ii) the definition of multiple objective functions, one for each output variable, (iii) the conditional robustness analysis (CRA) [22], in order to determine the influence of each model parameter on the observables.

We validate this new methodology on different ODE models of increasing complexity. Here we present the results of CRC applied to three models, two with *in silico* noisy data and one with experimental proteomic data. We also calibrate all the models using the PL approach, through the software Data2Dynamics (D2D) [23] and the standard ABC-SMC, through the ABC-SysBio software [19].

Moreover, in order to provide a reliable and complete comparison with the state of the art of this field, we also apply the Delayed Rejection Adaptive Metropolis (DRAM) algorithm through the MCMC toolbox [24, 25] to all the models presented in the Results section. DRAM combines two well-known strategies that are common in the MCMC schemes: adaptive Metropolis samplers and delayed rejection [24]. However, DRAM is not the only algorithm that improves the standard MCMC sampler. As shown in [26], many other similar and efficient methodologies are very popular in the literature. Since the underlying concept of these methodologies is the sampling from a proposal distribution, the common objective of this class of methods is to speed-up and improve the performance of the MCMC sampler. Some examples are the Multiple Try Metropolis (MTM) algorithm [27] and the Adaptive Gaussian Mixture Metropolis-Hastings [28, 29].

Our results show that CRC is successful in all examples and that its innovation features are critical when calibrating models with an high number of parameters and output variables. Especially when compared to ABC-SMC, CRC does not remain stuck in intermediate iterations because of its reduced computational cost and it is able to estimate parameter values of a model with an higher accuracy due to the definition of objective functions specific for each output variable. Finally, it

also introduces the concept of conditional robustness that is different from standard identifiability.

2. MATERIALS AND METHODS

2.1. ODE Model

Consider a deterministic ordinary nonlinear dynamical system:

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}), \quad \mathbf{x}(0) = x_0, \quad \mathbf{x} \in \mathbb{R}^n, \quad \boldsymbol{\theta} \in \Theta$$
$$\mathbf{y}(t) = h(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\gamma}), \quad \mathbf{y} \in \mathbb{R}^m, \quad \boldsymbol{\gamma} \in \mathbb{R}^s, \quad \mathbf{u} \in \mathbb{R}^g$$

where $\mathbf{x}$ is the state space vector, $\mathbf{u}$ denotes the external input vector, and $\mathbf{y}$ denotes the output responses of the system, i.e., the observables, which are usually derived from experimentally observed data. The vector $\boldsymbol{\theta}$ denotes the dynamical system parameters, taking values in the parameter space $\boldsymbol{\Theta}$, a subset of the positive orthant $\mathbb{R}^q_{>0}$. The vector function f is indeed defined over the following sets: $f(\cdot): \mathbb{R}^n \times \mathbb{R}^g \times \mathbb{R}^q \longrightarrow \mathbb{R}^n$. The observation function $h(\cdot): \mathbb{R}^n \times \mathbb{R}^g \times \mathbb{R}^s \longrightarrow \mathbb{R}^m$ maps the state variables to the vector of observable quantities $\mathbf{y}$. Usually, not all states of the system can be directly measured, so that it is common to have $m < n$. Vector γ contains scaling and offset parameters when measurements of the observables are performed. Setting $\mathbf{p} = \{\theta, \gamma, \mathbf{x}(0)\}$, $\mathbf{p} \in \mathbb{R}^{q+s+n}$, the model is completely determined. We assume that the parameter vector $\mathbf{p}$ is constant over time. We assume that data are collected at discrete time points $t_j \in [t_0, t_k]$ and thus the generic data measure can be written as:

$$y^*_{ij} = y_{ij} + \tau_{ij}, \quad i = 1, \ldots, m \tag{1}$$

where y_{ij} is the measurement and τ_{ij} is the unknown measurement error.

2.2. General Theory

The parameter vector of a mathematical model can be considered as a random variable $\mathbf{P}: \mathbb{P} \longrightarrow \mathbb{R}^{(q+s+n)}$, where $\mathbb{P}$ is the set of all possible vectors in the parameter space and $\mathbb{R}^{(q+s+n)}$ is the measurable space. Thus, a given vector $\mathbf{p}$ in the parameter space is one of the possible realizations of $\mathbf{P}$. Let $f_{\mathbf{P}}(\mathbf{p})$ denote the prior distribution. Our goal is to approximate the target posterior distribution, $f_{\mathbf{P}|\mathbf{y}}(\mathbf{p}) \propto f(\mathbf{y}|\mathbf{p}) f_{\mathbf{P}}(\mathbf{p})$, where $f(\mathbf{y}|\mathbf{p})$ is the likelihood density that describes the model. This is the pdf of the output variables of the model when parameters are distributed according to the prior $f_{\mathbf{P}}(\mathbf{p})$. To this purpose, we develop CRC, a variant of the ABC-SMC iterative algorithm. As it is well-established in this class of techniques, at the beginning of each iteration z, CRC samples parameters from a proposal distribution $q^z(\mathbf{p})$. However, differently to standard ABC approaches, we generate a fixed number N_S of samples from $q^z(\mathbf{p})$ along the different iterations. Then the fitting between observed and simulated data is measured through the computation of a distance function. CRC defines, at each iteration, multiple distance functions $d_{i,\mathbf{p}}$, each one associated with a single component of the output function, without the employment of any summary statistic. The distance functions defined above are the objective functions that CRC reduces iteration by iteration until a convergence criterion is met. Since parameter vector $\mathbf{P}$ is a random variable, $d_{i,\mathbf{p}}$ can be considered a realization of the random variable D_i that describes the distribution of the distance function, corresponding to the *i-th* output. The random variable D_i models the distribution of the error between simulated and real data when the parameter are sampled. Accordingly, at each iteration we define a set of thresholds ϵ^z_i that specify the maximum accepted level of agreement between each observable and the corresponding simulated data. At each iteration we obtain different parameter sets P_{S,ϵ^z_i}. Each set contains only those parameters that yield the values of a specific distance function under the corresponding threshold. Then, all these sets are intersected in order to obtain a single parameter set, $P_{S,\boldsymbol{\epsilon}^z}$, that ensures the compliance with all the thresholds. $P_{S,\boldsymbol{\epsilon}^z}$ contains samples of the approximate posterior distribution $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$. As for other ABC methods, if at the end of the z-th iteration a predefined stopping criterion is not satisfied, another iteration of CRC is performed, sampling from a new proposal distribution. Since the region of interest of the approximate posterior distribution is the one with highest probability, the proposal for the next iteration, $q^{z+1}(\mathbf{p})$, is centered on the mode of $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$. In order to increase the frequency of N_S samples in this region, the boundaries of $q^{z+1}(\mathbf{p})$ are tighter than those of $q^z(\mathbf{p})$. The algorithm terminates when the thresholds are sufficiently small. The output of CRC is $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^\zeta}}(\mathbf{p})$ where ζ is the number of the last iteration and its mode $\mathbf{p}^\zeta_m$ is the vector that reproduces the observed data with the highest probability.

2.3. CRC Algorithm

Figure 1 sums up the main steps of the CRC algorithm.

1. Sampling the parameter space and the posterior distribution. Generate a predefined number of samples N_S from the proposal distribution $q^z(\mathbf{p})$. If $z = 1$ the proposal distribution is the prior $f_{\mathbf{P}}(\mathbf{p})$. Let P_S be the set of parameter samples generated from $q^z(\mathbf{p})$. For each sample $\mathbf{p} \in P_S$, simulate the model in order to compute a dataset $\mathbf{y} = [y_i(t_j)]$, $i = 1, \ldots, m$ and $j = 0, \ldots, k$. This dataset contains the samples of the likelihood function.

2. Computation of the distance functions and pdf estimation. For each $\mathbf{p} \in P_S$, as many distance functions as observables are computed, i.e., $d_{i,\mathbf{p}}(y_i, y^*_i)$, $\forall i = 1, \ldots, m$, where y^*_i is the i-th variable of the experimental dataset and y_i is the sequence of simulated values over time for that variable. Then the associated densities are estimated using a kernel density approach. Let denote with $f_{D_i}(d_{i,\mathbf{p}})$ the pdf of D_i, where D_i is a transformation from the random variable $\mathbf{P}$ and whose realizations are given by $d_{i,\mathbf{p}}$.

3. Parameter sets identification. For each distance function we define a threshold $\epsilon^z_i \geq 0$ which is the quartile of level α of $f_{D_i}(d_{i,\mathbf{p}})$ and we obtain the following subset:

$$T(\epsilon^z_i, d_{i,\mathbf{p}}) = \{d_{i,\mathbf{p}} \leq \epsilon^z_i : \int_0^{\epsilon^z_i} f_{D_i}(d_{i,\mathbf{p}}) \mathrm{d}d_i = \alpha\}. \tag{2}$$

Therefore $T(\epsilon^z_i, d_{i,\mathbf{p}})$ induces the subset $P_{S,\epsilon^z_i} \in P_S$ defined as:

$$P_{S,\epsilon^z_i} = \{p \in P_S : d_{i,\mathbf{p}} \in T(\epsilon^z_i, d_{i,\mathbf{p}})\}. \tag{3}$$

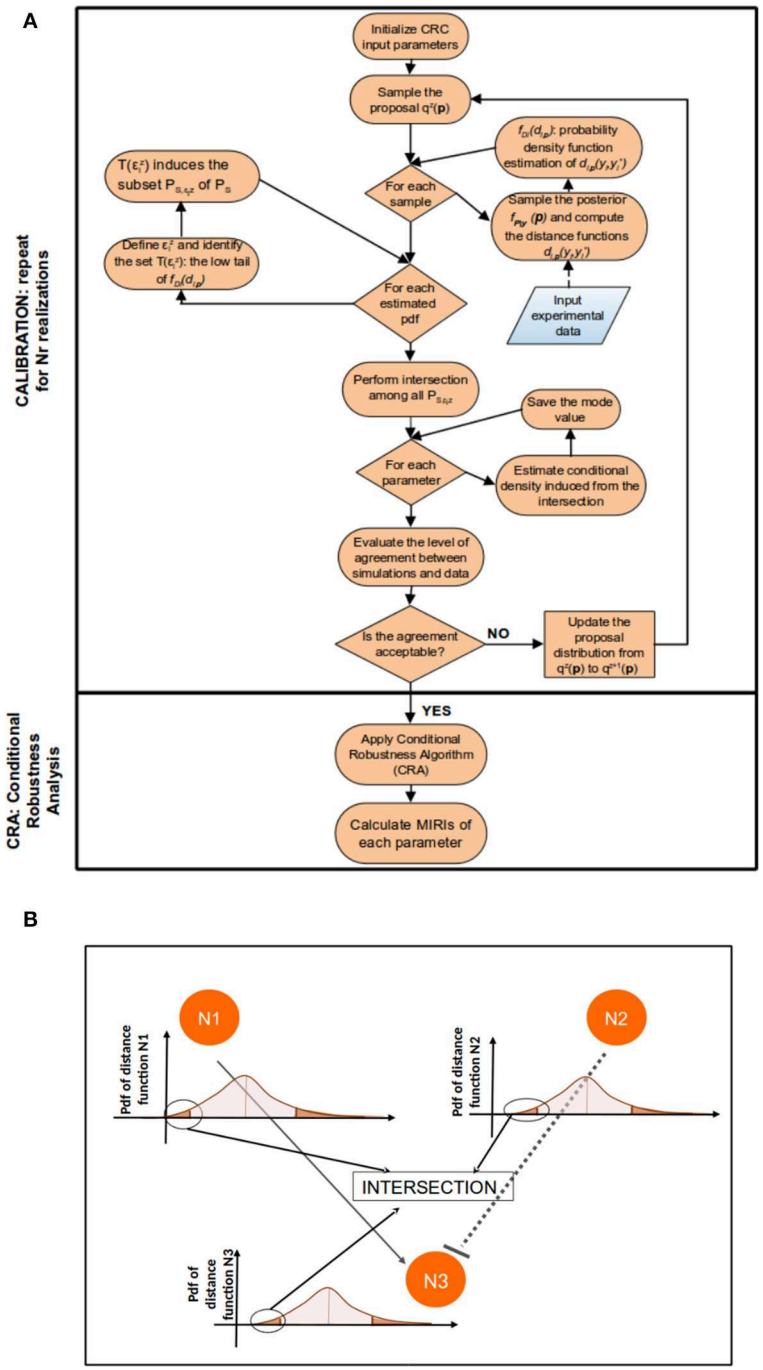

FIGURE 1 | CRC algorithm. **(A)** The flowchart of CRC is divided in two main phases: model calibration and robustness analysis. **(B)** Detail of steps 3–5 of CRC. N1, N2, and N3 represent three model observables that interact among each other. For each observable, a distance function $d_{i,\boldsymbol{p}}$ is computed and the corresponding pdf $f_{D_i}(d_{i,\boldsymbol{p}})$ is estimated. We are interested only in the low tail of each pdf, since it is the region where the model observables are closer to the experimental data. Then, all the low tails are intersected among each other to identify P_{S,ϵ^z}.

4. Intersection of Parameter Sets. Let denote $\boldsymbol{\epsilon}^z = \{\epsilon_1^z, \epsilon_2^z, \ldots, \epsilon_i^z, \ldots, \epsilon_m^z\}$ the set of thresholds corresponding to each observable at iteration z. We select the parameter samples that satisfy the conditions specified in the previous step simultaneously for all observables. This implies the definition of the following subset of P_S:

$$P_{S,\boldsymbol{\epsilon}^z} = \{\bigcap_{i=1}^{m} P_{S,\epsilon_i^z}\}. \tag{4}$$

Thus, in the parameter space, the accepted samples belong to the approximate posterior distribution $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$. **Figure 1B** clarifies the meaning of Equation (4): conceptually, in this step we identify and intersect the low tails of m distance function distributions, $f_{D_i}(d_{i,\mathbf{p}})$, in order to obtain, in the parameter space, the joint conditional density $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$. If the values of thresholds in $\boldsymbol{\epsilon}^z$ satisfy the stopping criterion, the algorithm terminates and the output is $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$. Otherwise go to step 6.

5. Update the proposal distribution. From $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$ we select the mode vector $\mathbf{p}_m^z$. The proposal distribution of the subsequent iteration, $q^{z+1}(\mathbf{p})$, is defined as:

$$q^{z+1}(\mathbf{p}) := q^z(\mathbf{p}; \mathbf{p}_m^z, U^{z+1}, L^{z+1}), \tag{5}$$

where $q^z(\mathbf{p}; \mathbf{p}_m^z, U^{z+1}, L^{z+1})$ is the proposal distribution at the current iteration z.

Note that while the shape of the proposal distribution does not change over the different iterations, the mean value and the upper and lower boundaries of the proposal do. The mean value is updated according to the mode $\mathbf{p}_m^z$ and the upper and lower boundaries, respectively U^{z+1} and L^{z+1}, are shrunk through the following formula:

$$U^{z+1} = \frac{U^z + k_{U,1}}{k_{U,2}}, \qquad L^{z+1} = \frac{L^z + k_{L,1}}{k_{L,2}} \tag{6}$$

where $k_{U,1}$, $k_{U,2}$, $k_{L,1}$, and $k_{L,2}$ are constants set by the user. They regulate the percentage variation of the lower and upper boundaries from the mode $\mathbf{p}_m^z$. Once the new proposal distribution is defined, restart from step 1.

In **Table 1**, the pseudocode associated to each step of the algorithm is presented. In the next section there is a detailed explanation of how the different tuning parameters have to be properly set.

TABLE 1 | CRC algorithm: pseudocode of a generic iteration z of the algorithm.

Inputs: ODE model, Experimental data

z:= Current iteration number
if z=1 **then**
  $q^z(\mathbf{p}) := f_P(\mathbf{p})$
else
  update the proposal distribution $\rightarrow q^z(\mathbf{p}) := q^{z-1}(\mathbf{p}; \mathbf{p}_m^{z-1}, U^z, L^z)$
end if
generate $P_S \rightarrow N_S$ samples from $q^z(\mathbf{p})$ in parameter space $\mathbb{P}$

for each $\mathbf{p}$ in P_S **do**
  integrate the model $\rightarrow \mathbf{y} = [y_i(t_j)]$, $i = 1, \ldots, m$ and $j = 0, \ldots, k$
  distance functions computation $\rightarrow d_{i,\mathbf{p}}(y_i, y_i^*), \forall i = 1, \ldots, m$
end for

for each y_i, $i = 1$ **to** m **do**
  density estimation of $D_i \rightarrow f_{D_i}(d_{i,\mathbf{p}})$
  threshold definition $\rightarrow \epsilon_i^z \geq 0$
  identification of $T(\epsilon_i^z, d_{i,\mathbf{p}}) \rightarrow \{d_{i,\mathbf{p}} \leq \epsilon_i^z : \int_0^{\epsilon_i^z} f_{D_i}(d_{i,\mathbf{p}})\mathrm{d}d_i = \alpha\}$
  identification of $P_{S,\epsilon_i^z} \rightarrow \{p \in P_S : d_{i,\mathbf{p}} \in T(\epsilon_i^z, d_{i,\mathbf{p}})\}$
end for

generate $P_{S,\boldsymbol{\epsilon}^z} \rightarrow \{\bigcap_{i=1}^m P_{S,\epsilon_i^z}\}$
joint conditional density estimation $\rightarrow f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$
select the mode vector of $f_{P|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p}) \rightarrow \mathbf{p}_m^z$

if stopping criterion is satisfied **then**
  terminate: $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$ and its mode $\mathbf{p}_m^z$ are the results found by the algorithm
else
  update U^z, L^z and the proposal distribution accordingly
  increase z and restart from the beginning
end if

2.3.1. Tuning Parameters

In this section, we discuss the setting of the tuning parameters of the proposed algorithm. First of all, it is necessary to choose a sampling technique for the parameter space. The objective is to generate a fixed number N_S of samples that optimally cover the entire parameter space defined by the proposal distribution. To this purpose, we use Latin Hypercube Sampling (LHS) because it divides the multidimensional parameter space in $N_S^{|\mathbf{p}|}$ regions and guarantees that each region is represented by a single sample [30, 31]. Since several studies indicate that log-transforming the parameters usually yields a better performance, we use as sampling schema the Logarithmic Latin Hypercube Sampling (LoLHS) [32]. A comparison of the results of CRC using LHS and LoLHS can be found in [33]. As for the number of samples N_S, it is fixed in advance taking into account the dimension of the parameter space and the number of observables of the model.

Then, at each iteration the choice of the threshold values strictly depends on two constraints. First of all, in order to approximate the posterior distribution, tolerances have to be set so that $\epsilon_i^z \leq \epsilon_i^{z-1}$. According to [22, 34], to generate a reliable non-parametric estimation of the conditional density, at least 1,000 samples of $P_{S,\boldsymbol{\epsilon}^z}$ are necessary. Thus, tolerances ϵ_i^z are chosen in order to ensure that the cardinality of $P_{S,\boldsymbol{\epsilon}^z}$ is at least 1,000.

Moreover, different combinations of tolerances may lead to the fulfillment of this condition. Therefore, as a guideline, threshold values can be chosen so that the number of accepted samples for each distance function $d_{i,\mathbf{p}}(y_i, y_i^*)$ is similar as much as possible for all output variables. As the number of iteration increases, thresholds progressively shift toward zero and, as in the standard ABC-SMC method [35], this guarantees that the approximate posterior distribution $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$ evolves toward the desired posterior distribution $f_{\mathbf{P}|\mathbf{y}}(\mathbf{p})$.

The constraints explained above regarding the choice of the thresholds are also influenced by N_S. For a given set of thresholds $\boldsymbol{\epsilon}^z$, the higher the value of N_S the higher is the cardinality of $P_{S,\boldsymbol{\epsilon}^z}$. Thus, increasing N_S, it is more likely to reach lower threshold values that satisfy $|P_{S,\boldsymbol{\epsilon}^z}| > 1,000$. However, N_S has also great impact on the computational cost of CRC and, for these reasons, its choice is a trade-off between the accuracy of the posterior estimation and the efficiency of CRC.

Another tuning parameter of CRC is the definition of the distance function that is used to measure the level of agreement between simulations and experimental data.

Here, we propose two different distance functions that can be used as objective functions when running CRC. The first one, called Absolute Distance Function (ADF), is the *l-norm* sum, over the whole time points set, of the distance between simulated and

real data. Equation (7) formalizes it:

$$ADF_i = \sum_{j=1}^{k} ||y_i(t_j) - y^*_{ij}||_l \quad i = 1, \ldots, m. \tag{7}$$

As an alternative, Equation (8) normalizes the error between simulated and real data with the corresponding point in the dataset. Moreover, the summation is divided by the number of available time points, thus Equation (8) represents the mean percentage error on each time point.

$$ANDF_i = \frac{1}{k} \sum_{j=1}^{k} \frac{||y_i(t_j) - y^*_{ij}||_l}{y^*_{ij}} \quad i = 1, \ldots, m. \tag{8}$$

In this paper we set $l = 1$, thus we use an l_1 norm, because it is robust to measurement noise, i.e., outlier-corrupted data [36]. The criterion that guides the user in the choice of the best distance function regards the range of variation of the available data. For a given absolute error (Equation 7), the corresponding percentage error (Equation 8) produces a bias. Thus, Equation (8) is preferred when the model and data are normalized.

Finally, as explained above, parameters $k_{U,1}$, $k_{U,2}$, $k_{L,1}$, and $k_{L,2}$ in Equation (6) determine the percentage shrinkage of the lower and upper boundaries of the proposal distribution. Iteration by iteration the percentage distance of the lower and upper boundaries of $f_{\mathbf{P}}(\mathbf{p})$ from the mode value is reduced.

In CRC the number of necessary iterations is not known a priori because the algorithm terminates when the stopping criterion is satisfied.

2.4. Conditional Robustness Analysis

The goal of the second phase of CRC is to perform a conditional robustness analysis (CRA) in order to identify which parameters mostly influence the output variables behavior [22]. To this purpose, we apply the conditional robustness algorithm in [22, 37]. Starting from $\mathbf{p}^{\zeta}_m$, i.e., the mode in the parameter space obtained in the last CRC iteration, we sample the parameter space using LHS and choosing as evaluation functions the distance functions $d_{i,\mathbf{p}}$ $\forall i = 1, \ldots, m$ previously defined during the calibration process. For each distance function, we define two thresholds $\eta_i^L \geq 0$ and $\eta_i^U \geq 0$ which are the quartiles of level β and λ respectively, obtaining the following subsets:

$$T(\eta_i^L, d_{i,\mathbf{p}}) = \{d_{i,\mathbf{p}} \leq \eta_i^L : \int_0^{\eta_i^L} f_{D_i}(d_{i,\mathbf{p}}) \mathrm{d}d_i = \beta\}, \tag{9}$$

$$T(\eta_i^U, d_{i,\mathbf{p}}) = \{d_{i,\mathbf{p}} \geq \eta_i^U : \int_{\eta_i^U}^{+\infty} f_{D_i}(d_{i,\mathbf{p}}) \mathrm{d}d_i = \lambda\}. \tag{10}$$

Thus, $T(\eta_i^L, d_{i,\mathbf{p}})$ and $T(\eta_i^U, d_{i,\mathbf{p}})$ contain values of the distance functions $d_{i,\mathbf{p}}$ that are, respectively smaller and larger than the defined thresholds η_i^L and η_i^U. Therefore, $T(\eta_i^L, d_{i,\mathbf{p}})$ and $T(\eta_i^U, d_{i,\mathbf{p}})$ induce two subsets in the parameter space P_S:

$$P_{S,\eta_i^L} = \{p \in P_S : d_{i,\mathbf{p}} \in T(\eta_i^L, d_{i,\mathbf{p}}))\}, \tag{11}$$

$$P_{S,\eta_i^U} = \{p \in P_S : d_{i,\mathbf{p}} \in T(\eta_i^U, d_{i,\mathbf{p}})\}. \tag{12}$$

The resulting conditioning sets are:

$$P_{S,\eta^L} = \{\bigcap_{i=1}^{m} P_{S,\eta_i^L}\}, \quad P_{S,\eta^U} = \{\bigcap_{i=1}^{m} P_{S,\eta_i^U}\}. \tag{13}$$

P_{S,η^L} and P_{S,η^U} define two regions in the parameter space, whose samples belong to the distributions $f_{\mathbf{P}|P_{S,\eta^L}}(\mathbf{p})$ and $f_{\mathbf{P}|P_{S,\eta^U}}(\mathbf{p})$.

The two conditional densities are employed in the calculation of the Moment Independent Robustness Indicator (MIRI) [22, 34] according to the following formula:

$$\boldsymbol{\mu} = \int |f_{\mathbf{P}|P_{S,\eta^L}}(\mathbf{p}) - f_{\mathbf{P}|P_{S,\eta^U}}(\mathbf{p})| \mathrm{d}\mathbf{p}. \tag{14}$$

Vector $\boldsymbol{\mu}$ contains MIRI values of all the components of parameter vector $\mathbf{p}$. Due to its definition, the MIRI value of each parameter is included in the interval $[0, 2]$. MIRIs measure the level of intersection between the two pdfs included in the calculus: the higher the resulting value and the more well separated are the conditional densities. Parameters with higher values of the MIRI have a major impact on output variables because there is a larger shift between the two conditional densities used for MIRI calculation. This means that different ranges of values for parameters with an high MIRI value lead the model observables to completely different behaviors. On the other hand, parameters with a low MIRI value do not affect the observables because their conditional densities overlap.

2.5. Benchmarking Algorithms

In this section, we give a brief overview of three methods against which CRC is compared: Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC), Profile Likelihood (PL), and Delayed Rejection Adaptive Metropolis (DRAM).

2.5.1. ABC-SMC

Since the CRC algorithm is a novel version of the standard ABC-SMC, in the current section we keep the mathematical notation consistent for those variables that have the same meaning.

ABC methods are a set of Bayesian methods for parameter estimation and model selection. They can be used to evaluate posterior distributions without having to calculate likelihoods. These methods are based on a comparison between observed and simulated data and thus are particularly useful when the likelihood function is too costly to evaluate [35]. ABC approaches usually proceed by: (i) sample a parameter vector $\mathbf{p}$ from a proposal distribution $q(\mathbf{p})$; (ii) simulate a dataset $\mathbf{y}(t)$ from the model having parameters $\mathbf{p}$; (iii) calculate a summary statistic $\mathbf{s}$ of the observables $\mathbf{y}$ of the model; (iv) calculate a distance function $d_{\mathbf{p}}(\mathbf{s},\mathbf{s}^*)$ between simulated and experimental data and accept $\mathbf{p}$ only if $d_{\mathbf{p}}(\mathbf{s}, \mathbf{s}^*) \leq \epsilon$ where ϵ is the tolerance level. Usually the distance function is defined on the full dataset. The simplest ABC algorithm is the ABC *rejection sampler* that implements the points described above in a single iteration, thus sampling the parameters only from the prior $f_{\mathbf{P}}(\mathbf{p})$ [38]. An improvement of the first version of ABC algorithm is ABC based on Markov chain Monte Carlo (ABC-MCMC algorithm), that exploits a Markov chain to explore the parameter space [39]. Conversely, ABC-SMC is a particular class of ABC algorithms based on SMC

sampling, which seeks to uncover an approximation of the true posterior distribution in a sequential manner through a series of intermediate distributions. It proceeds with the following steps:

- define a tolerance schedule $\epsilon^1 > \epsilon^2 > \ldots \epsilon^T \geq 0$;
- at the first iteration, sample parameter values, called particles, from a prior distribution, until N particles are obtained for which the distance is smaller than ϵ^1; for each particle is then computed a weight;
- for the further iterations, a particle is sampled from the previous population and perturbed with a perturbation kernel, until N accepted particles are selected. Weights are calculated for all accepted particles;
- the same procedure described above is repeated until N particles are selected in the last population.

The last population is the approximation of the posterior distribution $f_{\mathbf{P}|\mathbf{y}}(\mathbf{p})$. The ABC-SMC method is fully implemented in the ABC-SysBio software, a Python package for parameter estimation and model selection in the ABC framework [19]. A variant of the standard ABC-SMC algorithm is its population-based version, ABC-PMC [21]. The main feature of ABC-PMC algorithms is the adaptive calculation of the distance function and of the threshold at each iteration, based on the previous iteration's simulations.

2.5.2. Comparison Between ABC-SMC and CRC

Since CRC is a novel version of ABC-SMC, in this section we highlight the main differences between the two algorithms.

First of all, in the ABC-SMC algorithm, both the number of iterations and the threshold schedule have to be chosen in advance before running the whole procedure. Moreover, the total number of samples generated at each iteration is not known. The framework of the algorithm requires to specify the number of parameter samples, N, whose corresponding distance functions are below the threshold of the current iteration. Thus, for each iteration it is not known how many times the model will be integrated and, as the threshold is decreased, more samples are required to meet the constraint. The combination of these two aspects of the algorithm does not guarantee that the desired level of agreement between experimental and simulated data will be reached in a reasonable amount of time. In CRC, on the other hand, the approach is overturned. The number of samples N_S, that are generated at each iteration, is fixed and, on the contrary, the thresholds are not fixed in advance but they are chosen dynamically iteration by iteration. This guarantees that the time to perform each iteration is limited and at the end of each iteration a user defined stopping criterion is evaluated in order to decide whether or not to start again from the beginning.

Another relevant difference between the two algorithms is that CRC defines a distance function for each output variable without the employment of any summary statistic. ABC-SMC, on the contrary, defines a unique distance function regardless of the number of output variables of a model and it is often used with summary statistics. The effect of defining a single distance function has a great impact in the accuracy of model calibration especially when the model has an high number of parameters and/or output variables.

Finally, in the ABC-SMC framework, once a particle is sampled from a population it is perturbed with a kernel before simulating the model and if that particle is accepted a corresponding weight is computed (mathematical details are in [35]). As a consequence, in ABC-SMC, the way each population is updated iteration by iteration strictly depends on the kernel, that also influences the speed of convergence. In CRC, on the other hand, it is not used any perturbational kernel and at the end of each iteration, once the set $P_{S,\boldsymbol{\epsilon}^z}$ is identified, the empirical conditional density $f_{\mathbf{P}|P_{S,\boldsymbol{\epsilon}^z}}(\mathbf{p})$ is estimated. The proposal distribution of the next iteration is updated using the mode of the distribution cited above and, through Equation (6), the percentage of variation from the mode of each parameter is shrunk.

2.5.3. Profile Likelihood (PL)

PL is a framework for parameter estimation and uncertainty analysis that belongs to the frequentist class. As in the Bayesian approach, the starting point is the likelihood function that needs to be estimated. Assuming independent additive Gaussian noise with constant variance, maximizing the likelihood means minimizing the Residual Sum of Squares (RSS), commonly denoted as $\chi^2(\mathbf{p})$ where $\mathbf{p}$ is the estimated parameter vector. An optimization algorithm is commonly used to perform estimation. It can be deterministic, stochastic or hybrid [12]. In the class of stochastic algorithms, examples of commonly used methods are the so-called Genetic Algorithms (GA) and Simulated Annealing (SA). GA are iterative searching methods based on a natural selection process that mimics biological evolution. Starting from an initial population, at each iteration, a new population is generated until it evolves toward the optimal solution [40]. SA mimics the process of heating a material and then slowly lowering the temperature. It is a particular application of the Metropolis-Hastings algorithm. At each iteration a new point is generated and accepted according to an acceptance function, based on the temperature parameter [41]. Among deterministic algorithms, nonlinear least-squares (*lsqnonlin*) is considered one of the fastest and most reliable for common problems in Systems Biology [11, 13]. LHS can be used for setting initial parameters in a multi-start approach in order to avoid local optima as much as possible. For a performance analysis of optimization procedures see [11]. Once an optimal parameter set is obtained, it is important to assess the influence of all parameters on model behavior. The confidence interval $[\sigma_i^-, \sigma_i^+]$ of a parameter is a measure of its identifiability. It means that the true value of a parameter is located within this interval with a given probability α. To compute confidence intervals, PL is one the most employed algorithms. It is based on the following formula:

$$\chi^2_{PL}(p_i) = min_{j\neq i}[\chi^2(\mathbf{p})], \tag{15}$$

which means that, for each parameter p_i, the function χ^2 is reoptimized with respect to all parameters $p_{j\neq i}$. The process starts from the best fit and stops when the fit becomes unacceptable or a certain stopping criterion is met. At the end of it, a profile for each parameter is obtained. A parameter is declared structural non-identifiable if it has a flat profile, while it is practical non-identifiable if it has a minimum but the profile flattens out in one

of the directions of χ^2 (increasing or decreasing direction of p_i). In all the other cases, a parameter is considered identifiable [14].

The PL methodology works by transforming the parameter space from linear to logarithmic. Moreover, it has different tuning parameters that can be set by the user. It is possible to choose the lower and upper bounds for the parameters, the maximum number of steps along the profile for each parameter as well as the maximum and minimum step size. Moreover, in the likelihood function, the measurement noise is modeled as normal or log-normal distribution and it can be fitted simultaneously to the dynamical model.

2.5.4. DRAM

DRAM algorithm is an improved and modified version of the standard Metropolis-Hastings (MH) algorithm [24]. This algorithm belongs to the Markov Chain Monte Carlo (MCMC) methods, which approximate the posterior distribution of the parameter vector through a Markov chain [42]. One of the advantages of such methods is the possibility of sampling from an arbitrary pdf known up to a normalizing constant. DRAM is a strategy to combine the Delayed Rejection (DR) and Adaptive Metropolis (AM) estimators. When a candidate parameter sample is rejected, DRAM finds a new point exploiting also the information about the rejected one. In DR, given p_n as the current position of the chain, the first proposal move, $h_{n,1}$ is performed as in MH and thus it is accepted with probability:

$$\alpha_{n,1}(p_n, h_{n,1}) = min\left(1, \frac{f(h_{n,1}|\mathbf{y})q_{n,1}(h_{n,1}, p_n)}{f(p_n|\mathbf{y})q_{n,1}(p_n, h_{n,1})}\right), \quad (16)$$

where $q_{n,1}(\cdot)$ is the proposal distribution. The probability distribution $f(h_{n,1}|\mathbf{y})$ is computed as $f(h_{n,1}|\mathbf{y}) = f(\mathbf{y}|h_{n,1})\pi(h_{n,1})$, where $f(\mathbf{y}|h_{n,1})$ is the likelihood and $\pi(h_{n,1})$ is the prior function. In case of rejection of the proposed sample, the second move depends both on the current state of the chain and on the rejected sample. This expedient gives the name "Delayed" to the DR algorithm. The delayed mechanism can be iterated and interrupted at any stage. The mathematical details of the algorithm are presented in [43]. As in DR, also AM takes into account the history of the chain when the covariance matrix of the proposal distribution is updated. After an adaptation period n_0, AM assumes a Gaussian proposal distribution centered at the current state of the chain p_n and updates the covariance according to the following formula:

$$C_n = \begin{cases} C_0 & n \leq n_0 \\ s_d Cov(p_0, \ldots, p_{n-1}) + s_d \epsilon I_d & n \geq n_0, \end{cases} \quad (17)$$

where C_0 is the initial covariance, s_d depends on the dimension d of the parameter vector, ϵ is a constant chosen very small and I_d the d-dimensional identity matrix. In [24], the authors propose one of the ways to combine the two approaches explained above. DRAM includes AM in the DR framework as follows:

- at the first DR stage the proposal is adapted as in AM, where the covariance matrix C_n^1 is estimated from the collected samples of the chain;
- at the i-th stage $C_n^i = \gamma_i C_n^1$, where γ_i is a freely chosen scale factor.

TABLE 2 | Features of the models.

Model	Total parameters	Unknown parameters	States	Outputs	Data points
M1	4	2	2	2	8
M2	11	7	5	2	22
M3	93	53	40	16	96

To estimate the initial model parameters, DRAM performs the function *fminsearch*() when the model has a single observable. Otherwise, initial parameter values have to be set by the user by trial and error or according to the prior knowledge.

The initial estimate of the error variance needs to be set by the user. Then, it can be estimated as an extra model parameter, by setting a prior distribution for it. A convenient choice is the conjugate inverse chi-squared distribution, in case of Gaussian error model.

The combination of DR and AM improves the efficiency and the efficacy of both techniques reciprocally. On the one hand, AM guarantees an acceptable level of efficiency of DR even when there is not a good proposal distribution while, on the other hand, DR accelerates the adaptation process.

RESULTS

We test our novel proposed algorithm in three different models: Lotka-Volterra model (M1), EpoR system (M2), and signaling pathway of p38MAPK in multiple myeloma (MM) (M3). **Table 2** synthesizes the models features. M1 is characterized by an oscillatory behavior of both output variables, M2 is used in synthetic biology and contains initial conditions and scale factors to estimate while M3 is a high-dimensional model based on experimental proteomics data. All the simulations were performed on a Intel Core i7-4700HQ CPU, 2.40 GHz 8, 16 GB memory, Ubuntu 16.04 LTS (64 bit).

2.6. Lotka-Volterra Model (M1)

2.6.1. Model Description

The first model is the classical Lotka-Volterra model, which describes the interaction between the prey species x_1 and the predator species x_2 through the parameters a and b:

$$\begin{aligned} \dot{x}_1(a,t) &= ax_1(t) - x_1(t)x_2(t), \quad x_1(0) = 1, \\ \dot{x}_2(b,t) &= bx_1(t)x_2(t) - x_2(t), \quad x_2(0) = 0.5, \\ y(a,b,t) &= [x_1(t), x_2(t)]. \end{aligned} \quad (18)$$

The model parameter vector to estimate is $\mathbf{p} = [a, b]$, $\mathbf{p} \in \mathbb{R}^2$. Both nominal values of parameters are set to 1. The observables are both variables x_1 and x_2.

To calibrate the model, we generate an *in silico* noisy dataset in the same way described by [35], i.e., sampling eight values of the output variables at the same specified time points and adding Gaussian noise $\mathcal{N}(0, (0.5^2))$ (**Table S1**). Simulating the measurement noise with Gaussian noise is standard practice in mathematical modeling [5, 12, 35].

TABLE 3 | Tuning parameters and results of CRC in model M1.

Iteration (z)	L^z	U^z	$\mathbf{p}_m^z$	$\epsilon_{x_1}^z$	$\epsilon_{x_2}^z$	MSE_{x_1}	MSE_{x_2}
1	0.1	10	[0.37, 0.96]	7.3	6	0.71	0.54
2	0.55	5.5	[0.81, 1.26]	5.8	4	0.38	0.2
3	0.775	3.25	[0.84, 1.11]	5	3.4	0.35	0.19
4	0.8875	2.125	[0.98, 1.07]	4	3	0.22	0.16
5	0.9437	1.5625	[1.01, 1.07]	3.4	2.7	0.2	0.14
6	0.9718	1.2813	[1.06, 1.06]	3.1	2.7	0.19	0.13

The second and third column show, respectively, the lower and upper boundaries of $q^z(\mathbf{p})$. The fourth column reports the mode vector $\mathbf{p}_m^z$ computed at the end of each iteration and used to center the proposal distribution of the subsequent iteration. In the fifth and sixth columns the threshold schedule is reported and in the last two columns the resulting MSE for each observable is shown.

2.6.2. CRC Results

The tuning parameters of CRC are set as follows:

- the number of fixed samples in the parameter space is set to $N_S = 10^4$ for each iteration;
- Equation (7) $\forall i = 1, \ldots 2$ is chosen as distance function between experimental and simulated data;
- the prior distributions for a and b are taken to be log-uniform: $a,\ b \sim log - U(0.1, 10)$;
- $k_{U,1} = k_{L,1} = 1$ and $k_{U,2} = k_{L,2} = 2$ in all iterations.

According to Equation (7), the distance between the noisy dataset and the nominal solution is 3.09 for species x_1 and 2.6 for species x_2. Thus, the objective of the calibration is to obtain realizations of D_{x_1} and D_{x_2} that are close, respectively, to 3.09 and to 2.6. For this reason, the algorithm terminates when the two corresponding thresholds become sufficiently close to these two values. In order to do that, we perform six iterations of CRC. In the sixth iteration $\epsilon_{x_1}^6$ and $\epsilon_{x_2}^6$ are very close to the reference values presented above (3.09 and 2.6). This demonstrates that CRC reaches the desired level of agreement between simulated and experimental data. **Table 3** sums up the tuning parameters and the results of CRC along the six iterations.

Figures 2A,B display the time behavior of output variables x_1 and x_2 when parameters belong to the final subset P_{S,ϵ^6}. The model simulations are shown together with the noisy dataset, proving both the validity and robustness of the solution. **Figure 2C** shows how the subset of accepted particles P_{S,ϵ^z} changes during the execution of CRC.

CRC is fast in model calibration since the time simulation decreases from 288 s for the first iteration until 202 s for the last one. After model calibration, the CRA is applied in order to compute MIRIs for the parameters. We perturb the mode vector $\mathbf{p}_m^\zeta = [1.06, 1.06]$ generating 10^4 samples of the parameter space. The lower and upper boundaries of the sampling are fixed equal to 0.1 and 10 and the probabilities β and λ are both fixed to 0.1, following the guidelines in [22]. For both parameters, we obtain that MIRIs have values around 1, meaning that a and b have approximately the same influence on the two output variables (**Figure S3**).

2.6.3. PL Results

First of all, we estimate parameter values using three different optimization algorithms, available in the software D2D [13]: *lsqnonlin*, genetic algorithms (GA), and simulated annealing (SA). In this example, both the default *lsqnonlin* and GA correctly fit the model yielding the same results, while SA totally fails in parameter estimation. The default method *lsqnonlin* estimates parameter a equal to 1.07 and parameter b equal to 1.05. Using these parameter values, the MSE is equal, respectively, to 0.19 for x_1 and to 0.13 for x_2. In order to evaluate the identifiability of model parameters and to assess their confidence intervals, we also calculate the PL. All tuning parameters of the algorithm are left to their default values. PL estimates as identifiable both model parameters (**Figure S5**). PL employs few seconds for parameter estimation and less than a minute for identifiability analysis of both parameters.

2.6.4. ABC-SMC Results

The application of the standard ABC-SMC method to the M1 model and the results obtained are comprehensively explained in [35]. In five steps the procedure converges to the considered threshold. In the 5-th iteration, parameter a has a median of 1.05 and a 95% interquartile range of [1, 1.12] whiifble parameter b has a median of 1 and a 95% interquartile range of [0.87, 1.11] [35]. When parameters are equal to their median values, MSE_{x_1} is 0.28 and MSE_{x_2} is 0.19.

2.6.5. DRAM Results

We apply DRAM to the M1 model varying the initial parameter values in the interval [0, 10], the initial error variance in the interval [0.01, 1] and the corresponding prior weight in the interval [1, 5]. The error variance is then estimated by setting *options.updatesigma* = 1. Moreover, in order to perform a reliable comparison with CRC, we set the number of simulations equal to 10^4 and the boundaries of the parameters between [0, 10]. In all cases, the chains are stable and close to the nominal parameter values after one run. For the sake of brevity, we show only the results for initial error variance set to 0.1, prior weight set to 3 and initial parameter values equal to 0, 5, and 10. The time employed to apply the algorithm on the M1 model is about 3 minutes. Detailed results of DRAM are in **S1 File**.

2.7. EpoR System (M2)

2.7.1. Model Description

The ODE model presented in this section is taken from the Erythropoietin Receptor (EpoR) [44]. The model represents the catalysation of a substrate S by an enzyme E that is activated via two steps by an external ligand L [45]. This reaction cascade produces a product P whose dynamical behavior is the purpose of the model prediction. Generally the concentration over the time of the product P cannot be measured directly. Let denote with $\mathbf{p} = [k_1, k_2, k_3, init_E, init_S, scale_E, scale_S], \mathbf{p} \in \mathbb{R}^7$, the set of parameter to estimate. The nominal values of model parameters are $\mathbf{p} = [0.1, 0.1, 0.1, 10, 5, 4, 2]$. The equations of the model, the corresponding initial conditions and the observables together with the dataset used for model calibration are reported in **S1 File**.

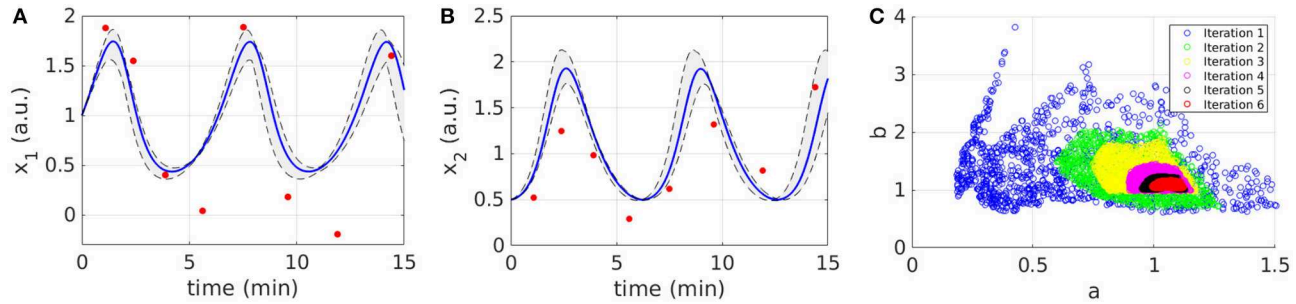

FIGURE 2 | Results of CRC for model M1. **(A,B)** The red line is the time behavior of output variables when the parameter vector is equal to $\mathbf{p}_m^6$ (see **Table 3**); red dots are the experimental data; the gray area represents the uncertainty in the temporal behavior of observables when parameters vary between the 2.5th and 97.5th percentile of their corresponding conditional pdfs. **(C)** Scatter plot of the sample distribution of $f_{a|P_{S,\epsilon^z}(a)}$ (x-axis) vs. $f_{b|P_{S,\epsilon^z}(b)}$ (y-axis) in all iterations.

TABLE 4 | CRC parameters for model M2.

Iteration (z)	L^z	U^z	ϵ_1^z	ϵ_2^z
1	0.01	100	51	31
2	0.02	50	50	31
3	0.04	25	46	30.9
4	0.08	12.5	40	30.3
5	0.16	6.25	33.5	29.1
6	0.32	3.125	25	27.5
7	0.64	1.5625	15.5	22
8	0.82	1.2813	13.4	10
9	0.91	1.1406	13	5.75

The first column reports the iteration number z, the second and the third ones the boundaries of the proposal distribution $q^z(\mathbf{p})$ in each iteration and the fourth and the fifth ones the values of the two thresholds ϵ_1^z and ϵ_2^z.

2.7.2. CRC Results

The prior distributions for all the model parameters are supposed log-uniform with the lower and upper boundaries set equal to $L^1 = 0.01$ and $U^1 = 100$. The number of fixed samples in the parameter space is $N_S = 10^5$. We choose Equation (7) as distance function to evaluate the error between nominal and noisy data for the outputs of the model. According to the selected distance function, the errors between the nominal data points and the experimental ones are equal to 12.78 for y_1 and 5.6 for y_2. They represent the target thresholds to reach at the end of the last iteration in order to assert the success of CRC. To this purpose, we perform nine iterations of CRC in order to make the two thresholds close enough to their corresponding target values. **Table 4** shows the boundaries of the proposal distribution in each iteration. These values are obtained by setting $k_{U,1} = k_{L,1} = 0$, $k_{U,2} = 2$, and $k_{L,2} = 0.5$ for the first seven iterations and $k_{U,1} = k_{L,1} = 1$ and $k_{U,2} = k_{L,2} = 2$ for the eighth and ninth iterations. **Table 4** also shows the obtained thresholds for each performed iteration of CRC.

In the ninth iteration, ϵ_1^9 and ϵ_2^9 are very similar to the target values presented above. This proves that CRC estimates a parameter vector that guarantees the desired level of agreement between simulated and experimental data. The mode vector in output from CRC is $\mathbf{p}_m^\zeta = [0.11, 0.02, 0.08, 34.93, 2.99, 1.12, 3.37]$. The model simulation using as parameter vector the mode $\mathbf{p}_m^\zeta$ has an MSE of 18.24 and 0.32 for y_1 and y_2 respectively. **Figures 3A,B** show the time behavior of both output variables when the parameter vector is set equal to the mode $\mathbf{p}_m^\zeta$.

Moreover, regions in gray are the confidence bands of observables when parameter values are chosen between the 2.5-th and 97.5-th percentile of their corresponding conditional pdfs $f_{\mathbf{p}|P_{S,\epsilon^9}}(\mathbf{p})$. In **S1 File**, additional details of $f_{D_{y_1}}(d_{y_1,\mathbf{p}})$ and $f_{D_{y_2}}(d_{y_2,\mathbf{p}})$ and the estimated conditional pdfs of parameters are reported. CRC is quite fast since it employs about 8 min (527 s) to complete one iteration.

Once the model has been correctly calibrated, we perform a robustness analysis in order to find those parameters that most affect the behavior of the output variables. We perturb the mode vector $\mathbf{p}_m^\zeta$ with Linear LHS using 10^5 samples. The lower and upper boundaries of the sampling are fixed equal to 0.01 and 100 respectively. Using the guidelines reported in [22] we fix the level of probabilities β and λ to 0.1. In **Figure 3C** the resulting MIRIs are shown. MIRIs corresponding to initial conditions parameters and scale factors are close to their maximum value and are much higher than those of the kinetic ones. This means that initial conditions and scale factors have major impact on observables compared to the kinetic parameters. We repeat the entire procedure ten times, obtaining ten independent realizations in order to ensure the invariance of results.

2.7.3. PL Results

The calculation of PL for M2 is presented in [45] where the authors reported that PL takes < 1 min per parameter on a 1.8 GHz dual core machine. The parameter vector estimated through the PL is $\hat{p} = [0.087, 0.019, 0.37, 10.05, 4.97, 4.027, 2.1]$. The MSE obtained through the PL approach is 10.06 and 0.3 for y_1 and y_2, respectively. As regards the identifiability analysis, according to the PL approach, parameter k_2 is classified as structurally non-identifiable, parameter k_3 is practically non-identifiable and the others are assessed to be identifiable. More details of the PL results are provided in **S1 File**.

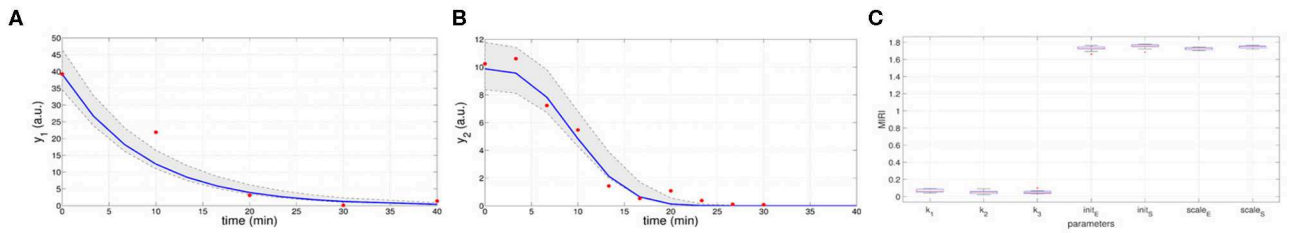

FIGURE 3 | Results of CRC for model M2. **(A,B)** Time behavior of output variables. Red dots are the noisy experimental data [45]; blue lines are the simulations of the observables of the model when the parameter vector is set equal to the mode $\mathbf{p}_m^c$; gray regions are the confidence bands when parameter values are chosen between the 2.5-th and 97.5-th percentile of their corresponding conditional pdfs $f_{\mathbf{P}|P_{S,\epsilon^9}}(\mathbf{p})$. **(C)** Boxplots of the MIRIs in output from the CRA for the seven model parameters and all the 10 independent realizations.

2.7.4. ABC-SMC Results

ABC-SMC input parameters are set in order to resemble those of CRC. The distance function is defined as:

$$d((y_1, y_2), (y_1^*, y_2^*)) = \sum_{i=1}^{2} \sum_{j=1}^{11} |y_i(t_j) - y_{ij}^*| = \sum_{i=1}^{2} ADF_i. \quad (19)$$

Under the hypothesis of parameters having a prior uniform distribution in [0, 100], we try to perform nine iterations of ABC-SMC. Thus, we set $f_{\mathbf{P}}(\mathbf{p}) = U(0, 100)$ and $z = 1, \ldots, 9$. The thresholds for all the iterations are chosen as the sum, over y_1 and y_2, of the two corresponding thresholds obtained from the application of CRC (**Table 4**). At each iteration, we select 1,000 particles of the parameter space under the desired threshold. The algorithm could not come to an end in a reasonable time and it finds parameter samples only until the 7-th iteration. In order to quantify the precision of the ABC-SMC approach, we perform simulations of the model, setting the parameter vector equal to the median of the ABC-SMC results. The MSE using these parameter values is 103 and 20 for y_1 and y_2, respectively. Further results of ABC-SMC are shown in **S1 File**.

2.7.5. DRAM Results

As in model M1, we run DRAM varying the initial error variance between [0.001,1] and the corresponding prior weight between [1, 5], and then we sample and estimate the error variance. Initial conditions were chosen close to nominal parameter values, i.e., [0.5, 0.5, 0.5, 8, 7, 5, 3]. As in CRC, the number of simulations is set to 10^5, the number of run performed is nine and the parameter boundaries are [0, 100]. We run DRAM using, in one case, a Gaussian prior for all parameters and, in the other case, a lognormal prior. DRAM results are fairly stable against the different values of the initial error variance and the prior distribution. The chains of the first three parameters, k_1, k_2, and k_3, equally span all the values in the interval [0, 100] and accordingly the pdfs for those parameters are almost uniformly distributed, meaning that this parameters are classified as non identifiable. On the other hand, the chains corresponding to the other model parameters converge and are stable around specific values. This is also clearer from the corresponding pdfs that show a peak around the estimated value. Here, we report the results obtained in the final run with a lognormal prior, a prior weight set to 1 and an initial error variance that varies between 0.001 and 1. DRAM employs about 15 min to complete one run. In **S1 File**, figures of DRAM results are provided.

2.8. Multiple Myeloma Model (M3)

2.8.1. Model Description

M3 is the ODE model proposed in [46]. The mathematical model is defined to help study the roles that various p38 MAPK isoforms play in MM. It has 40 ODEs, built using only the law of mass action, and 53 kinetic parameters (**S1 File**). According to Peng et al. [46], data associated to the model are the output of a Reverse Phase Protein Array (RPPA) experiment [2], where MM cell lines were analyzed to detect the activity of proteins active in various p38 MAPK signaling pathways. RPPA was performed on the following cell lines: four different RPMI 8226 MM cell sublines with stable silenced expression of $p38\alpha$, $p38\beta$, $p38\gamma$, and $p38\delta$, respectively, as well as an RPMI 8226 MM stable cell subline transfected with empty vector as the negative control. Cells were treated with arsenic trioxide (ATO), bortezomib (BZM) or their combination. RPPA analyzed a total of 80 samples and measured 153 proteins at six different time points. The proteins whose phosporylation level was included both in the pathway and in the experiment are 16. All RPPA data are normalized based on the initial concentration value. For this reason, initial conditions of proteins in the ODE model are all set to 1, i.e., $\mathbf{x}(0)=\mathbf{1}$. Parameters to estimate are only the kinetic ones, i.e., $\mathbf{p} \in \mathbb{R}^{53}$. In [46], available data for model calibration belongs to the $p38\delta$ knockdown cell line treated with BZM. The corresponding RPPA dataset is presented in **Table S13**.

2.8.2. CRC Results

For model M3, we set tuning parameters of CRC in the following way:

- the number of parameter samples N_S is equal to 10^6;
- for each output variable, Equation (8) is chosen as distance function;
- the prior of each kinetic parameter is supposed to be $log - U(0.1, 10)$;
- $k_{U,1} = k_{L,1} = 1$ and $k_{U,2} = k_{L,2} = 2$ in all iterations.

Six iterations of CRC are performed, using the threshold schedule reported in **Table 5**. At the end of the process, the maximum error between simulated and experimental data is only of the 16%.

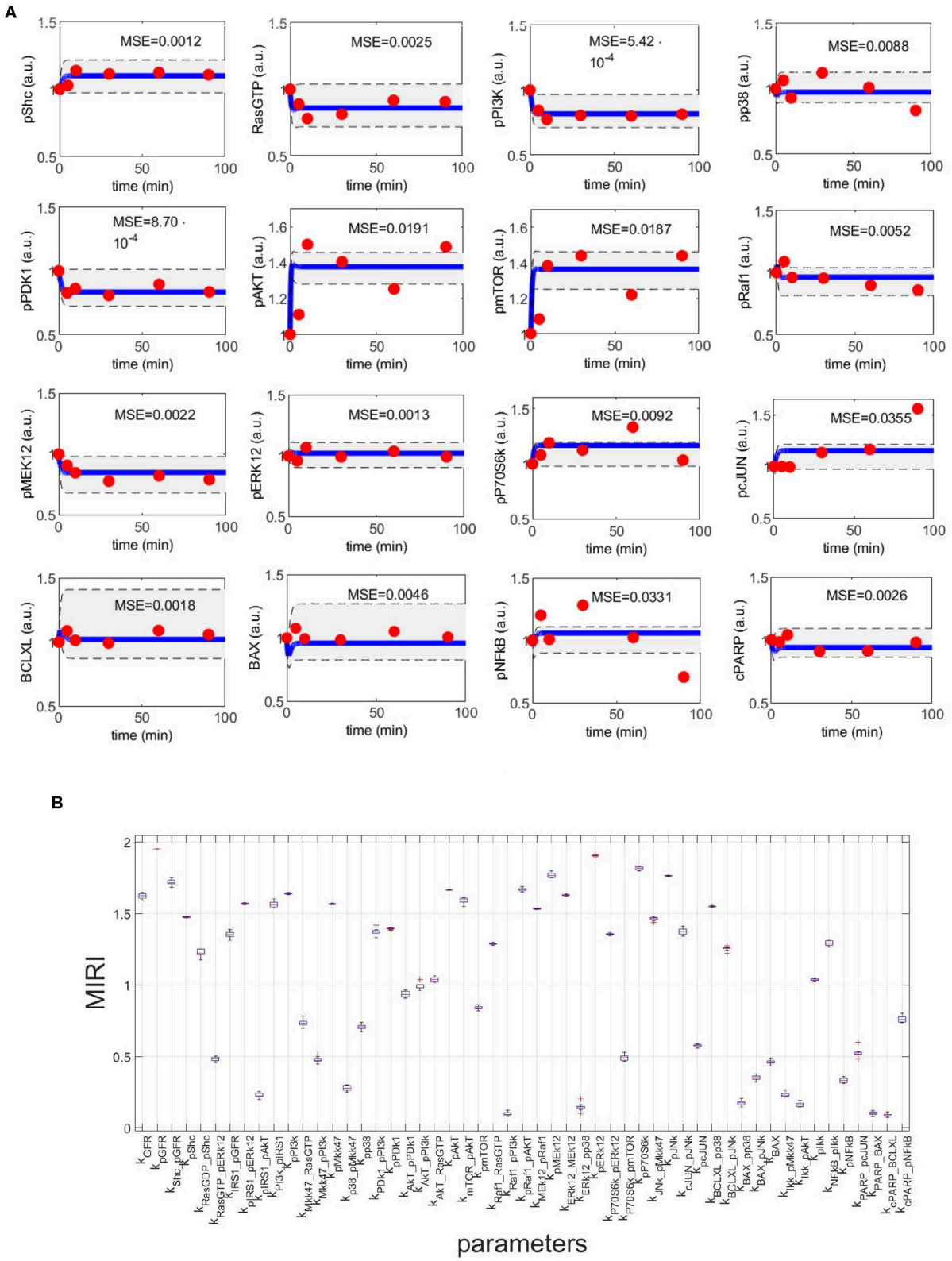

FIGURE 4 | CRC results for model M3. **(A)** The blue line is the time behavior of observables when the parameter vector is equal to $\mathbf{p}_m^6$ (see **Table S25**); red dots are the RPPA data [46]; the gray area reproduces the variation of the temporal behavior when parameters vary between the 2.5-th and 97.5-th percentile of their corresponding conditional pdfs $f_{\mathbf{P}|P_{S,\epsilon^6}}(\mathbf{p})$. On the top of each plot, the MSE between the associated model simulation and the data is reported. **(B)** Boxplot of MIRIs for the estimated parameter values, along the 10 independent realizations performed.

TABLE 5 | Threshold schedule of distance functions for model M3.

	Iteration (z)					
	1	2	3	4	5	6
RasGTP	0.7	0.4	0.3	0.2	0.1	0.08
pPI3K	0.7	0.4	0.3	0.2	0.1	0.06
pp38	0.7	0.4	0.3	0.15	0.13	0.09
pPDK1	0.8	0.5	0.3	0.2	0.1	0.06
pAKT	0.3	0.2	0.15	0.12	0.095	0.088
pmTOR	0.5	0.3	0.2	0.13	0.1	0.088
pRaf1	0.7	0.4	0.2	0.15	0.1	0.08
pMEK12	0.7	0.4	0.3	0.2	0.1	0.08
pERK12	0.6	0.4	0.2	0.1	0.05	0.04
pP70S6k	0.5	0.3	0.2	0.15	0.1	0.08
pcJUN	0.6	0.3	0.2	0.15	0.12	0.12
BCLXL	1	0.7	0.5	0.3	0.2	0.15
BAX	0.8	0.6	0.5	0.35	0.2	0.15
pNFkB	0.6	0.4	0.25	0.2	0.18	0.16
cPARP	0.7	0.4	0.3	0.2	0.15	0.12

The distance functions are $d_{i,\mathbf{p}}(y_i, y_i^*)$, $\forall i = 1, \ldots, 16$.

Figure 4A shows the time behavior of output variables at the end of the calibration, along with experimental data points. The figure proves that the algorithm is successful in finding a robust solution (**S1 File**). As regards the application of CRA, we perturb the parameter space in the interval having as lower and upper boundaries 0.01 and 100, respectively and centered around the final mode vector $\mathbf{p}_m^6$, generating 10^6 parameter samples. In **Figure 4B**, MIRIs are presented for each parameter. As expected, not all parameters have a strong impact on the observables since the number of output variables is small compared to all kinetic parameters. For instance, parameters k_{Raf1_pPI3K}, k_{ERK12_pp38}, k_{BAX_pp38}, k_{IKK_pAKT}, k_{PARP_BAX}, k_{cPARP_BCLXL} have all a MIRI value under 0.2. On the other hand, about a fifth of the total number of parameters has MIRI values above 1.6 (e.g., parameters k_{GFR}, k_{pGFR}, k_{Shc_pGFR}, k_{pPI3K}, k_{pAKT}, k_{pRaf1_pAkt}, k_{pMEK12}, k_{ERK12_MEK12}, k_{pERK12}, $k_{pP70S6K}$, k_{pJNK}), meaning that they have high influence on the outputs. We repeat the entire procedure ten times, obtaining ten independent realizations in order to ensure the invariance of results.

2.8.3. PL Results

First of all, through the D2D software, we calibrate the model using *lsqnonlin* as optimization algorithm. To avoid local optima, we execute a sequence of $n = 100$ fits using LHS. Moreover, we also estimate model parameters using GA and SA. While GA is not able to correctly reproduce experimental data, SA successfully estimates the time behavior of output variables. To compute confidence intervals, the maximum number of sampling steps in both the increasing and decreasing direction of each parameter is set to 200 while all the other tuning parameters of the method are set to their default values. The upper and lower boundaries of the parameter prior are set respectively to 10^{-5} and 10^3. Results of parameter estimation and identifiability analysis are shown in **S1 File**. The algorithm assesses that all parameters are identifiable except for the following parameters: k_{P70S6K_pERK12}, k_{cPARP_BCLXL}, k_{pJNK} that are practically non-identifiable and k_{pIRS1_pAKT} is structurally non-identifiable. Nevertheless, the results obtained are not so reliable since some parameters have a confidence interval of only a single value (e.g., k_{IRS1_pGFR}) while others have an estimated value outside the corresponding confidence region (e.g., k_{PDK1_pPI3K}). The PL algorithm employs less than one minute for parameter estimation and about 40 min for identifiability analysis of all parameters.

2.8.4. ABC-SMC Results

Using the ABC-SysBio software, we fix ABC-SMC parameters as follows:

- the distance function is defined as:

$$d(y, y^*) = \frac{1}{6}\sum_{i=1}^{16}\sum_{j=1}^{6}\frac{|y_i(t_j) - y_{ij}^*|}{y_{ij}^*} = \sum_{i=1}^{16} ANDF_i; \qquad (20)$$

- the number of iterations is set to 6;
- the threshold fixed in each iteration is equal to the sum of all thresholds fixed in CRC in the same iteration (see **Table S29**);
- the number of accepted particles at each iteration is set to 1,000;
- all parameters to estimate are supposed to have a uniform prior distribution: U(0.1,10);
- all the other parameters are left to their default values.

The application of ABC-SMC to the M3 model was very time consuming and, after 10 days, it had not converged yet. Thus, results are available only until the 4-th iteration (see **S1 File**).

2.8.5. DRAM Results

As in the previous examples, we run DRAM setting the number of simulations equal to those of CRC (10^6) and the number of run to six. The initial error variance, that is estimated, and the corresponding prior weight are set equal to 0.1 and 10, respectively. Initial values of parameters are all equal to 1 and the parameter boundaries are [0, 10]. We run DRAM using both a uniform and lognormal prior for all parameters. DRAM employs about 3 h to complete one run and, in the end, most parameters have a uniform distribution. In **S1 File**, we provide figures of DRAM results after six run with a lognormal prior. We show the chains and pdfs of the parameters and the time simulations of the output variables of the model when parameters are equal to mean of the corresponding chains.

3. DISCUSSION

Here we present a novel Bayesian approach for parameter estimation of mathematical models that is used to fit *omics* data in Systems Biology applications. The availability of high-throughput data with the need to calibrate high dimensional models using computational feasible algorithms, makes CRC a useful and innovative procedure in the overview of the Bayesian parameter estimation and robustness analysis. Our algorithm modifies the standard ABC-SMC in order to increase the efficiency and the reliability of the estimated parameter vector. Moreover, CRC

TABLE 6 | Comparison between parameter estimation algorithms.

	CRC	ABC-SMC	PL	DRAM
Class	Bayesian	Bayesian	Frequentist	Bayesian
Output	Mode of the final approximate	Median of the final	Optimal	Mean values
parameter vector	posterior distribution	population	parameter set	of the MCMC chains
Uncertainty	Posterior distribution	Histogram	Confidence interval	Posterior distribution
Prior information	Prior distribution	Prior distribution	None	Initial parameter vector
Objective function	One for each observable	One for all observables	One for all observables	One for all observables
Robustness analysis	Yes	No	No	No
Models successfully calibrated	M1, M2, and M3	M1	M1 and M2	M1
Computational cost	M1: $\sim$ 5 min (per iteration)	M1: $\sim$ 5 min	M1: $<$ 1 min	M1: $\sim$ 3 min (per run)
	M2: $\sim$ 8 min (per iteration)	M2: Not converged	M2: $<$ 5 min	M2: $\sim$ 15 min (per run)
	M3: $\sim$ 70 min (per iteration)	M3: Not converged	M3: $\sim$ 40 min	M3: $\sim$ 180 min (per run)

presents many distinctive improvements as compared to other algorithms of the ABC-SMC family, such as ABC-PMC and Adaptive-ABC [21].

We validated this new methodology in three ODE models, each one with specific features, in order to demonstrate the flexibility and reliability of our approach. In addition, we compared CRC results with those obtained by methods representing the state of the art of this field, i.e., the standard ABC-SMC, PL, and DRAM. **Table 6** summarizes the comparison between CRC and the other benchmarking algorithms tested.

First of all, we tested all the calibration procedures in the Lotka-Volterra model. We showed that all algorithms performed well, but with some differences. CRC returns a reliable and robust solution. Compared with ABC-SMC, we performed one more iteration but the computational burden was almost irrelevant since each iteration took about 5 min to complete. CRC also finds a more precise solution and generates a remarkable minor number of particles for sampling the parameter space. PL succeeds in fitting the data when *lsqnonlin* and GA are used. Then, through confidence intervals, it classifies both parameters as identifiable, in accordance with MIRI values. DRAM, after the initial adaptation period, finds acceptable points and a good mixing of the chain, regardless of the choice for the tuning parameter values.

Next we compared the results of CRC in the model presented in [44, 45]. In this example, CRC finds an alternative solution of the parameter vector compared to that of PL. PL fits the data properly and fast and identifiability results are in accordance with MIRI values. ABC-SMC fails in the calibration procedure and it cannot go beyond the 7-th iteration, proving that it cannot reach an error as low as the one of CRC. As regards DRAM, the final results are in agreement with those of CRC. The chains of parameters are not significantly affected by the variation of the tuning parameters and, for most of them, the chains converged generating a peak in the corresponding pdf. For the first three kinetic parameters the estimated posterior pdfs are uniformly distributed since the chains equally span all the interval, meaning that they are non identifiable. As for CRC, DRAM finds a different solution for the parameter vector. However, the parameter vector estimated by DRAM is not able to produce reliable time behaviors of the output variables and as a consequence the experimental data feed to DRAM are not well-recapitulated by the time behavior of the observables. This is mainly due to the fact that in the given dataset there are many missing values and DRAM filtered out all the observations with at least one missing value. So the poor results in terms of time behavior are not due to a lack of the algorithm itself but mainly because it cannot deal with missing values, contrary to the other benchmarking algorithms presented in this paper.

Finally, the last model is an high-dimensional ODE model calibrated on real experimental data [46]. CRC was able to find a set of parameter vectors that fit well experimental data. In addition, robustness analysis highlights that about half of the parameters influences most output variables. PL is successful in model calibration but computation of confidence intervals gives confounding results which do not allow a reliable comparison with MIRI values. ABC-SMC fails in model calibration because it remains blocked in the 5-th out of 6 iterations. Also DRAM does not find a reliable solution since all the parameter chains are not stable and, after 10^6 simulations, they span the interval [0, 10] almost uniformly. Moreover, compared to CRC, it has an higher computational cost.

In summary, the main disadvantage of the standard ABC-SMC method is the time necessary to complete a simulation which increases with the model dimension. The PL method is fast in model calibration even for high dimensional models since it implements an optimization algorithm. However, the returned solution does not contain any information on the distribution of parameters since it represents a single point in the parameter space. Moreover, as shown in model M3, it may return improper results in the computation of parameter profiles. As regards DRAM, its results are highly affected by the initial values of the parameters, which must be set from the beginning. This point is crucial since in Systems Biology models most parameter values are unknown and cannot be measured experimentally.

CRC is able to identify a stable and precise solution in all test models, mainly because of some of distinctive features. One of its main innovations is the use of a fixed number of points for sampling the parameter space, which is initially chosen by the user and does not change throughout iterations. As a result, the model is always integrated N_S times in each iteration. Since most of the computational cost of an iteration of CRC is

given by the integration of the model, CRC guarantees a limited computational cost through the different iterations. On the other hand, in other ABC-SMC methods, the computational burden is substantial because the number of samples at each iteration is not known *a priori* but strictly depends on the threshold value. Since the threshold usually decreases at each step, the number of generated samples could increase together with the simulation time. However, compared to the frequentist approach, CRC has always an higher computational cost since it requires multiple subsequent iterations and multiple integrations of the model in order to converge toward the final solution.

Moreover, another significant innovation introduced is the definition of an objective function for each output variable. This allows a model calibration that takes equally into account all experimental endpoints. On the other hand, the other techniques evaluate only a single and unique objective function, which includes information about all observables. Even if the introduction of multiple objective functions improves the accuracy and the performances of CRC, it requires multiple thresholds to be chosen by the user, at each iteration. As a consequence, different combinations of the values of thresholds could guarantee the fulfillment of the two constraints explained in section 2.3.1 ($\epsilon_i^z \leq \epsilon_i^{z-1}$ and $|P_{S,\boldsymbol{\epsilon}^z}| > 1000$). In order to overcome this drawback of CRC, it is possible to implement an optimization strategy step that automatically computes, at each iteration, the minimum value of each threshold that satisfies the constraints explained above and potentially further policies defined by the user. This disadvantage of CRC becomes more relevant in models with an high number of output variables to calibrate.

Finally, we also analyzed the robustness of model parameters in a new way, taking inspiration from the CRA presented in [22]. This algorithm is based on the concept of robustness proposed by Kitano [47], which defines it as the property of a system to maintain its status against internal and external perturbations. We employed CRA in order to quantify the robustness of the model observables against the simultaneous perturbation of the parameters.

Robustness analysis is useful for applications in cancer drug discovery aimed at finding which node of a network could be identified as novel potential drug target. Moreover, the concept of robustness is slightly different from that of identifiability introduced with the PL approach. A parameter that is declared identifiable should have an high MIRI value since it has great impact on the outputs behavior. On the other hand, if a parameter is non-identifiable it is impossible to understand its influence on the observables dynamical response, without performing our robustness analysis. While ABC-SMC evaluates parameter identifiability only through histograms of final parameter values and DRAM computes the parameter posterior distribution, CRC estimates conditional parameter densities, and performs robustness analysis through the MIRI indicator that quantifies the influence of each parameter on the behavior of interest. Indeed, the higher the MIRI value the higher the impact of the parameter on the entire set of observables. All the innovations introduced with CRC are important for a successful calibration of high dimensional nonlinear models in Systems Biology applications based on *omics* data.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://github.com/fortunatobianconi/CRC.

AUTHOR CONTRIBUTIONS

FB, CA, and LT conceived and designed the methods. FB, CA, LT, and PV designed computational benchmark experiments and wrote the paper. CA and LT performed computational analysis. All authors reviewed the manuscript.

FUNDING

This study was supported by the Italian Association for Cancer Research (AIRC) grant (15713/2014).

ACKNOWLEDGMENTS

This manuscript has been released as a Pre-Print at [48].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2020.00025/full#supplementary-material

S1 File | For each model, equations and data are shown together with additional tables and figures of the results.

REFERENCES

1. Ludovini V, Bianconi F, Siggillino A, Piobbico D, Vannucci J, Metro G, et al. Gene identification for risk of relapse in stage I lung adenocarcinoma patients: a combined methodology of gene expression profiling and computational gene network analysis. *Oncotarget*. (2016) **7**:30561–74. doi: 10.18632/oncotarget.8723
2. Ludovini V, Chiari R, Tomassoni L, Antonini C, Baldelli E, Baglivo S, et al. Reverse phase protein array (RPPA) combined with computational analysis to unravel relevant prognostic factors in non- small cell lung cancer (NSCLC): a pilot study. *Oncotarget*. (2017) **8**:83343–53. doi: 10.18632/oncotarget.18480
3. Motta S, Pappalardo F. Mathematical modeling of biological systems. *Brief Bioinformatics*. (2012) **14**:411–22. doi: 10.1093/bib/bbs061
4. Bartocci E, Lió P. Computational modeling, formal analysis, and tools for systems biology. *PLoS Comput Biol*. (2016) **12**:e1004591. doi: 10.1371/journal.pcbi.1004591
5. Lillacci G, Khammash M. Parameter estimation and model selection in computational biology. *PLoS Comput Biol*. (2010) **6**:e1000696. doi: 10.1371/journal.pcbi.1000696
6. Machado D, Costa RS, Rocha M, Ferreira EC, Tidor B, Rocha I. Modeling formalisms in systems biology. *AMB Express*. (2011) **1**:45. doi: 10.1186/2191-0855-1-45

7. Kitano H. Systems biology: a brief overview. *Science.* (2002) **295**:1662–64. doi: 10.1126/science.1069492
8. Alon U. *An Introduction to Systems Biology : Design Principles of Biological Circuits*. Boca Raton, FL: Chapman & Hall; CRC Press (2007).
9. Penas DR, González P, Egea JA, Doallo R, Banga JR. Parameter estimation in large-scale systems biology models: a parallel and self-adaptive cooperative strategy. *BMC Bioinformatics.* (2017) **18**:52. doi: 10.1186/s12859-016-1452-4
10. Fröhlich F, Kaltenbacher B, Theis FJ, Hasenauer J. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput Biol.* (2017) **13**:e1005331. doi: 10.1371/journal.pcbi.1005331
11. Degasperi A, Fey D, Kholodenko BN. Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *NPJ Syst Biol Appl.* (2017) **3**:20. doi: 10.1038/s41540-017-0023-2
12. Vanlier J, Tiemann CA, Hilbers PAJ, van Riel NAW. Parameter uncertainty in biochemical models described by ordinary differential equations. *Math Biosci.* (2013) **246**:305–14. doi: 10.1016/j.mbs.2013.03.006
13. Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE.* (2013) **8**:e74335. doi: 10.1371/journal.pone.0074335
14. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*. (2009) **25**:1923–9. doi: 10.1093/bioinformatics/btp358
15. Thomaseth K, Saccomani MP. Local identifiability analysis of nonlinear ODE models: how to determine all candidate solutions. *IFAC-PapersOnLine*. (2018) **51**:529–34. doi: 10.1016/j.ifacol.2018.03.089
16. Raue A, Karlsson J, Saccomani MP, Jirstrand M, Timmer J. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics.* (2014) **30**:1440–8. doi: 10.1093/bioinformatics/btu006
17. Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinformatics.* (2006) **8**:109–16. doi: 10.1093/bib/bbm007
18. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian computation. *PLoS Comput Biol.* (2013) **9**:e1002803. doi: 10.1371/journal.pcbi.1002803
19. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc.* (2014) **9**:439–56. doi: 10.1038/nprot.2014.025
20. Brooks S. Markov chain Monte Carlo method and its application. *J R Stat Soc Ser D.* (1998) **47**:69–100. doi: 10.1111/1467-9884.00117
21. Prangle D, et al. Adapting the ABC distance function. *Bayesian Anal.* (2017) **12**:289–309. doi: 10.1214/16-BA1002
22. Bianconi F, Baldelli E, Luovini V, Petricoin EF, Crinò L, Valigi P. Conditional robustness analysis for fragility discovery and target identification in biochemical networks and in cancer systems biology. *BMC Syst Biol.* (2015) **9**:70. doi: 10.1186/s12918-015-0216-5
23. Raue A, Steiert B, Schelker M, Kreutz C, Maiwald T, Hass H, et al. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics.* (2015) **31**:3558–60. doi: 10.1093/bioinformatics/btv405
24. Haario H, Laine M, Mira A, Saksman E. DRAM: efficient adaptive MCMC. *Stat Comput.* (2006) **16**:339–54. doi: 10.1007/s11222-006-9438-0
25. Laine M. *MCMC Toolbox for Matlab*. (2018). Available online at: https://mjlaine.github.io/mcmcstat/
26. Martino L. A review of multiple try MCMC algorithms for signal processing. *Digital Signal Process.* (2018) **75**:134–52. doi: 10.1016/j.dsp.2018.01.004
27. Liu JS, Liang F, Wong WH. The multiple-try method and local optimization in Metropolis sampling. *J Am Stat Assoc.* (2000) **95**:121–34. doi: 10.1080/01621459.2000.10473908
28. Luengo D, Martino L. Fully adaptive gaussian mixture metropolis-hastings algorithm. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC (2013). p. 6148–52.
29. Giordani P, Kohn R. Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *J Comput Graph Stat.* (2010) **19**:243–59. doi: 10.1198/jcgs.2009.07174
30. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics.* (2000) **42**:55–61. doi: 10.1080/00401706.2000.10485979
31. Stein M. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*. (1987) **29**:143–51. doi: 10.1080/00401706.1987.10488205
32. Wyss GD, Jorgensen KH. *A Users Guide to LHS: Sandias Latin Hypercube Sampling Software.* Albuquerque, NM: Sandia National Labs. (1998).
33. Bianconi F, Antonini C, Tomassoni L, Valigi P. An application of Conditional Robust Calibration (CRC) to ordinary differential equations (ODEs) models in computational systems biology: a comparison of two sampling strategies. *IET Syst Biol.* (2019) 14:107–19. doi: 10.1049/iet-syb.2018.5091
34. Bianconi F, Baldelli E, Ludovini V, Crinò L, Perruccio K, Valigi P. Robustness of complex feedback systems: application to oncological biochemical networks. *Int J Control.* (2013) **86**:1304–21. doi: 10.1080/00207179.2013.800646
35. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface.* (2009) **6**:187–202. doi: 10.1098/rsif.2008.0172
36. Maier C, Loos C, Hasenauer J. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics.* (2016) **33**:718–25. doi: 10.1093/bioinformatics/btw703
37. Bianconi F, Antonini C, Tomassoni L, Valigi P. CRA toolbox: software package for conditional robustness analysis of cancer systems biology models in MATLAB. *BMC Bioinformatics.* (2019) **20**:385. doi: 10.1186/s12859-019-2933-z
38. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics.* (2002) **162**:2025–35.
39. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA*. (2003) **100**:15324–8. doi: 10.1073/pnas.0306899100
40. McCall J. Genetic algorithms for modelling and optimisation. *J Comput Appl Math*. (2005) **184**:205–22. doi: 10.1016/j.cam.2004.07.034
41. Ingber L. Adaptive simulated annealing (ASA): lessons learned. *Control Cybern.* (1996) **25**:33–54.
42. Robert CP, Richardson S. Markov chain Monte Carlo methods. In: *Discretization and MCMC Convergence Assessment*. New York, NY: Springer (1998). p. 1–25.
43. Green PJ, Mira A. Delayed rejection in reversible jump Metropolis–Hastings. *Biometrika.* (2001) **88**:1035–53. doi: 10.1093/biomet/88.4.1035
44. Becker V, Schilling M, Bachmann J, Baumann U, Raue A, Maiwald T, et al. Covering a broad dynamic range: information processing at the erythropoietin receptor. *Science* (2010) **328**:1404–8. doi: 10.1126/science.1184913
45. Maiwald T, Timmer J, Kreutz C, Klingmüller U, Raue A. Addressing parameter identifiability by model-based experimentation. *IET Syst Biol.* (2011) **5**:120–30. doi: 10.1049/iet-syb.2010.0061
46. Peng H, Peng T, Wen J, Engler DA, Matsunami RK, Su J, et al. Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach. *Bioinformatics.* (2014) **30**:1899–907. doi: 10.1093/bioinformatics/btu133
47. Kitano H. Towards a theory of biological robustness. *Mol Syst Biol.* (2007) **3**:137. doi: 10.1038/msb4100179
48. Bianconi F, Tomassoni L, Antonini C, Valigi P. A new Bayesian methodology for nonlinear model calibration in Computational Systems Biology. *BioRxiv [Preprint].* (2019). doi: 10.1101/633180

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

TECHNOLOGY AND CODE
published: 20 July 2020
doi: 10.3389/fgene.2020.00730

ABC-GWAS: Functional Annotation of Estrogen Receptor-Positive Breast Cancer Genetic Variants

*Mohith Manjunath[1,2], Yi Zhang[2,3], Shilu Zhang[4], Sushmita Roy[4,5], Pablo Perez-Pinera[2,3,6,7] and Jun S. Song[1,2,7]**

[1] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [3] Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [4] Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, WI, United States, [5] Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, United States, [6] The Carle Illinois College of Medicine, Champaign, IL, United States, [7] Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Urbana, IL, United States

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Dylan Glubb,
QIMR Berghofer Medical Research Institute, The University of Queensland, Australia
Jun Zhong,
National Cancer Institute (NCI),
United States

***Correspondence:**
Jun S. Song
songj@illinois.edu

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: 18 February 2020
Accepted: 16 June 2020
Published: 20 July 2020

Citation:
Manjunath M, Zhang Y, Zhang S, Roy S, Perez-Pinera P and Song JS (2020) ABC-GWAS: Functional Annotation of Estrogen Receptor-Positive Breast Cancer Genetic Variants. Front. Genet. 11:730. doi: 10.3389/fgene.2020.00730

Over the past decade, hundreds of genome-wide association studies (GWAS) have implicated genetic variants in various diseases, including cancer. However, only a few of these variants have been functionally characterized to date, mainly because the majority of the variants reside in non-coding regions of the human genome with unknown function. A comprehensive functional annotation of the candidate variants is thus necessary to fill the gap between the correlative findings of GWAS and the development of therapeutic strategies. By integrating large-scale multi-omics datasets such as the Cancer Genome Atlas (TCGA) and the Encyclopedia of DNA Elements (ENCODE), we performed multivariate linear regression analysis of expression quantitative trait loci, sequence permutation test of transcription factor binding perturbation, and modeling of three-dimensional chromatin interactions to analyze the potential molecular functions of 2,813 single nucleotide variants in 93 genomic loci associated with estrogen receptor-positive breast cancer. To facilitate rapid progress in functional genomics of breast cancer, we have created "Analysis of Breast Cancer GWAS" (ABC-GWAS), an interactive database of functional annotation of estrogen receptor-positive breast cancer GWAS variants. Our resource includes expression quantitative trait loci, long-range chromatin interaction predictions, and transcription factor binding motif analyses to prioritize putative target genes, causal variants, and transcription factors. An embedded genome browser also facilitates convenient visualization of the GWAS loci in genomic and epigenomic context. ABC-GWAS provides an interactive visual summary of comprehensive functional characterization of estrogen receptor-positive breast cancer variants. The web resource will be useful to both computational and experimental biologists who wish to generate and test their hypotheses regarding the genetic susceptibility, etiology, and carcinogenesis of breast cancer. ABC-GWAS can also be used as a user-friendly educational resource for teaching functional genomics. ABC-GWAS is available at http://education.knoweng.org/abc-gwas/.

Keywords: GWAS, breast cancer, functional characterization, variant annotation, web resource

INTRODUCTION

Genome-wide association studies (GWAS) have implicated thousands of genetic variants in various complex traits, including diseases (MacArthur et al., 2017). However, only a few studies to date have been successful in characterizing the underlying molecular mechanisms that govern how genetic variations affect molecular interactions (Musunuru et al., 2010; Cowper-Sal lari et al., 2012; Bauer et al., 2013; Huang et al., 2014; Smemo et al., 2014; Gallagher et al., 2017; Zhang et al., 2018b). Studying the molecular function of a typical GWAS locus presents several key challenges (Gallagher and Chen-Plotkin, 2018). First, most of the variants found through GWAS are located in non-coding regions of the human genome; as a result, the precise link between a non-coding variant and some target protein's function is not immediately clear. Second, GWAS variants may indirectly correlate with a phenotype through a complex gene regulatory network involving multiple target genes, unknown causal variants, and transcription factors (TFs). For example, a reported GWAS variant may simply be genetically linked to another proximal variant that itself directly perturbs the binding affinity of a TF and changes the expression of a distal target oncogene or tumor suppressor forming a chromatin loop with the causal variant. In such cases, there is the additional complexity of having to dissect how different components of a gene regulatory network are altered and function together to modulate a trait. Finally, functional characterization of GWAS loci must be carried out in the right cell type representing the phenotype in question; however, one often lacks a complete set of data in genomic, epigenomic, and transcriptomic contexts in the cell type of interest or even faces a difficulty in determining the right cell type. Therefore, there is an urgent need for comprehensive and easily accessible resources that integrate information from heterogeneous large-scale datasets to facilitate rapid functional characterization of GWAS findings and ultimately contribute toward the development of therapeutic preventions and interventions.

Building on the public catalog of GWAS variants (MacArthur et al., 2017), there are currently a few databases providing functional annotation of disease variants. The GRASP database annotates GWAS results by summarizing millions of single nucleotide variant-phenotype associations from 1,390 GWAS studies through correlations such as expression quantitative trait loci (eQTLs), metabolite QTLs, and methylation QTLs (Leslie et al., 2014). Similarly, GWASdb curates trait-associated single nucleotide polymorphisms (SNPs) with detailed functional annotations including eQTL and disease ontology terms (Li et al., 2016). Phenoscanner is a curated database containing variant-phenotype associations of several types such as disease, methylation, gene expression, and protein levels (Staley et al., 2016). More recently, Qtlizer provides associations of variants with gene expression levels and protein abundance using published QTLs (Munz et al., 2019). In the context of cancer, PancanQTL provides a comprehensive list of cis- and trans-eQTLs, including GWAS-related eQTLs, in 33 cancer types (Gong et al., 2018). These web resources have specific advantages, such as having a detailed annotation of GWAS SNPs and/or a list of potential target genes found through eQTL analysis. However, these resources do not perform an in-depth integrative analysis of a specific cancer type using state-of-the-art information about cell type-specific epigenetic landscape, chromatin contact interactions, and TF binding affinity, required for a complete functional characterization of GWAS loci.

Most studies investigating breast cancer GWAS variants have so far focused only on eQTL analysis to find genes correlated with a variant genotype, while only few have pursued a systematic analysis of causal variants and target genes through chromatin structure and TFs (Cowper-Sal lari et al., 2012; French et al., 2013; Li et al., 2013; Ghoussaini et al., 2014, 2016; Darabi et al., 2015; Dunning et al., 2016; Michailidou et al., 2017; Zhang et al., 2018b; Zhang Y. et al., 2019). This paper presents ABC-GWAS, an interactive database containing our comprehensive analysis of 70 manually curated estrogen receptor-positive (ER+) breast cancer GWAS loci and 23 additional ER+ breast cancer loci from a recent fine mapping study (Fachal et al., 2020). The set of 70 loci was obtained from the literature on breast cancer GWAS (Turnbull et al., 2010; Michailidou et al., 2013, 2015). Utilizing large-scale multi-omics datasets such as the Cancer Genome Atlas (TCGA) and the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium, 2012), our analysis pipeline includes eQTL analyses for identifying putative target genes, causal variant prioritization utilizing relevant epigenomic datasets, motif and expression correlation analyses for identifying putative TFs, and three-dimensional chromatin contact predictions for assessing long-distance enhancer-gene interactions. ABC-GWAS aggregates and organizes these results, not readily available in other existing databases, via a user-friendly web interface, making them easily accessible to researchers for additional analysis or experimental validation. It features an embedded genome browser that includes histone modification, chromatin interaction, and TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) tracks for further exploration of the GWAS locus and linked non-coding variants of interest. ABC-GWAS also shows the average DNA copy number information in TCGA breast cancer samples at each GWAS locus. Our resource thus provides useful practical results and conceptual approaches to the functional genomics community in general and breast cancer researchers in particular.

MATERIALS AND METHODS

TCGA Data and Genotype Imputation

The processed RNA-seq expression data in RSEM (RNA-Seq by Expectation-Maximization) units for 794 ER+ breast cancer patients were obtained from the TCGA Genomic Data Commons (GDC) Legacy Archive (Grossman et al., 2016). The germline genotypes of 788 patients in birdseed format for TCGA-BRCA (Breast Invasive Carcinoma) patients were also obtained from the TCGA Data Portal. The copy number segmentation data for 693 patients in hg19 coordinates were retrieved from the GDC Legacy Archive (Grossman et al., 2016). For genotype imputation of the raw genotypes in birdseed format, confidence score greater than 0.1 was used to mark the probed genotypes as missing,

which was then imputed along with the non-probed SNPs. We used the Michigan Imputation Server for imputation (Das et al., 2016), choosing the Haplotype Reference Consortium (HRC) r1.1 2016 as a reference panel (Loh et al., 2016a), Eagle v2.3 for phasing (Loh et al., 2016b), and EUR population as the quality control option. Imputed genotypes were retained if the minor allele frequency (MAF) exceeded 0.005 and estimated imputation accuracy (R^2) exceeded 0.4.

Credible Causal Variants in 23 Additional GWAS Loci

We obtained the full list of credible causal variants (CCVs) from Fachal et al. (2020) and then selected the variants that are single-nucleotide variants, associated with ER+ breast cancer (column ERpos = 1), and have posterior probability of being causal greater than zero (column PP_ERpos > 0). We further removed SNPs that did not pass the quality control tests in the Michigan Imputation Server or for which genotypes could not be imputed confidently in the TCGA data. Finally, excluding 227 CCV SNPs already present in the list of 2,510 SNPs that were in high linkage disequilibrium ($r^2 > 0.8$, 1000 Genomes Phase 3, EUR population) with the reported GWAS SNPs in the 70 manually curated regions yielded 303 CCVs with non-zero posterior probability of being causal in ER+ breast cancers. The 303 CCVs resided in 32 GWAS regions, and 23 of these regions differed from the 70 manually curated regions. ABC-GWAS thus contains the analysis of 530 CCVs out of the 1,238 CCVs reported for ER+ breast cancer.

Genome Browser

The WashU EpiGenome Browser source code was obtained from their GitHub repository (Li et al., 2019; WashU, 2019). The browser uses hg19 coordinates. The JavaScript files from the source code were used to generate the tracks in the embedded browser of ABC-GWAS. The tracks included TF ChIP-seq peaks publicly available in ReMap 2018 database (Cheneby et al., 2018), ENCODE DNase-seq signals, and ESR1, GATA3, and FOXA1 ChIP-seq signals in MCF-7 and T-47D cell lines, POLR2A, CTCF, and ESR1 ChIA-PET interactions, and chromatin interaction predictions in MCF-7 cell line. CTCF is known to play an important role in defining the activity of ESR1 in ER+ breast cancer (Carroll et al., 2006; Chan and Song, 2008). The above datasets were downloaded from the corresponding sources and integrated into our server (**Supplementary Table 1**).

Chromatin Interaction Predictions

To predict SNP-associated interactions, we applied HiC-Reg (Zhang S. et al., 2019), a tool for predicting Hi-C contact counts between pairs of genomic loci from their one-dimensional regulatory signals such as histone modification data, TF ChIP-seq, and chromatin accessibility. We obtained ChIP-seq datasets for 10 histone marks and TFs, and DNase-seq datasets in five cell lines from ENCODE (**Supplementary Material**). HiC-Reg can be trained using cell-line-specific datasets for a cell line with available high-resolution (5 kb) Hi-C data, e.g., the five human cell lines available from Rao et al. (2014). Once trained, HiC-Reg takes as input the genomic features associated with a pair of regions and predicts the chromatin contact count for that pair. We used the method to make predictions in the MCF-7 cell line by training eight different models at 5 kb resolution (**Supplementary Material**). To interpret our results, we averaged the predictions across eight models and displayed the resulting contact count profile associated with each SNP on ABC-GWAS.

eQTL Analysis

To identify candidate target genes for each GWAS SNP, we scanned all genes within 4 Mb centered at the SNP by constructing a multivariate linear regression model with the expression level of each gene as the response variable and the genotype of the GWAS SNP and the copy number (CN) of the gene as predictors (Zhang et al., 2018b; Zhang Y. et al., 2019). The processed gene expression levels in RSEM units were transformed as $\log_2(RSEM + 1)$. The patients with ER+ breast cancer based on TCGA clinical information were retained for subsequent analysis. The genotypes of each GWAS SNP were encoded as the number of risk alleles based on the risk allele information from the NHGRI GWAS catalog (MacArthur et al., 2017). The tumor copy number segmentation values were transformed into gene copy number by taking gene length-weighted average and using $CN = 2 \times 2^{\{segmentation\}}$. We then performed multivariate linear regression and selected genes with mean RSEM larger than 1 and genotype p-value less than 0.05 as candidate target genes for each breast cancer GWAS SNP. On the website, a violin plot using plotly.js is displayed to show the distribution of a candidate target gene's mRNA expression as a function of the GWAS SNP's genotype status (Plotly, 2018).

ENCODE Data

ChIP-seq files for 715 TFs and histone marks in 231 cell lines and tissues were obtained from the ENCODE website (Davis et al., 2018). The locations of the breast cancer risk variants along with the high LD SNPs were then intersected with the peaks of each TF or histone mark in every cell line using bedtools (Quinlan and Hall, 2010). A list of TFs, relevant cell lines, and distance of the SNP from peak center were then tabulated for display.

Motif Analysis

Position weight matrices (PWM) for TFs were obtained from several public databases included in the MotifDb and motifbreakR packages on Bioconductor (Coetzee et al., 2015; Shannon, 2017). The public databases included Jaspar 2018 (Khan et al., 2018), HOCOMOCO (Kulakovskiy et al., 2018), hPDI (Xie et al., 2010), Jolma (Jolma et al., 2013), cisbp (Weirauch et al., 2014), UniPROBE (Hume et al., 2015), Swiss Regulon (Pachkov et al., 2013), HOMER (Heinz et al., 2010), ENCODE motifs (Kheradpour and Kellis, 2014), and FactorBook (Wang et al., 2012). TRANSFAC matrices were also added to the above list (Wingender, 2008). In the first step, the motifbreakR package was used to get possible motif disruptions by candidate SNPs with a p-value threshold of 10^{-3}. We then used our previously developed random mutation model to test the significance of difference in motif scores for the two sequences carrying reference and alternative alleles (Zhang et al., 2018b). The motif disruptions that passed the permutation test p-value threshold of

0.05 were denoted as significant and subsequently included in the ABC-GWAS database.

Correlation Analysis

The list of putative TFs from motif analysis was filtered by removing TFs whose log-transformed mean expression levels across TCGA ER+ breast cancer patients were less than 1 (mean $\log_2(RSEM + 1) < 1$). For each putative target gene from eQTL analysis and TFs passing the expression cut-off threshold, we computed the Pearson correlation coefficient between the expression levels of the target gene and TF across TCGA ER+ breast cancer primary tumor samples, stratifying the patients into three genotype groups: homozygous-risk, heterozygous, and homozygous-alternative. We reasoned that for a good candidate TF, the correlation should be strongest in the homozygous genotype group preserving the TF motif and weakest in the homozygous genotype group disrupting the motif.

RESULTS

Analysis Pipeline for Prioritization of Functional Candidates

We applied the analysis pipeline from our previous work (Zhang et al., 2018b), summarized in **Figure 1**, on a list of manually curated ER+ breast cancer GWAS variants and all SNPs in high linkage disequilibrium (LD) with the GWAS variants, as well as an additional 303 credible causal variants (CCVs) with non-zero posterior probability of being causal in ER+ breast cancers (Fachal et al., 2020; section "Materials and Methods"). The basic framework performs various genomic analyses outlined below to infer how a GWAS variant or a linked SNP changes the binding affinity of a TF in a regulatory region, which in turn alters the transcription of a target gene. In our analysis, linked SNPs residing in accessible open chromatin sites with activating histone modifications (H3K4me1 and H3K27ac) are prioritized as candidate causative SNPs. The genotypes and gene expression data were obtained from TCGA, where the non-probed SNPs' genotypes were imputed using the Michigan imputation server (Das et al., 2016; section "Materials and Methods"). We gathered various heterogeneous datasets from high-throughput experimental techniques such as DNase I hypersensitive sites sequencing (DNase-seq) for prioritization of candidate causal variants, ChIP-seq for TF binding evidence, and chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) and RNA-seq for target gene prioritization in breast cancer samples or cell lines (section "Materials and Methods"; **Supplementary Table 1**). In order to assess how a SNP may perturb a TF's binding affinity and consequently modulate a target gene's expression, we performed eQTL analysis, motif analysis, and TF vs. target gene expression correlation analysis to determine a list of candidate (SNP, target gene, TF) triplets (section "Materials and Methods").

ABC-GWAS User Interface

ABC-GWAS is divided into several modules for interactive data exploration. In the query module, the user first selects a GWAS SNP of interest from the list of 70 SNPs which represent the best reported variants in the manually curated implicated loci, after which a list of high LD ($r^2 > 0.8$, 1000 Genomes Phase 3, EUR population) SNPs of the queried GWAS SNP is populated (**Figure 2**). A table containing the list of GWAS studies implicating the selected SNP in breast cancer is shown on the

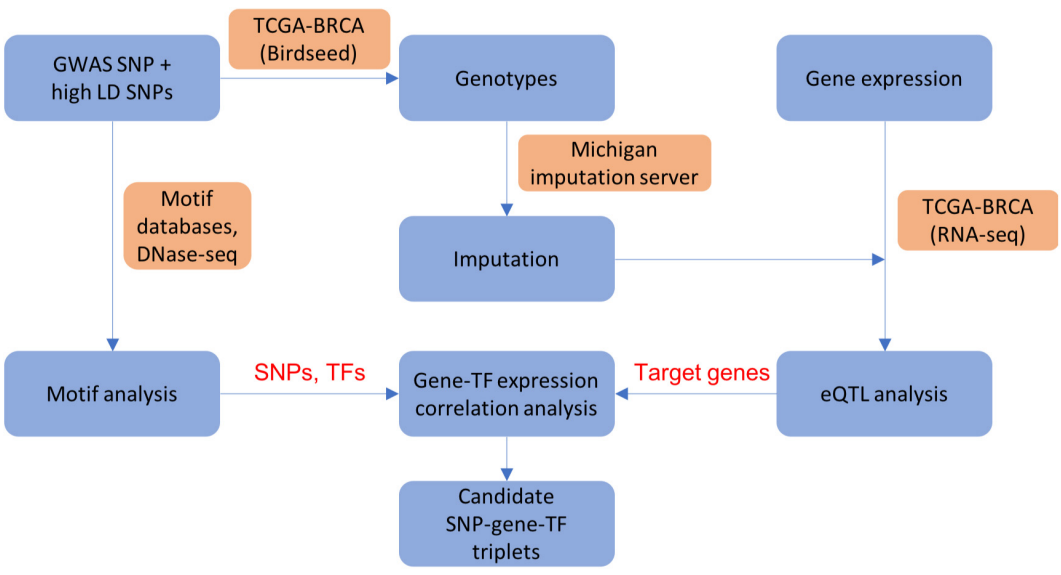

FIGURE 1 | A flowchart showing integrative analysis pipeline used in ABC-GWAS. For each GWAS locus, we perform eQTL analysis, motif analysis, and gene-TF expression correlation analysis to obtain candidate (SNP, gene, TF) triplets. Blue boxes show analysis steps, orange boxes indicate data and/or resources used in the analysis, and text in red shows intermediate results.

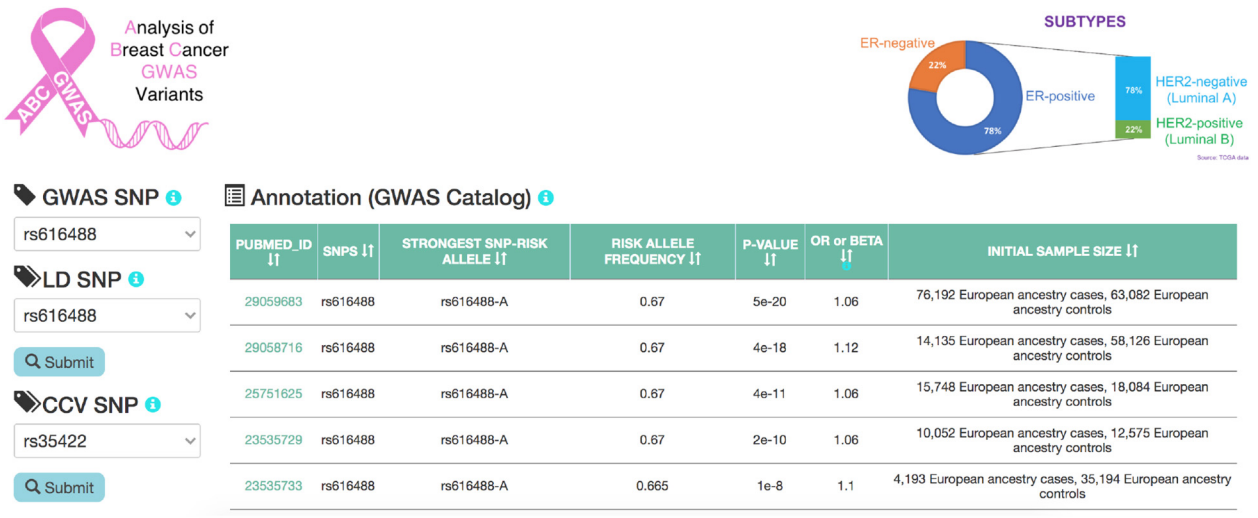

FIGURE 2 | A snapshot of the homepage of ABC-GWAS. Selecting a GWAS SNP using the left-hand-side drop-down menu populates the table on the right with relevant GWAS publications. Upon selecting an LD SNP or a CCV and clicking on the "Submit" button, various tabs on the bottom containing analysis modules are loaded.

right-hand side of the query module (**Figure 2**). Alternatively, the user may choose one of the additional 303 CCVs, not found in the list of all high LD SNPs. After submitting a high LD or CCV SNP as the query variant, all the analysis tabs below the query module get updated. The first tab contains an embedded genome browser showing ChIP-seq, DNase-seq, and ChIA-PET sequencing tracks around the queried SNP locus (**Figure 3A**; section "Materials and Methods"). The second tab displays predicted chromatin interactions in the MCF-7 breast cancer cell line, showing significant interactions between the queried LD SNP location and nearby gene promoters (**Figure 3B**; section "Materials and Methods", and **Supplementary Material**); this track is not available for the 303 CCVs. The third tab consists of two modules. One module shows the average DNA copy number around the queried GWAS SNP location using the TCGA copy number segmentation data for normal and tumor samples (**Figure 3C**; section "Materials and Methods"). The other module checks whether the queried SNP is a CCV (Fachal et al., 2020); when available, a list of likely target genes of the queried SNP obtained from the same study is also displayed. The fourth tab summarizes our eQTL analysis results for the selected GWAS SNP or CCV using the genotypes and RNA-seq data from TCGA breast cancer samples (section "Materials and Methods"). A table containing significant eQTL results and a violin plot of the target gene's expression stratified into genotype groups are displayed. The fifth tab shows a table of all ENCODE ChIP-seq peaks that intersect the queried SNP (section "Materials and Methods"). The peaks are categorized based on whether the experiment is for a TF or histone modification. The results can also be filtered to show peaks occurring only in breast tissue or breast cancer-related cell lines. The last tab contains two modules showing putative TFs, the binding activities of which are predicted to be affected by the given SNP, as assessed by motif analysis (section "Materials and Methods") and expression correlation analysis (section "Materials and Methods"). A motif logo with the nucleotide perturbed by the SNP is available for each of the putative TFs. In the "Expression correlation" tab, the putative TFs from motif analysis are further prioritized based on the expression correlation between each TF and eQTL target genes. Pearson correlation coefficients are displayed as a heatmap with the putative TFs along the rows and genotype groups along the columns.

Case Study (Validated Result From the Literature): (rs4784227, *TOX3*, FOXA1)

Cowper-Sal lari et al. (2012) analyzed the functional mechanism of the GWAS SNP rs4784227 and proposed it to be a causal regulatory SNP targeting the gene *TOX3*. Furthermore, the risk allele rs4784227-T was shown to increase the binding affinity of the pioneer factor FOXA1, resulting in a fivefold decrease in *TOX3* gene expression. Here, we sought to verify the reported mechanism at the rs4784227 locus using the results from our database. **Figure 4A** shows a snapshot of the genomic region around rs4784227 from the embedded genome browser. The MCF-7 DNase tracks in **Figure 4A** clearly indicate that the GWAS SNP is located within open chromatin region. Furthermore, the "ReMap 2018 Peaks" track, which represent TF binding peak locations collected from ENCODE and Gene Expression Omnibus (GEO) datasets (Barrett et al., 2013; Cheneby et al., 2018), showed several TF binding sites, supporting that this SNP is likely a causal SNP. The eQTL results showed a negative correlation between the risk allele rs4784227-T and the mRNA level of *TOX3* in TCGA breast cancer samples (**Figure 4B**). Our motif analysis results further suggested FOXJ3 as one of the top candidate TFs (**Figure 4C**); given the similarity of FOXJ3 and FOXA1 motifs (q-value = 0.0098), as predicted by

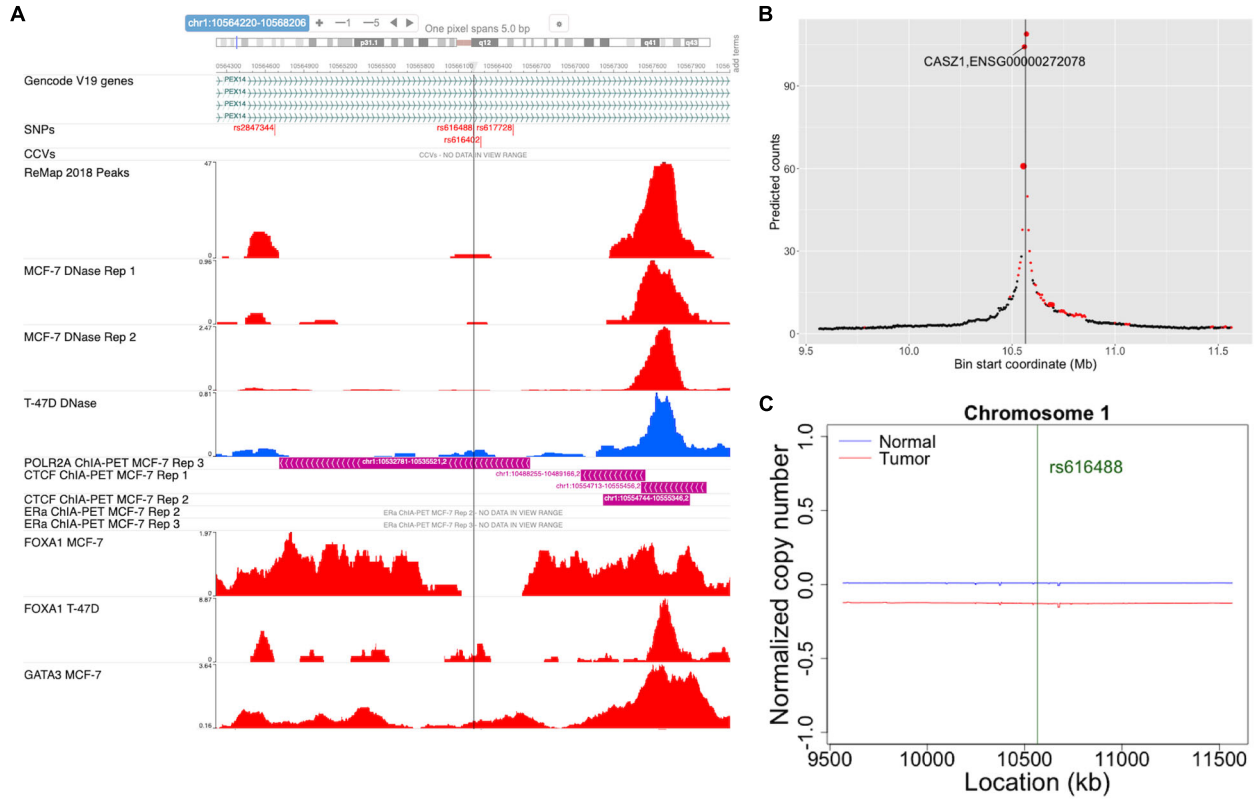

FIGURE 3 | Snapshots of some of the ABC-GWAS analysis modules. **(A)** Embedded genome browser showing the queried LD SNP location (vertical line) and ChIP-seq tracks in MCF-7 and T-47D cell lines. The "SNPs" track shows the locations of the GWAS SNPs and high LD SNPs, while the "CCVs" track shows the locations of the credible causal variants and their GWAS lead SNPs. **(B)** The plot shows average predicted chromatin contact counts across several HiC-Reg models in MCF-7 as a function of genomic location centered at the queried LD SNP (vertical line). The predictions for which at least one model shows significance ($q < 0.05$) is filled in red. The size of the markers is proportional to the number of models showing significance. **(C)** Normalized DNA copy number for normal (blue) and tumor (red) samples from TCGA in a 2 Mb window centered at the GWAS SNP location.

the Tomtom motif comparison tool from MEME web resource (Gupta et al., 2007; Bailey et al., 2009), our overall results were thus consistent with the findings of Cowper-Sal lari et al. (2012).

Case Study (Novel): (rs1250003, *ZMIZ1*, GATA)

The SNP rs704010, residing within an intron of the gene *ZMIZ1*, was reported to be associated with increased breast cancer risk in Turnbull et al. (2010), and this association was subsequently verified in later studies (Michailidou et al., 2013, 2015, 2017; Zhang et al., 2018a). **Figure 5A** shows a snapshot of the locus from the embedded genome browser. Among the 12 high LD SNPs shown in the first track, we identified rs1250003 to be the only SNP residing within an open chromatin region in MCF-7 and also to a lesser extent in T-47D, as shown by the DNase tracks. This candidate SNP rs1250003 was located about 5 kb from the GWAS SNP and in high LD with the GWAS SNP (r^2 = 0.99, 1000 Genomes Phase 3, EUR population). We also found that in the European population (1000 Genomes, Phase 3), rs1250003 was in perfect LD with two SNPs (rs1250008, rs1250009) previously reported to be CCVs (Fachal et al., 2020). Several TFs relevant to breast cancer – such as ESR1, FOXA1, and GATA3 – were found to bind near the SNP, as shown by the corresponding ChIP-seq tracks, indicating an important regulatory role of the SNP. The genotype status of rs704010 significantly correlated with the mRNA level of *ZMIZ1* ($p = 7.7 \times 10^{-4}$) (**Figure 5B**). POLR2A ChIA-PET track in MCF-7 further showed a chromatin-looping interaction between the SNP location and the promoter of *ZMIZ1* (**Figure 5A**). A significant interaction was also computationally predicted between the two loci in MCF-7 (**Figure 5C**). Our integrative analysis thus implicated *ZMIZ1* to be the top candidate target gene for the locus. Finally, we found GATA family binding motifs to be significantly disrupted by the SNP (**Figure 5D**), consistent with the ChIP-seq data. Thus, a quick analysis based on ABC-GWAS found the triplet (rs1250003, *ZMIZ1*, GATA) to be a novel putative functional mechanism behind the GWAS SNP rs704010 for increasing risk for breast cancer.

DISCUSSION

We demonstrated the capability of ABC-GWAS to find known, as well as novel, functional mechanisms of breast cancer GWAS

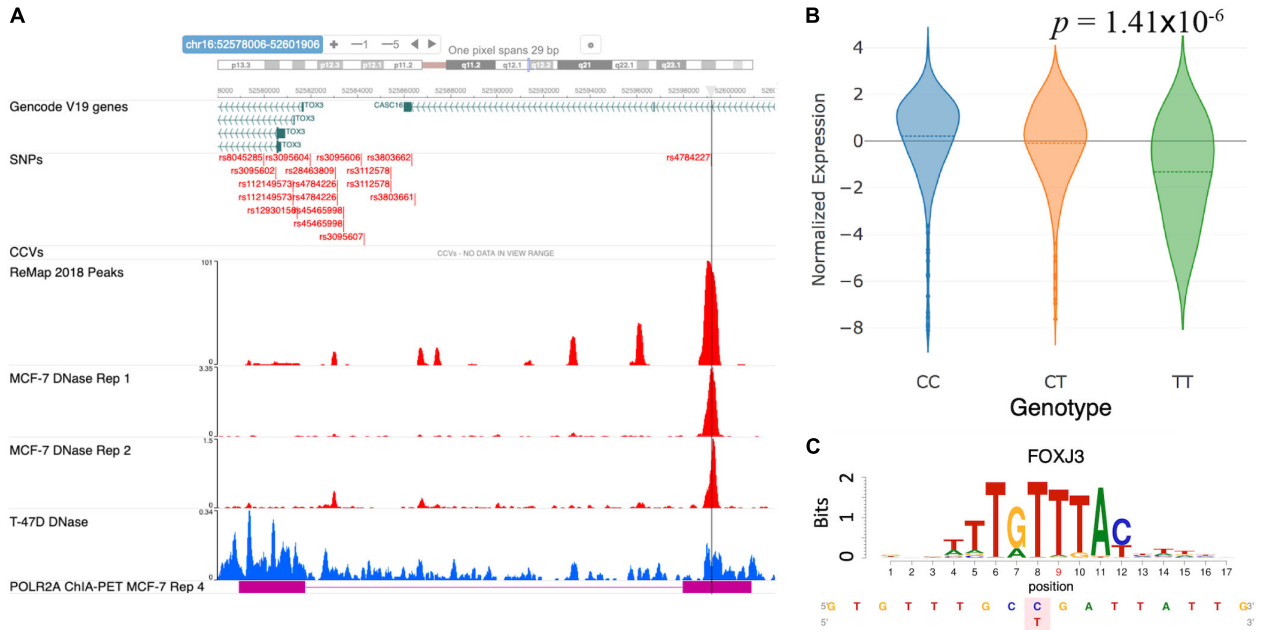

FIGURE 4 | Case study of GWAS locus at *TOX3*: (rs4784227, *TOX3*, FOXA1). **(A)** Embedded genome browser showing the SNP rs4784227 within an open chromatin region (MCF-7 DNase tracks) and several TF peaks ("ReMap 2018 Peaks"). **(B)** eQTL plot showing significant correlation between *TOX3* expression and rs4784227 genotypes. **(C)** Predicted TF candidate FOXJ3 motif from Kheradpour and Kellis (2014).

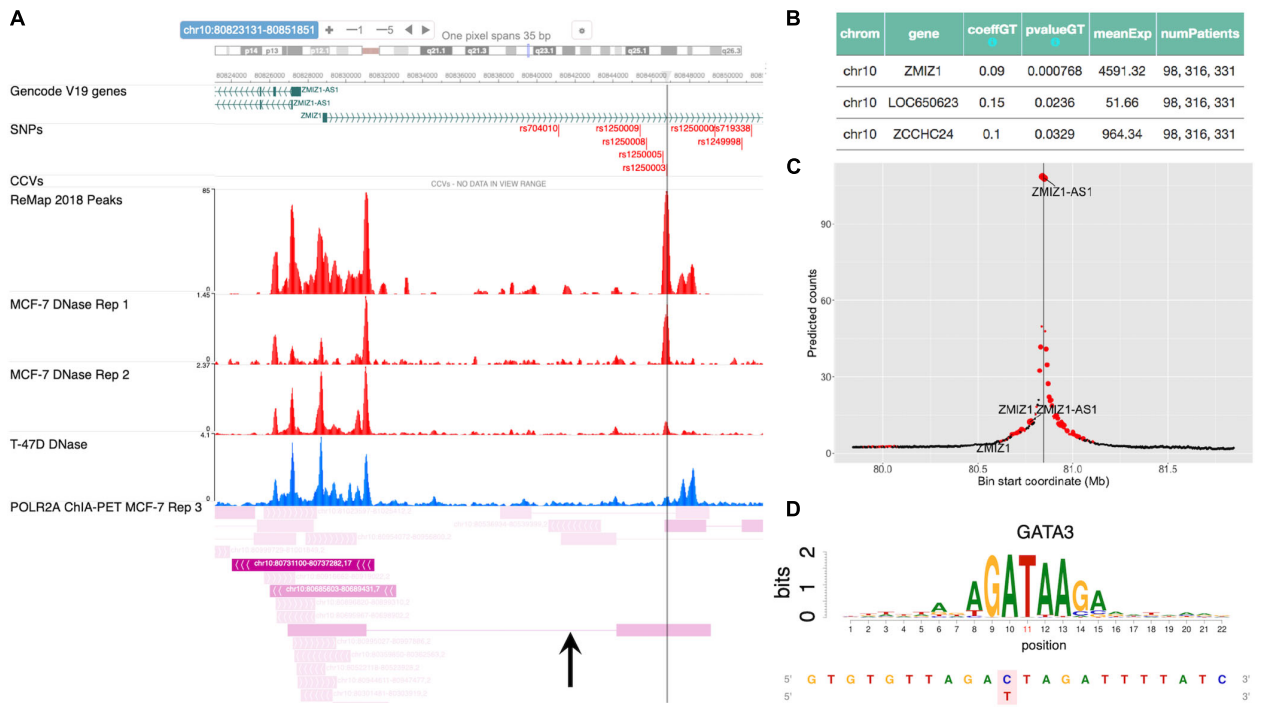

FIGURE 5 | Case study of GWAS locus at *ZMIZ1*: (rs1250003, *ZMIZ1*, GATA). **(A)** Embedded genome browser showing the LD SNP rs1250003 within an open chromatin region (MCF-7 DNase track) and several TF peaks ("ReMap 2018 Peaks"). Black arrow indicates a chromatin looping interaction between the SNP locus and the *ZMIZ1* promoter. **(B)** eQTL results showing the top significant target gene of the GWAS SNP to be *ZMIZ1*. **(C)** Average predicted chromatin contact counts across several HiC-Reg models in MCF-7, showing the *ZMIZ1* promoter as one of the loci significantly interacting with the LD SNP. **(D)** Predicted TF candidate GATA motif from Kheradpour and Kellis (2014).

loci. The computational and organizational framework of ABC-GWAS can be readily extended to other cancers. Once a (SNP, target gene, TF) triplet is identified through ABC-GWAS, several molecular experiments can be performed to validate the prediction. For example, the genotype of the predicted causative SNP could be modified through CRISPR-Cas9 base editors to study its effect on target gene expression (Komor et al., 2016). ChIP-quantitative polymerase chain reaction (ChIP-qPCR) is one way to measure how the SNP's genotype status modulates the binding affinity of the predicted TF. ABC-GWAS thus provides a valuable resource, currently not available in other databases, for functional characterization of GWAS results. ABC-GWAS currently contains analysis results for only a predetermined set of SNPs, and a useful future extension could allow our integrative analysis pipeline to be performed on any genetic variant of interest chosen by the user. Another informative feature could be to provide a pathway analysis of candidate target genes and transcription factors in the context of breast cancer biology.

ABC-GWAS is an interactive web resource containing results from an integrative functional analysis of ER+ breast cancer variants. We combined data from TCGA, ENCODE, and several motif databases to create a comprehensive resource that includes an embedded genome browser with relevant tracks in breast cancer cell lines and several modules describing results from eQTL, motif, and expression correlation analyses. Using our resource, we have verified the known functional mechanism of a genetic variant regulating the gene *TOX3* and also proposed a novel mechanism targeting the *ZMIZ1* locus. ABC-GWAS aims to take GWAS discoveries to the next level by providing a one-stop resource for in-depth functional analyses critical for interpreting and prioritizing GWAS variants. We thus hope that our resource will help both experimental and computational researchers accelerate breast cancer research.

SOFTWARE AVAILABILITY

1. Project name: ABC-GWAS.
2. Project home page: http://education.knoweng.org/abc-gwas/.
3. Operating system(s): Platform independent.
4. Programming language: HTML, JavaScript, R, Python.
5. Other requirements: JavaScript supporting web browser.
6. License: GNU GPL v3.0.

DATA AVAILABILITY STATEMENT

The datasets analyzed in the current study are available in the ENCODE project (https://www.encodeproject.org/) and the TCGA repository (http://cancergenome.nih.gov/) through GDC (https://portal.gdc.cancer.gov/projects) and dbGaP (https://www.ncbi.nlm.nih.gov/gap).

ETHICS STATEMENT

The usage of NIH controlled-access datasets was approved by the NCBI dbGaP.

AUTHOR CONTRIBUTIONS

JS contributed to the conception and design of the project. MM, YZ, and SZ contributed to the data curation and analysis. MM, YZ, SZ, SR, and JS contributed to the methodology. MM and YZ developed the resource and visualization tools. SR, PP-P, and JS supervised the research and secured the funding. MM and JS wrote the original draft. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by funds from the National Institutes of Health (NIH) R01CA163336, NIH U54GM114838, and the Grainger Engineering Breakthroughs Initiative to JS, Planning Grant from the Cancer Center at Illinois to PP-P and JS, and NIH R01-HG010045-01, NIH U54 AI117924, and UW Madison Vilas fellowship to SR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00730/full#supplementary-material

REFERENCES

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37(Suppl._2), W202–W208.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bauer, D. E., Kamran, S. C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* 342, 253–257. doi: 10.1126/science.1242088

Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., et al. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* 38, 1289–1297. doi: 10.1038/ng1901

Chan, C. S., and Song, J. S. (2008). CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Res.* 68, 9041–9049. doi: 10.1158/0008-5472.CAN-08-2632

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 46, D267–D275. doi: 10.1093/nar/gkx1092

Coetzee, S. G., Coetzee, G. A., and Hazelett, D. J. (2015). motifbreakR: an R/bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* 31, 3847–3849. doi: 10.1093/bioinformatics/btv470

Cowper-Sal lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., et al. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* 44, 1191–1198. doi: 10.1038/ng.2416

Darabi, H., McCue, K., Beesley, J., Michailidou, K., Nord, S., Kar, S., et al. (2015). Polymorphisms in a putative enhancer at the 10q21.2 breast cancer risk locus regulate NRBF2 expression. *Am. J. Hum. Genet.* 97, 22–34. doi: 10.1016/j.ajhg.2015.05.002

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Dunning, A. M., Michailidou, K., Kuchenbaecker, K. B., Thompson, D., French, J. D., Beesley, J., et al. (2016). Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat. Genet.* 48, 374–386. doi: 10.1038/ng.3521

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Fachal, L., Aschard, H., Beesley, J., Barnes, D. R., Allen, J., Kar, S., et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat. Genet.* 52, 56–73. doi: 10.1038/s41588-019-0537-1

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., et al. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.* 92, 489–503. doi: 10.1016/j.ajhg.2013.01.002

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS Era: from association to function. *Am. J. Hum. Genet.* 102, 717–730. doi: 10.1016/j.ajhg.2018.04.002

Gallagher, M. D., Posavi, M., Huang, P., Unger, T. L., Berlyand, Y., Gruenewald, A. L., et al. (2017). A dementia-associated risk variant near TMEM106B alters chromatin architecture and gene expression. *Am. J. Hum. Genet.* 101, 643–663. doi: 10.1016/j.ajhg.2017.09.004

Ghoussaini, M., Edwards, S. L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., et al. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat. Commun.* 4:4999. doi: 10.1038/ncomms5999

Ghoussaini, M., French, J. D., Michailidou, K., Nord, S., Beesley, J., Canisus, S., et al. (2016). Evidence that the 5p12 variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through FGF10 and MRPS30 regulation. *Am. J. Hum. Genet.* 99, 903–911. doi: 10.1016/j.ajhg.2016.07.017

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., et al. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res.* 46, D971–D976. doi: 10.1093/nar/gkx861

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375, 1109–1112. doi: 10.1056/NEJMp1607591

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8:R24.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Huang, Q., Whitington, T., Gao, P., Lindberg, J. F., Yang, Y., Sun, J., et al. (2014). A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat. Genet.* 46, 126–135. doi: 10.1038/ng.2862

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 43, D117–D122. doi: 10.1093/nar/gku1045

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-binding specificities of human transcription factors. *Cell* 152, 327–339. doi: 10.1016/j.cell.2012.12.009

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 46:D1284. doi: 10.1093/nar/gkx1188

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* 42, 2976–2987. doi: 10.1093/nar/gkt1249

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424. doi: 10.1038/nature17946

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 46, D252–D259. doi: 10.1093/nar/gkx1106

Leslie, R., O'Donnell, C. J., and Johnson, A. D. (2014). GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* 30, i185–i194. doi: 10.1093/bioinformatics/btu273

Li, D., Hsu, S., Purushotham, D., Sears, R. L., and Wang, T. (2019). WashU epigenome browser update 2019. *Nucleic Acids Res.* 47, W158–W165. doi: 10.1093/nar/gkz348

Li, M. J., Liu, Z., Wang, P., Wong, M. P., Nelson, M. R., Kocher, J. P., et al. (2016). GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* 44, D869–D876. doi: 10.1093/nar/gkv1317

Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., et al. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 152, 633–641. doi: 10.1016/j.cell.2012.12.034

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, A. Y., Finucane, H. K., et al. (2016a). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443–1448. doi: 10.1038/ng.3679

Loh, P. R., Palamara, P. F., and Price, A. L. (2016b). Fast and accurate long-range phasing in a UK biobank cohort. *Nat. Genet.* 48, 811–816. doi: 10.1038/ng.3571

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* 47, 373–380. doi: 10.1038/ng.3242

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* 45, 353–361. doi: 10.1038/ng.2563

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94. doi: 10.1038/nature24284

Munz, M., Wohlers, I., Simon, E., Busch, H., Schaefer, A. S., and Erdmann, J. (2019). Qtlizer: comprehensive QTL annotation of GWAS results. *bioRxiv* [Preprint]. doi: 10.1101/495903

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466, 714–719. doi: 10.1038/nature09266

Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E., and van Nimwegen, E. (2013). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.* 41, D214–D220. doi: 10.1093/nar/gks1145

Plotly (2018). *Plotly JavaScript Graphing Library*. Available online at: https://plot.ly/javascript/ (accessed September 11, 2018).

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Shannon, P. (2017). *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. R Package Version 1.18.0*.

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gomez-Marin, C., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375. doi: 10.1038/nature13138

Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* 32, 3207–3209. doi: 10.1093/bioinformatics/btw373

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., et al. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* 42, 504–507. doi: 10.1038/ng.586

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22, 1798–1812. doi: 10.1101/gr.139105.112

WashU (2019). *EpiGenome Gateway - WashU EpiGenome Browser*. Available online at: https://github.com/epgg/eg (accessed March 19, 2019).

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* 9, 326–332. doi: 10.1093/bib/bbn016

Xie, Z., Hu, S., Blackshaw, S., Zhu, H., and Qian, J. (2010). hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics* 26, 287–289. doi: 10.1093/bioinformatics/btp631

Zhang, S., Chasman, D., Knaack, S., and Roy, S. (2019). In silico prediction of high-resolution Hi-C interaction matrices. *Nat. Commun.* 10:5449. doi: 10.1038/s41467-019-13423-8

Zhang, Y., Manjunath, M., Yan, J., Baur, B. A., Zhang, S., Roy, S., et al. (2019). The cancer-associated genetic variant Rs3903072 modulates immune cells in the tumor microenvironment. *Front. Genet.* 10:754. doi: 10.3389/fgene.2019.00754

Zhang, Y., Manjunath, M., Zhang, S., Chasman, D., Roy, S., and Song, J. S. (2018a). Abstract 1220: integrative genomic analysis discovers the causative regulatory mechanisms of a breast cancer-associated genetic variant. *Cancer Res.* 78, 1220–1220.

Zhang, Y., Manjunath, M., Zhang, S., Chasman, D., Roy, S., and Song, J. S. (2018b). Integrative genomic analysis predicts causative Cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084. *Cancer Res.* 78, 1579–1591. doi: 10.1158/0008-5472.CAN-17-3486

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 21 July 2020
doi: 10.3389/fonc.2020.00900

Genetic Alterations and Transcriptional Expression of m^{6}A RNA Methylation Regulators Drive a Malignant Phenotype and Have Clinical Prognostic Impact in Hepatocellular Carcinoma

*Gui-Qi Zhu[1,2†], Lei Yu[3†], Yu-Jie Zhou[4,5†], Jun-Xian Du[6†], Shuang-Shuang Dong[1,2], Yi-Ming Wu[7], Ying-Hong Shi[1,2], Jian Zhou[1,2], Jia Fan[1,2] and Zhi Dai[1,2]**

[1] State Key Laboratory of Genetic Engineering, Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai, China, [2] Key Laboratory of Carcinogenesis and Cancer Invasion, Fudan University, Ministry of Education, Shanghai, China, [3] Department of Liver Surgery and Transplantation, Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai, China, [4] Key Laboratory of Gastroenterology and Hepatology, Ministry of Health, Shanghai, China, [5] Division of Gastroenterology and Hepatology, Renji Hospital, School of Medicine, Shanghai Institute of Digestive Disease, Shanghai Jiao Tong University, Shanghai, China, [6] Department of General Surgery, Zhongshan Hospital, Fudan University, Shanghai, China, [7] Department of Urology, The Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University, Wenzhou, China

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Jun Zhong,
National Cancer Institute (NCI), United States
Tania Lee Slatter,
University of Otago, New Zealand

***Correspondence:**
Zhi Dai
dai.zhi@zs-hospital.sh.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 27 July 2019
Accepted: 07 May 2020
Published: 21 July 2020

Citation:
Zhu G-Q, Yu L, Zhou Y-J, Du J-X, Dong S-S, Wu Y-M, Shi Y-H, Zhou J, Fan J and Dai Z (2020) Genetic Alterations and Transcriptional Expression of m^{6}A RNA Methylation Regulators Drive a Malignant Phenotype and Have Clinical Prognostic Impact in Hepatocellular Carcinoma. Front. Oncol. 10:900. doi: 10.3389/fonc.2020.00900

Background: N6-methyladenosine (m^{6}A) RNA methylation, associated with cancer initiation and progression, is dynamically regulated by the m^{6}A RNA regulators. However, its role in liver carcinogenesis is poorly understood.

Methods: Three hundred seventy-one hepatocellular carcinoma (HCC) patients from The Cancer Genome Atlas database with sequencing and copy number variations/mutations data were included. Survival analysis was performed using Cox regression model. We performed gene set enrichment analysis to explore the functions associated with different HCC groups. Finally, we used a machine-learning model on selected regulators for developing a risk signature (m^{6}Ascore) The prognostic value of m^{6}Ascore was finally validated in another two GEO datasets.

Results: We demonstrated that 11 m^{6}A RNA regulators are significantly differentially expressed among 371 HCC patients stratified by clinicopathological features (P<0.001). We then identified two distinct HCC clusters by applying consensus clustering to m^{6}A RNA regulators. Compared with the cluster2 subgroup, the cluster1 subgroup correlates with poorer prognosis ($P < 0.001$). Moreover, the cell cycle, splicesome and notch signaling pathway are significantly enriched in the cluster1 subgroup. We further derived m^{6}Ascore, using four m^{6}A regulators, predicting HCC prognosis well at three (AUC = 0.7) or 5 years (AUC=0.7) in validation. The prognostic value of m^{6}Ascore also was validated successfully in two GEO datasets ($P < 0.05$). Finally, we discovered that mutations and copy number variations of m^{6}A regulators, conferring worse survival, are strongly associated with TP53 mutations in HCC.

Conclusions: We find a significant relationship between the alterations and different

expressions causing increased m^{6}A level and worse survival, especially in TP53-mutated HCC patients. Genetic alterations of m^{6}A genes might cooperate with TP53 and its regulator targets in the HCC pathogenesis. Our m^{6}Ascore may be applied in the clinical trials for patient stratification in HCC.

Keywords: RNA modification, m^{6}A, hepatocellular carcinoma, TP53 mutation, prognosis

INTRODUCTION

Hepatocellular carcinoma (HCC) is the fifth leading cause of malignant cancer and the third most common reason for cancer-specific death worldwide (1, 2). The HCC mortality is often high because of metastasis and postsurgical liver recurrence. Hence, effective treatments are eagerly awaited to hold back additional metastases to improve the disappointing HCC outcomes (2). However, advanced HCC with recurrence or low response to chemotherapy have low overall survival nowadays. The lack of effective interventions and high mortality of HCC demand a better understanding of the cancer molecular mechanism.

N6-methyladenosine (m^{6}A), first described in 1974, the most abundant form of internal messenger RNA (mRNA) modification in higher eukaryotes, has become of great interest in recent years (3). It is known to play vital part in regulating gene expression, splicing in cellular biology, and cellular protein translation (3–5). The m^{6}A regulators comprise "writers" such as methyltransferase like 3 (METTL3), WT1-associated protein (WTAP), METTL14, RNA binding motif protein 15 (RBM15), zinc finger CCCH domain-containing protein 13 (ZC3H13) and KIAA1429 (also known as VIRMA). "Readers" such as YTH domain-containing 1 (YTHDC1), YTH N6-methyladenosine RNA binding protein 2 (YTHDF2), YTH domain-containing 1 (YTHDC2), heterogeneous nuclear ribonucleoprotein C (HNRNPC), YTH N6-methyl-adenosine RNA binding protein 1 (YTHDF1) and 'erasers' such as fat mass- and obesity-associated protein (FTO) and α-ketoglutarate-dependent dioxygenase alkB homolog 5 (ALKBH5) (6–12). m^{6}A dysregulation regulated by knockdown of genes could result in decreased cell proliferation, cell death and developmental defects (3).

In recent years, increasing amounts of evidence showed that genetic changes and dysregulated expression of m^{6}A RNA methylation genes are correlated with malignant phenotype closely in different types of cancer (7, 13–17). For example, knockout of METTL3 disturbs embryonic stem cell differentiation (18). Depletion of erasers, such as FTO and ALKBH5, can result in obesity and dysregulation of spermatogenesis (3, 9). Knockdown of m^{6}A methyltransferase can cause regulation of the TP53 signaling pathway relevant to tumorigenesis (4). Recently, the overexpression of METTL3 result in HCC tumor progression by repressing SOCS2 through the m^{6}A-YTHDF2 axis in HCC (19). Additionally, down-regulation of METTL14 as a dismal prognostic factor for HCC overall survival (20). Therefore, it is surprising that the profile of genetic alterations affecting m^{6}A regulatory genes and gene expression of corresponding m^{6}A genes have not been explored in HCC.

Hence, in our study, we systematically evaluated the genetic alterations and expression of 13 widely reported m^{6}A RNA regulators with RNA sequencing data from The Cancer Genome Atlas (TCGA) (n = 377) datasets. We analyzed the alteration spectrum and expression of every m^{6}A modification regulator with regards to different clinicopathological factors, including survival.

Abbreviations: HCC, hepatocellular carcinoma; OS, overall survival; CI, confidence interval; HR, hazard risk; m^{6}A, N6-methyladenosine RNA methylation; CNV, copy number variations; METTL3, methyltransferase like 3; WTAP, WT1-associated protein; RBM15, RNA binding motif protein 15; ZC3H13, zinc finger CCCH domain-containing protein 13; YTHDC1, YTH domain-containing 1; YTHDF1, YTH N6-methyl-adenosine RNA binding protein 1; HNRNPC, and heterogeneous nuclear ribonucleoprotein C; FTO, fat mass- and obesity-associated protein; ALKBH5, α-ketoglutarate-dependent dioxygenase alkB; TCGA, The Cancer Genome Atlas; GISTIC, Genomic Identification of Significant Targets in Cancer algorithm; DFS, disease-free survival; PCA, principal component analysis; GSEA, gene set enrichment analysis.

MATERIALS AND METHODS

Patient Data

The datasets GSE14520 and GSE63898 were downloaded from the expression database GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) (21, 22). GSE14520 included a total of 488 samples, 241 samples were paired non-tumor samples, while the other 247 samples were HCC samples. Platform Information was [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array for 43 samples, and [HT_HG-U133A] Affymetrix HT Human Genome U133A Array for the other 445 samples. GSE63898 included 228 HCC and 168 cirrhotic samples, and the platform was [HG-U219] Affymetrix Human Genome U219 Array for all samples. The clinicopathological, mutation, deletion, amplification, copy number variation and/or survival data from HCC patients are available via the cBioportal (23), the TCGA data portal and/or reported in a previous publication (24). Of the 424 HCC patients in the TCGA cohort, matched mutation, deletion, amplification and copy number variation (CNV) data are available for 366 patients via cBioportal (23). We therefore included only these patients in our genetic alteration analyses. In addition, of the 424 HCC patients included in our gene expression analysis, corresponding complete clinical information are available for 236 patients from the TCGA cohort.

Selection of m^{6}A RNA Methylation Regulators

We first collected 16 m^{6}A RNA methylation regulators from published literature (7, 12, 16), and we retrieved the m^{6}A genes with available gene expression from the HCC TCGA

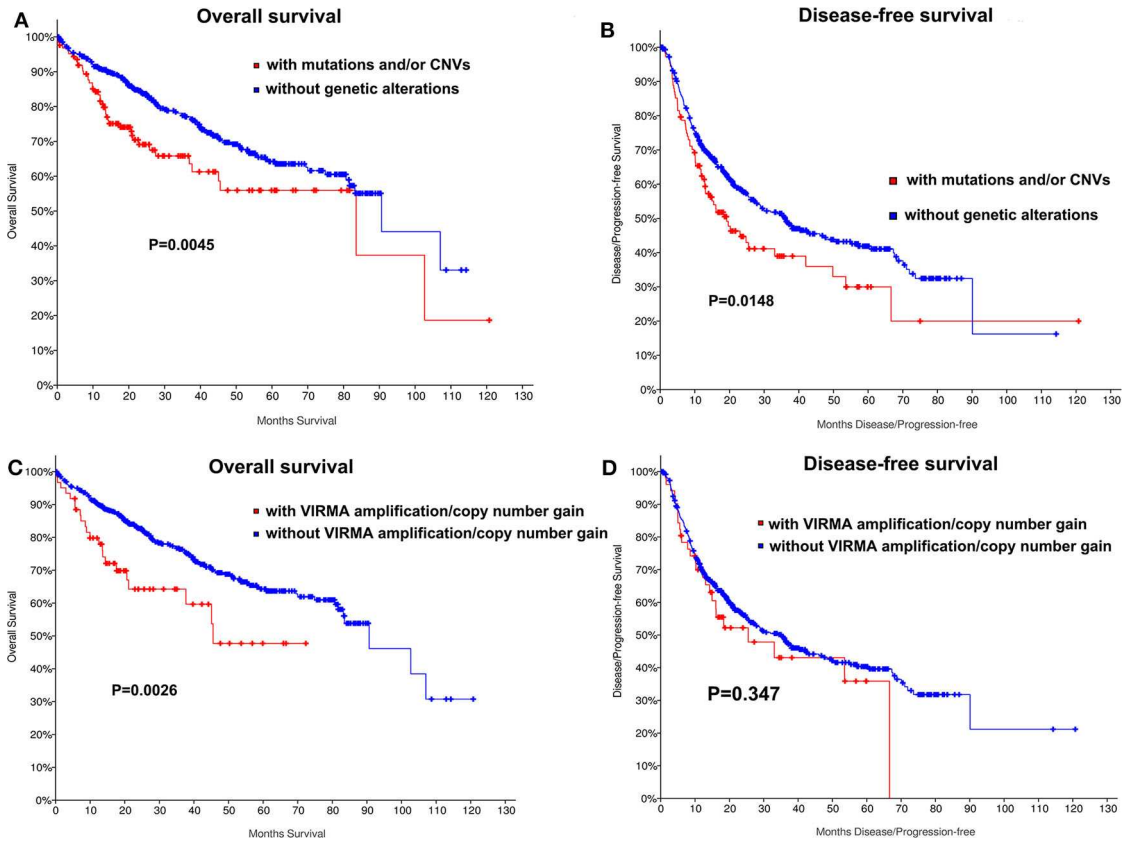

FIGURE 1 | Genetic alterations of m^{6}A regulatory genes in HCC. **(A)** Percentage of HCC samples with mutations or CNVs of the m^{6}A regulators based on the data from TCGA. **(B)** Events of copy number gain or loss of m^{6}A regulatory genes in HCC samples. **(C)** Relationships between Putative copy-number alterations and mRNA for FTO. **(D)** Relationships between Putative copy-number alterations and mRNA for ALKBH5.

cohort. This brought a list of 13 m^{6}A regulators. Finally, the genetic alterations and expression of m^{6}A RNA methylation regulators in HCC with different clinicopathological factors were compared systematically.

Statistical Analysis

For gene expression analysis, to investigate the function of m^{6}A RNA methylation gene regulators in HCC, we clustered the HCC into different groups with "ConsensusClusterPlus" (resample rate of 80%, 50 iterations and Pearson correlation, http://www.bioconductor.org/). We used PCA with the R package for R v3.5.1 to explore the gene expression profiles among various HCC patient subgroups. We performed GO and KEGG pathway analyses with the Database for Annotation, Visualization, and Integrated Discovery to annotate differentially expressed genes in various HCC subgroups. We also analyzed interactions among m^{6}A RNA methylation regulators from the STRING database (http://www.string-db.org/). We finally applied gene set enrichment analysis (GSEA) to evaluate the functions associated with different HCC subgroups. To determine the prognostic value of m^{6}A RNA methylation genes, we therefore performed several univariate Cox regression analyses of their gene expression in the TCGA cohort. To this end, we confirmed nine genes that were significantly associated with patient survival ($P < 0.05$), which we then selected for further functional analysis and development of a risk signature (m^{6}Ascore) with the machine learning algorithm model (25). Finally, four m^{6}A regulatory genes and their coefficients were determined by R package "Coxboost" within the training set (All patients were randomly divided 75% for training and 25% for validation) by the machine learning model (25). The m^{6}Ascore for the signature was calculated using the formula:

$$\mathrm{m^6Ascore} = \sum_{i=1}^{n} \mathrm{Coef_i} * \mathrm{x_i},$$

where $\mathrm{Coef_i}$ is the coefficient, and $\mathrm{x_i}$ is the relative z-score-transformed gene expression of selected genes. We used the above formula to calculate a risk score for each HCC patient in the training (75% of TCGA patients) and internal validation (25% of TCGA patients) datasets.

Additionally, the loss and gain levels of CNVs have been confirmed using segmentation analysis and Genomic Identification of Significant Targets in Cancer algorithm (GISTIC) for CNV. To explore the clinical significance of the CNV or mutation, this TCGA cohort was divided into two HCC groups: "with mutation and/or CNV of 10 m^{6}A regulatory" and

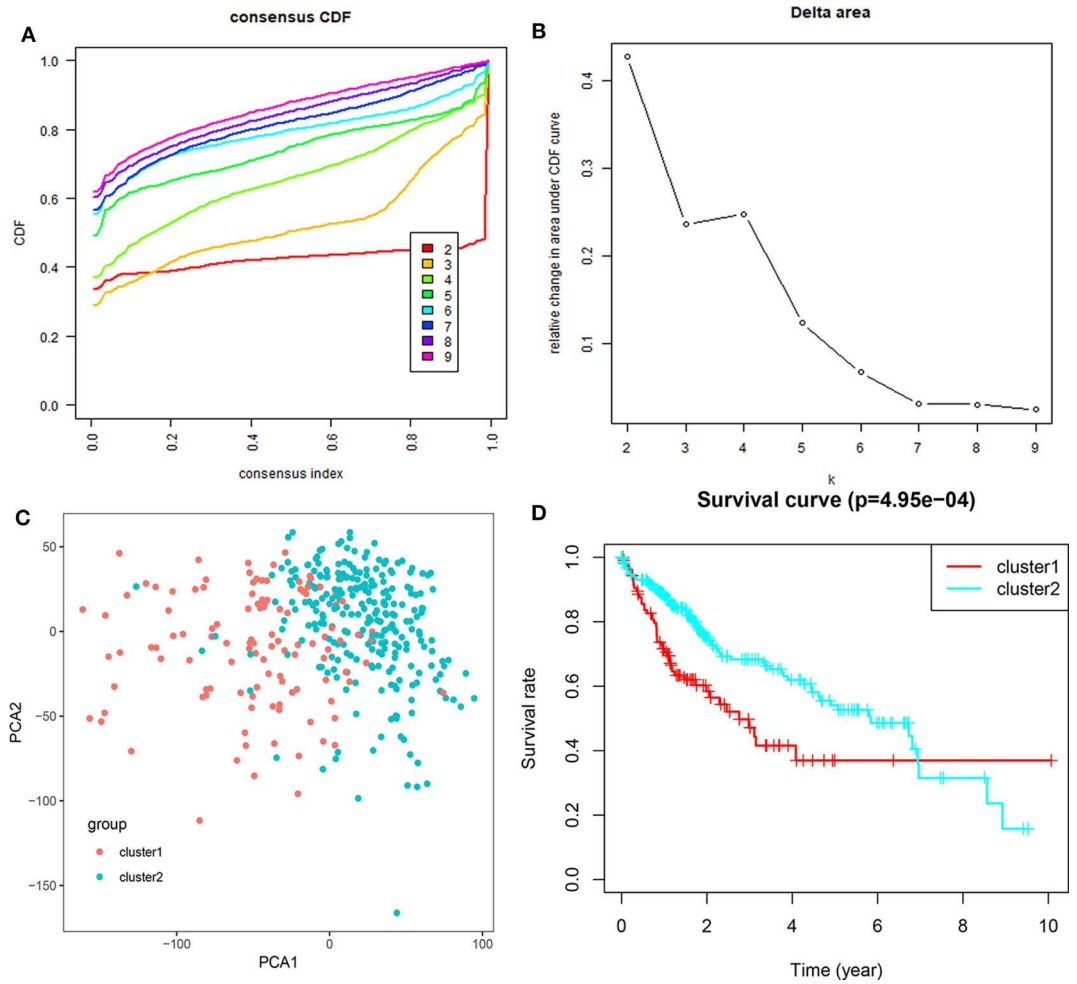

FIGURE 2 | Kaplan-Meier curves for overall and disease-free survival of TCGA HCC patients by the presence and absence of **(A,B)** mutation and/or CNVs of m^{6}A regulatory genes, **(C,D)** amplification/copy number gain of VIRMA.

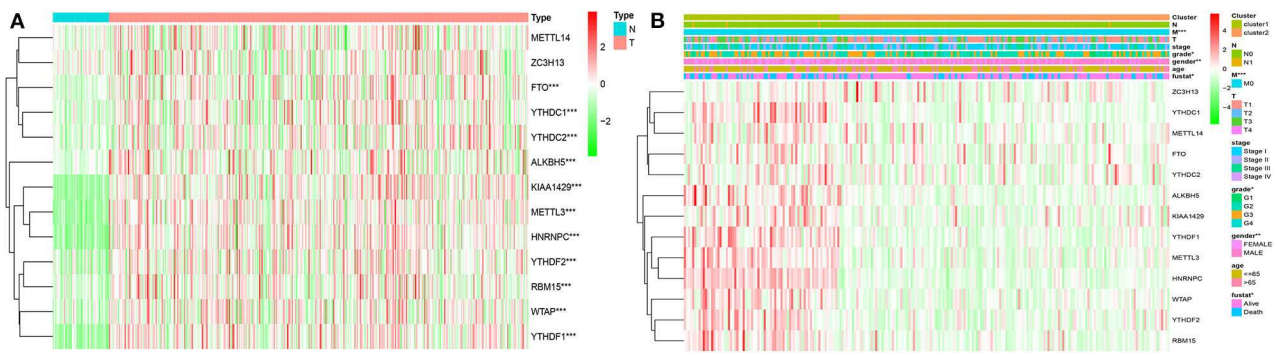

FIGURE 3 | Expression of m^{6}A RNA methylation regulators with different clinicopathological features in HCC. **(A)** stratified by tumor and normal samples; **(B)** stratified by cluster1/2 groups. *P < 0.01; **P < 0.001; ***P < 0.0001.

"without CNV and mutation." We calculated the gene expression from RNA-Seq V2 RSEM release, and used log scale before analyzing the associations between gene expression and CNVs.

Categorical variables applying the chi-square test or Fisher's exact test were compared. A Fisher-Freeman-Halton test was performed for contingency tables that are larger than 2 × 2.

We used the Mann-Whitney *U*-test for comparing continuous variables that were not distributed normally. The Kaplan-Meier analysis and the log-rank test were used to estimate the distribution of overall (OS) and disease-free survival (DFS) and to compare differences between survival curves, respectively. We then performed multivariate analyses by applying the Cox proportional hazards regression models. Variables with $P < 0.05$ (log-rank test) in univariate analysis for OS were included in the further models. All statistical analyses were conducted by using the SPSS statistical package, version 24.0 (IBM, Corp.). In addition, because some comparisons were made using limited data, the statistical comparisons were correlated for multiple testing by R software. For all tests, statistical differences were considered as significant at $P < 0.05$ (two-sided).

RESULTS

Genetic Alterations of m^{6}A Regulators Predict Poorer Survival in HCC

Mutations of m^{6}A regulatory genes were found in 8.5% (31/366) of HCC (**Figure 1A**). We identified gene variation in copy number in 23.0% (84/366) of patients (**Table S1**, detailed information in **Table S2**). A comparable frequency of copy number loss which is measured as shallow deletion by using GISTIC ($n = 33$) and copy number gain ($n = 46$) of m^{6}A genes (**Figure 1B**). Copy number loss of VIRMA is the most frequent among the HCC cohort (59/366, 16%; **Figure S1**). Notably, 6.3% (23/366) of HCC patients obtained copy number loss and gain of more than one m^{6}A genes simultaneously (**Table S3**). In seven of those 23 cases, an amplification of an m^{6}A writer or m^{6}A reader was investigated with a shallow or deep deletion of m^{6}A eraser genes concomitantly (**Table S3**), which indicated a potential synergistic change of m^{6}A enzymes that may result in increased levels of m^{6}A RNA modification. Shallow deletions of ALKBH5 and FTO conferred reduced mRNA expression of these m^{6}A genes significantly (**Figures 1C,D**). While copy number gain of METTL3, VIRMA, YTHDF1 and YTHDF2 conferred a significant increase in its RNA expression (**Figure S2**). Hence, copy number gain and shallow deletion might lead to the decreased and increased mRNA expression of m^{6}A genes.

Genetic Changes of m^{6}A Genes Were Correlated to Clinicopathological and Molecular Characteristics

It was determined whether mutations or copy number variations (CNVs) of m^{6}A regulators are correlated with HCC clinical and molecular characteristics. Mutations or CNVs of METTL14, METTL3, VIRMA, RBM15, ZC3H13, WTAP, YTHDC1, YTHDC2, YTHDF1, YTHDF2, HNRNPC, FTO, and ALKBH5 as a subgroup were significantly correlated with lower albumin ($P = 0.038$, **Table S1**), and poorer tumor stage in HCC ($P < 0.0043$, **Table S1**). In addition, we observed a significant increase in the status of TP53 mutations ($P < 0.012$, **Table S1**), TERT mutations ($P = 0.018$, **Table S1**), and ARID1A mutations ($P = 0.047$, **Table S1**) in HCC patients obtaining genetic changes of m^{6}A genes. These above molecular characteristics also were correlate with mutations of m^{6}A regulators alone, except for TERT mutations (**Table S1**). However, TERT ($P = 0.011$) and ARID1A mutations (0.0037), except for TP53 mutations (26) were also correlate with CNVs of m^{6}A regulators alone (**Table S1**), which might be because of the small sample sizes for mutations and CNVs associated with most mutated HCC genes.

We also determined whether shallow or deep deletion of VIRMA is correlate with the clinical and molecular characteristics. Consistent with our results for all m^{6}A regulators, amplification/copy number gain of VIRMA was significantly correlated with poorer clinical stage and the presence of TP53 and TERT mutation in this HCC cohort ($P = 0.04$, $P = 0.027$, respectively; **Table S4**). TP53 and TERT mutations were also present in HCC patients with amplification/copy number gain of VIRMA (**Table S4**). Kaplan-Meier analyses evaluating the impact of genetic changes for m^{6}A regulators on OS and DFS in HCC patients were performed. As a group, HCC patients with the mutation of any of m^{6}A regulators had a worse OS ($P = 0.004$) and DFS ($P = 0.0148$, **Figures 2A,B**). In addition, unfavorable OS were prominent in HCC patients who had amplification/copy number gain of VIRMA (**Figures 2C,D**).

Of all clinical and molecular characteristics regarded as the *de novo* HCC cohort, higher grade T classification ($T>0$), higher grade of stage and ARID1A, TP53 mutations and genetic alterations of any m^{6}A genes were associated with poorer OS in univariate Cox analysis significantly (**Table S5**). Therefore, we investigated the impact of m^{6}A gene mutations or CNVs on the HCC outcome with poor clinical characteristics. Changes of m^{6}A genes as a subgroup were correlated with poorer OS in HCC patients regardless of stage, TP53, and ARID1A mutations (**Table S6**). These genetic alterations did not confer a worse OS in patients with higher grade T classification ($T>0$), tumor grade or ARID2, TERT mutations (**Table S6**). We then identified the HCC patients' survival according to whether they showed combined TP53 mutations and genetic changes of m^{6}A genes. Almost half of the patients with mutated TP53 (40%, **Table S1**) had ≥1 genetic change of m^{6}A gene. We further explored the gene expression of m^{6}A regulatory genes between wild-type and mutated TP53. Interestingly, we found the m^{6}A eraser genes showed significantly lower expression in mutated TP53 group than the wild-type group, but the m^{6}A writer and reader genes showed significant higher expression in TP53 mutated group. It indicates that higher m^{6}A levels among HCC patients were parallel with the mutation rates of TP53 during the carcinogenesis or initiation in HCC. The group of TP53 mutated patients had poorer OS than those patients not obtaining any of these genetic changes (**Table S7**).

A Strong Correlation Between Genetic Alterations of m^{6}A Genes and TP53 Mutations

Since mutations, amplifications, deletions, and/or CNVs of m^{6}A genes were relatively restricted to patients with wild-type TERT and ARID1A (88.5%, **Table S1**), we then identified whether these genetic changes affect OS stratified by the status of TERT or ARID1A mutation. Poorer OS were seen in HCC patients

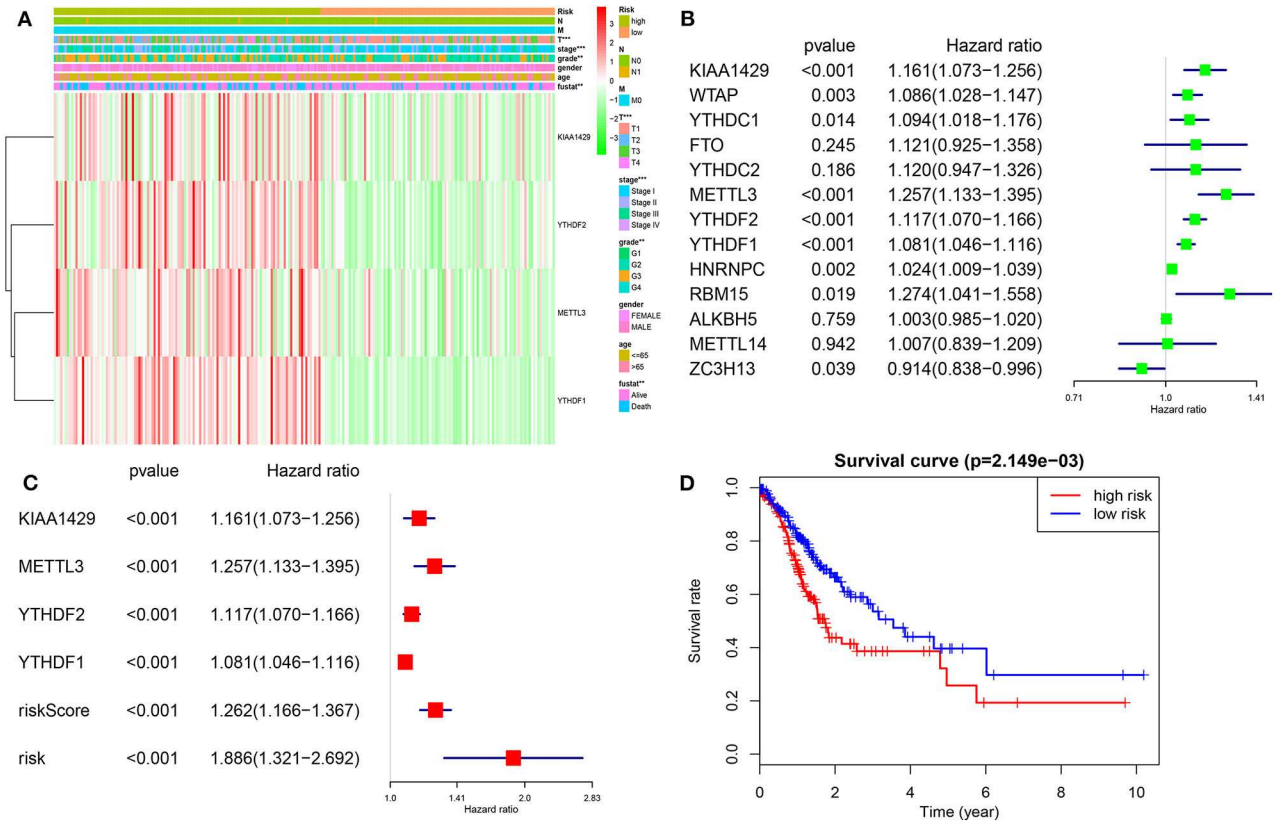

FIGURE 4 | Differential clinicopathological features and overall survival of HCC in the cluster1/2 subgroups. **(A)** Consensus clustering cumulative distribution function (CDF) for $k = 2$–10. **(B)** Relative change in area under CDF curve for $k = 2$–10. **(C)** Principal component analysis of the total RNA expression profile in the TGGA dataset. **(D)** Kaplan–Meier overall survival (OS) curves for HCC patients according to cluster1/2.

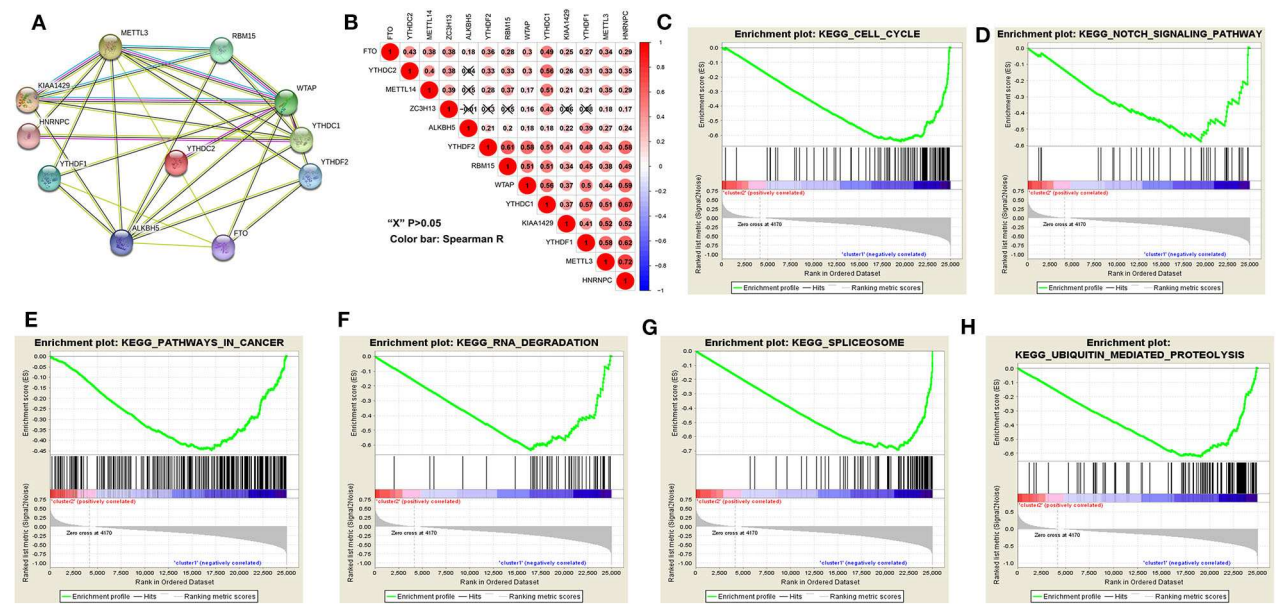

FIGURE 5 | Interaction among m^{6}A RNA methylation regulators and pathway analysis of HCC in cluster1/2 subgroups. **(A)** The m^{6}A modification-related interactions among the 11-m^{6}A RNA methylation regulators. **(B)** Spearman correlation analysis of the 13-m^{6}A modification regulators. **(C–H)** GSEA revealed that genes with higher expression in cluster1 subgroup were enriched for hallmarks of malignant tumors.

with wild-type TERT who had more than one genetic change of m^{6}A genes ($P = 0.001$, **Table S7**). Notably, these patients also have worse OS ($P < 0.05$) compared to those patients who had mutated TERT but had no genetic changes of m^{6}A genes (**Table S7**). Genetic changes of m^{6}A genes as a subgroup were also significantly associated with a worse OS in wild-type ARID1A patients ($P = 0.009$, **Table S7**). A combination of molecular analysis of m^{6}A genes might be valuable to identify a worse outcome in HCC patients who had neither been classified as "poor risk" because of the presence of mutated TERT (27), nor better survival outcome conferred by ARID1A mutations (28, 29), particularly in the TP53 wild-type HCC patients. From a multivariate Cox analysis including variables correlated with worse survival, genetic changes of m^{6}A genes, as a subgroup was not an independent factor for OS (**Table S8**). Genetic changes of m^{6}A genes, however, predicted poorer OS independently (HR = 1.8; 95% CI, 1.2–4.4; $P = 0.010$) when the variable TP53 mutation was excluded from the original model (**Table S8**).

Consensus Clustering of m^{6}A RNA Methylation Regulators Identified Two Clusters of HCC With Distinct Clinical Outcomes and Characteristics

Considering the important biological functions of each m^{6}A RNA genes in the HCC tumorigenesis, we explored the mRNA expression of each m^{6}A regulators between tumor and normal tissues systematically. The expression level of each m^{6}A RNA methylation regulator between tumor and normal are presented as heatmaps (**Figures 3A,B**), showing that the expressions of most m^{6}A RNA methylation regulators are significantly higher in tumor than normal group (**Figure 3** and **Figure S3**), except for METTL14 and ZC3H13 (but the expression still trends higher among tumor group than normal). We then validated these regulatory genes in our 69-paired HCC tumor and normal tissues by RT-PCR, which showed the same results as in the TCGA dataset (data were not shown).

Based on the expression similarity of m^{6}A RNA regulators, k =3 seemed to be a suitable selection with clustering stability increasing from $k = 2$ to 10 in the TCGA datasets (**Figure 4**; **Figure S4**). However, we found that 116 out of 371 HCC patients clustered into one of these two groups in the TCGA dataset (**Figure 3B**). Hence, we compared the clinical characteristics of these two groups clustered by $k = 2$, including cluster1 and cluster2 (**Figure 4D**). The cluster2 subgroup is significantly correlated with no metastasis ($P < 0.0001$), lower grade ($P < 0.01$), and lower stage ($P < 0.0001$; **Figure 3B**). The cluster1 subgroup mainly contains HCC with higher grade and clinical stage at diagnosis. In addition, we observed a shorter OS in the cluster1 subgroup than the cluster2 group (median survival: 76.1 months vs. 90.1 months) (**Figure 4D**).

Categories Identified by Consensus Clustering Are Closely Correlate to the Malignancy of HCC

The above results indicated that the clustering results were closely associated with the malignancy of the HCC. To better comprehend the interactions among these m^{6}A RNA methylation regulators, which is significantly differentially expressed in tumor and normal tissues, we also analyzed the interaction (**Figure 5A**) and correlation (**Figure 5B**) among these gene regulators. WTAP, METTL3 and ALKBH5 seems to be the hub gene of the "writers" and "erasers," and WTAP's interactions or co-expressions with RMB15, YTHDC1, YTHDF1, YTHDF2, YTHDC2, METTL3, KIAA1429, are validated by experimental data and by text mining in the String database (**Figure 5A**). In addition, METTL3's interactions or co-expressions with ALKBH5, WTAP, YTHDF1, RBM15, KIAA1429, YTHDC1, HNRNPC and FTO. The expression of WTAP was significantly associated with the "readers" of YTHDC1, YTHDF2, HNRNPC and YTHDF2 in HCC (**Figure 5B**). The expression of METTL3 was also significantly associated with "readers" of HNRNPC and YTHDC1 ($P < 0.05$). All these 11 differentially expressed m^{6}A regulatory genes correlated each other, suggesting complicated mechanism underlying each interaction groups in HCC (**Figure 3A**), but the expressions of YTHDF2 were not significantly correlated with YTHDC2 and METTL14 in HCC (**Figure 5B**). Moreover, the expressions of all m^{6}A regulatory genes were positively correlated with each other in HCC (**Figure 5B**). These findings were consistent with the expression levels of WTAP, METTL3, RBM15, KIAA1429, YTHDF2, YTHDF1, FTO, and ALKBH5 being positively correlated with the increasing HCC malignancy.

We then further applied principal component analysis (PCA) to compare the transcriptional profile between cluster1 and cluster2 groups. The findings revealed a clear difference between them (**Figure 4C**). We determined genes that were upregulated significantly (Score (d) for SAM > 8, fold change >2, and normalized $P < 0.01$) or downregulated [Score(d) for SAM < -8, fold change < 0.5, and normalized $P < 0.01$] in the cluster1 group, and then annotated their functions by pathway analysis for biological processes (**Figures 5C–H**).

The gene set enrichment analysis (GSEA) showed that cell cycle, notch signaling pathway, splicesome, ubiquitin mediated proteolysis and pathways in cancer were significantly associated with the cluster1 (**Figures 5C–H**, **Table S9**), while for cluster2, GSEA results showed metabolism-related pathways, including primary bile acid biosynthesis, drug metabolism and tryptophan metabolism (**Table S10**). All of these results showed that the two clusters determined by consensus clustering are closely associated with the carcinogenesis of HCC.

Prognostic Value of m^{6}A RNA Methylation Genes and a Risk Signature Built by Four-Selected m^{6}A RNA Regulators

We then investigated the prognostic value of m^{6}A RNA regulators in HCC. We conducted a univariate Cox analysis on the gene expressions in the TCGA dataset (**Figure 6**). The results revealed that nine out of thirteen genes are significantly associated with OS ($P < 0.05$). Among these nine genes (**Figure 6B**), all the KIAA1429, WTAP, METLL3, YTHDC1, YTHDF2, YTHDF1, HNRNPC, RBM15, and ZC3H13 are risky genes with HR > 1. Furthermore, to better predict the HCC

survivals with m^{6}A RNA regulators, we applied the machine-learning model (Coxboost regression) to the nine prognosis-associated genes (**Figure S3B**) in the 75% TCGA dataset, which was used as a training cohort (**Figure 6A**). Four genes, including METTL3, KIAA1429, YTHDF2 and YTHDF1, were then selected to construct the risk signature according to the minimum criteria, and the coefficients derived from the Coxboost algorithm were applied to calculate the risk score for both the training dataset (75% TCGA) and the validation dataset (25% TCGA). To evaluate the prognostic value of the seven-m^{6}A risk signature, we divided the HCC patients in the training set ($n = 278$) and validation set ($n = 93$) into low- and high-risk groups according to the median value of risk score and saw the significant differences in OS between the two clusters (both $P = 0.002$; **Figure 6A**). In addition, the prognostic value of m6Ascore are also prominent in another two GEO datasets (all $P < 0.05$; **Figure S7**). The heatmap shows the expression of the four selected m^{6}A RNA regulators in the two groups including high- or low-risk HCC patients in the TCGA dataset (**Figure 6A**). The significant differences between the high and low risk groups with regards to metastasis ($P < 0.001$), tumor grade ($P < 0.01$) and clinical stage ($P < 0.001$) can be seen. The multivariate analysis showed the m^{6}Ascore signature is an independent factor for OS in HCC patients (HR= 1.886, 95%CI 1.321–2.692, $P < 0.001$; **Figure 6C**). The ROC curve analysis showed that the m^{6}Ascore can predict overall survival very well at 2,000 days (AUC = 0.70), 3-year survival (AUC = 0.69), cluster1/2 subgroups (AUC = 0.87) and TP53 mutations (AUC = 0.68; **Figure S5**). Furthermore, we explored whether m^{6}Ascore could discriminate distinct survival stratified by TP53 mutations. The results showed m^{6}Ascore can predict overall survival at 2,000 days well in TP53 mutations (AUC = 0.70) and TP53 wild-type (AUC = 0.61). These findings showed that the risk scores calculated by that signature could accurately predict HCC patient clinical outcomes and characteristics, especially for the cluster1/2 groups.

DISCUSSION

It remains a major challenge that identifying new molecular biomarkers tutors the evolvement of anti-HCC treatments. Our findings favored an evident association between genetic changes of m^{6}A genes and TP53 mutation. One is confounding the other in predicting the prognosis of an HCC patient, which indicates that both might be complementary in the HCC pathogenesis or maintenance. The molecular biomarkers to identify tumor subtypes and patients prognosis still demand continuous refinement (30, 31). Regarding that, the m^{6}A modification to mRNA owned wide biological functions; its impairment might be correlated with the progression of HCC. The current WHO classification emphasized epigenetic modifiers during the process of HCC clonal evolution as being mutated (30, 32, 33). Novel genetic subgroups embrace gene mutations that encode TP53 and epigenetic modifiers (27, 30, 32, 33).

Our research is the first to identify some clinical associations and effect of genetic changes influencing m^{6}A genes in HCC. Although one previous study has showed that some m^{6}A regulators, such as METTL3 and YTHDF1 were upregulated in HCC, and they were independent poor prognostic factors (34), we not only demonstrated that the expression of combined m^{6}A regulators genes is also closely correlated with the prognosis of HCC, but also showed that a remarkable correlation between genetic changes of those m^{6}A regulators as a whole group and the status of TP53 mutations (**Table S1**). More importantly, genetic changes of m^{6}A regulators correlated with poorer clinical prognosis in HCC patients, even though this might be confounded by the unfavorable effects of the status of TP53 mutations on HCC survival (35). It has been revealed that the loss of METTL3 lead to alternative mRNA splicing and mRNA expression changes of more than 20 genes which are involved in the TP53 signaling pathway including MDM2, and P21 in HCC (4, 35). It is reasonable that genetic changes of m^{6}A genes, TP53, or its regulator/downstream-molecular targets result in complementary pathways to the HCC pathogenesis. Hence, further studies in larger HCC cohorts could help confirm our results and spur future research into the functional effects of m^{6}A RNA modification in HCC and its association with carcinogenesis pathways, especially for TP53 signaling.

Because our study has showed genetic changes of m^{6}A genes, giving a deeper insight into their mechanism and link to HCC tumorigenesis pathways, the expressions of regulatory genes associated with clinical characteristics and its prognostic value have not been explored. Hence, we firstly determined two HCC subgroups, cluster1/2, by consensus clustering according to the mRNA expression of m^{6}A regulators. The subgroups of cluster1/2 not only affect the HCC prognosis but also were closely associated with functional processes and cancer signaling pathways. Additionally, we established a novel prognostic risk signature with four m^{6}A RNA regulators, stratifying the OS with HCC into high- and low-risk groups.

Because of the tumor tissue specificity of the "writers," "erasers," and "readers," these genes involved in m^{6}A dysregulation would be diverse in different cancers (32, 36). Among the m^{6}A RNA methylation regulators, the writer METTL3 is often highly expressed in tumors and contributes to HCC tumorigenesis (19), which is consistent with our results; while METTL14, down regulated in HCC, acts as an unfavorable prognostic factor for HCC (8, 19). The reader YTHDF2 and the eraser FTO promotes cancer cell proliferation in pancreatic cancer and glioma (17, 37, 38). These results indicated that upregulation or downregulation of any m^{6}A methylation regulators are associated with deregulated RNAs in cancers, and the same m^{6}A regulators might have different biological functions in various cancers (9, 20, 37, 39).

In our HCC cohort, the frequency of genetic changes of the 13-m^{6}A genes was much higher than that showed in AML (7), suggesting that the dysregulation of m^{6}A may play a vital role in HCC carcinogenesis. Additionally, there was a high frequency of concurrent genetic changes of two regulatory genes, suggesting that m^{6}A writer gene and reader gene might play a synergistic role during the process of RNA m^{6}A modification (14, 20, 37, 40). In TP53-mutated samples, the expression of writer and reader

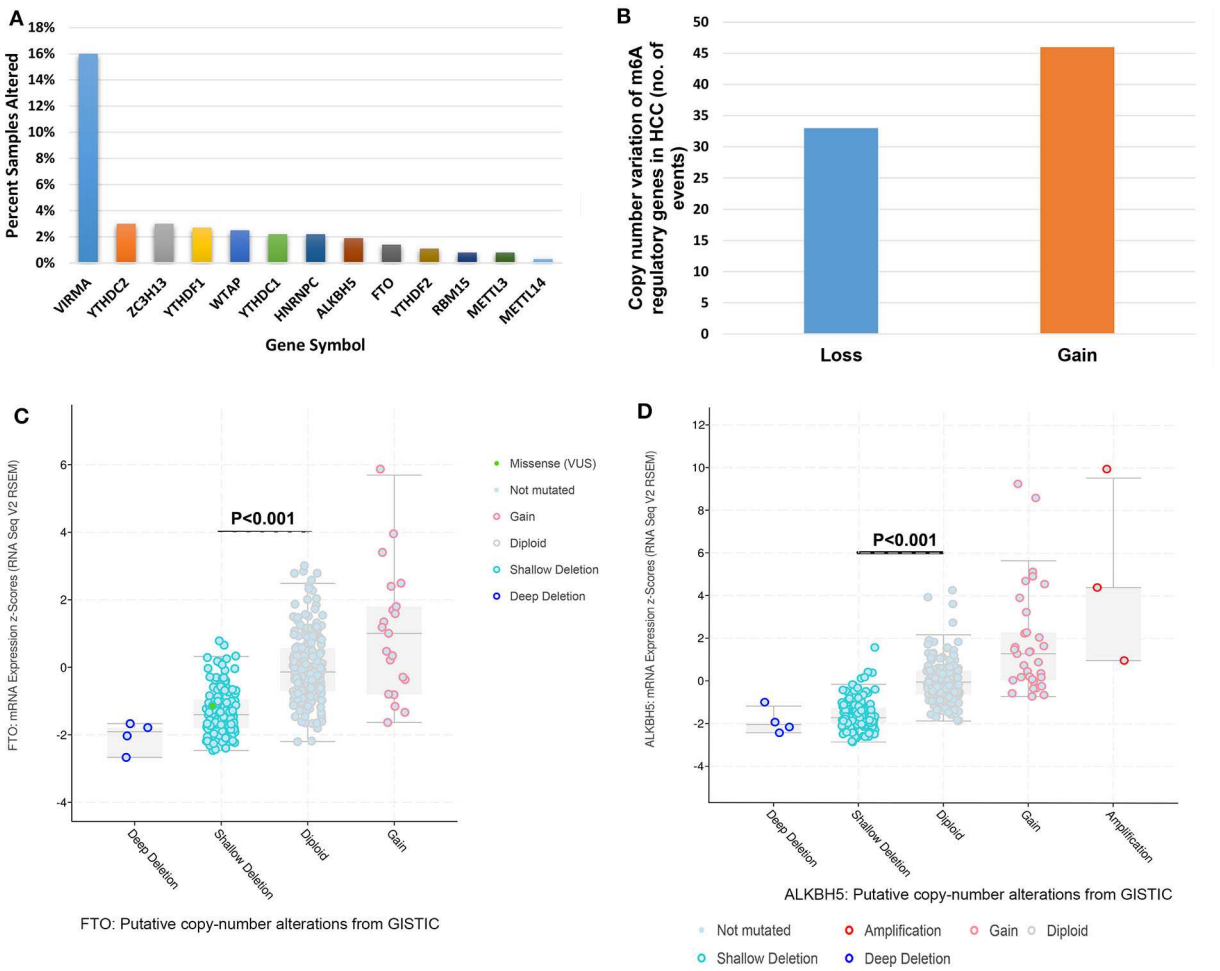

FIGURE 6 | Risk signature with four m^{6}A RNA methylation regulators. **(A)** Heatmap and clinicopathologic features of the two groups (risk low/high) defined by the m^{6}A RNA methylation regulators consensus expression. **(B)** The process of building the signature containing 13 m^{6}A RNA methylation regulators. The hazard ratios (HR), 95% confidence intervals (CI) calculated by univariate Cox regression. **(C)** The process of building the signature containing four m^{6}A RNA methylation regulators calculated by machine learning model. **(D)** Kaplan–Meier overall survival (OS) curves for HCC patients according to m^{6}Ascore. *$P < 0.01$; **$P < 0.001$; ***$P < 0.0001$.

genes, such as METTL3, RBM15, YTHDF1, and YTHDF2, were higher than wild-type samples, while the eraser genes, FTO and ALKBH5 were lower than the wild-type group (**Figure S6**). It indicates that the levels of m^{6}A may correlated with the rate of TP53 mutations in HCC.

Unlike the CNVs in AML, most of the CNVs in writer and reader genes lead to the gain of function with up-regulation of the corresponding genes, while CNVs of the eraser genes were mainly gaining function resulting in down-regulation of the relevant genes. Regarding the opposite effect on m^{6}A status for those two gene groups, these genetic changes increased the m^{6}A level in HCC. Consistent with our findings, many researches on other cancers, like colorectal and pancreatic cancer (6, 20, 31, 36, 38, 40, 41) have also observed the up-regulated m^{6}A level. This could be explained by the associations between m^{6}A and cellular differentiation pathways controlling cancer stem cell fate (29, 31, 42).

We also comprehensively analyzed the expression of all m^{6}A RNA regulators in HCC with different clinical characteristics. As an m^{6}A methylation writer, the expression of METLL3 was increased in tumor group, higher tumor grade and stage. WTAP expression was significantly increased in higher-grade and metastasis. For the m^{6}A methylation readers, the expression of HNRNPC, YTHDF1, and YTHDF2 was also significantly increased in higher tumor grade and stage. Interestingly, the expression of FTO was decreased in no metastasis, lower tumor grade and stage. Taken together, the expression of m^{6}A RNA regulators is closely correlated with malignant clinical characteristics in HCC. These results are also helpful for establishing new therapeutic methods through characterizing the expression of individual m^{6}A regulators in HCC, as chemicals targeting m^{6}A methylation are regarded as a novel method of cancer treatment (43).

Whether the expression level of m^{6}A RNA methylation genes could be applied as a prognostic biomarker plays a vital role in the research of cancers. In this study, our HCC prognostic signature derived using four m^{6}A RNA methylation regulators was found to stratify the OS for TP53 mutations and cluster1/2 subgroups. A

similar scenario was also seen in the multivariate Cox regression analysis. This might be caused by the strong association between the m^{6}Ascore and TP53 status.

CONCLUSIONS

Overall, our results systematically revealed the genetic changes, mRNA expression, potential biological function, and clinical prognostic value of m^{6}A RNA methylation regulators in HCC. We observed a remarkable relationship between the genetic changes and different expressions lead to increased m^{6}A level and poorer clinical prognosis. It is reasonable that genetic changes of m^{6}A modifiers, TP53, or its regulator/downstream spots contribute in complementary pathways to the pathogenesis of HCC. Additionally, the expressions of m^{6}A RNA methylation regulators are associated with the increased expression levels of genes significantly enriched in the biological processes and cancer signaling pathways that facilitate the HCC malignant progression. Finally, our research confers vital evidence for future investigation of the role of mRNA m^{6}A methylation in HCC.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by TCGA database. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

G-QZ, LY, Y-JZ, J-XD, Y-MW, and ZD designed the study. G-QZ, Y-MW, J-XD collected and downloaded data. LY, Y-JZ did the statistical analyses. Y-HS and S-SD prepared figures. G-QZ, LY, JF, JZ, S-SD, and Y-HS reviewed the results, interpreted data, and wrote the manuscript. All authors have made an intellectual contribution to the manuscript and approved the submission.

FUNDING

This work was supported by the National Natural Science Fund of China (Grant numbers: 81672330; 81871916).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00900/full#supplementary-material

Figure S1 | Genetic alteration spectrum of m^{6}A regulatory genes for HCC.

Figure S2 | Relationships between putative copy-number alterations and m^{6}A genes' mRNA expression.

Figure S3 | Gene expression of m^{6}A regulatory enzymes and booting step by machine learning model.

Figure S4 | Consensus clustering cumulative distribution function (CDF) for k = 2–10.

Figure S5 | ROC curves with AUCs of prognostic predictors built by m^{6}Ascore calculated by m^{6}A regulatory genes.

Figure S6 | Gene expression of m^{6}A regulatory genes between TP53 mutated and wild-type sample.

Figure S7 | Validation of prognostic value of m^{6}A score in two GEO datasets.

Table S1 | Clinical and molecular characteristics of TCGA HCC patients according to the mutation and/or copy number variation status of genes encoding m^{6}A regulatory enzymes.

Table S2 | Genetic alterations of 13 m^{6}A regulatory genes in HCC.

Table S3 | HCC samples with a point mutation, deep deletion, amplification, shallow deletion and/or copy number gain of one or more genes encoding m^{6}A regulatory enzymes. *Examples of potentially synergistic changes that may increase RNA m^{6}A levels.

Table S4 | Clinical and molecular characteristics of HCC patients with an amplification or copy number gain of the gene encoding an m^{6}A writer, VIRMA.

Table S5 | Univariate analysis for clinicopathological and molecular features.

Table S6 | Subgroup analysis for alterations of m^{6}A regulatory genes as a group associated with overall survival.

Table S7 | Univariate analysis for overall survival of patients stratified by the status of m^{6}A regulatory gene alterations in addition to (A) TP53, (B) TERT, and (C) ARID1A mutation status.

Table S8 | Multivariate analysis for overall survival in HCC patients.

Table S9 | Pathway analysis for cluster1 in HCC.

Table S10 | Pathway analysis for cluster2 in HCC.

REFERENCES

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin.* (2011) 61:69–90. doi: 10.3322/caac.20107
2. Villanueva A. Hepatocellular carcinoma. *N Eng J Med.* (2019) 380:1450–62. doi: 10.1056/NEJMra1713263
3. Liu N, Pan T. N6-methyladenosine-encoded epitranscriptomics. *Nat Struc Mol Biol.* (2016) 23:98–102. doi: 10.1038/nsmb.3162
4. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, et al. Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature.* (2012) 485:201–6. doi: 10.1038/nature11112
5. Lin X, Chai G, Wu Y, Li J, Chen F, Liu J, et al. RNA m(6)A methylation regulates the epithelial mesenchymal transition of cancer cells and translation of snail. *Nat Commun.* (2019) 10:2065. doi: 10.1038/s41467-019-09865-9
6. Ding C, Zou Q, Ding J, Ling M, Wang W, Li H, et al. Increased N6-methyladenosine causes infertility is associated with FTO expression. *J Cell Physiol.* (2018) 233:7055–66. doi: 10.1002/jcp.26507
7. Kwok CT, Marshall AD, Rasko JE, Wong JJ. Genetic alterations of m(6)A regulators predict poorer survival in acute myeloid leukemia. *J Hematol Oncol.* (2017) 10:39. doi: 10.1186/s13045-017-0410-6
8. Scholler E, Weichmann F, Treiber T, Ringle S, Treiber N, Flatley A, et al. Interactions, localization, and phosphorylation of the m(6)A generating METTL3-METTL14-WTAP complex. *RNA.* (2018) 24:499–512. doi: 10.1261/rna.064063.117
9. Tang C, Klukovich R, Peng H, Wang Z, Yu T, Zhang Y, et al. ALKBH5-dependent m^{6}A demethylation controls splicing and stability of long 3′-UTR

mRNAs in male germ cells. *Proc Natl Acad Sci USA.* (2018) 115:E325–33. doi: 10.1073/pnas.1717794115
10. Wojtas MN, Pandey RR, Mendel M, Homolka D, Sachidanandam R, Pillai RS. Regulation of m(6)a transcripts by the 3′–>5′ RNA helicase YTHDC2 is essential for a successful meiotic program in the mammalian germline. *Mol Cell.* (2017) 68:374–87.e12. doi: 10.1016/j.molcel.2017.09.021
11. Wu R, Yao Y, Jiang Q, Cai M, Liu Q, Wang Y, et al. Epigallocatechin gallate targets FTO and inhibits adipogenesis in an mRNA m(6)A-YTHDF2-dependent manner. *Int J Obesity.* (2018) 42:1378–88. doi: 10.1038/s41366-018-0082-5
12. Yang Y, Hsu PJ, Chen YS, Yang YG. Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. *Cell Res.* (2018) 28:616–24. doi: 10.1038/s41422-018-0040-8
13. Cui Q, Shi H, Ye P, Li L, Qu Q, Sun G, et al. m(6)A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Rep.* (2017) 18:2622–34. doi: 10.1016/j.celrep.2017.02.059
14. Dai D, Wang H, Zhu L, Jin H, Wang X. N6-methyladenosine links RNA metabolism to cancer progression. *Cell Death Dis.* (2018) 9:124. doi: 10.1038/s41419-017-0129-x
15. Pan Y, Ma P, Liu Y, Li W, Shu Y. Multiple functions of m(6)A RNA methylation in cancer. *J Hematol Oncol.* (2018) 11:48. doi: 10.1186/s13045-018-0590-8
16. Wang S, Sun C, Li J, Zhang E, Ma Z, Xu W, et al. Roles of RNA methylation by means of N(6)-methyladenosine (m(6)A) in human cancers. *Cancer Lett.* (2017) 408:112–20. doi: 10.1016/j.canlet.2017.08.030
17. Zhang C, Cheng W, Ren X, Wang Z, Liu X, Li G, et al. Tumor purity as an underlying key factor in glioma. *Clin Cancer Res.* (2017) 23:6279–91. doi: 10.1158/1078-0432.CCR-16-2598
18. Liu J, Eckert MA, Harada BT, Liu SM, Lu Z, Yu K, et al. m(6)A mRNA methylation regulates AKT activity to promote the proliferation and tumorigenicity of endometrial cancer. *Nat Cell Biol.* (2018) 20:1074–83. doi: 10.1038/s41556-018-0174-4
19. Chen M, Wei L, Law CT, Tsang FH, Shen J, Cheng CL, et al. RNA N6-methyladenosine methyltransferase-like 3 promotes liver cancer progression through YTHDF2-dependent posttranscriptional silencing of SOCS2. *Hepatology.* (2018) 67:2254–70. doi: 10.1002/hep.29683
20. Weng H, Huang H, Wu H, Qin X, Zhao BS, Dong L, et al. METTL14 inhibits hematopoietic stem/progenitor differentiation and promotes leukemogenesis via mRNA m(6)A modification. *Cell Stem Cell.* (2018) 22:191–205.e9. doi: 10.1016/j.stem.2017.11.016
21. Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* (2010) 70:10202–12. doi: 10.1158/0008-5472.CAN-10-2607
22. Villanueva A, Portela A, Sayols S, Battiston C, Hoshida Y, Mendez-Gonzalez J, et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. *Hepatology.* (2015) 61:1945–56. doi: 10.1002/hep.27732
23. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery.* (2012) 2:401–4. doi: 10.1158/2159-8290.CD-12-0095
24. Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods.* (2014) 11:599–600. doi: 10.1038/nmeth.2956
25. Singal AG, Mukherjee A, Elmunzer BJ, Higgins PD, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am J Gastroenterol.* (2013) 108:1723–30. doi: 10.1038/ajg.2013.332
26. Pezzuto F, Buonaguro L, Buonaguro FM, Tornesello ML. Frequency and geographic distribution of TERT promoter mutations in primary hepatocellular carcinoma. *Infect Agents Cancer.* (2017) 12:27. doi: 10.1186/s13027-017-0138-5
27. Lee HW, Park TI, Jang SY, Park SY, Park WJ, Jung SJ, et al. Clinicopathological characteristics of TERT promoter mutation and telomere length in hepatocellular carcinoma. *Medicine.* (2017) 96:e5766. doi: 10.1097/MD.0000000000005766
28. Bai Y, Yang C, Wu R, Huang L, Song S, Li W, et al. YTHDF1 regulates tumorigenicity and cancer stem cell-like activity in human colorectal carcinoma. *Front. Oncol.* (2019) 9:332. doi: 10.3389/fonc.2019.00332
29. Otto JE, Kadoch C. A two-faced mSWI/SNF subunit: dual roles for ARID1A in tumor suppression and oncogenicity in the liver. *Cancer cell.* (2017) 32:542–3. doi: 10.1016/j.ccell.2017.10.014
30. Tan PS, Nakagawa S, Goossens N, Venkatesh A, Huang T, Ward SC, et al. Clinicopathological indices to predict hepatocellular carcinoma molecular classification. *Liver Int.* (2016) 36:108–18. doi: 10.1111/liv.12889
31. Barbieri I, Tzelepis K, Pandolfini L, Shi J, Millan-Zambrano G, Robson SC, et al. Promoter-bound METTL3 maintains myeloid leukaemia by m(6)A-dependent translation control. *Nature.* (2017) 552:126–31. doi: 10.1038/nature24678
32. Bou-Nader M, Caruso S, Donne R, Celton-Morizur S, Calderaro J, Gentric G, et al. Polyploidy spectrum: a new marker in HCC classification. *Gut.* (2019) 69:355–64. doi: 10.1136/gutjnl-2018-318021
33. Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouze E, Blanc JF, et al. Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. *J Hepatol.* (2017) 67:727–38. doi: 10.1016/j.jhep.2017.05.014
34. Zhou Y, Yin Z, Hou B, Yu M, Chen R, Jin H, et al. Expression profiles and prognostic significance of RNA N6-methyladenosine-related genes in patients with hepatocellular carcinoma: evidence from independent datasets. *Cancer Manag Res.* (2019) 11:3921–31. doi: 10.2147/CMAR.S191565
35. Lin KT, Ma WK, Scharner J, Liu YR, Krainer AR. A human-specific switch of alternatively spliced AFMID isoforms contributes to TP53 mutations and tumor recurrence in hepatocellular carcinoma. *Genome Res.* (2018) 28:275–84. doi: 10.1101/169029
36. Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, et al. Stem cells. m^{6}A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science.* (2015) 347:1002–6. doi: 10.1126/science.1261417
37. Su R, Dong L, Li C, Nachtergaele S, Wunderlich M, Qing Y, et al. R-2HG exhibits anti-tumor activity by targeting FTO/m(6)A/MYC/CEBPA signaling. *Cell.* (2018) 172:90–105.e23. doi: 10.1016/j.cell.2017.11.031
38. Zhang S, Zhao BS, Zhou A, Lin K, Zheng S, Lu Z, et al. m(6)A demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program. *Cancer cell.* (2017) 31:591–606.e6. doi: 10.1016/j.ccell.2017.02.013
39. Vu LP, Pickering BF, Cheng Y, Zaccara S, Nguyen D, Minuesa G, et al. The N(6)-methyladenosine (m(6)A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat Med.* (2017) 23:1369–76. doi: 10.1038/nm.4416
40. Song H, Feng X, Zhang H, Luo Y, Huang J, Lin M, et al. METTL3 and ALKBH5 oppositely regulate m(6)A modification of TFEB mRNA, which dictates the fate of hypoxia/reoxygenation-treated cardiomyocytes. *Autophagy*. (2019) 2019:1–19. doi: 10.1080/15548627.2019.1586246
41. Wu Y, Yang X, Chen Z, Tian L, Jiang G, Chen F, et al. m(6)A-induced lncRNA RP11 triggers the dissemination of colorectal cancer cells via upregulation of Zeb1. *Mol Cancer.* (2019) 18:87. doi: 10.1186/s12943-019-1014-2
42. Chen J, Wang C, Fei W, Fang X, Hu X. Epitranscriptomic m^{6}A modification in the stem cell field and its effects on cell death and survival. *Am J Cancer Res.* (2019) 9:752–64.
43. Deng X, Su R, Weng H, Huang H, Li Z, Chen J. RNA N(6)-methyladenosine modification in cancers: current status and perspectives. *Cell Res.* (2018) 28:507–17. doi: 10.1038/s41422-018-0034-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fimmu.2020.01583

Large-Scale Structure-Based Prediction of Stable Peptide Binding to Class I HLAs Using Random Forests

*Jayvee R. Abella[1], Dinler A. Antunes[1], Cecilia Clementi[2,3] and Lydia E. Kavraki[1]**

[1] Department of Computer Science, Rice University, Houston, TX, United States, [2] Center for Theoretical Biological Physics, Rice University, Houston, TX, United States, [3] Department of Chemistry, Rice University, Houston, TX, United States

Prediction of stable peptide binding to Class I HLAs is an important component for designing immunotherapies. While the best performing predictors are based on machine learning algorithms trained on peptide-HLA (pHLA) sequences, the use of structure for training predictors deserves further exploration. Given enough pHLA structures, a predictor based on the residue-residue interactions found in these structures has the potential to generalize for alleles with little or no experimental data. We have previously developed APE-Gen, a modeling approach able to produce pHLA structures in a scalable manner. In this work we use APE-Gen to model over 150,000 pHLA structures, the largest dataset of its kind, which were used to train a structure-based pan-allele model. We extract simple, homogenous features based on residue-residue distances between peptide and HLA, and build a random forest model for predicting stable pHLA binding. Our model achieves competitive AUROC values on leave-one-allele-out validation tests using significantly less data when compared to popular sequence-based methods. Additionally, our model offers an interpretation analysis that can reveal how the model composes the features to arrive at any given prediction. This interpretation analysis can be used to check if the model is in line with chemical intuition, and we showcase particular examples. Our work is a significant step toward using structure to achieve generalizable and more interpretable prediction for stable pHLA binding.

Keywords: structural modeling, random forests, machine learning, HLA-I, peptide binding, docking, immunopeptidomics, antigen presentation

OPEN ACCESS

Edited by:
George Bebis, University of Nevada, Reno, United States

Reviewed by:
Lin Wang, Stanford Healthcare, United States
Tin Nguyen, University of Nevada, Reno, United States

***Correspondence:**
Lydia E. Kavraki
kavraki@rice.edu

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: *31 January 2020*
Accepted: *15 June 2020*
Published: *22 July 2020*

Citation:
Abella JR, Antunes DA, Clementi C and Kavraki LE (2020) Large-Scale Structure-Based Prediction of Stable Peptide Binding to Class I HLAs Using Random Forests. Front. Immunol. 11:1583. doi: 10.3389/fimmu.2020.01583

1. INTRODUCTION

Class I Major histocompatibility complexes (MHCs), also known as Human Leukocyte Antigens (HLAs) for humans, are the major players in the endogenous peptide presenting pathway. In this pathway, HLA receptors are loaded with intracellular peptides of length 9–11 amino acids (1). If the binding is stable enough, the resulting peptide-HLA (pHLA) complexes will end up traveling from the endoplasmic reticulum (ER) all the way to the cell surface (2). Surveilling T-cells can inspect the pHLAs and engage an immune response, particularly when the inspected cell is diseased and is presenting immunogenic peptides (i.e., peptides capable of triggering T-cells for being somewhat unusual relative to self peptides). This mechanism is one of the ways personalized immunotherapy has been used to attack tumor cells, through for example, finding T-cells that can target tumor-specific peptides being presented by the patient's HLAs (3).

Inside the ER, peptides are loaded onto the HLA, which is in a peptide-receptive state (4). Thus, peptides that make it to this stage are "trial" bound to the HLA, and only stable complexes make it to the cell surface. Experimentally, mass spectrometry (MS) can be used in combination with acid elution to identify the peptides that are found on the cell surface. Therefore, peptides found with MS can be assumed to be stable (5). However, in the context of personalized immunotherapy, being able to run a fast and accurate computational screening for stable pHLAs would save time and reduce costs, thus prioritizing resources and contributing to better outcomes.

Most of the current methods are based on building a model to predict binders/non-binders using peptide sequences (6–8). NetMHC is the most popular method that trains a neural network to classify binders/non-binders using a dataset of experimental binding affinity measurements (7). It was later shown that binding stability (half-life) is a better measure for immunogenicity over binding affinity, and NetMHCstab was developed using a dataset of experimental half-life measurements (9). While stability may be a more relevant quantity to immunogenicity, more data is available for binding affinity (10), and so NetMHC remained a popular method for binder/non-binder classification. That is until the rise of availability of MS data (11), which can provide a plentiful source of peptide binders directly eluted from different cell types (12, 13). Methods such as NetMHCpan have also used MS data, along with data from experimental binding assays, to enhance the prediction of binding affinity (14). Peptides found with MS can be assumed to have somewhat high binding affinities, since low affinity binders are probably lost during the steps preceding peptide elution. In addition, peptides found with MS can also be assumed to be stable binders, since the pHLA complex must have been stable enough to make it to the cell surface. Hence, methods that use MS data for training are implicitly predicting both affinity and stability. In this work, a large portion of our dataset is derived from MS data for training a predictor of HLA-binders from structure.

Prediction models are typically built on a per-allele basis, so that the only information required for training are peptide sequences known to bind to that specific allele (7, 8). Approaches, such as NetMHC, are then restricted to alleles which have sufficient experimental data. In this work, we are interested in developing a *single* model to classify binders from non-binders across any allele. Such models are also known as pan-allele models and are trained using all of the available pHLA data as one dataset. Thus, pan-allele models have the potential to generalize across alleles and provide accurate predictions for a given pHLA even if little or no experimental data exists for the particular HLA. Sequence-based methods, such as NetMHCstabpan and NetMHCpan, have been developed for this purpose (14, 15). However, pHLA structures could also form the basis for generalizable models, which could work for any allele. If the stability of a pHLA complex is most directly influenced by the chemical interactions found in the structure, then a machine learning algorithm can be used to map these interactions to stability. The information that is specific for a given allele is implicitly encoded in the interactions found in the structure, so any pHLA structure can be treated in the same way during training (i.e., no sequence or structural alignment is required). In other words, machine learning models developed in this formulation are automatically pan-allele.

Another benefit of the model described in this work, which combines the use of structural information with less complex machine learning methods, is greater interpretability. While sequence-based methods such as NetMHCstabpan and NetMHCpan produce highly accurate predictions, these models are not particularly interpretable since they rely on neural networks. Neural networks arrive at a particular output through repeated, non-linear operations, starting from the input features. Thus, it is difficult to analyze the contribution of a particular feature toward any given prediction. However, machine learning models with less complexity, such as random forests, allow more interpretation (16). In turn, the ability to assess the contributions of particular features and mapping these contributions back to the input pHLA structures can be a powerful tool for checking whether the model is in line with chemical intuition.

Although the use of structural information to create generalizable HLA-binding prediction methods has been pursued by many groups in the past, these efforts have been greatly impaired by the computational difficulty in modeling pHLA structures (17). In addition, the massive number of possible combinations involving different HLA alleles and peptide sequences is significantly greater than the number of pHLA crystal structures determined experimentally (e.g., less than 1,000 structures available in the PDB at the time of this writing). Therefore, the development of structure-based binding prediction methods requires large-scale modeling of pHLA complexes. Unfortunately, previously available approaches for generating pHLA structures either do so in a simplistic manner (e.g., peptide threading) (18) or require running for long times per structure, which renders large-scale modeling infeasible (19–22).

Once a 3D structure has been generated for a given pHLA (e.g., through some type of sampling), it is usually passed to a scoring function, which is a sum of energy-related terms aimed at quantifying the binding strength. The weights of these scoring functions can be optimized for pHLAs, (21) or even for specific HLA alleles (23). For instance, structural features based on energy-related terms from the Rosetta scoring function (along with sequence features) were used as input for machine learning, and applied to a training set of 1,000 structures for a single MHC allele (24). Alternatively, simulations of pHLA structures have also produced accurate binding predictions (25). Methods based on molecular dynamics, such as PB/GBSA, have been used to assess binding strength (26). Monte Carlo approaches, such as the one available in the Rosetta package, have also been used to characterize peptide binding profiles for a given allele (27). Unfortunately, simulation approaches are even more computationally expensive than aforementioned modeling methods, also preventing their use in a large scale. Therefore, more research is needed in using a purely machine learning approach to map structures onto binding strength predictions, which will likely be enabled by the availability of large datasets of pHLA structures.

In this work, we use pHLA structures to predict stable binding. Ideally, a dataset of experimental half-life measurements (like in (15)) that spans multiple alleles would be used here in a regression framework, but such datasets are not readily available or easy to produce. Thus, we rely only on MS data for our source of stable peptide binders, and work within a classification framework (i.e., classifying peptides as stable binders or non-binders). Then, starting from pHLA sequences, we perform large-scale structural modeling. We have recently developed a new method to model pHLA structures called the Anchored Peptide-MHC Ensemble Generator (APE-Gen) (28). APE-Gen has the ability to rapidly generate native-like conformations of pHLA complexes, by leveraging the conserved positioning of the peptide's so called "anchor residues" to particular pockets of the HLA binding cleft. With the development of APE-Gen, we can now use machine learning on pHLA structures on a scale that has not been reached before. The rest of this paper is organized as follows. In the next section, we will describe our approach (**Figure 1**) of (i) generating pHLA structures, (ii) extracting simple features based on pHLA interactions, and (iii) training a random forest model to classify binders from non-binders. Finally, we show that our model produces high values for the area under the receiver operating characteristic curve (AUROC) in validation tests, and showcase the greater interpretability of the results as compared to neural network approaches. The generated dataset of pHLA structures provides new opportunities to build improved structure-based models to assess pHLA binding, and our model can serve as a benchmark for future models.

2. METHODS

2.1. Generating Peptide-HLA Structures

The dataset of pairs of peptide sequences and HLA alleles was obtained from two different databases. The list of stable binders to a given HLA allele (i.e., positive labels) was taken from a dataset curated by the authors of ForestMHC (16). Their curated dataset is derived from multiple sources, including the SysteMHC Atlas (11), which is a database of eluted peptides from MS experiments. For the polyallelic samples, they used MixMHCpred (29) to deconvolute the samples to a specific allele among a set of multiple well-defined alleles (polyallelic samples). They discarded samples which could come from alleles that MixMHCpred did not support. Note that only 9-mers were considered by ForestMHC (16), so our method was also only trained on 9-mers. However, in principle, other n-mers could be considered by our method as well.

A list of unstable peptides (i.e., negative labels) was obtained from a curated dataset of experimental binding assays, mostly coming from IEDB (30), which was prepared by the authors of MHCFlurry (8). This dataset differs from the previous, as there is an associated value representing the binding affinity measurement of the peptide to the HLA. Since we are interested in finding negative labels, we applied a threshold on the binding affinity with the assumption that low affinity binding implies unstable binding. All pHLA pairs with binding affinity measurements greater than 20,000 nM were extracted from the MHCFlurry dataset. Thresholds are typically set to 500 nM, where peptides with affinity values below this threshold are predicted to be strong enough binders to be presented by the corresponding HLA in the cell surface. Since there do exist peptides that are presented with affinity values greater than 500 nM, we applied a conservative threshold in order TO have more confidence that our dataset of non-binders consists of peptides that are not presented by the corresponding HLA.

Finally, APE-Gen (28) is used to model all of the peptide sequences bound to a given HLA allele. HLAs which did not have a crystal structure available in the PDB were modeled with MODELLER (31), using the corresponding HLA sequence from IMGT (32), and a structural template of an HLA allele from the same supertype classification (33). This is possible due to the conserved tertiary structure of HLA molecules (34), and the fact that alleles within the same supertype share similar peptide binding characteristics (33). Briefly, APE-Gen runs rounds of the following three steps: anchor alignment, backbone reconstruction, and side-chain addition with energy minimization. The list of flexible HLA residues is derived from a list of known important residues for peptide binding (33). A single round of APE-Gen is used per pHLA complex, taking

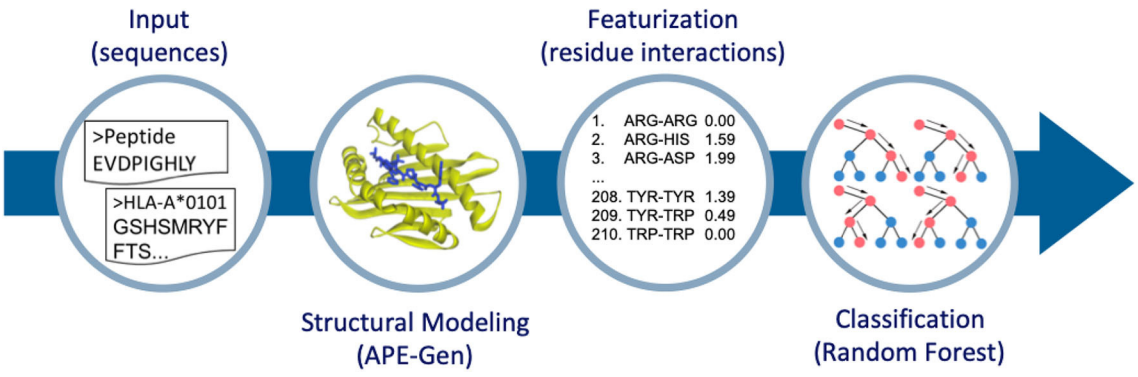

FIGURE 1 | Overview of the method to go from sequence to structure-based features for classification. APE-Gen is used to model pHLA structures, then featurization is done by extracting the residue-residue interactions between peptide and HLA. The final random forest model is trained on these structure-based features.

approximately 2 min to model per complex across 6 cores on an Intel Xeon Platinum 8160. The anchor constraint was changed to 4 Angstroms (from the default 2 Angstroms), since it is expected that the anchor interactions of non-binders will be more unstable when undergoing the energy minimization step of APE-Gen. All other parameters are kept at their default (28). For a given pHLA, APE-Gen generates an ensemble of pHLA conformations. This ensemble may include some additional information that is missed when only analyzing a single conformation (35). Therefore, we considered two datasets for training. The first is to simply take the best scoring conformation per pHLA, according to the default scoring function used in APE-Gen, which is Vinardo (36). The second is to take the whole ensemble of each pHLA, pooling every conformation into the dataset. The median number of conformations generated per pHLA is 18. Note that the number of conformations generated per pHLA is not constant due to filtering steps done within APE-Gen (**Figure S1**). APE-Gen is open-source and available at https://github.com/KavrakiLab/APE-Gen.

2.2. Featurization

Each pHLA conformation generated with APE-Gen is then transformed into a feature vector containing information on the residue-residue interactions between the peptide and HLA. The feature vector for a given conformation contains 210 elements, representing the total number of possible pairings between the 20 amino acids, including interactions between two residues of the same type. Each element represents the amount of a particular type of interaction (for example, between alanines and leucines) found in the conformation. The amount of interaction is quantified as the sum of some function of the residue-residue distances, which is defined as the distance in Angstroms between the nearest two heavy atoms computed using the MDTraj Python package (37). Intuitively, such a function should have a high value for low residue-residue distances and monotonically decrease as the residue-residue distance increases in order to represent the amount of interaction. In this work, we consider three functions: the reciprocal, reciprocal squared, and a sigmoid function (**Figure S2**). The sigmoid function was chosen such that a value of 0.5 occurs at 5 Angstroms. Residue-residue contacts are usually defined within the range of 4.5-5 Angstroms (38).

As an example, assume the function is simply the reciprocal of the residue-residue distance. Furthermore, assume that the fourth element in the feature vector represents interactions between an arginine and aspartic acid. The order of the interactions may be chosen arbitrarily, but is fixed across all pHLA structures. In this scenario, we start by measuring all the distances between arginines in the peptide and aspartic acids in the HLA, as well as between arginines in the HLA and aspartic acids in the peptide. Then, the value for the fourth element in the feature vector is computed as the sum of the reciprocal values of the measured distances. Note that in this implementation peptide-peptide and HLA-HLA interactions were ignored, since interactions between peptide and HLA are expected to have a more direct contribution to stability. With this featurization process, small values in the feature vector represent little interaction for the particular residue-residue pair. Values of exactly zero indicate that the corresponding interaction was not found in the conformation. Conversely, large values represent instances where there was significant residue-residue contact (i.e., low residue-residue distances) found in the conformation. Note that in our construction, only simple, homogenous features based on residue-residue distances between the peptide and HLA are extracted for the model, as opposed to more complex, heterogenous features that were based on energy terms from a scoring function (24).

2.3. Model Selection

Models were chosen with five-fold cross validation using the area under the receiver operating characteristic curve (AUROC) as the main metric. The receiver operating characteristic curve plots the true positive rate and false positive rate across different thresholds of the output probability, where a random guess would produce an AUROC of 0.5 and a perfect classifier produces an AUROC of 1.0. We tried three different classifiers, namely logistic regression, gradient boosting, and random forest, across a variety of parameterizations and featurization functions. For each model type, we also tested whether the use of the whole ensemble of conformations improved the final AUROC score. The implementation of the models and analysis is done using Scikit-Learn (39).

3. RESULTS

3.1. Generalizability

3.1.1. Random Forest Was the Most Robust Model

The final dataset consisted of 155,562 pHLA structures across 99 different alleles, which is to-date the largest dataset of modeled pHLA structures. Within this dataset, 43 alleles have available experimental data on both binders and non-binders. The identity of all modeled pHLAs in this dataset can be found in the **Supplementary Material**. In total, about 300,000 CPU-hours were required to generate the dataset. There is an approximately 70:30 binders/non-binders ratio across the two sources of data, so class weights were adjusted for all models given the imbalance of class labels. The five-fold cross validation results of the three classifiers tested can be found in **Table 1**. These results relate to the use of the best parameters found for each type of model, across the three different featurization types. Results for all tested parameters can be found in **Tables S1**–**S3**. We find that across the parameters tested, logistic regression performs the worse, while random forest and gradient boosting classifiers give the most robust results as the AUROC values are consistently high with little variation. The overall best performing model was based on random forests (average AUROC: 0.978) with an ensemble of 1,000 decision trees, which use about 7 features ($\log_2 210$) and Gini impurity to determine the quality of a split.

3.1.2. Ensemble Dataset With Sigmoid Featurization Improves Performance

With the random forest model, we tested whether training with the whole ensemble of conformations produced by APE-Gen could further increase the AUROC. This dataset consists of

TABLE 1 | Average AUROC values from five-fold validation tests across different classifiers and different featurizations.

Model	Feat	AUROC
rf	$1/d$	0.978 (0.000)
rf	$1/d^2$	0.976 (0.001)
rf	sig	0.975 (0.001)
gb	$1/d$	0.970 (0.002)
gb	$1/d^2$	0.970 (0.001)
gb	sig	0.977 (0.001)
lr	$1/d$	0.875 (0.003)
lr	$1/d^2$	0.880 (0.002)
lr	sig	0.882 (0.001)

Only the best parameters per classifier are shown. rf stands for random forest, gb stands for gradient boosting, and lr stands for logistic regression. Average AUROC values are reported along with standard deviations. Random forest classifiers produce the most robust models.

TABLE 2 | Average AUROC values from five-fold validation tests across different featurizations and different datasets.

Feat	Data	AUROC
$1/d$	Single	0.978 (0.000)
$1/d^2$	Single	0.976 (0.001)
sig	Single	0.975 (0.001)
$1/d$	Ensemble	0.987 (0.001)
$1/d^2$	Ensemble	0.988 (0.000)
sig	Ensemble	0.990 (0.000)

Average AUROC values are reported along with standard deviations. The best model uses sigmoid-based features trained on the ensemble-enriched dataset.

2,825,185 data points with an average of about 18 conformations per pHLA. All of the conformations for a given pHLA are pooled together with the same appropriate label. Across the three different featurization types, a random forest model was trained on the ensemble-enriched dataset of pHLA conformations. Therefore, when testing on an unseen pHLA, APE-Gen is first run to generate an ensemble of conformations. Each featurized conformation is then classified with the model and the output probabilities are averaged to produce the final output. The five-fold cross validation results using the random forest model across the featurization and dataset types are presented in **Table 2**. We find that across all the different configurations, the best performing random forest model uses the sigmoid-based featurization and the ensemble-enriched dataset with an average AUROC of 0.990. We also found that the ensemble-enriched dataset improves the performance of the other types of models (**Table S4**) with the gradient boosting model (with sigmoid featurization) also achieving a high average AUROC of 0.982.

Sigmoid-based features perform best since higher values are achieved at distances where residue-residue contacts are typically defined (**Figure S2**). The positive effect of the ensemble may be due to two reasons. First, interactions that are present in multiple conformations for a given pHLA could be an indication for stable interactions, which are now present in the data that is used to train the model. Second, APE-Gen produces on average about 5 more conformations for a true binder than it does for a true non-binder (**Figure S1**). The additional conformations for binders could turn into bias that the model has learned from.

3.1.3. Final Model Is Competitive With Sequence-Based Approaches on Leave-One-Allele-Out Tests

While our random forest model achieves a high average AUROC on standard five-fold cross validation tests, a tougher test for generalizability would be to partition the train/test split based on the HLA allele. A method that can perform well for tests on unseen allele data would be valuable for cases where pHLA binding prediction is to be done for rarer alleles, with little to no experimental data available. To simulate this scenario, we set aside data related to all the associated examples for a given HLA allele (i.e., both positive and negative examples). We then trained a random forest model using the same procedure described above on the rest of the data. The same procedure for training and testing was then repeated for each one of the HLA alleles in the dataset. This validation setup is called "leave-one-allele-out," and has been used before in testing NetMHCpan (14). We performed the validation setup across the set of 43 alleles for which there are both positive and negative examples in our dataset, and compared our approach to 5 sequence-based methods: NetMHCstabpan 1.0, NetMHCpan 4.0, NetMHC 4.0, MHCFlurry 1.4.3, and MixMHCpred 2.0.2.

The distribution of AUROC values across all alleles tested can be seen in **Figure 2**, and corresponding AUROC values can be found in the **Supplementary Material**. Our method achieves a median AUROC of 0.985 which is greater than NetMHCstabpan (0.969) and competitive with NetMHCpan (0.989). Additionally, the overall distributions of AUROC values shows that our method is more robust (smaller variations) than the other methods, achieving AUROC values greater than 0.9 for all but three alleles (namely HLA-B*39:06, HLA-C*04:01, and HLA-C*14:02). The one exception was MixMHCpred, which achieves the highest median AUROC (0.993) and good robustness. This result is not too surprising since this method was used in the construction of the positive labels (16). Despite having lower median AUROCs against some methods, the difference was never more than 0.01.

We also note that the comparison with NetMHCpan is not particularly fair since there is overlap between the data used to train NetMHCpan and the allele-based validation sets discussed here. In fact, 46% of the data from this work is included in the training set for NetMHCpan. However, this set could not be removed since the overlap is largely on our set of negative labels. Therefore, removing them would complicate the interpretation of AUROC values, and AUROC values cannot even be computed when there are no negative labels.

Other models were also tested in the leave-one-allele-out framework using both the single conformation and ensemble-enriched datasets. The difference in average AUROC between the random forests model (AUROC: 0.985) and the gradient

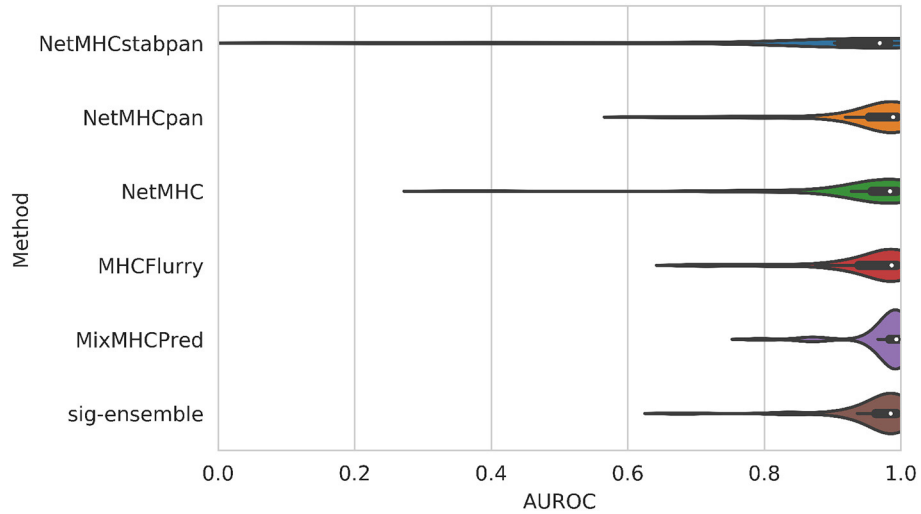

FIGURE 2 | Comparison of AUROC values for leave-one-allele-out experiments. Our structural method based on using sigmoid featurization and ensemble-enriched dataset achieves competitive results with sequence-based methods.

boosting model (AUROC: 0.939) is significant unlike the five-fold cross-validation case. This difference can be qualitatively seen in **Figure S3**: the distribution shape for the random forests model (ensemble-enriched) reveals the higher concentration of samples closer to 1.0. We speculate that the higher performance of random forests may be due to its robustness in overfitting due to sampling random subsets of data in the training process. We note that our tuning process could be made more exhaustive, and it is conceivable that gradient boosting could perform just as well as random forests for this data. However, it is unclear as to whether gradient boosting could achieve the same interpretability capabilities as random forests.

Finally, we can also compare the performance on a per-allele basis by comparing the AUROC of our method against the others for a given allele. Against NetMHCstabpan, our method improves the AUROC by 0.116 (mean across alleles), while against NetMHCpan, our method improves the AUROC by 0.010. When compared to methods that were trained on a per-allele basis, our method improves the AUROC by 0.018 for NetMHC and 0.010 for MHCFlurry, but is lower against MixMHCpred (0.013). Nevertheless, our method can achieve high AUROCs across alleles in a manner that is competitive against sequence-based methods on average with improved robustness.

3.2. Interpretability

The fact that our model is based on random forest also offers a significant interpretability advantage. For instance, we can compute an importance value for each feature of the random forest model by finding the mean decrease in impurity across all the decision trees in the ensemble. We found that most of the top features were hydrophobic interactions with an average importance value of 0.5% over the global average of 0.4%. The least important features include interactions that are less frequent in the dataset, such as ASN-ASN, HIS-HIS, and MET-MET. The full set of feature importance values can be found in the **Figure S3**.

While computing feature importances provides a global view of what the model is learning, we can also inspect how the model arrives at a prediction for a particular example (40). The prediction output of the random forest model, $P(x)$, for a particular example x is a probability to be in the positive class (which is thresholded by 0.5 to classify as stable binder/non-binder). The output can be decomposed as

$$P(x) = \text{bias} + \sum_{j=1}^{210} \text{contri}_j(x) \tag{1}$$

where the bias term reflects the ratio of positive examples in the data, which is 0.5 in this work because the classes were reweighted from the model. The interesting quantity is $\text{contri}_j(x)$, which is the contribution of feature j toward the prediction output. This equation tells us that the contributions are then combined in a linear manner reflecting how the decision trees split on a single feature at a time. The contribution values can be positive (contribute toward the stable binding) or negative (contribute toward unstable binding). Furthermore, since each residue-residue interaction was added to the corresponding element in the feature vector in a linear manner, we can decompose the contribution values further across every possible residue-residue contact in the original pHLA structure. While the contribution values are computed in a non-linear way based on the values of the other features across the training dataset, we can still inspect the features that greatly contribute to the prediction for a given example and test if they are in line with chemical intuition. The feature contributions are computed with the Python package treeinterpreter (40).

As an example, we model the structure of the peptide, EVDPIGHLY, bound to HLA-A*01:01. This is a peptide that has

been used as a target for T-cell-based immunotherapy against melanoma (41). Our model correctly predicts that this peptide is a stable binder, so we analyze the feature contributions leading to this prediction. The anchor residues for this allele are in position 2 (VAL) and position 9 (TYR), and we find that anchor-related interactions account for 26% of the positive contributions. However, our model is finding a significant positive contribution from other interactions. The feature with the largest position contribution is the ASP-ARG interaction (13%). In fact, position 3 is an aspartic acid (ASP), and interactions involving position 3 have the largest total positive contribution (32%). Interestingly, it is known that aspartic acid is a "preferred" residue in position 3 for peptides binding to HLA-A*01:01 (30).

When we model a destructive mutation on the anchor residue in position 2, from VAL to TRP, our model predicts that the new peptide is unstable. As expected, the feature contributions indicate that 42% of the negative contributions come from the TRP in position 2. Thus, the model is indeed using the interactions introduced by this mutation.

Our models are publicly available, alongside the ability to do the interpretation analysis presented in this section. The interpretation analysis has been automated to be able to produce summarized results as well as the raw data. This data contains more information than presented in this section, including a decomposition of the contribution values across each peptide-HLA residue-residue pair. The structural modeling with APE-Gen, classification with random forest, and interpretation analysis can be done for any pHLA of interest, and is available as an easy to use Docker image (https://hub.docker.com/r/kavrakilab/apegen/tags) with a tutorial found in https://github.com/KavrakiLab/pHLA-RFclassifier-from-structure.

4. DISCUSSION

In this work, we performed large-scale modeling of pHLA conformations, which is used to train an interpretable, structure-based classifier for pHLA binding prediction. With APE-Gen as the enabling technology, we generated a dataset of pHLA conformations that is the largest of its kind, opening the door for machine learning to be performed on top of pHLA conformations. We investigated various featurizations that are solely based on simple, homogenous conformational features (i.e., peptide-HLA, residue-residue distances). We show on our dataset that our model achieves competitive AUROCs against sequence-based methods. Additionally, our model based on random forest offers an interpretability advantage over approaches based on neural networks.

Note that while our dataset of structures is large with respect to structural modeling efforts (e.g., over 150,000 different pHLAs), this number becomes small when compared to the number of sequences that sequence-based methods have been trained on (e.g., about 3 million for NetMHCpan). Additionally, this work has only been tested on 9-mer ligands, but other n-mers do of course exist as binders to a significant extent (42). It should also be noted that our source of positive labels was dependent on the accuracy of MixMHCpred. In order to push the accuracy of our classifier, we need to include all of the available high-quality experimental data for training, which should increase our confidence in the final model. Our classifier does not have any inherent limitation on the peptide length, as APE-Gen can model other n-mers, and the featurization process is also not specific to 9-mers. Future work can focus on modeling more pHLAs, including the proper modeling of longer peptides by APE-Gen.

Despite the efficiency of APE-Gen, the step of modeling a new structure still takes a few minutes. Modeling structures takes significant computational resources, and reaching the scale of training data that sequence-based methods train on requires at least an order of magnitude more computational time. This makes our structure-based classifier slower than a sequence-based method for unseen pHLAs, and currently requires high performance computational resources to make large peptide screenings viable. However, the modeling of pHLA structures would only have to be performed once. Thus in the future, we can try to alleviate this burden by creating a database of previously modeled pHLAs, so that the classifier can skip the modeling step for all previously modeled complexes.

The use of structure may be the reason that we achieve high AUROCs on our dataset despite the relatively small dataset size. Models based on sequence are supposed to infer structural information, like the interactions between peptide and HLA, in order to get to accurate binding predictions. Our construction feeds this information directly to the model, which may be the key for generalizability. In fact, we can confirm when the model is properly using interaction information because our model was based on random forests. Our model can be made transparent, and we can understand *why* the model reached any given prediction. For "black-box" methods like neural networks, the best that can be done would be to try identifying patterns among the highest scoring samples. A list of random peptides could be run through the neural network for a given allele, and then the top scoring peptides can be analyzed for any noticeable peptide binding motif. For any given peptide, one might guess how the neural network arrived to the prediction by reasoning back to the peptide binding motif. This route is indirect at best, since it is extremely difficult to interrogate the neural network into revealing what leads to a particular prediction, which is an inherent problem of this methodology. There is no way of knowing that the reason a peptide was classified as a non-binder was because the model learned to penalize when a TRP exists in the peptide sequence along with a TYR in the HLA sequence, for example; a potentially spurious association with no obvious biochemical reason for affecting the binding. Our random forest model does reveal such information on a *per prediction basis*, as demonstrated in the Results section. For any given prediction, correct or not, we can see how the model composes the features into its final output, and check if it is in line with chemical intuition. This can even be useful for suggesting the kind of additional data needed for training when analyzing an example that was incorrectly classified.

We would like to make it clear however, that the goal of this work is not to produce a method for pHLA binding prediction that will replace the gold-standard methods, such as NetMHC and NetMHCpan, which are available as a public webservers for

rapid prediction. Many challenges remain as mentioned in this section. The contribution of this work reveals that for the pHLA binding prediction task, structure-based methods can work as a proof-of-concept. The time investment spent in doing the structural modeling enables the benefit of added interpretability. The residue-residue interactions present between peptide and HLA can be directly extracted as simple features for the model. Additionally, the random forests model can highlight how the features are composed to form the output of any given pHLA. When combined together, one instantly has a link to relate the binding prediction back to each individual peptide-HLA residue-residue interaction for further analysis. Such a capability can be valuable as a complement to sequence-based approaches. For instance, it could be used after epitope discovery efforts, providing more detailed analysis of binding for peptides that are strong candidates as targets for vaccine development, T-cell-based immunotherapy, or as the potential triggers for autoimmune reactions. The obtained structural information could be used to lead peptide optimization efforts, or to provide a molecular basis for the presentation of unusual HLA-binders. As we continue to push the accuracy of our method, our results and dataset of pHLA structures can be used as a benchmark for a new generation of structure-based methods for HLA binding prediction and epitope discovery.

DATA AVAILABILITY STATEMENT

The dataset of structural features, based on sigmoid featurization, can be found inside a Docker image (https://hub.docker.com/r/kavrakilab/apegen/tags) with a tutorial found in https://github.com/KavrakiLab/pHLA-RFclassifier-from-structure. The dataset of pHLA structures and the other featurized datasets are available upon request.

AUTHOR CONTRIBUTIONS

JA conceived the project idea, generated the data, and developed the model. JA, DA, CC, and LK designed the experiments, analyzed the results, wrote, and edited the paper.

FUNDING

This work was supported by a training fellowship from the Gulf Coast Consortia on the Training Program in Biomedical Informatics, National Library of Medicine T15LM007093. This work has also been supported in part by Cancer Prevention and Research Institute of Texas (CPRIT) through Grant award RP170508, through a Fellowship from the Computational Cancer Biology Training Program (RP170593), Einstein Foundation Berlin (Einstein Visiting Fellowship to CC), the National Science Foundation (CHE-1740990, CHE-1900374, and PHY-1427654 to CC), and the Welch Foundation (grant C-1570 to CC).

ACKNOWLEDGMENTS

The authors thank the Texas Advanced Computing Center for allowing access to the *Stampede2* supercomputer (MCB180164), as well as the Center for Research Computing for allowing access to the *NOTS* computing cluster.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.01583/full#supplementary-material

REFERENCES

1. Hewitt EW. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*. (2003) 110:163–9. doi: 10.1046/j.1365-2567.2003.01738.x
2. Garstka MA, Fish A, Celie PH, Joosten RP, Janssen GM, Berlin I, et al. The first step of peptide selection in antigen presentation by MHC class I molecules. *Proc Natl Acad Sci USA*. (2015) 112:1505–10. doi: 10.1073/pnas.1416543112
3. Filley AC, Henriquez M, Dey M. CART immunotherapy: development, success, and translation to malignant gliomas and other solid tumors. *Front Oncol*. (2018) 8:453. doi: 10.3389/fonc.2018.00453
4. Mage MG, Dolan MA, Wang R, Boyd LF, Revilleza MJ, Robinson H, et al. The peptide-receptive transition state of MHC class I molecules: insight from structure and molecular dynamics. *J Immunol*. (2012) 189:1391–9. doi: 10.4049/jimmunol.1200831
5. Purcell AW, Ramarathinam SH, Ternette N. Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc*. (2019) 14:1687–707. doi: 10.1038/s41596-019-0133-y
6. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLoS Comput Biol*. (2018) 14:e1006457. doi: 10.1371/journal.pcbi.1006457
7. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*. (2003) 12:1007–17. doi: 10.1110/ps.0239403
8. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst*. (2018) 7:129–32. doi: 10.1016/j.cels.2018.05.014
9. Jorgensen KW, Rasmussen M, Buus S, Nielsen M. NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology*. (2014) 141:18–26. doi: 10.1111/imm.12160
10. Harndahl M, Rasmussen M, Roder G, Dalgaard Pedersen I, Sorensen M, Nielsen M, et al. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur J Immunol*. (2012) 42:1405–16. doi: 10.1002/eji.201141774
11. Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaino JA, et al. The SysteMHC atlas project. *Nucleic Acids Res*. (2018) 46:D1237–47. doi: 10.1093/nar/gkx664
12. Bassani-Sternberg M, Coukos G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr Opin Immunol*. (2016) 41:9–17. doi: 10.1016/j.coi.2016.04.005
13. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity*. (2017) 46:315–26. doi: 10.1016/j.immuni.2017.02.007
14. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide-MHC class i interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. (2017) 199:3360–8. doi: 10.4049/jimmunol.1700893

15. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, et al. Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. *J Immunol*. (2016) 197:1517–24. doi: 10.4049/jimmunol.1600582
16. Boehm KM, Bhinder B, Raja VJ, Dephoure N, Elemento O. Predicting peptide presentation by major histocompatibility complex class I: an improved machine learning approach to the immunopeptidome. *BMC Bioinformatics*. (2019) 20:7. doi: 10.1186/s12859-018-2561-z
17. Antunes DA, Devaurs D, Moll M, Lizee G, Kavraki LE. General prediction of peptide-MHC binding modes using incremental docking: a proof of concept. *Sci Rep*. (2018) 8:4327. doi: 10.1038/s41598-018-22173-4
18. Knapp B, Giczi V, Ribarics R, Schreiner W. PeptX: using genetic algorithms to optimize peptides for MHC binding. *BMC Bioinformatics*. (2011) 12:241. doi: 10.1186/1471-2105-12-241
19. Antunes DA, Moll M, Devaurs D, Jackson KR, Lizee G, Kavraki LE. DINC 2.0: A new protein-peptide docking webserver using an incremental approach. *Cancer Res*. (2017) 77:e55–7. doi: 10.1158/0008-5472.CAN-17-0511
20. Rigo MM, Antunes DA, Vaz de Freitas M, Fabiano de Almeida Mendes M, Meira L, Sinigaglia M, et al. DockTope: a web-based tool for automated pMHC-I modelling. *Sci Rep*. (2015) 5:18413. doi: 10.1038/srep18413
21. Kyeong HH, Choi Y, Kim HS. GradDock: rapid simulation and tailored ranking functions for peptide-MHC Class I docking. *Bioinformatics*. (2018) 34:469–76. doi: 10.1093/bioinformatics/btx589
22. Liu T, Pan X, Chao L, Tan W, Qu S, Yang L, et al. Subangstrom accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. *J Chem Inf Model*. (2014) 54:2233–42. doi: 10.1021/ci500393h
23. Logean A, Sette A, Rognan D. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg Med Chem Lett*. (2001) 11:675–9. doi: 10.1016/S0960-894X(01)00021-X
24. Aranha MP, Spooner C, Demerdash O, Czejdo B, Smith JC, Mitchell JC. Prediction of peptide binding to MHC using machine learning with sequence and structure-based feature sets. *Biochim Biophys Acta Gen Subj*. (2020) 1864:129535. doi: 10.1016/j.bbagen.2020.129535
25. Antunes DA, Abella JR, Devaurs D, Rigo MM, Kavraki LE. Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes. *Curr Top Med Chem*. (2018) 18:2239–55. doi: 10.2174/1568026619666181224101744
26. Wan S, Knapp B, Wright DW, Deane CM, Coveney PV. Rapid, precise, and reproducible prediction of peptide-MHC binding affinities from molecular dynamics that correlate well with experiment. *J Chem Theory Comput*. (2015) 11:3346–56. doi: 10.1021/acs.jctc.5b00179
27. Yanover C, Bradley P. Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proc Natl Acad Sci USA*. (2011) 108:6981–6. doi: 10.1073/pnas.1018165108
28. Abella JR, Antunes DA, Clementi C, Kavraki LE. APE-Gen: a fast method for generating ensembles of bound peptide-MHC conformations. *Molecules*. (2019) 24:881. doi: 10.3390/molecules24050881
29. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol*. (2017) 13:e1005725. doi: 10.1371/journal.pcbi.1005725
30. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res*. (2019) 4:D339–43. doi: 10.1093/nar/gky1006
31. Webb B, Sali A. Protein structure modeling with MODELLER. *Methods Mol Biol*. (2017) 1654:39–54. doi: 10.1007/978-1-4939-7231-9_4
32. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA database. *Nucleic Acids Res*. (2019). 48:D948–55. doi: 10.1093/nar/gkz950
33. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. *BMC Immunol*. (2008) 9:1. doi: 10.1186/1471-2172-9-1
34. Khan JM, Ranganathan S. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res*. (2010) 6:S2. doi: 10.1186/1745-7580-6-S1-S2
35. Fodor J, Riley BT, Borg NA, Buckle AM. Previously hidden dynamics at the TCR-peptide-MHC interface revealed. *J Immunol*. (2018) 200:4134–45. doi: 10.4049/jimmunol.1800315
36. Quiroga R, Villarreal MA. Vinardo: a scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS ONE*. (2016) 11:e0155183. doi: 10.1371/journal.pone.0155183
37. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J*. (2015) 109:1528–32. doi: 10.1016/j.bpj.2015.08.015
38. Yao XQ, Momin M, Hamelberg D. Establishing a framework of using residue-residue interactions in protein difference network analysis. *J Chem Inf Model*. (2019) 59:3222–8. doi: 10.1021/acs.jcim.9b00320
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. (2011) 12:2825–30.
40. Saabas A. Treeinterpreter (2015). Available online at: https://github.com/andosa/treeinterpreter
41. Raman MC, Rizkallah PJ, Simmons R, Donnellan Z, Dukes J, Bossi G, et al. Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Sci Rep*. (2016) 6:18851. doi: 10.1038/srep18851
42. Gfeller D, Guillaume P, Michaux J, Pak HS, Daniel RT, Racle J, et al. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J Immunol*. (2018) 201:3705–16. doi: 10.4049/jimmunol.1800914

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 August 2020
doi: 10.3389/fgene.2020.00905

Identification of *KIF18B* as a Hub Candidate Gene in the Metastasis of Clear Cell Renal Cell Carcinoma by Weighted Gene Co-expression Network Analysis

Huiying Yang[1†], Yukun Wang[2†], Ziyi Zhang[3] and Hua Li[1]*

[1] Department of Nephrology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China,
[2] Department of Urology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China,
[3] Department of Endocrinology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

Background: Clear cell renal cell carcinoma (ccRCC) is a common type of fatal malignancy in the urinary system. As the therapeutic strategies of ccRCC are severely limited at present, the prognosis of patients with metastatic carcinoma is usually not promising. Revealing the pathogenesis and identifying hub candidate genes for prognosis prediction and precise treatment are urgently needed in metastatic ccRCC.

Methods: In the present study, we conducted a series of bioinformatics studies with the gene expression profiles of ccRCC samples from Gene Expression Omnibus (GEO) and the cancer genome atlas (TCGA) database for identifying and validating the hub gene of metastatic ccRCC. We constructed a co-expression network, divided genes into co-expression modules, and identified ccRCC-related modules by weighted gene co-expression network analysis (WGCNA) with data from GEO. Then, we investigated the functions of genes in the ccRCC-related modules by enrichment analyses and built a sub-network accordingly. A hub candidate gene of the metastatic ccRCC was identified by maximal clique centrality (MCC) method. We validate the hub gene by differentially expressed gene analysis, overall survival analysis, and correlation analysis with clinical traits with the external dataset (TCGA). Finally, we explored the function of the hub gene by correlation analysis with targets of precise therapies and single-gene gene set enrichment analysis.

Results: We conducted WGCNA with the expression profiles of GSE73731 from GEO and divided all genes into 8 meaningful co-expression modules. One module is proved to be positively correlated with pathological stage and tumor grade of ccRCC. Genes in the ccRCC-related module were mainly enriched in functions of mitotic cell division and several proverbial tumor related signal pathways. We then identified *KIF18B* as a hub gene of the metastasis of ccRCC. Validating analyses in external dataset observed the up-regulation of *KIF18B* in ccRCC and its correlation with worse outcomes. Further analyses found that the expression of KIF18B is related to that of targets of precise therapies.

OPEN ACCESS

Edited by:
Panayiotis V. Benos,
University of Pittsburgh, United States

Reviewed by:
Alessandro Giuliani,
Istituto Superiore di Sanità (ISS), Italy
Silvia Liu,
University of Pittsburgh, United States

***Correspondence:**
Hua Li
h_li@zju.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *17 October 2019*
Accepted: *21 July 2020*
Published: *20 August 2020*

Citation:
Yang H, Wang Y, Zhang Z and Li H (2020) Identification of KIF18B as a Hub Candidate Gene in the Metastasis of Clear Cell Renal Cell Carcinoma by Weighted Gene Co-expression Network Analysis. Front. Genet. 11:905. doi: 10.3389/fgene.2020.00905

Conclusion: Our study proposed *KIF18B* as a hub candidate gene of ccRCC for the first time. Our conclusion may provide a brand-new clue for prognosis evaluating and precise treatment for ccRCC in the future.

Keywords: clear cell renal cell carcinoma, weighted gene co-expression network analysis, enrichment analysis, maximal clique centrality, survival analysis, precise therapies, sing-gene gene set enrichment analysis

INTRODUCTION

Renal cell carcinoma (RCC) is one of the top 10 prevalent malignancies and makes up approximately 2–3% of all cancers (Ljungberg et al., 2015). Clear cell renal cell carcinoma (ccRCC) is the most familiar histological subtype of RCC (Ochocki et al., 2018), the pathogenesis of which is still far from clear. As approximately one third of ccRCC patients were diagnosed with distant metastasis (Gupta et al., 2008) and the disease has low insensitivity toward traditional chemotherapy or radiotherapy, metastasis accounts for about 90% of ccRCC-related mortality (Chaffer and Weinberg, 2011). Nonetheless, precise treatments, such as targeted therapy (Campbell et al., 2018) and immunological therapy (Albiges et al., 2019), have shown relatively satisfactory effects in the treatment of metastatic ccRCC. Hence, it has become an urgent mission to identify novel hub candidate genes behind the mechanism of the metastasis, which may provide valuable targets for precise therapies.

At present, the pathogenesis of ccRCC has been partially clarified. The complete loss mutation through genetic and/or epigenetic mechanisms of the von Hippel-Lindau (VHL) tumor suppressor gene is regarded as the earliest and most significant oncogenic factor in ccRCC. The loss mutation of VHL leads to aberrant accumulation of hypoxia-inducible factors (HIF) even if the tissue microenvironment is adequately oxygenated, which results in abnormal activation of HIF targeting genes and then regulates the processes of angiogenesis, glycolysis, and apoptosis. The genetic diversity in ccRCC provides the substrate in ccRCC, and the selection upon the substrate enables the tumor to adapt to pressures and metabolic demands. Except for the mechanism in genetic field, analyses of gene expression, metabolic, and immunological status of ccRCC have given important mechanistic and clinical insights in ccRCC as well (Hsieh et al., 2017).

Thanks to contemporary breakthroughs of biological techniques, bioinformatics analyses have become new approaches for uncovering the pathogenesis of diseases. Besides studies about genetic sequence and mutations, researches focusing on gene expression levels have attracted more attention. Among various means for expression profile analysis, weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) stands out because of its superiorities in identifying hub candidate genes involved in diseases. WGCNA could divide genes with similar expression patterns into several biologically meaningful co-expression modules, analyze the relationship between gene modules and clinical traits, and finally evaluate the significance of genes in trait-related modules and excavate the hub candidate genes underlying the mechanism of diseases.

WGCNA have been widely used in various medical fields, such as tumor (Xiang Z. et al., 2019), neurological and psychiatric disorders (Katrinli et al., 2019; Tang and Liu, 2019), chronic disease (Chen et al., 2019), and infectious diseases (Bando et al., 2019). What's more, most of the conclusions drawn from WGCNA can be further confirmed by bioinformatics analyses or biological experiments, which guarantees the high reliability of WGCNA. WGCNA has been used for screening hub genes in ccRCC as well, and more efforts are needed for exploring novel hub genes blamed for metastasis or could act as potential targets for precise treatment.

Our study constructed a weighted co-expression network with the expression profiles of ccRCC tissues and related co-expression modules with clinical traits. Then we analyzed the main functions of genes in the trait-related module by enrichment analyses and successfully identified a hub candidate gene of ccRCC. Finally, we validated the reliability and clinical significance of the hub gene and explored its functions with an external dataset. We expect that our study could make a contribution to explain the pathogenesis of the metastasis of ccRCC and provide a potential target for treatment.

MATERIALS AND METHODS

Data Collection and Pre-processing

The overall design and procedures are described in a flow chart (**Figure 1**).

We searched Gene Expression Omnibus (GEO)[1] database with the keyword "clear cell renal cell carcinoma" and decided on dataset GSE73731 (Wei et al., 2017) for hub gene extracting as it has relative large sample size and detailed clinical information (gender, pathological stage, and tumor grade included). Samples without information of clinical traits were excluded from analyses containing clinical information (details of samples with complete clinical information are available in **Supplementary Table S1**). After downloading the raw data (already log2 transformed), we carried out probe annotation with microarray platform file under R environment. Probes matching more than one gene were discarded, and average values were taken for genes detected by more than one probes. Before WGCNA, we conducted sample clustering with hierarchical clustering method and excluded outlier samples accordingly. We filtered non-varying genes in the whole expression profile by variance as they are deemed as noise and may result in adverse effects to WGCNA.

For the validation of hub gene, we obtained the expression profile of ccRCC from the database of the cancer genome atlas

[1] https://www.ncbi.nlm.nih.gov/geo/

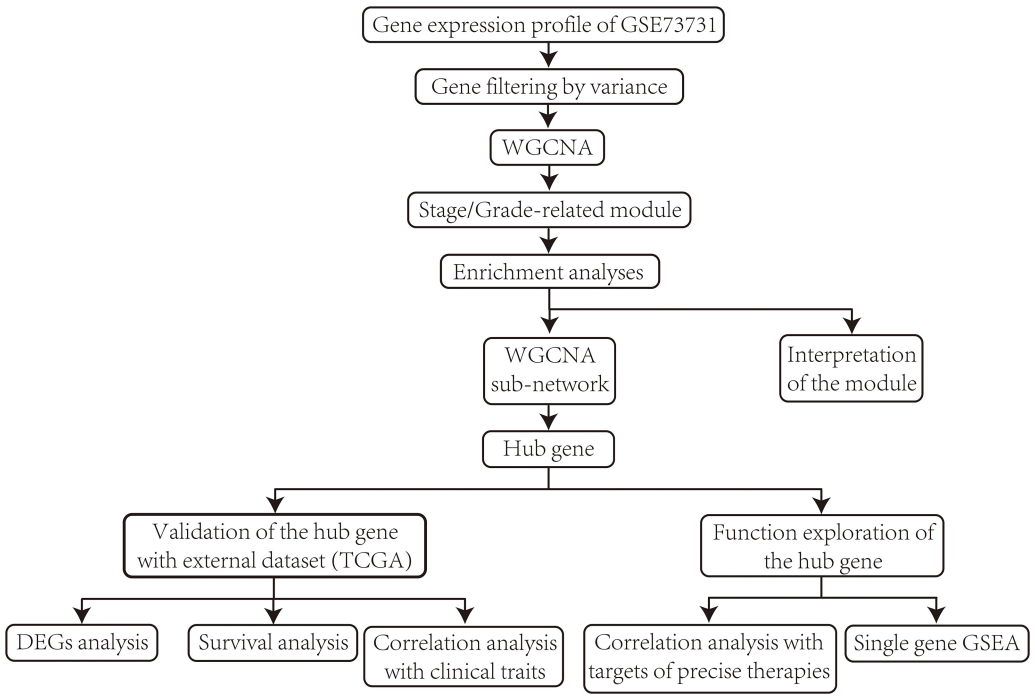

FIGURE 1 | Flow chart of data collection, data pre-processing, data analyzing, identification, validation, and function exploration of hub gene.

(TCGA)[2]. TCGA included the expression profile and clinical information of 72 normal control and 539 ccRCC patients.

Construction of Weighted Co-expression Network and Division of Co-expression Modules

We conducted WGCNA with the WGCNA package (Langfelder and Horvath, 2008) under R environment.

Firstly, we constructed a Pearson's correlation matrix of all pairwise genes by Pearson's correlation analysis. Secondly, we converted the Pearson's correlation matrix into an adjacency matrix (scale-free network) by a β the power operation (β value was known as the soft-thresholding value). To decide the most appropriate β value, we calculated the scale-free fit index and mean connectivity for each supposed β from 1 to 20. As higher scale-free fit index represents better coincidence with scale-free network and higher mean connectivity means better connection of the whole network, we referred to both of the indexes and decided the β value with scale-free fit index bigger than 0.85 as well as highest mean connectivity as the proper one. Then, we transformed the adjacency matrix into a topological overlap matrix (TOM) by calculating the topological overlap between pairwise genes, by which we could take indirect correlations into consideration as well as reduce noise and spurious correlations. Finally, we used the average linkage hierarchical clustering based on the TOM-based dissimilarity measure to divide genes into several co-expression modules, so that genes with co-expression relationships were gathered in the same module and genes expressed separately were divided. Modules of high similarity (higher than 0.75) were merged together.

Identification of Clinically Meaningful Modules

The clinical traits of GSE73731 contain gender, pathological stage, and tumor grade. To excavate hub genes related with the advancement and metastasis of ccRCC, we mainly aimed at modules of positive correlation with the trait of stage and grade. We conducted module-trait correlation analysis by Spearman's correlation analysis between module eigengene (ME, the first principal component of a given module) and clinical traits (stages I–IV was represented as 1, 2, 3, 4, and so do G1–G4). Modules of significant correlations with traits of pathological stage or tumor grade are defined as ccRCC-related modules, and genes in such modules were extracted for subsequent hub gene extraction.

We introduce the conceptions of Gene significance (GS) and Module Membership (MM) (Langfelder and Horvath, 2008) here. GS represents the correlation of the expression level of a gene and a clinical trait, and MM means the Pearson's correlation of the expression level of a certain gene and the module eigengene. An ideal clinic-related module is supposed to contain genes of high correlation between GS and MM.

Enrichment Analyses on ccRCC-Related Module

Gene Ontology (GO) (Lu et al., 2008) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG)

[2]https://portal.gdc.cancer.gov/

(Ogata et al., 1999) pathway enrichment analysis could reveal the biological processes and signal pathways in which certain gene cluster is involved. We performed enrichment analyses on genes in ccRCC-related modules and displayed the results with the clusterProfiler package (Yu et al., 2012) under R environment. The criteria for enriched terms were set as $p < 0.01$ and Benjamin-Hochberg adjusted $p < 0.05$. We mainly focus on the category of biological process (BP) among the results of GO enrichment analysis.

Identification of Hub Gene

After GO enrichment analysis on trait-related module, we extracted the genes from the most statistically significant GO term and built a sub-network with the weighted correlations among them. Then, we employed Maximal Clique Centrality (MCC) with cytohubba (Chin et al., 2014), a plug-in of cytoscape (Shannon et al., 2003), to assess the centrality of each gene in the sub-network. Genes with top MCC values are deemed as potential hub candidate genes related with the metastasis of ccRCC.

Validation of Hub Gene With Differentially Expressed Gene Analysis

To verify whether the hub gene is significantly up-regulated or down-regulated in ccRCC compared with normal control, we conducted differentially expressed gene (DEG) analysis on decided hub gene by Wilcoxon test method with limma package under R environment (Ritchie et al., 2015). DEGs analysis was conducted with the data obtained from TCGA. The cut-off criteria of DEGs were set as $p < 0.01$ and $|\log FC| > 0.5$.

Validation of Hub Gene With Survival Analysis

For validating whether the hub gene could affect the survival of ccRCC patients, we conducted overall survival (OS) analysis on the data obtained from TCGA with the survival package under R environment. Data from TCGA contains the expression profiles and follow-up information of 539 ccRCC samples. Samples with follow-up time less than 90 days were excluded. All samples were divided into two groups of high-expression or low-expression depending on the expression level of the hub gene. Then we conducted OS analysis with Kaplan-Meier method with a two-sided log-rank test to explore the difference of OS between the two groups.

Validation of Hub Gene by Analyzing the Relationship Between the Expression Level of Hub Gene and Clinical Traits

We divided all ccRCC samples in TCGA into high-expression and low-expression group by the median of the expression level of the hub gene. Then, we conducted correlation analysis between the expression groups and clinical traits (age, gender, tumor grade, pathological stage, T stage, and distant metastasis) to confirm the validity of our hub gene. The analysis was conducted by chi-square test under R environment. The criterion for statistical significance was set as $p < 0.01$.

Exploration of the Correlation Relationships Between the Expression Level of Hub Gene and Targets for Precise Therapies

Precise treatment, including immunotherapy and targeted therapy, is a novel approach for treating patients without opportunities for surgery or acting as a supplementary treatment before/after surgeries. Recently, more and more precise therapies were admitted for treating ccRCC by the Food and Drug Administration (FDA) (Barata and Rini, 2017). We analyzed the correlation relationships between the expression level of our hub gene and the targets of precise treatments by Pearson's correlation analysis to validate the significance of the hub gene in ccRCC and estimate the potential capacity of predicting the therapeutic effects with our hub gene. The targets of precise therapies are listed as follows: programmed cell death 1 (*PD1*), programmed cell death ligand 1 (*PDL1*), vascular Endothelial Growth Factor Receptor (*VEGFR1*), Fms-like tyrosine kinase 3 (*FLT3*), vascular Endothelial Growth Factor Receptor 3 (*VEGFR3*), mammalian target of rapamycin (*mTOR*), platelet-derived growth factor receptor alpha (*PDGFRA*), platelet-derived growth factor receptor beta (*PDGFRB*), *KIT* proto-oncogene (*KIT*), ret proto-oncogene (*RET*), and *MET* protooncogene (*MET*).

Function Analysis of the Hub Gene by Single-Gene Gene Set Enrichment Analysis

For exploring the biological function of hub gene in ccRCC, we conducted single-gene gene set enrichment analysis (GSEA) (Subramanian et al., 2005) on the hub gene with the data from TCGA. All samples were divided into high-expression and low-expression groups by the median of the hub gene, and GSEA was conducted to explore the up-regulated and down-regulated signal pathways in different groups. The criteria for statistical significance were set as $p < 0.01$ and $FDR < 0.25$.

RESULTS

Data Pre-processing

For GSE73731, we got an expression profile of 22,320 genes after probe annotation and reserved 113 ccRCC samples after sample clustering and outlier sample exclusion; 5,580 genes were adopted after gene filtering. The sample clustering tree depicted a satisfying result (**Figure 2**). Samples of similar pathological stages or tumor grades were gathered together, and the distance between samples with higher pathological stages or tumor grades and samples with lower pathological stages or tumor grades were relatively far.

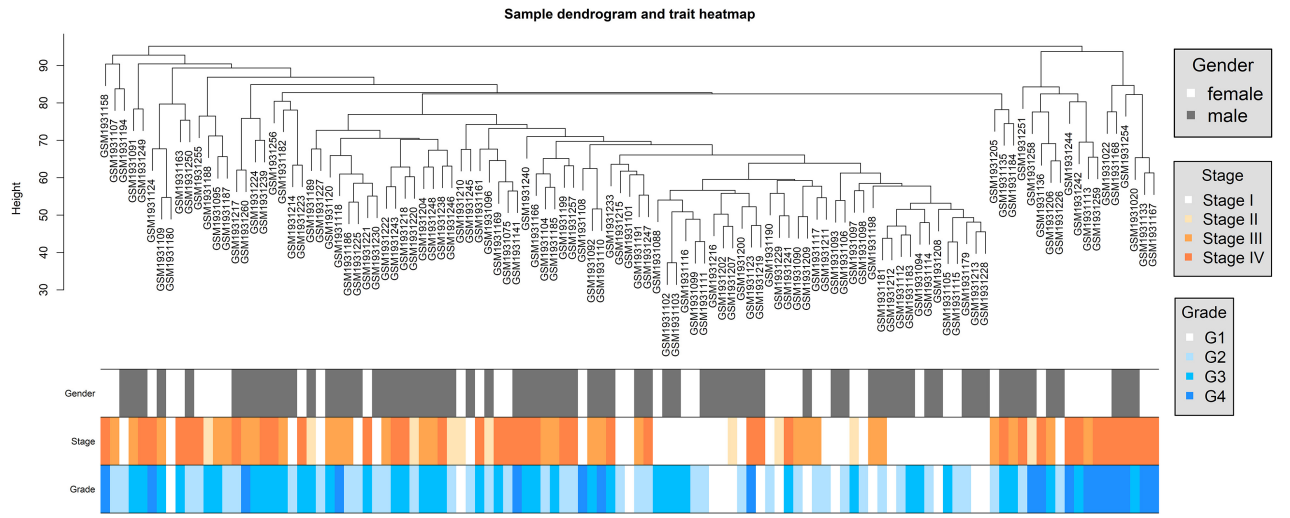

FIGURE 2 | Dendrogram of sample clustering and heatmap of clinical traits of all ccRCC samples in dataset of GSE73731.

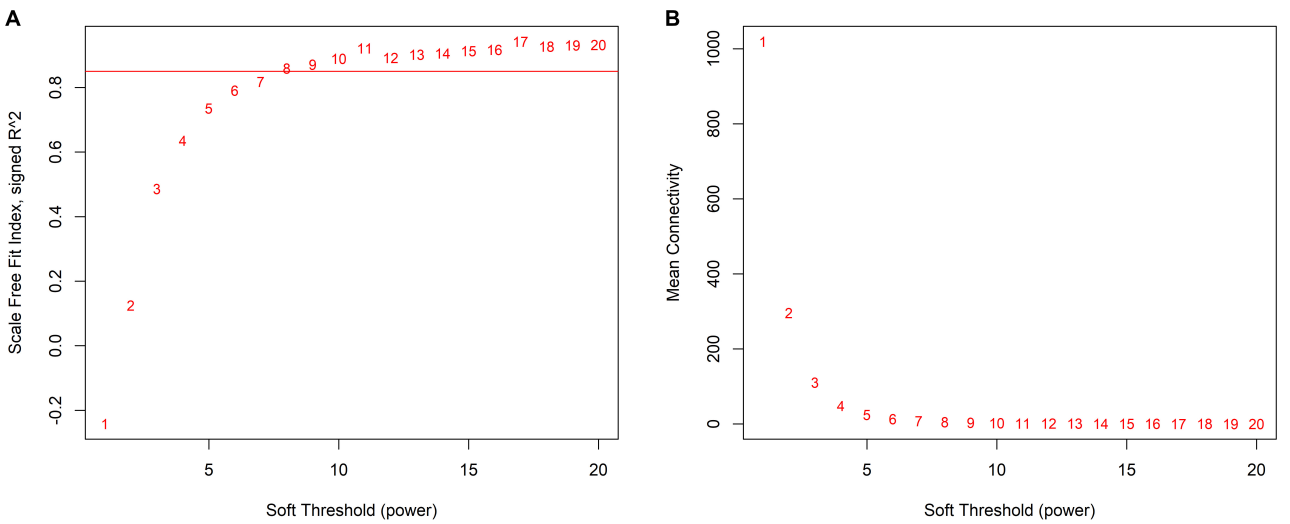

FIGURE 3 | Determination of appropriate soft-thresholding values (β). **(A)** The scale-free fit index of supposed β value from 1 to 20. **(B)** The mean connectivity of supposed β value from 1 to 20.

Weighted Gene Co-expression Network Construction and Module Division

We decided 8 as the proper β value (**Figure 3**) and converted the expression matrix into a topological overlap matrix according to the method mentioned before. Then, 9 co-expression modules were divided from all 5,580 genes and distinguished with colors (**Figure 4A**). The brown, black, magenta, blue, turquoise, pink, green, yellow, and gray modules contained 711, 534, 77, 1,037, 1,394, 82, 408, 672, and 665 genes, respectively. The gray module contained genes that couldn't be divided into any co-expression modules (all modules and corresponding genes are available in **Supplementary Table S2**).

In order to verify the accuracy of the module division, we mapped an adjacency heatmap of all analyzed genes (**Figure 4B**). The results indicated that genes showed stronger co-expression relationships with genes from the same module and weaker relationships with genes in other modules, which proved the preciseness of the module division.

Module-Trait Relationship Analysis and Identification of ccRCC-Related Module

We calculated the correlation coefficients and corresponding statistical significance between module eigengenes and clinical traits and showed the results with a heatmap (**Figure 5A**). We found that the brown module was positively related to clinical trait of pathological stage and tumor grade of ccRCC simultaneously. That is to say, the up-regulation of genes in

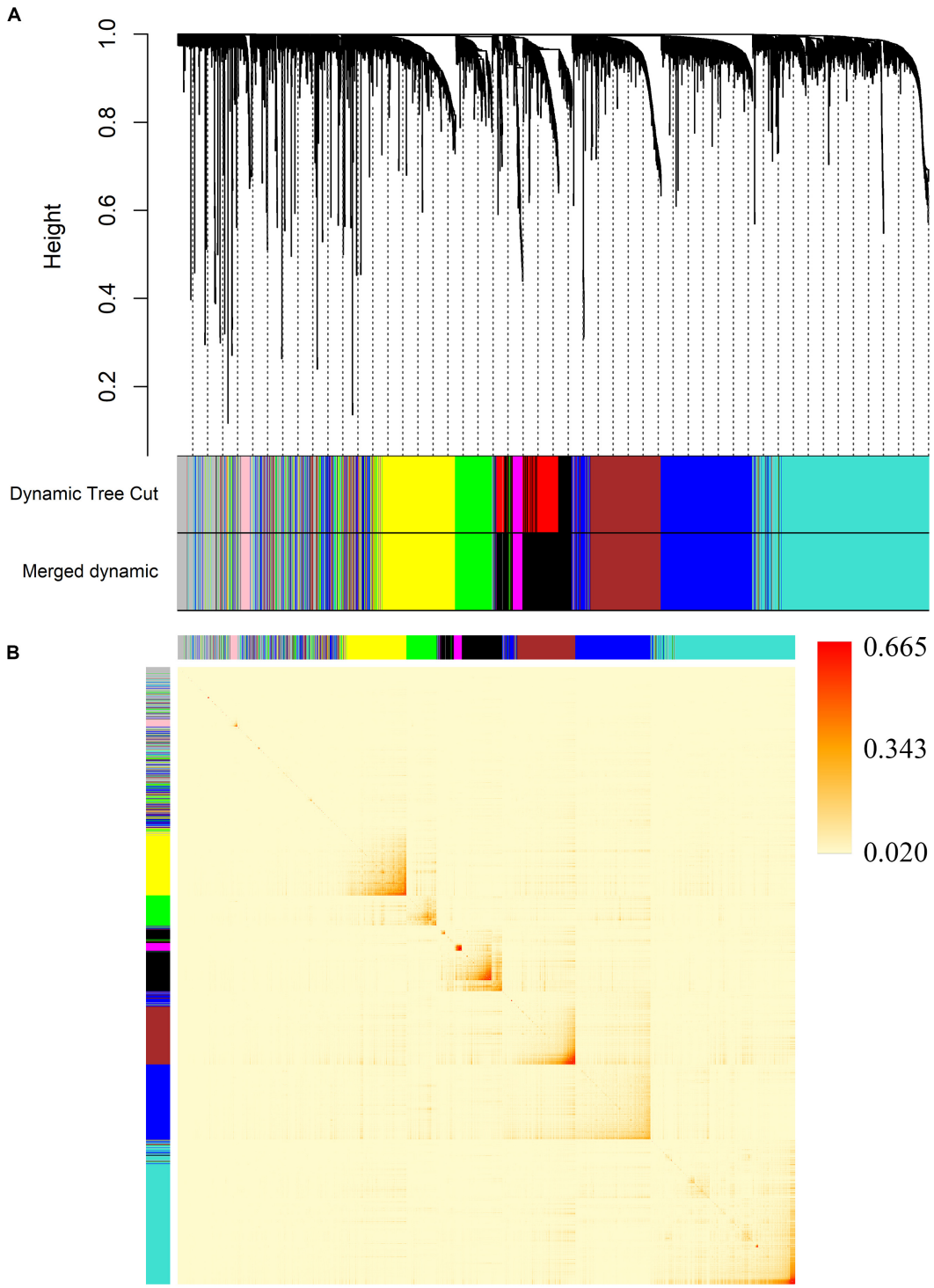

FIGURE 4 | Division of co-expression gene modules and an adjacency heatmap of genes. **(A)** Dendrogram of all 5,580 genes clustered by average linkage hierarchical clustering based on the TOM-based dissimilarity measure. **(B)** The adjacency heatmap of 5,580 genes. The color intensity represents the weighted correlation coefficients between pairwise genes.

the brown module may have grate efforts on the metastasis of ccRCC. We renamed the brown module as the ccRCC-related module.

Further, we evaluated the correlation of GS and MM in the ccRCC-related module to find high correlation coefficients (cor = 0.54, p = 4.6E-55 and cor = 0.65, p = 1.4E-86,

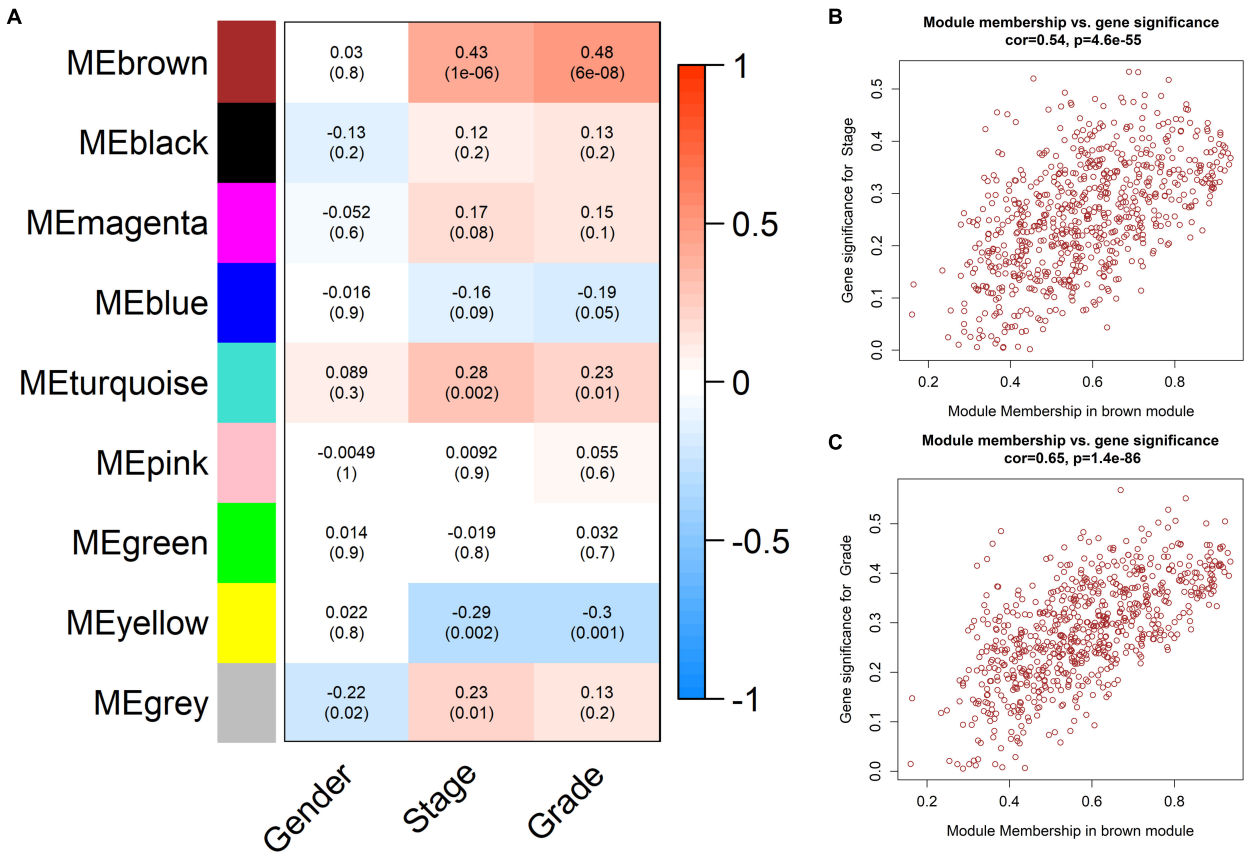

FIGURE 5 | Determination of the ccRCC-related module. **(A)** Heatmap of trait-module correlations. The color of each cell reflects the correlation coefficients (red for positive correlation and blue for negative correlation). The statistical significances of correlation relationships were noted in the cells. The brown module was thus identified as a ccRCC-related module. **(B)** Scatter plot displaying correlation between Gene significance (GS) and Module Membership (MM) of pathological stage of genes in the brown module. **(C)** Scatter plot displaying correlation between Gene significance (GS) and Module Membership (MM) of tumor grade of genes in the brown module.

Figures 5B,C), so that the ccRCC-related module was regarded as an appropriate module for subsequent analyses and hub gene extraction.

GO and KEGG Pathway Enrichment Analyses of the ccRCC-Related Module

We carried out GO and KEGG pathway enrichment analyses on genes in ccRCC-related module to find out the mainly enriched biological processes and signal pathways. **Figure 6** represents the top 10 terms of GO-BP and KEGG enrichment analyses (all enriched terms and the interpretations of the top-10 terms are available in **Supplementary Table S3**).

For GO-BP enrichment analysis (**Figure 6A**), the majority of the terms were about the procedure of mitotic cell division, such as "mitotic nuclear division" (gene count = 50, p = 1.11E-23), "sister chromatid segregation" (gene count = 46, p = 1.60E-22), and "nuclear division" (gene count = 59, p = 8.46E-22).

Results of KEGG pathway enrichment analysis (**Figure 6B**) were similar to that of GO-BP enrichment analysis. Genes in the brown module were mainly enriched in signal pathways about cell division, such as "Cell cycle" (gene count = 24, p = 2.82E-11) and "DNA replication" (gene count = 7, p = 0.0003), which are directly connected with the excessive proliferation of cells in tumor. Meanwhile, several items have already been proved to participate in the occurrence and development of multiple tumors, such as "IL-17 signaling pathway" (gene count = 15, p = 1.90E-06), "p53 signaling pathway" (gene count = 12, p = 1.46E-05), and "NF-kappa B signaling pathway" (gene count = 13, p = 0.0001).

Excavation of the Hub Gene

We extracted genes as well as their weighted co-expression coefficients from the GO term of "mitotic nuclear division" (the most significantly enriched GO-BP term) and constructed a sub-network of the co-expression network. With the weighted correlations among genes, we analyzed the centrality of genes in the sub-network with MCC method (only top 500 correlations were concerned, **Supplementary Table S4**). Genes with higher MCC values were regarded as connecting more closely with others and playing

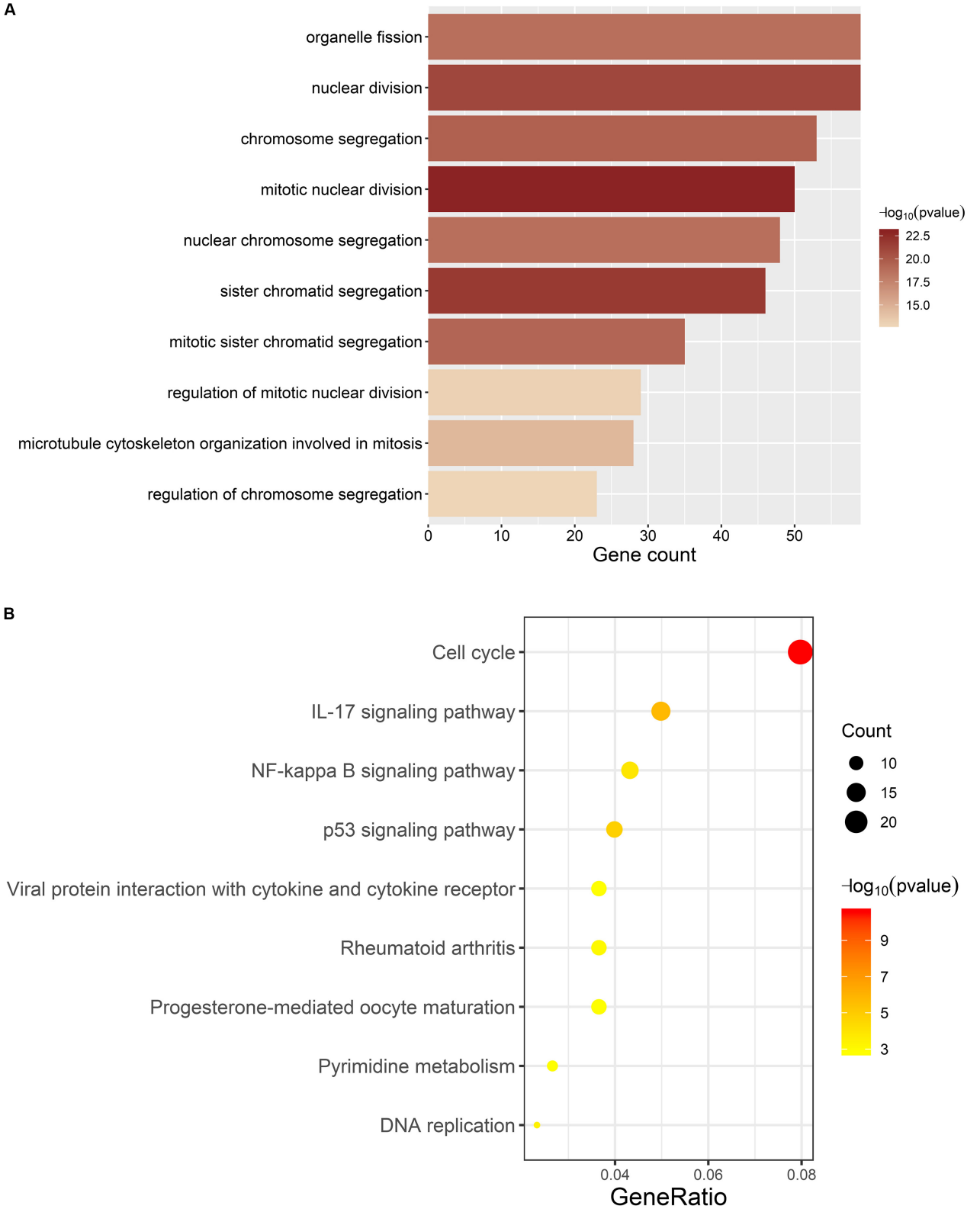

FIGURE 6 | Results of enrichment analyses. **(A)** Top 10 GO-BP terms. The length and color of barplot displays the enriched gene count and the statistical significance for each term respectively. **(B)** All KEGG terms. The size and color of the dot represents the enriched gene count and statistical significance for each term, respectively. Gene ratio means the ratio of enriched gene count and all genes involved in KEGG enrichment analysis.

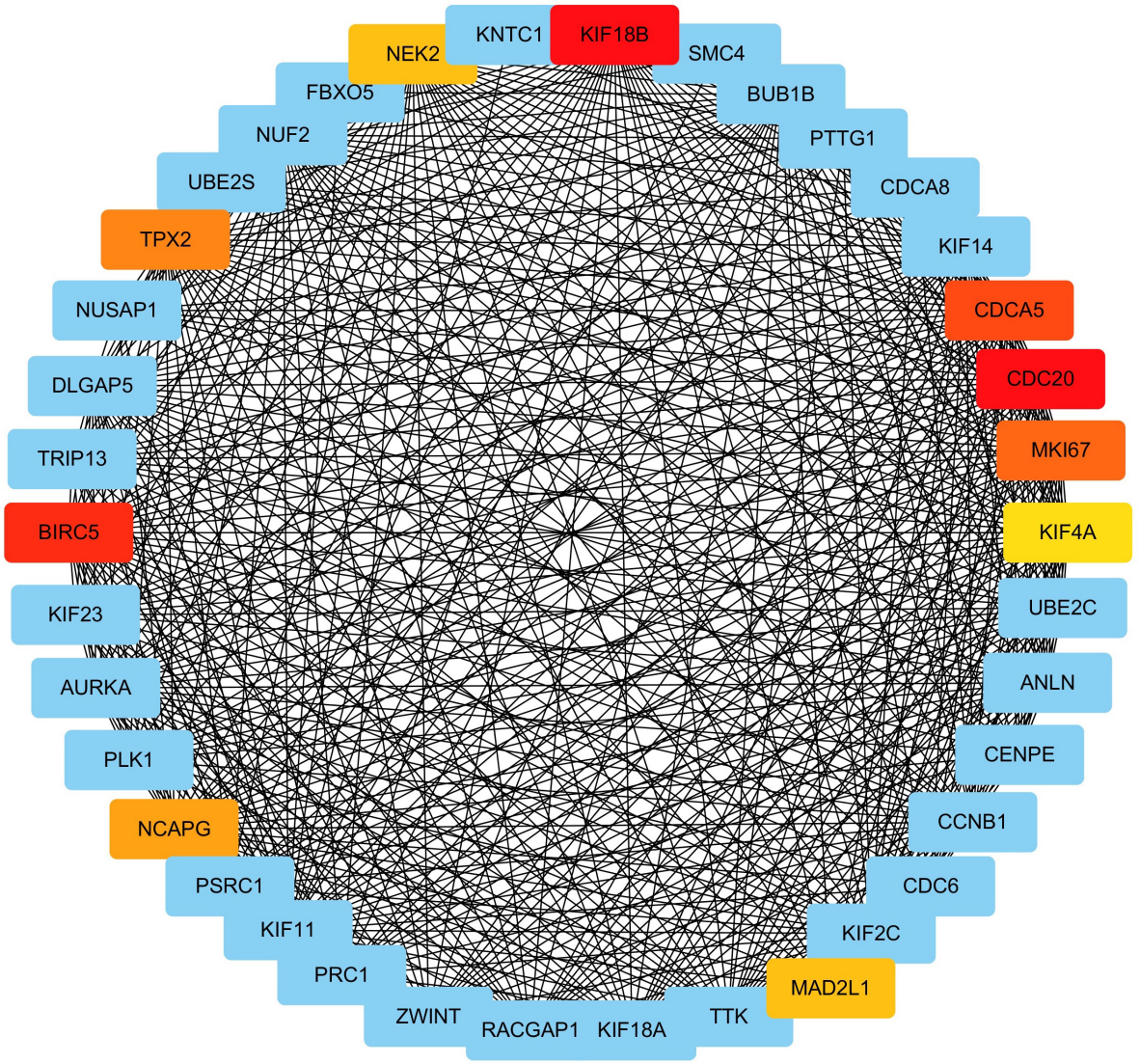

FIGURE 7 | Identification of hub gene by MCC method. The MCC values of genes in the GO term "mitotic nuclear division" were calculated. Genes with top 10 MCC values were colored with red and yellow color and other genes was colored with blue. Among genes with top-10 MCC values, red means relative bigger MCC value and yellow means relative smaller MCC values, and the same color means the same MCC values.

crucial roles in the co-expression relationship. The results are displayed in **Figure 7** and the top 10 central genes are highlighted.

The top 10 genes were *KIF18B*, *BIRC5*, *CDC20*, *CDCA5*, *MKI67*, *TPX2*, *NCAPG*, *MAD2L1*, *NEK2*, and *KIF4A*. *KIF18B* had the highest MCC value and is deemed as the hub gene of ccRCC as a result.

Validation of KIF18B by Differentially Expressed Gene Analysis

We explored the differentially expressed level of *KIF18B* between ccRCC and normal control. The results illustrated that *KIF18B* was significantly up-regulated in ccRCC compared with control (logFC = 2.724, p = 8.05E-35, **Figure 8A**).

Validation of KIF18B by Survival Analysis

We validated by survival analysis that ccRCC patients with higher expression level of *KIF18B* would have obviously worse prognostic outcomes and shorter overall survival. The 5-year survival rate of the high-expression group and the low-expression group was 50 and 75%, respectively (p = 9.48E-7, **Figure 8B**).

Exploration of the Relationship Between the Expression of KIF18B and Clinical Traits

We analyzed the correlation between different expression group of *KIF18B* and clinical traits by chi-square test. The results verified that different expression group of *KIF18B* is significantly related to tumor grade (p = 2.43E-05), pathological stage

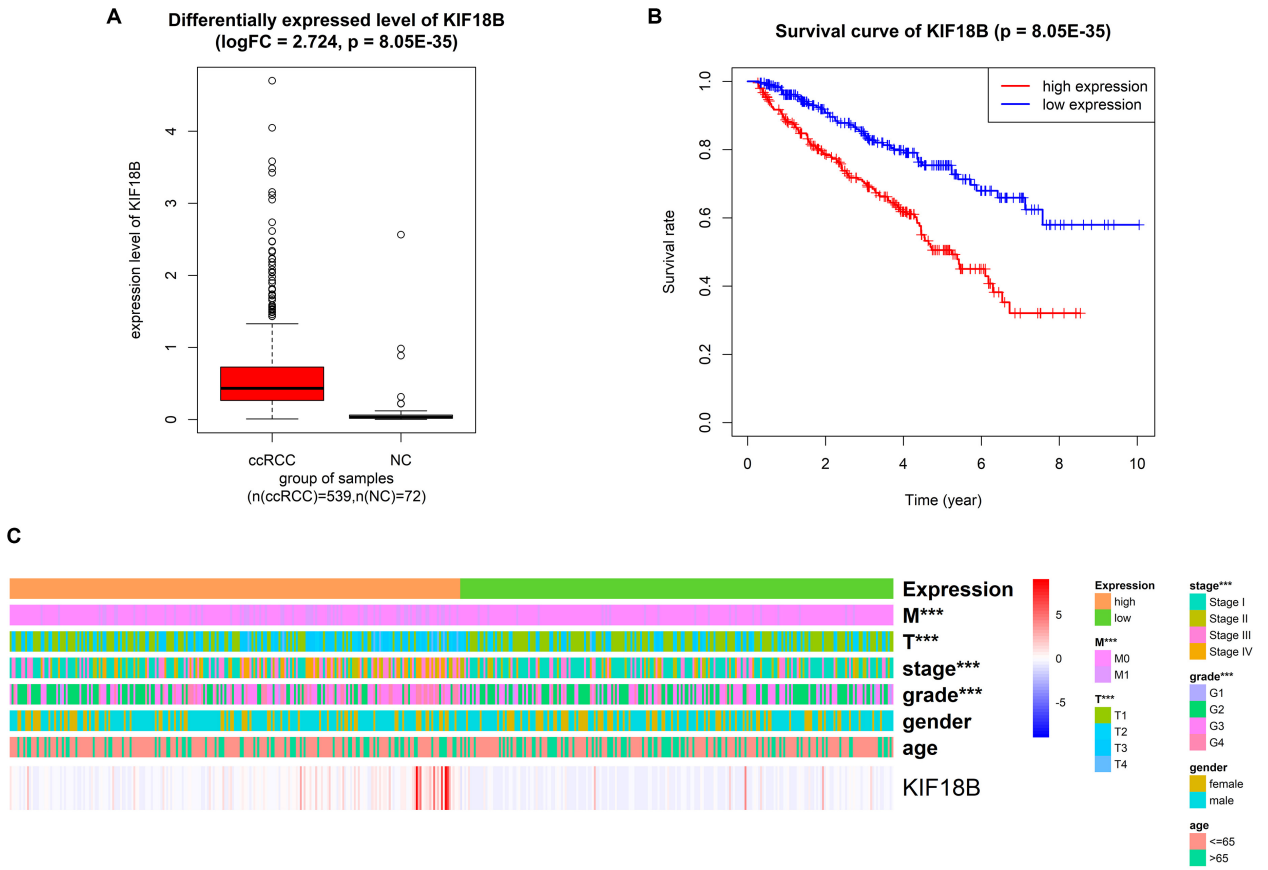

FIGURE 8 | Validation of the reliability of *KIF18B* with external dataset (TCGA). **(A)** Differentially expressed gene analysis of *KIF18B* between ccRCC and normal control. **(B)** Overall survival analysis on *KIF18B*. **(C)** Correlation analysis between expression level of *KIF18B* and clinical traits (***$p < 0.001$).

(p = 6.394E-06), T stage (p = 3.178E-06), and distant metastasis (p = 2.043E-05), but it isn't related to gender and age (**Figure 8C**; results of chi-square test are given in **Supplementary Table S5**).

Correlation Analysis Between the Expression of KIF18B and Targets of Precise Treatments

The results of correlation analysis showed that the expression level of *KIF18B* is positively related to the expression level of *PDL1* (cor = 0.3788, p = 6.149E-11, **Figure 9A**) and is negatively related to the expression level of *VEGFR3* (cor = −0.381, p = 1.441E−10, **Figure 9B**), which hints that patients with high-expression of *KIF18B* might respond better to treatments targeting *PDL1* and worse to treatments targeting *VEGFR3*.

Single-Gene GSEA on KIF18B

Figure 9C exhibits the top-5 up-regulated and top-5 down-regulated signal pathways in the high-expression group of *KIF18B* (all significant enriched terms are given in **Supplementary Table S6**). The results indicated that *KIF18B* is involved in the biological processes of "Base excision repair," "DNA replication," "Mismatch repair," and "homologous recombination" in ccRCC.

DISCUSSION

In the current study, we selected the gene expression profile of GSE73731 from GEO database for conducting WGCNA and extracting the hub gene. For data processing, we removed the outlier samples after sample clustering and filtered genes of low variance. We conducted WGCNA and divided all genes into 8 meaningful co-expression modules. After trait-module correlation analysis, the brown module was identified as a key module of positive correlation with higher pathological stage and tumor grade in ccRCC, which means that genes in this module were related with the development and metastasis of ccRCC. Subsequent GO and KEGG enrichment analyses indicated that genes in the ccRCC-related module were mainly enriched in biological function of cell cycle, cell proliferation, tumor metastasis, and material metabolism, which agreed with the characteristics of tumor well. To screen out the hub gene of ccRCC from the ccRCC-related module, we pulled out the genes in the top GO-BP term and constructed a sub-network of WGCNA. Then, *KIF18B* was confirmed as the hub candidate gene with MCC method. We then validated the reliability of *KIF18B* with the external dataset obtained from TCGA database. DEGs analysis found that *KIF18B* was up-regulated in ccRCC

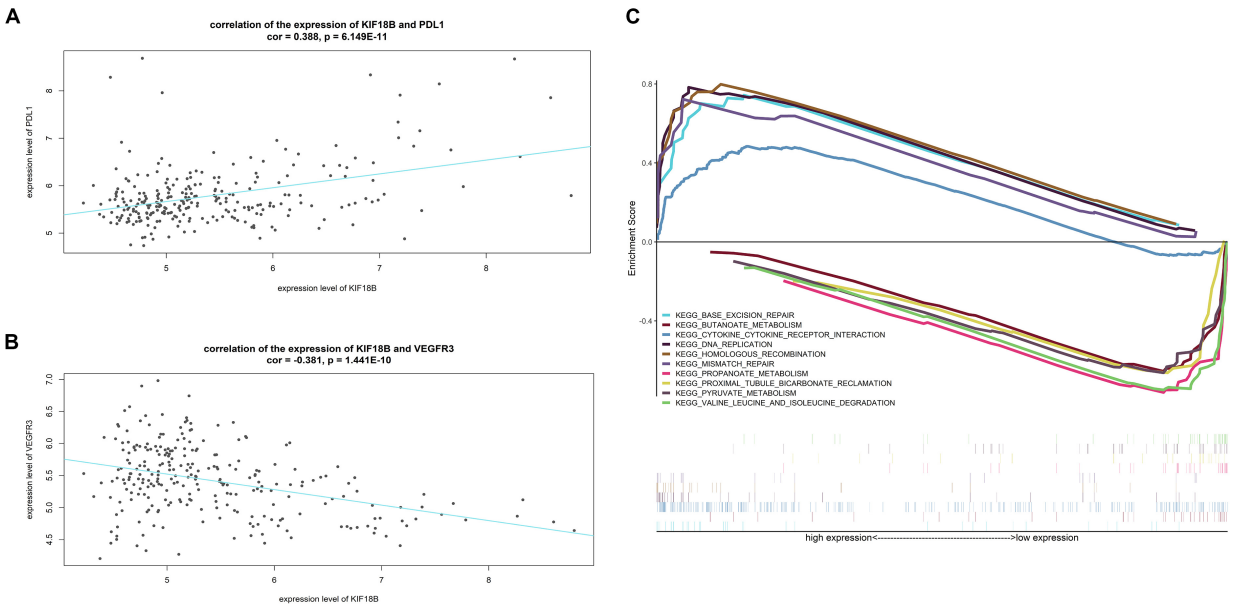

FIGURE 9 | Function exploration of *KIF18B*. **(A)** Correlation analysis between the expression level of *KIF18B* and *PDL1*. **(B)** Correlation analysis between the expression level of *KIF18B* and *VEGFR3*. **(C)** Single-gene GSEA analysis of *KIF18B*.

patients compared with normal control and OS analysis revealed a worse prognostic outcome in patients with high expression level of *KIF18B*. Correlation analysis between the expression of KIF18B and clinical traits proved that the expression of *KIF18B* is significantly related to tumor grade, pathological stage, T stage, and distant metastasis. Finally, in order to explore the function and clinical significance of *KIF18B*, we analyzed the correlation between the expression level of *KIF18B* and targets of precise therapies and found the expression level of *KIF18B* is correlated with that of *PDL1* and *VEGFR3*. Meanwhile, single-gene GSEA revealed that *KIF18B* is mainly involved in DNA replication and mutation in ccRCC.

WGCNA is a powerful bioinformatics tool for extracting hub genes participating in the pathogenesis and affecting the prognosis of tumor. Most of the top 10 hub genes identified by MCC method in our study have already been proved as hub candidate genes of ccRCC. Numerous studies agreed that the overexpression of *BIRC5* means more advanced pathological stage, severer metastasis, and shortened overall survival (Pu et al., 2017). Studies concerning the effects of *CDC20* (Yuan et al., 2017), *KIF4A* (Wei et al., 2019), *NEK2* (Arai et al., 2015), *TPX2* (Wei et al., 2019), and *NCAPG* (Wei et al., 2019) on ccRCC have drawn similar conclusions. Most importantly, the overexpression of *MKI67* is closely related with the overall survival, pathological stage, and Fuhrman grade in ccRCC, so that it is regarded as a biomarker for the disease (Xie et al., 2017). As *KIF18B* was identified as a hub gene together with the above-mentioned genes by MCC method and WGCNA showed apparent co-expression relationships among *KIF18B* and these hub genes, the conclusions propose by other researchers demonstrate the high credibility of our results indirectly. After identification of *KIF18B*, we validated it with an external dataset obtained from TCGA to confirm the reliability of our conclusion. Finally, we explored the functions and the relationship of *KIF18B* with targets of precise therapies to discover the clinical value of our hub gene. In brief, *KIF18B* is worthy of deeper research as a hub candidate gene in ccRCC.

KIF18B is a protein-coding gene that encodes kinesin family member 18B, which is a member of over 40 different kinds of kinesin proteins (Hirokawa et al., 2009). Kinesin functions with dynein as motor proteins to carry out microtubule-regulated movement in many vital biological processes such as cell division and cargo transport (Stout et al., 2011; Wu et al., 2006). *KIF18B* mainly locates in the nucleus and its expression is cell cycle-dependent. Researches have shown that the expression level of *KIF18B* is remarkably elevated at late Second-Gap/Metaphase in cell cycles, demonstrating that *KIF18B* may act as an important mitosis-regulating motor protein (Lee et al., 2010). Studies have shown that *KIF18B* plays an important role during the process of Metaphase by regulating the movement of chromosomes from the spindle poles toward the spindle equator (Rath and Kozielski, 2012).

As it hasn't been long since researchers noticed the importance of *KIF18B* for maintaining normal biological functions, there are relatively few studies about the relation between the disordered expression of *KIF18B* and diseases. Yet, several studies have shown close connections between *KIF18B* and cancer.

Wu et al. (2018) observed significant up-regulation of *KIF18B* in cervical cancer compared with normal control, and the up-regulation is positive correlated with the size of the primary tumor and tumor grade. The invasion capacities of the tumor cells were weakened after the knockdown of *KIF18B* by siRNAs *in vitro*. On the contrary, the overexpression of *KIF18B* promotes the proliferation, invasion, and migration of cancer cells.

They hypothesized that *KIF18B* may function through *Wnt*/β-catenin pathway and verified the down-regulation of *C-myc*, β-catenin, and phosphorylated *GSK3*β after the knockdown of *KIF18B*. Moreover, the volumes and weight of the tumor were obviously reduced in mouse model after down-regulation treatment of *KIF18B*. The results have uncovered that *KIF18B* acts as a potential oncogene in cervical cancer.

Another research (Xiang X. H. et al., 2019) reveals that *KIF18B* is down-regulated in senescent cells and abnormally up-regulated in hepatocellular carcinoma cells, and the overexpression of *KIF18B* was a risk factor for poorer survival. Other reports have drawn similar conclusions in lung adenocarcinoma (Zhang et al., 2019) and bladder cancer (Pan et al., 2019). Accordingly, we suppose the high utilization value of *KIF18B* as a biomarker for prognosis evaluating and a specific target for precise treatment in ccRCC.

At the end of the article, we'd like to enumerate several limitations and future directions of our research. Firstly, the whole study was carried out on the basis of public databases (GEO and TCGA), so that we are attempting to collect some samples by ourselves and validate the results in more external datasets. Secondly, we will attempt to explore the effect of *KIF18B* on its related genes by analyzing the receptor-ligand relationships of *KIF18B* by further analyses of single-cell RNA-seq. Finally, the function of *KIF18B* should be further validated by not only bioinformatics analyses but also experiments in cells or animals. We'll devote ourselves to conduct those studies to make our conclusions more complete.

CONCLUSION

In summary, our study identified and validated *KIF18B* as a hub candidate gene of ccRCC by WGCNA and a series of systematic bioinformatics analyses. *KIF18B* may act as a potential biomarker for prognosis prediction, precise treatment in the future. Our conclusions provided novel insights for uncovering the mechanism of ccRCC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73731. Data obtained from TCGA database is available at: https://portal.gdc.cancer.gov/.

AUTHOR CONTRIBUTIONS

HY and YW conceived, designed, and conducted the study, as well as wrote the manuscript. ZZ selected and pre-processed the data. All authors reviewed the manuscript and participated in the language modification.

FUNDING

This study was supported by the Natural Science Foundation of Zhejiang Province (Grant No. LQ19H050012). The funders had no role in how the studies were conducted or interpreted.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00905/full#supplementary-material

REFERENCES

Albiges, L., Powles, T., Staehler, M., Bensalah, K., Giles, R. H., Hora, M., et al. (2019). Updated european association of urology guidelines on renal cell carcinoma: immune checkpoint inhibition is the new backbone in first-line treatment of metastatic clear-cell renal cell carcinoma. *Eur. Urol.* 76, 151–156. doi: 10.1016/j.eururo.2019.05.022

Arai, E., Gotoh, M., Tian, Y., Sakamoto, H., Ono, M., Matsuda, A., et al. (2015). Alterations of the spindle checkpoint pathway in clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. *Int. J. Cancer* 137, 2589–2606. doi: 10.1002/ijc.29630

Bando, S. Y., Iamashita, P., Silva, F. N., Costa, L. D. F., Abe, C. M., Bertonha, F. B., et al. (2019). Dynamic gene network analysis of Caco-2 Cell response to shiga toxin-producing escherichia coli-associated hemolytic-uremic syndrome. *Microorganisms* 7:195. doi: 10.3390/microorganisms7070195

Barata, P. C., and Rini, B. I. (2017). Treatment of renal cell carcinoma: current status and future directions. *CA Cancer J. Clin.* 67, 507–524.

Campbell, M. T., Bilen, M. A., Shah, A. Y., Lemke, E., Jonasch, E., Venkatesan, A. M., et al. (2018). Cabozantinib for the treatment of patients with metastatic non-clear cell renal cell carcinoma: a retrospective analysis. *Eur. J. Cancer* 104, 188–194. doi: 10.1016/j.ejca.2018.08.014

Chaffer, C. L., and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science* 331, 1559–1564. doi: 10.1126/science.1203543

Chen, R., Ge, T., Jiang, W., Huo, J., Chang, Q., Geng, J., et al. (2019). Identification of biomarkers correlated with hypertrophic cardiomyopathy with co-expression analysis. *J. Cell Physiol.* 234, 21999–22008. doi: 10.1002/jcp.28762

Chin, C. H., Chen, S., Wu, H., Ho, C., Ko, M., Lin, C., et al. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8(Suppl. 4):S11. doi: 10.1186/1471-2105-9-11

Gupta, K., Miller, J. D., Li, J. Z., Russell, M. W., and Charbonneau, C. (2008). Epidemiologic and socioeconomic burden of metastatic renal cell carcinoma (mRCC): a literature review. *Cancer Treat. Rev.* 34, 193–205. doi: 10.1016/j.ctrv.2007.12.001

Hirokawa, N., Noda, Y., Tanaka, Y., and Niwa, S. (2009). Kinesin superfamily motor proteins and intracellular transport. *Nat. Rev. Mol. Cell Biol.* 10, 682–696. doi: 10.1038/nrm2774

Hsieh, J. J., Purdue, M. P., Signoretti, S., Swanton, C., Albiges, L., Schmidinger, M., et al. (2017). Renal cell carcinoma. *Nat. Rev. Dis. Primers* 3:17009.

Katrinli, S., Lori, A., Kilaru, V., Carter, S., Powers, A., Gillespie, C. F., et al. (2019). Association of HLA locus alleles with posttraumatic stress disorder. *Brain Behav. Immun.* 81, 655–658. doi: 10.1016/j.bbi.2019.07.016

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9:559. doi: 10.1186/1471-2105-9-559

Lee, Y. M., Kim, E., Park, M., Moon, E., Ahn, S., Kim, W., et al. (2010). Cell cycle-regulated expression and subcellular localization of a kinesin-8 member human KIF18B. *Gene* 466, 16–25. doi: 10.1016/j.gene.2010.06.007

Ljungberg, B., Bensalah, K., Canfield, S., Dabestani, S., Hofmann, F., Hora, M., et al. (2015). EAU guidelines on renal cell carcinoma: 2014 update. *Eur. Urol.* 67, 913–924.

Lu, Y., Rosenfeld, R., Simon, I., Nau, G. J., and Bar-Joseph, Z. (2008). A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.* 36:e109. doi: 10.1093/nar/gkn434

Ochocki, J. D., Khare, S., Hess, M., Ackerman, D., Qiu, B., Daisak, J. I, et al. (2018). Arginase 2 suppresses renal carcinoma progression via biosynthetic cofactor pyridoxal phosphate depletion and increased polyamine toxicity. *Cell Metab.* 27, 1263–1280.e6. doi: 10.1016/j.cmet.2018.04.009

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., et al. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34.

Pan, S., Zhan, Y., Chen, X., Wu, B., and Liu, B. (2019). Identification of biomarkers for controlling cancer stem cell characteristics in bladder cancer by network analysis of transcriptome data stemness indices. *Front. Oncol.* 9:613. doi: 10.3389/fnagi.2019.00613

Pu, Z., Wu, G., and Wang, Q. (2017). Clinicalpathological and prognostic significance of survivin expression in renal cell carcinoma: a meta-analysis. *Oncotarget* 8, 19825–19833. doi: 10.18632/oncotarget.15082

Rath, O., and Kozielski, F. (2012). Kinesins and cancer. *Nat. Rev. Cancer* 12, 527–539. doi: 10.1038/nrc3310

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Stout, J. R., Yount, A. L., Powers, J. A., Leblanc, C., Ems-McClung, S. C., Walczak, C. E., et al. (2011). Kif18B interacts with EB1 and controls astral microtubule length during mitosis. *Mol. Biol. Cell* 22, 3070–3080. doi: 10.1091/mbc.e11-04-0363

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tang, R., and Liu, H. (2019). Identification of temporal characteristic networks of peripheral blood changes in Alzheimer's Disease based on weighted gene co-expression network analysis. *Front. Aging Neurosci.* 11:83. doi: 10.3389/fnagi.2019.00083

Wei, W., Lv, Y., Gan, Z., Zhang, Y., Han, X., Xu, Z., et al. (2019). Identification of key genes involved in the metastasis of clear cell renal cell carcinoma. *Oncol. Lett.* 17, 4321–4328.

Wei, X., Choudhury, Y., Lim, W. K., Anema, J., Kahnoski, R. J., Lane, B., et al. (2017). Recognizing the continuous nature of expression heterogeneity and clinical outcomes in clear cell renal cell carcinoma. *Sci. Rep.* 7:7342.

Wu, X., Xiang, X., and Hammer, J. A. III (2006). Motor proteins at the microtubule plus-end. *Trends Cell Biol.* 16, 135–143. doi: 10.1016/j.tcb.2006.01.004

Wu, Y., Wang, A., Zhu, B., Huang, J., Lu, E., Xu, H., et al. (2018). KIF18B promotes tumor progression through activating the Wnt/beta-catenin pathway in cervical cancer. *Oncol. Targets Ther.* 11, 1707–1720. doi: 10.2147/ott.s157440

Xiang, X. H., Yang, L., Zhang, X., Ma, X., Miao, R., Gu, J., et al. (2019). Seven-senescence-associated gene signature predicts overall survival for Asian patients with hepatocellular carcinoma. *World J. Gastroenterol.* 25, 1715–1728. doi: 10.3748/wjg.v25.i14.1715

Xiang, Z., Li, J., Song, S., Wang, J., Cai, W., Hu, W., et al. (2019). A positive feedback between IDO1 metabolite and COL12A1 via MAPK pathway to promote gastric cancer metastasis. *J. Exp. Clin. Cancer Res.* 38:314.

Xie, Y., Chen, L., Ma, X., Li, H., Gu, L., Gao, Y., et al. (2017). Prognostic and clinicopathological role of high Ki-67 expression in patients with renal cell carcinoma: a systematic review and meta-analysis. *Sci. Rep.* 7:44281.

Yu, G., Wang, L., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

Yuan, L., Chen, L., Qian, K., Qian, G., Wu, C., Wang, X., et al. (2017). Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC). *Genom Data* 14, 132–140. doi: 10.1016/j.gdata.2017.10.006

Zhang, L., Zhu, G., Wang, X., Liao, X., Huang, R., Huang, C., et al. (2019). Genomewide investigation of the clinical significance and prospective molecular mechanisms of kinesin family member genes in patients with lung adenocarcinoma. *Oncol. Rep.* 42, 1017–1034.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 21 August 2020
doi: 10.3389/fphys.2020.00830

Quantifying Glioblastoma Drug Response Dynamics Incorporating Treatment Sensitivity and Blood Brain Barrier Penetrance From Experimental Data

Susan Christine Massey[1], *Javier C. Urcuyo*[1], *Bianca Maria Marin*[2], *Jann N. Sarkaria*[2] *and Kristin R. Swanson*[1,3,4]*

[1] Precision Neurotherapeutics Innovation Program, Mayo Clinic, Phoenix, AZ, United States, [2] Department of Radiation Oncology, Mayo Clinic, Rochester, MN, United States, [3] Department of Neurological Surgery, Mayo Clinic, Phoenix, AZ, United States, [4] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, United States

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Hermann Frieboes,
University of Louisville, United States
Dominic G. Whittaker,
University of Nottingham, United Kingdom

***Correspondence:**
Kristin R. Swanson
swanson.kristin@mayo.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: 29 February 2020
Accepted: 22 June 2020
Published: 21 August 2020

Citation:
Massey SC, Urcuyo JC, Marin BM, Sarkaria JN and Swanson KR (2020) Quantifying Glioblastoma Drug Response Dynamics Incorporating Treatment Sensitivity and Blood Brain Barrier Penetrance From Experimental Data. Front. Physiol. 11:830. doi: 10.3389/fphys.2020.00830

Many drugs investigated for the treatment of glioblastoma (GBM) have had disappointing clinical trial results. Efficacy of these agents is dependent on adequate delivery to sensitive tumor cell populations, which is limited by the blood-brain barrier (BBB). Additionally, tumor heterogeneity can lead to subpopulations of cells with different sensitivities to anti-cancer drugs, further impacting therapeutic efficacy. Thus, it may be important to evaluate the extent to which BBB limitations and heterogeneous sensitivity each contribute to a drug's failure. To address this challenge, we developed a minimal mathematical model to characterize these elements of overall drug response, informed by time-series bioluminescence imaging data from a treated patient-derived xenograft (PDX) experimental model. By fitting this mathematical model to a preliminary dataset in a series of nonlinear regression steps, we estimated parameter values for individual PDX subjects that correspond to the dynamics seen in experimental data. Using these estimates as a guide for parameter ranges, we ran model simulations and performed a parameter sensitivity analysis using Latin hypercube sampling and partial rank correlation coefficients. Results from this analysis combined with simulations suggest that BBB permeability may play a slightly greater role in therapeutic efficacy than relative drug sensitivity. Additionally, we discuss recommendations for future experiments based on insights gained from this model. Further research in this area will be vital for improving the development of effective new therapies for glioblastoma patients.

Keywords: glioblastoma, blood–brain barrier, drug sensitivity, epidermal growth factor receptor (EGFR), parameter estimation

1. INTRODUCTION

Glioblastoma (GBM) is an aggressive primary brain cancer that is notoriously difficult to treat due to its diffuse infiltration into surrounding normal-appearing brain (Giese et al., 2003). These diffusely invading GBM cells cannot be completely resected surgically (Baldock et al., 2014), and are difficult to target with radiation therapy while sparing normal brain (Corwin et al., 2013).

As a result, clinicians rely on chemotherapy to treat the full extent of the tumor. However, chemotherapeutic efficacy can be limited in two main ways: there may be insufficient delivery across the blood–brain barrier (BBB), and the tumor may not be uniformly sensitive to the agent.

The BBB acts to keep pathogens and many toxins out of the sensitive brain tissue. Angiogenesis in dense tumor regions induces disruption of the BBB, potentially allowing chemotherapeutic drugs to "leak" into these tumor regions. Current dogma in neuro-oncology holds this as being largely sufficient to treat the tumor, but GBM cells invade beyond these regions into tissue where the BBB remains rather intact (Van Tellingen et al., 2015). Further, tissue interstitial pressure and drug properties such as lipophilicity and polarity may influence the delivery of drugs across angiogenesis-induced BBB "leaks" (Ningaraj, 2006). Due to these factors, it remains unclear whether the delivery of BBB-impermeable antineoplastic agents reaches adequate concentrations throughout the tissue to provide the anticipated therapeutic effect.

Drug insensitivity (which includes, but is not limited to resistance) of tumor subpopulations is also a key suspect behind unsuccessful molecularly-targeted therapy results (Wen and Kesari, 2008; Ene and Holland, 2015). GBMs frequently present with gene mutations and/or amplification for a number of targets, such as epidermal growth factor receptor (EGFR). However, due to the spatial heterogeneity of GBM, these targets may have been identified for a subpopulation that is predominant in the dense tumor core, but less common in the invading portions of the tumor. There may also be activated compensatory signaling pathways in some cells or tumor regions that confer reduced responsiveness to the drug. Thus, while therapies already exist for molecular targets identified in GBM (Nagane et al., 1996; Brennan et al., 2013; Eskilsson et al., 2017; Reardon et al., 2017; van den Bent et al., 2017), a significant proportion of tumor may be less sensitive to these drugs, potentially explaining why they have failed in clinical trials (de Groot et al., 2008; De Witt Hamer, 2010; Reardon et al., 2010; AbbVie, 2019). However, it has been difficult to separate the possible insensitivity related causes of drug failure from that of inadequate delivery across the BBB and distribution throughout the tumor, since the majority of these drugs were not developed specifically for brain.

In order to explore both the contributions of inadequate delivery of therapy across the BBB and drug insensitivity, we developed a minimal mathematical model based on experimental data from preclinical subjects treated with an EGFR-targeted antibody drug conjugate (ADC). First, we describe model development based on this data, which consists of two tumor subpopulations with high vs. low sensitivity to the ADC therapy, and steps to estimate parameter regimes via data-fitting. Next, we explore the global model parameter sensitivity to understand how these parameters impact model outcomes. Finally, we run model simulations for the data-derived parameter regimes to assess the relative contributions of drug distribution and sensitivity, and discuss how it might be useful in assessing results from future experiments comparing different tumor models or different drugs. Overall, our model suggests that the degree of drug exposure may be more impactful than the relative sensitivity to therapy between the tumor subpopulations. Thus, in order to improve treatment outcomes, it is critical to determine predictors of drug distribution in individual patients' tumors and surrounding brain tissue to ensure invading tumor cells are adequately exposed to the therapy.

2. METHODS

Our ordinary differential equation (ODE) model of tumor growth and treatment response accounts for both variable treatment exposure and differential sensitivity to treatment by different tumor subpopulations. Development of this model was informed by experimental observations, which were also used to determine relevant parameter regimes for running simulations.

2.1. Experimental Data

The form of our model was based on experimental data from testing an EGFR-targeted antibody drug conjugate (ADC) in a patient-derived xenograft (PDX) model of GBM (Marin et al., 2018). These experiments were performed in full accordance with the guidelines of the Mayo Clinic Institutional Animal Care and Use Committee. The GBM12 PDX line used in this model is derived from a primary GBM in a male patient, and is EGFR amplified, MGMT methylated, and IDH1 and IDH2 wildtype. Full detail regarding this PDX line is available from the Mayo Clinic Brain Tumor Patient-Derived Xenograft National Resource (https://www.mayo.edu/research/labs/translational-neuro-oncology/mayo-clinic-brain-tumor-patient-derived-xenograft-national-resource/), where the line was developed and is maintained. These cells were implanted intracranially into 10 female athymic nude mice, and into the flank of 10 additional mice. After tumors were established, a subset of the surviving mice in each group was treated with either 10 mg/kg of a sham control antibody (Ab-095; four mice in each), or with 5 mg/kg of ABT-414 (an ADC also known as depatuxizumab mafodotin; five mice in each), administered via tail vein injection every 7 days. Tumor growth was monitored via bioluminescent imaging (BLI, **Figure 1**). Since BLI flux is *linearly* correlated with tumor cell number (Hartung et al., 2014), this provided us with a close approximation of tumor cell populations across time. Importantly, the data from PDX tumors grown in the flank and brain allowed us to compare treatment effect in tumors with and without BBB impediments to drug distribution.

2.2. Treatment Exposure and Sensitivity Model

Our model consists of three coupled ordinary differential equations describing the dynamics of both cell populations (H,

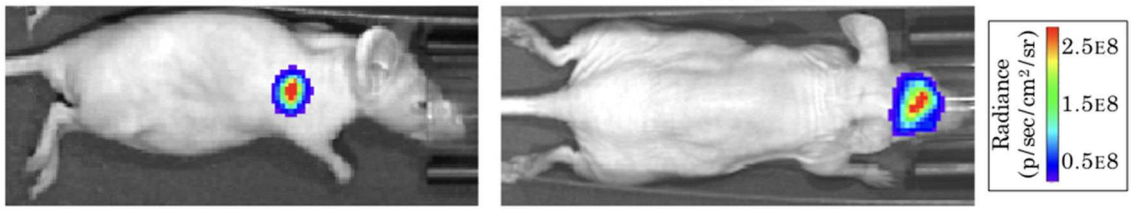

FIGURE 1 | Example bioluminescence images for patient-derived xenografts. Colors represent the BLI radiance (in photons/second/cm^2/steradian), which is related to the BLI flux (measured in photons per second) for the total area.

L) and the ADC (A):

$$\frac{dH}{dt} = \underbrace{\rho H}_{\text{proliferation}} - \underbrace{\gamma \mu_H A H}_{\text{drug-induced apoptosis}} \tag{1a}$$

$$\frac{dL}{dt} = \underbrace{\rho L}_{\text{proliferation}} - \underbrace{\gamma \mu_L A L}_{\text{drug-induced apoptosis}} \tag{1b}$$

$$\frac{dA}{dt} = \underbrace{\sum_{n=1}^{N} A_{\text{dose}}(n)\delta\,(t - 7n)}_{\text{drug dose given at time } t} - \underbrace{\lambda A}_{\text{drug decay}} \tag{1c}$$

where parameters and their definitions are outlined in **Table 1**, and their derivations can be found in section 2.3.

In the absence of the ADC, both highly sensitive (H) and less-sensitive (L) tumor populations grow exponentially, at proliferation rate ρ. However, the two populations differ in sensitivity to the ADC, A, which is captured by the drug-induced apoptosis rates μ_H and μ_L (for populations with high and low sensitivity, respectively). The terms for tumor cell death due to ADC are further modified by factor γ, which represents the proportion of cells exposed to ADC. We assume that the ADC is readily distributed to flank PDXs such that tumor cell exposure is high ($\gamma = 1$), but that the BBB limits this distribution for intracranial PDXs ($0 \leq \gamma \leq 1$). In order to capture the ADC dynamics, we let $A_{\text{dose}}(n)$ represent the nth dose, with doses administered every seven days, as noted by the dirac delta function $\delta(t - 7n)$. The ADC then decays at rate λ. These dynamics are schematized in **Figure 2**.

The model as described incorporates several assumptions that are useful to note explicitly. (1) While in theory there may be many groups of cells with varying levels of therapeutic sensitivity (some of which may even encapsulate resistant subpopulations), the model divides these into two main groups: those of relatively high therapeutic sensitivity and those with lower sensitivity to the therapy. (2) Any drug effect on proliferation rate is captured by the cell death rate parameter, as these effects are indistinguishable with the available data. (3) The effects of the tissue environment in the flank vs. the brain on tumor *growth* are encapsulated in environment-specific cellular proliferation rate parameter values (ρ_{flank} and ρ_{IC}, as described in section 2.3). (4) Relatedly, the cell death rate due to therapy for the sensitive tumor subpopulation, μ_H, is assumed to be the same intracranially as in the flank setting, only modified by drug exposure after crossing the BBB in intracranial tumors. That is, the model assumes that the only environmental effect on *treatment* is its distribution. (5) ADC and tumor subpopulations are well-mixed. In reality there is likely spatial variation in both flank and intracranial sites due to the tumor microenvironment and different blood vessel densities in particular tumor regions, but in the absence of spatially–resolved data, we assume well-mixedness and use an ODE model.

TABLE 1 | Model parameter definitions and values.

Symbol	Definition	Value range	Units
ρ	cellular proliferation rate	0.2–0.5	day^{-1}
μ_H	ADC-mediated high sensitivity cell kill rate	1–10	mg^{-1}day^{-1}
μ_L	ADC-mediated low sensitivity cell kill rate	$z\mu_H$	mg^{-1}day^{-1}
q	proportion of implanted cells with low sensitivity	0–1	unitless
z	relative sensitivity (μ_L/μ_H)	0–1	unitless
λ	rate of ADC decay	ln(2)/7	day^{-1}
γ	proportion of tumor exposed	0–1	unitless
A_{dose}	ADC given in a single dose	0.1	mg

Parameter value ranges were estimated through fitting the model to experimental data or parameters were confined to a value range by their theoretical meaning, except in the case of ADC parameters, as described in section 2.3.

This model can be solved analytically, as shown in **Appendix**. For simplification, at any given time t, $C(t)$ represents the total number of cells, calculated by the sum of high sensitivity $H(t)$ and low sensitivity $L(t)$ cells. This total cell number was used in section 2.3 for comparing with bioluminescence imaging data, which shows the total tumor cell population. The initial proportion of total implanted cells with low sensitivity is denoted by $q = L_0/C_0$. Similarly, the extent to which these cells L are less sensitive to the agent than the highly sensitive cells H is denoted by the relative sensitivity ratio $z = \mu_L/\mu_H$, which is bounded between 0 and 1 to ensure that μ_L is a fraction of μ_H in the regression-based parameterization in section 2.3. With these notational changes, we can then write the analytical solution (derived in **Appendix**) as

$$C(t) = C_0 e^{\rho t}\left(q e^{-\gamma \mu_H \int A(t)dt} + (1-q)e^{-\gamma z \mu_H \int A(t)dt}\right), \tag{2}$$

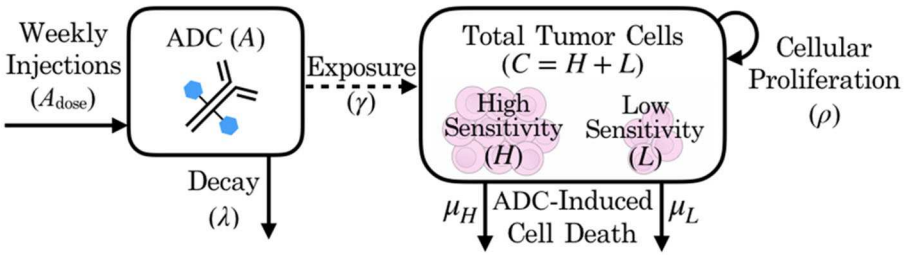

FIGURE 2 | Schematic of patient-derived xenograft response to an antibody drug conjugate, including key variables and parameters of the mathematical model.

where

$$\int A(t)dt = \sum_{n=1}^{N} 2^n A_{\text{dose}}(n) \frac{\left(e^{7n\lambda} - e^{\lambda t}\right)\theta(t-7n)}{\lambda}. \quad (3)$$

Using this solution (2), the model can be parameterized through comparison of simulations to time-series BLI data.

2.3. Data-Based Parameter Estimation

Most model parameters were unknown, with the exception of ADC-specific parameters: the timing of dose administration and dose amounts [$A_{\text{dose}}(n)$], as well as the half-life of the drug, which allowed us to solve for the drug decay rate (λ). Dose amounts were adjusted for the weight of each animal (5 mg/kg), so we applied the average initial animal weight of 20 g to obtain the constant ADC dose, $A_{\text{dose}} = 0.1$ mg used in simulations. All of the remaining model parameters were determined through several iterations of fitting the model via least squares regression to preliminary BLI data from an experiment. The various arms of the experiment included untreated and treated groups of subjects, as well as flank and intracranial tumor sites to separate out BBB influences. By fitting the model to these various subgroups, we were able to identify and estimate each of the parameters, as described below.

Step 1: Fit to untreated data to estimate growth rate, ρ. When fitting the model to untreated data, since the ADC is not injected ($A = 0$), the model's treatment components zero out and only an exponential growth function remains: $C(t) = C_0 e^{\rho t}$. Fitting this model function to untreated data via least squares regression with the `lsqcurvefit` function in MATLAB® (MATLAB Release 2018b, The MathWorks, Inc., Natick, Massachusetts, United States), we were able to obtain estimates of the tumor proliferation rate, ρ, and the number of viable implanted PDX cells, C_0 (**Figure 3**). (While a consistent number of cells are initially implanted for each subject, C_0 is in fact unique for each, as a variable number of cells die off, possibly due to an inability to establish themselves in the proper microenvironment for growth.) This yielded subject-specific values for ρ and C_0 (which were bound on the intervals [0, inf) and $[10^2, 10^{10}]$, respectively), and the mean ρ was recorded as the net proliferation rate for the cells of the particular PDX line used in the experiments grown in either the flank (ρ_{flank}) or intracranial (ρ_{IC}) setting. As noted in the model assumptions, these site–specific ρ parameter values from untreated tumors are used to help to account for microenivironmental effects on PDX growth in the two different locations that are independent of the BBB.

Step 2: Fit to treated flank data to estimate μ_H, z, and q. Using the estimated net proliferation rate ρ_{flank} from the previous step, we proceed to fit the treated data in the flank. We assume the estimate of ρ_{flank} remains the same in the treated case as untreated, since the microenvironment remains similar and any differences should be encapsulated in the treatment effect term. Additionally, since the tumor was injected in the flank, there is no BBB effect to limit the proportion of tumor exposed to the ADC, such that the exposure parameter $\gamma = 1$. The initial condition C_0 is fit using the initial untreated time point and the passed mean ρ_{flank} value. Pairing these with other known parameters (see **Table 1**), the only remaining three unknown parameters to be fit to the data are the cell death rates due to drug, μ_H and μ_L, for the two cell populations and the proportion of implanted cells that have low sensitivity, q. Using the definition $z = \mu_L/\mu_H$ and the analytical solution of the model (2), we can then apply a nonlinear least squares regression (again using `lsqcurvefit`) to fit subject-specific parameters for parameters μ_H, z, and q (bound on the intervals [1, 10], [0, 1] and $[10^{-10}, 10^{-2}]$, respectively).

Step 3: Fit to treated intracranial data to estimate γ, z, and q. Proceeding to fit the data from treated intracranial tumors, we apply the same approach to estimate parameters as in the flank, this time assuming that the estimate of ρ_{IC} from the untreated setting remains the same for the treated intracranial tumors due to a similar microenvironment. Again we determine the initial condition C_0 by fitting the untreated model with the passed mean ρ_{IC} value to the initial untreated time point. Because we assume that the cell death rate due to ADC for the highly sensitive tumor subpopulation (μ_H) is the same intracranially as in the flank setting, we pass the average μ_H value determined in Step 2 and estimate parameter γ, the fraction of tumor exposed to therapy, in addition to parameters z and q.

At the conclusion of these steps (summarized in **Figure 3**), all unknown model parameters had net and individual estimates. Note that no more than three parameters were fitted with any experimentally-derived data set, in order to reduce the potential for overfitting. To examine this further, initial parameter guesses for `lsqcurvefit` were selected from within the value ranges listed in **Table 1**, and we had mostly consistent convergence using low, middle, and high values (see **Supplementary Material**). Additionally, as we show in the **Figure 3** plots with dash–dot

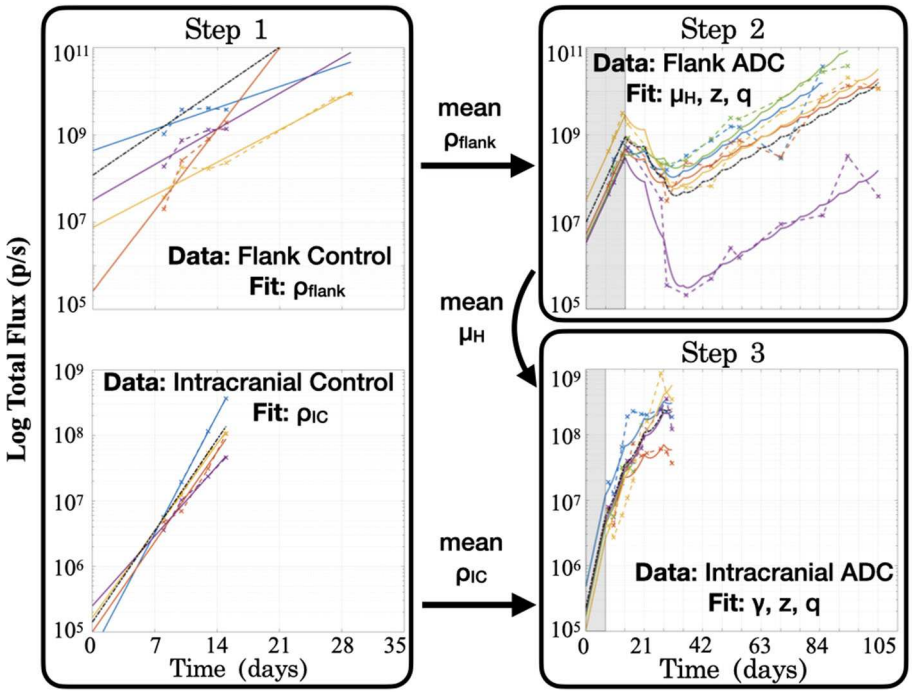

FIGURE 3 | Summary of the series of steps used to estimate model parameter values through fitting the model to different experimental data sets. **Step 1** consists of two separate fittings for the two different tumor locations, and the mean ρ values are passed for use in subsequent steps in the respective tumor sites, while the mean μ_H value from the flank found in **Step 2** is passed for use in **Step 3**. Shaded regions for Steps 2 and 3 indicate time prior to initiation of treatment with the ADC. Black dash–dot lines in each plot are simulations using the averaged fitted parameter values across all subjects within the group. It is worth noting that at each step, no more than three parameters are fitted to the data, in order to ensure identifiability and prevent overfitting.

lines, the simulations that result when using the averages of the fitted parameter values across all the subjects capture the dynamics of the data well. Thus, the averaged ρ_{flank}, ρ_{IC}, and μ_H values that we pass for later fitting steps correspond well with the group as a whole. Using these values then allowed us to run simulations in a reasonable range of parameter values, as well as to perform a model sensitivity analysis to understand how variability in these values affect model outcomes.

3. RESULTS

3.1. Parameter Sensitivity Analysis

Due to the uncertainty and variability in our parameter estimates, it was important to better characterize the effects of parameters on model results. To do this, we conducted a parameter sensitivity analysis via Latin hypercube sampling (LHS) and partial-rank correlation coefficients (PRCC) (McKay et al., 1979; Iman and Helton, 1988; Blower and Dowlatabadi, 1994). To perform the LHS analysis, we first drew 1,000 equiprobable samples for each unknown parameter, including the initial condition C_0, from a statistical distribution of values. These distributions were informed by our fits of the preliminary data when available; in the case of the unitless parameters, we assumed a uniform distribution on the interval $[0, 1]$. These samples were then randomly paired in a Latin hypercube scheme to run a series of 1,000 Monte Carlo simulations. Using these simulation results, we then computed PRCCs between each parameter and two different model outcomes across all time points: the total number of tumor cells and the fraction of tumor that has low sensitivity (**Figure 4**). PRCC are computed using partial correlation applied to value ranks, as opposed to the actual values of the parameters and model outcome. Partial correlation helps control for effects due to other covariates, and ranks are used to evaluate the associative relationship between high and low values of parameters and the model outcomes, rather than the values themselves. Thus, the PRCC values at a given time point indicate how closely a high model outcome value relates to a high or low parameter value given at that time in the simulation. A PRCC close to 1 indicates a strong association between high parameter value and high model outcome value, and a PRCC close to -1 indicates a strong association between a low parameter value and a high model outcome value. For PRCC values between -0.5 and 0.5, the association is considered to be weak. Further details for about this method and the code files used are available on GitHub: https://github.com/scmassey/model-sensitivity-analysis.

3.1.1. Total Tumor Population Depends Most Strongly on Proliferation Rate, Followed by Treatment Response Parameters

At early time points, particularly before the initiation of therapy, the tumor population is strongly positively correlated with both

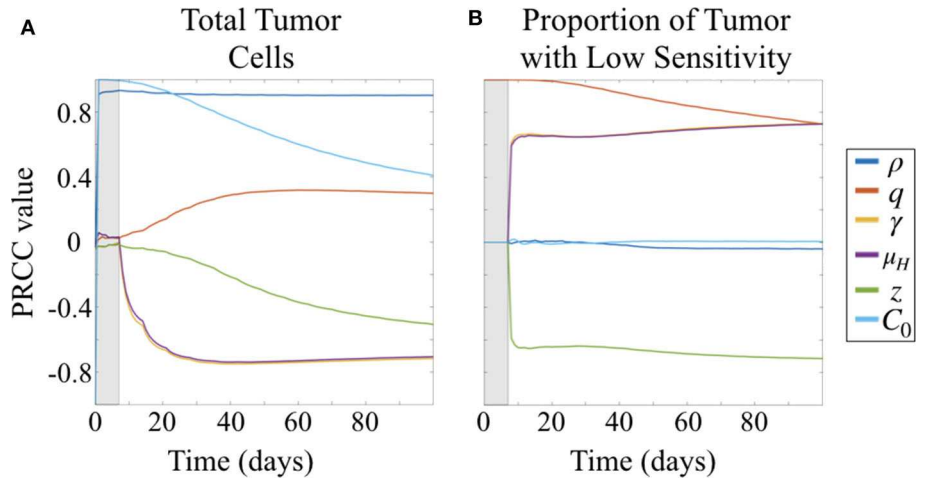

FIGURE 4 | Partial rank correlation coefficients (PRCC) of parameters with respect to **(A)** tumor cells and **(B)** fraction of cells with low sensitivity ($L/(H+L) = L/C$), visualized across simulation time.

the initial number of cells implanted, C_0 and proliferation rate ρ (**Figure 4A**). By 30 days, or after approximately three doses of therapy, the population remains strongly positively correlated with ρ but the effect of C_0 begins to wane. At the same time, drug sensitivity of the H cell population, μ_H, and exposure to drug, γ are strongly negatively correlated with total tumor cells. Relative sensitivity z, which determines the fraction of drug sensitivity in the L cell population, is also negatively correlated with total tumor, but less strongly, and only approaches a PRCC value of -0.5 after 100 days. This suggests that relative sensitivity z is less impactful than either the treatment response rate μ_H for the subpopulation with high ADC sensitivity or the degree of tumor exposure to ADC, γ.

Parameters γ and μ_H track together in the sensitivity analysis (overlapping lines in **Figure 4**). This is expected given our substitution $\mu_L = z\mu_H$, which results in the coefficient $-\gamma z\mu_H$ in the term describing drug induced apoptosis for the equation describing the L population (Equation 1b), mirroring that for the H population (Equation 1a), $-\gamma\mu_H$. Thus, sensitivity analysis is unable to compare the differential impacts of these two parameters, and highlights a potential parameter identifiability issue for our model. Since we had preliminary data in both the treated flank as well as the treated intracranial PDX settings, we were able to obtain parameter estimates for these by keeping $\gamma = 1$ in the flank setting, and assuming that μ_H is the same intracranially as in flank.

3.1.2. Less Sensitive Fraction of Tumor Driven by Initial Proportion of These Cells, Followed by Treatment Response Parameters

Prior to the initiation of therapy, only parameter q, the fraction of initially implanted cells that have low sensitivity, is correlated with the proportion of total tumor that has low sensitivity (**Figure 4B**). Once treatment is initialized, q remains highly positively correlated, and this correlation decreases slightly over time during the course of treatment.

Three other parameters show correlation with the fraction of tumor that has lower sensitivity following initiation of therapy, all of which involve drug response. Parameters γ and μ_H, representing the degree of tumor exposure to ADC and the ADC-induced cell kill rate of cells with high sensitivity, respectively, are both positively correlated and track together, while parameter z, representing the relative treatment sensitivity between the cells with higher and lower sensitivity, is negatively correlated with the fraction of tumor that has low sensitivity. Further, the PRCC values do not vary over the time of the simulation after treatment is initiated and sustained. These correlations are consistent with expectations from the behavior of the system described by the model.

3.2. Simulation Results

To more fully explore the effect of parameters on model predicted outcomes, we ran simulations for varied values of the parameters relating to treatment response: γ, μ_H, and z (degree of ADC exposure, the ADC response rate in cells with high sensitivity, and the relative sensitivity between the two tumor subpopulations, respectively). Codes used to run simulations and plot the results may be found on GitHub: https://github.com/scmassey/treatment-exposure-sensitivity-model.

3.2.1. Treatment Exposure Impacts Tumor Burden More Than Relative Sensitivity

Comparing simulation results across a range of values for parameters γ and z while holding μ_H fixed, we see that γ plays a larger influence on total tumor cells than does z. That is, looking across rows of drug exposure values γ, we see that for relative sensitivity $z > 0.6$, there is no variation in total tumor burden. For lower levels of drug exposure, this is even more pronounced, as tumor burden is quite high regardless of the relative sensitivity. This is consistent with the parameter sensitivity results of section 3.1.1, but shows the impact of this dynamic in greater detail.

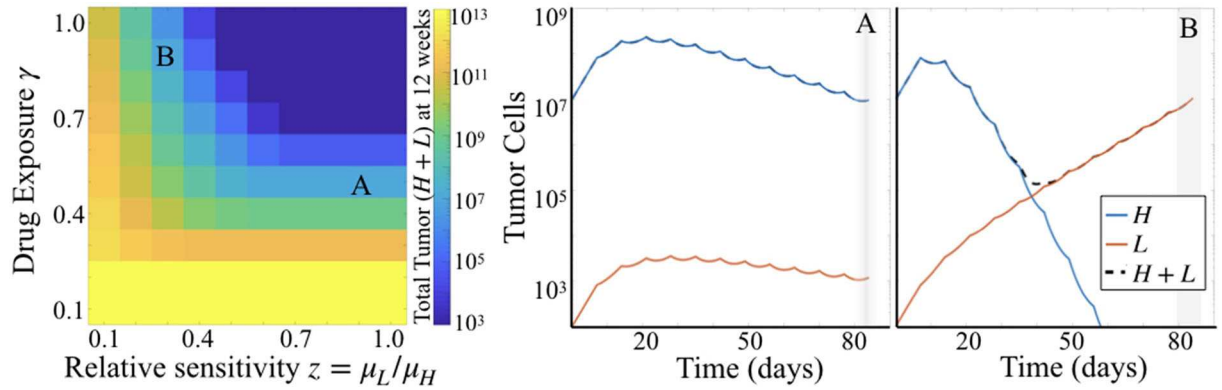

FIGURE 5 | Simulation results for varied γ and z. **(Left)** Heatmap shows total tumor cells at 84 days (12 weeks) post tumor initiation, across 10 values each of γ and z, for fixed $\mu_H = 5$. **(Right)** Two particular simulations corresponding to the labels (A,B) in the heatmap.

3.2.2. Different Subpopulation Proportions Can Yield Same Total Tumor Burden

Simulations also highlighted that there can be distinct differences in the dynamics of the two subpopulations of cells underlying predicted tumor burden (**Figure 5**). Looking at long time scales—in this case at 12 weeks or 84 days, the average survival time of the treated subjects—we observe the effect of an extended time of treatment in the simulations. Comparing two simulations with the same predicted tumor burden, we see that one simulation retains a large proportion of cells with high sensitivity (**Figure 5A**), while another is made up almost entirely of cells with low treatment sensitivity (**Figure 5B**). Thus, while the overall tumor may look similar at many points along its trajectory if sampled sparsely, one is about to be "uncontrolled" at later time points, while the other will stay relatively stable.

4. DISCUSSION

Failure of targeted therapies in glioblastoma can be attributed to many different causes, many of which are driven by various aspects of tumor heterogeneity. These include the potential mismatch of treatment to target beyond the center of the tumor and/or inadequate delivery of therapy to these cells invading outlying brain parenchyma. Often these have been investigated separately, focusing either on sensitivity (through optimizing targets or overcoming resistance) (Cloughesy et al., 2014), or engineering approaches for enhancing drug delivery (Liu et al., 2010; Van Tellingen et al., 2015). Infrequently, elements of both are combined (Stein et al., 2018), but even in those cases the relative impact of these upon treatment outcomes has not been compared. Thus, we created our Treatment Exposure and Sensitivity model describing tumor growth and treatment response incorporating both exposure and differential sensitivity to therapy, based on experimental data, to investigate the relationship between them. Through parameter sensitivity analysis and simulation, we found that both can contribute in similar ways, but exposure may have a greater impact overall. In particular, for simulated tumors that were given the same treatment responsiveness for the highly sensitive population, therapeutic exposure impacts tumor burden more than the relative sensitivity between the two tumor populations. Sensitivity analysis also revealed that parameter ρ is the strongest positive influence on total cell population, as expected, and distinguished between the effect of reduced therapeutic sensitivity (z) and exposure to therapy (γ) in reducing the total tumor population. Not only is there a difference in the magnitude of correlation between parameters γ and z with total tumor burden, there is also a difference in the temporal dynamics of the change in these correlations over time. The correlation coefficients between parameters and the fraction of the tumor with lower treatment sensitivity, however, is quite stable over time.

4.1. Limitations

Our model is relatively minimal by design, as the amount of available data constrains the number of model parameters we can fit. Thus, this model does not compare potential sources of differential treatment sensitivity (such as various mechanisms of resistance). Although tumor heterogeneity may actually provide for many populations with varying levels of sensitivity to therapy, as described in section 2.2 we assumed that these cluster toward more or less sensitive, reducing them to two. Larger data sets generated by similar studies in the future may support including more populations and additional mechanistic differences or interactions between them. Related to this, we were also limited in distinguishing BBB impacts from other microenvironmental effects when comparing data from tumors grown in the flank vs. the intracranial setting beyond the proliferation rate, ρ. For example, microenvironmental effects, such as vascular density, likely create regions of differential drug exposure (as well as regions of different tumor subpopulations). Experiments generating larger data sets and greater spatial detail would facilitate adding these features in future modeling efforts. Finally, while our present analysis of the model enables us to understand the overall dynamics between the parameters, we are not able to conclude anything with respect to efficacy of this particular drug among patients due to the use of a single PDX line. As we discuss

below, however, by using the parameter fitting method presented to estimate drug sensitivity and exposure across multiple PDX lines in future studies, it may be possible to gain insight in these factors across patients with this heterogeneous disease.

4.2. Model Recommendations for Future Experiments

Parameter sensitivity analysis of the model highlighted the necessity of having the flank and intracranial treatment groups for practical identifiability in obtaining estimates for therapeutic sensitivity (μ_H) and exposure (γ). This "tradeoff" between γ and μ_H was also observed in simulations and is expected given our model formulation. However, the emergence of this dynamic in the creation and parameter estimation of our model underscores that this relationship should be carefully considered in the design of experimental studies for new glioma therapies. Sensitivity analysis revealed that correlation coefficients between parameters and total tumor burden change most dramatically at earlier time points, and after approximately 50 days, change relatively little. This suggests that experiments conducted to examine the relationship between exposure and sensitivity to therapy should focus on collecting time course data more densely for the first 7 weeks as compared to longer times. As shown by simulations, there can be several parameterizations that fit tumor burden data at any single time which correspond to different proportions of cells with high and low sensitivity to treatment. Thus, time series data is essential for detecting differences in these and the contributions of the BBB (parameter γ) and relative sensitivity (z).

Because our Treatment Exposure and Sensitivity model is minimal and reduces mechanisms down to a few key parameters, it has great utility for fitting experimental data to estimate these parameters for individuals. Having these individual parameterizations is key to understanding the extent to which drug exposure and resistance each contributed to variations in outcome. In particular, quantifying drug sensitivity and exposure parameters for individual subjects within and between groups with additional therapies and PDX lines (better capturing interpatient, as well as intrapatient, heterogeneity) may be a promising avenue for future research. This will provide further insights for developing novel approaches to therapy optimization—including delivery—for individual patients.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by Mayo Clinic Institutional Animal Care and Use Committee.

AUTHOR CONTRIBUTIONS

SM and JU contributed to the software and analysis. SM, JU, and BM contributed to the investigation. SM contributed to the preparation writing of the original draft. JS and KS contributed to the funding acquisition. All authors contributed to the conceptualization, methodology, and the writing—review and editing.

FUNDING

This material was based upon work supported by the National Institutes of Health (U54CA210180).

ACKNOWLEDGMENTS

A previous version of this manuscript has been released as a pre-print at bioRxiv (Massey et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.00830/full#supplementary-material

REFERENCES

AbbVie (2019). *Abbvie Provides Update on Depatuxizumab Mafodotin (depatux-m), An Investigational Medicine for Newly Diagnosed Glioblastoma, an Aggressive Form of Brain Cancer [Press Release]*. Available online at: https://news.abbvie.com/news/press-releases/abbvie-provides-update-on-depatuxizumab-mafodotin-depatux-m-an-investigational-medicine-for-newly-diagnosed-glioblastoma-an-aggressive-form-brain-cancer.htm (accessed September 3, 2019).

Baldock, A. L., Ahn, S., Rockne, R., Johnston, S., Neal, M., Corwin, D., et al. (2014). Patient-specific metrics of invasiveness reveal significant prognostic benefit of resection in a predictable subset of gliomas. *PLoS ONE* 9:e99057. doi: 10.1371/journal.pone.0099057

Blower, S. M., and Dowlatabadi, H. (1994). Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. *Int. Stat. Rev.* 229–243. doi: 10.2307/1403510

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

Cloughesy, T. F., Cavenee, W. K., and Mischel, P. S. (2014). Glioblastoma: from molecular pathology to targeted treatment. *Annu. Rev. Pathol.* 9, 1–25. doi: 10.1146/annurev-pathol-011110-130324

Corwin, D., Holdsworth, C., Rockne, R. C., Trister, A. D., Mrugala, M. M., Rockhill, J. K., et al. (2013). Toward patient-specific, biologically optimized radiation therapy plans for the treatment of glioblastoma. *PLoS ONE* 8:e79115. doi: 10.1371/journal.pone.0079115

de Groot, J. F., Gilbert, M., Aldape, K., Hess, K. R., Hanna, T., Ictech, S., et al. (2008). Phase II study of carboplatin and erlotinib (tarceva, osi-774) in patients with recurrent glioblastoma. *J. Neuro-Oncol.* 90, 89–97. doi: 10.1007/s11060-008-9637-y

De Witt Hamer, P. C. (2010). Small molecule kinase inhibitors in glioblastoma: a systematic review of clinical studies. *Neuro-oncology* 12, 304–316. doi: 10.1093/neuonc/nop068

Ene, C. I., and Holland, E. C. (2015). Personalized medicine for gliomas. *Surg. Neurol. Int*. 6(Suppl. 1):S89. doi: 10.4103/2152-7806.151351

Eskilsson, E., Røsland, G. V., Solecki, G., Wang, Q., Harter, P. N., Graziani, G., et al. (2017). EGFR heterogeneity and implications for therapeutic intervention in glioblastoma. *Neuro-Oncology* 20, 743–752. doi: 10.1093/neuonc/nox191

Giese, A., Bjerkvig, R., Berens, M., and Westphal, M. (2003). Cost of migration: invasion of malignant gliomas and implications for treatment. *J. Clin. Oncol.* 21, 1624–1636. doi: 10.1200/JCO.2003.05.063

Hartung, N., Mollard, S., Barbolosi, D., Benabdallah, A., Chapuisat, G., Henry, G., et al. (2014). Mathematical modeling of tumor growth and metastatic spreading: validation in tumor-bearing mice. *Cancer Res*. 74, 6397–6407. doi: 10.1158/0008-5472.CAN-14-0721

Iman, R. L., and Helton, J. C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Anal*. 8, 71–90. doi: 10.1111/j.1539-6924.1988.tb01155.x

Liu, H.-L., Hua, M.-Y., Chen, P.-Y., Chu, P.-C., Pan, C.-H., Yang, H.-W., et al. (2010). Blood-brain barrier disruption with focused ultrasound enhances delivery of chemotherapeutic drugs for glioblastoma treatment. *Radiology* 255, 415–425. doi: 10.1148/radiol.10090699

Marin, B., Mladek, A., Burgenske, D., He, L., Hu, Z., Bakken, K., et al. (2018). Ddis-01. the antibody-drug conjugate ABT-414 demonstrates single-agent anti-cancer activity across a panel of GBM patient-derived xenografts. *Neuro-Oncology* 20(Suppl. 6):vi69. doi: 10.1093/neuonc/noy148.280

Massey, S. C., Urcuyo, J. C., Marin, B. M., Sarkaria, J., and Swanson, K. R. (2019). Quantifying glioblastoma drug response dynamics incorporating resistance and blood brain barrier penetrance from experimental data. *bioRxiv*. doi: 10.1101/822585

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21, 239–245. doi: 10.1080/00401706.1979.10489755

Nagane, M., Coufal, F., Lin, H., Bögler, O., Cavenee, W. K., and Huang, H.-J. (1996). A common mutant epidermal growth factor receptor confers enhanced tumorigenicity on human glioblastoma cells by increasing proliferation and reducing apoptosis. *Cancer Res*. 56, 5079–5086.

Ningaraj, N. S. (2006). Drug delivery to brain tumours: challenges and progress. *Expert Opin. Drug Deliv*. 3, 499–509. doi: 10.1517/17425247.3.4.499

Reardon, D. A., Desjardins, A., Vredenburgh, J. J., Gururangan, S., Friedman, A. H., Herndon, J. E., et al. (2010). Phase 2 trial of erlotinib plus sirolimus in adults with recurrent glioblastoma. *J. Neuro-Oncol*. 96, 219–230. doi: 10.1007/s11060-009-9950-0

Reardon, D. A., Lassman, A. B., van den Bent, M., Kumthekar, P., Merrell, R., Scott, A. M., et al. (2017). Efficacy and safety results of abt-414 in combination with radiation and temozolomide in newly diagnosed glioblastoma. *Neuro-Oncology* 19, 965–975. doi: 10.1093/neuonc/now257

Stein, S., Zhao, R., Haeno, H., Vivanco, I., and Michor, F. (2018). Mathematical modeling identifies optimum lapatinib dosing schedules for the treatment of glioblastoma patients. *PLoS Comput. Biol*. 14:e1005924. doi: 10.1371/journal.pcbi.1005924

van den Bent, M., Gan, H. K., Lassman, A. B., Kumthekar, P., Merrell, R., Butowski, N., et al. (2017). Efficacy of depatuxizumab mafodotin (abt-414) monotherapy in patients with EGFR-amplified, recurrent glioblastoma: results from a multi-center, international study. *Cancer Chemother. Pharmacol*. 80, 1209–1217. doi: 10.1007/s00280-017-3451-1

Van Tellingen, O., Yetkin-Arik, B., De Gooijer, M., Wesseling, P., Wurdinger, T., and De Vries, H. (2015). Overcoming the blood-brain tumor barrier for effective glioblastoma treatment. *Drug Resist. Updates* 19, 1–12. doi: 10.1016/j.drup.2015.02.002

Wen, P. Y., and Kesari, S. (2008). Malignant gliomas in adults. *New Engl. J. Med*. 359, 492–507. doi: 10.1056/NEJMra0708126

Disclaimer: The antibody drug conjugate ABT (also known as depatuxizumab mafodotin) was provided at no charge by AbbVie through its Investigator-Initiated studies program. The company had no role in the design of experiments, data collection, or analysis.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 28 August 2020
doi: 10.3389/fphys.2020.00976

Pinning Control for the p53-Mdm2 Network Dynamics Regulated by p14ARF

Oscar J. Suarez[1], Carlos J. Vega[1], Edgar N. Sanchez[1], Ana E. González-Santiago[2], Otoniel Rodríguez-Jorge[3], Alma Y. Alanis[4], Guanrong Chen[5] and Esteban A. Hernandez-Vargas[6]**

[1] Electrical Engineering Department, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Guadalajara, Mexico, [2] Biomedical Sciences Department, Centro de Investigación Multidisciplinario en Salud, Universidad de Guadalajara, Tonalá, Mexico, [3] Biochemistry and Molecular Biology Department, Instituto de Investigaciones Básicas y Aplicadas, Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico, [4] Computer Sciences Department, Universidad de Guadalajara, Guadalajara, Mexico, [5] Electrical Engineering Department, City University of Hong Kong, Hong Kong, China, [6] Frankfurt Institute for Advanced Studies, Frankfurt, Germany

p53 regulates the cellular response to genotoxic damage and prevents carcinogenic events. Theoretical and experimental studies state that the p53-Mdm2 network constitutes the core module of regulatory interactions activated by cellular stress induced by a variety of signaling pathways. In this paper, a strategy to control the p53-Mdm2 network regulated by p14ARF is developed, based on the pinning control technique, which consists into applying local feedback controllers to a small number of nodes (pinned ones) in the network. Pinned nodes are selected on the basis of their importance level in a topological hierarchy, their degree of connectivity within the network, and the biological role they perform. In this paper, two cases are considered. For the first case, the oscillatory pattern under gamma-radiation is recovered; afterward, as the second case, increased expression of p53 level is taken into account. For both cases, the control law is applied to p14ARF (pinned node based on a virtual leader methodology), and overexpressed Mdm2-mediated p53 degradation condition is considered as carcinogenic initial behavior. The approach in this paper uses a computational algorithm, which opens an alternative path to understand the cellular responses to stress, doing it possible to model and control the gene regulatory network dynamics in two different biological contexts. As the main result of the proposed control technique, the two mentioned desired behaviors are obtained.

Keywords: p53, Mdm2, p14ARF, pinning control, computational modeling

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Jaewhan Song,
Yonsei University, South Korea
Abdessamad Zerrouqi,
Medical University of Warsaw, Poland

****Correspondence:***
Edgar N. Sanchez
edgar.sanchez@cinvestav.mx
Esteban A. Hernandez-Vargas
vargas@fias.uni-frankfurt.de

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *23 August 2019*
Accepted: *17 July 2020*
Published: *28 August 2020*

Citation:
Suarez OJ, Vega CJ, Sanchez EN, González-Santiago AE, Rodríguez-Jorge O, Alanis AY, Chen G and Hernandez-Vargas EA (2020) Pinning Control for the p53-Mdm2 Network Dynamics Regulated by p14ARF. Front. Physiol. 11:976. doi: 10.3389/fphys.2020.00976

1. INTRODUCTION

Gene regulatory networks play key roles in every process of life, including cell cycle, metabolism, signal transduction, cell communication, and cellular differentiation. These complex biological networks use large amounts of data, necessary for modeling, analyzing, and controlling. Mathematical and computational methods are very helpful approaches for constructing network models at molecule level to predict cell behavior under normal conditions or pathological ones. Network topology and interactions between nodes (representing molecules, proteins, genes, mRNA, and others), and edges (establishing regulatory properties) describe the network dynamical behavior (Bolouri and Davidson, 2002). Different mathematical models have been developed for

studying gene regulatory networks, which can be divided into four classes (De Jong, 2002): the first ones are logical models, which describe regulatory networks qualitatively, namely, Boolean networks (Kauffman, 1969; Akutsu et al., 1999; Wang et al., 2012), probabilistic Boolean networks (Shmulevich et al., 2002a,b), Bayesian networks (Friedman et al., 2000; Rau et al., 2010), and Petri nets (Chaouiya, 2007; Karlebach and Shamir, 2008); the second ones are defined by continuous models such as ordinary differential equations (Chen et al., 1999; Szallasi et al., 2006; Cao et al., 2012), and the S-system formalism (Kikuchi et al., 2003; Wang et al., 2010); the third ones are single-molecule level models (Cai et al., 2006; Elf et al., 2007; Selvin and Ha, 2008), which account for interactions among molecules; and the last ones are hybrid models combining different formulations like discrete-time and continuous-time frameworks (Ahmad et al., 2006; Fromentin et al., 2010). The continuous-time approach consists in connecting a group of dependent variables to biochemical reaction kinetics. In this case, it is essential to assume that molecules have constant concentrations with respect to cellular compartments, in which their variations are continuous functions of time (Chen et al., 1999; Szallasi et al., 2006; Cao et al., 2012). This approach is adopted in the present paper.

On the other hand, control theory has rapidly developed for complex networks (Wang and Chen, 2002; Sorrentino et al., 2007; Liu and Barabási, 2016). Recently, research focuses on the important issue of how to incorporate control techniques for biological systems and networks (Nowzari et al., 2016; Vinayagam et al., 2016; Gao et al., 2017; Jiao et al., 2018; Papatsenko et al., 2018; Wang et al., 2019), such as pinning control for gene regulatory networks (Lin et al., 2014; Chen et al., 2016; Yue et al., 2017; Li et al., 2018; Burbano et al., 2019). In Lin et al. (2014), a Boolean network model to reproduce the two-phase dynamics of the p53 network in response to DNA damage is developed. In particular, two types of Boolean attractors are presented; the first one is an apoptosis attractor and the second one is a repair attractor. Based on this model, practical control schemes for steering into the apoptosis attractor in presence of DNA damage by pinning the state of a single node or perturbing the weight of a single link are applied. In Chen et al. (2016), an Autonomous Boolean Control Networks (ABCNs) for designing and analyzing the therapeutic intervention strategies are introduced. An important issue in therapeutic intervention is to design a control sequence steering an ABCN from an undesirable location (implying a diseased condition) to a desirable one (corresponding to a healthy condition). Based on this motivation, pinning control strategy is proposed for steering an ABCN from any given condition to the desired one in the shortest time. Cluster synchronization of the coupled genetic regulatory network, represent with ABCN model is investigated in Yue et al. (2017) with a directed topology and using the event-based strategy and pinning control; for this network, a synthetic regulatory network analogous to that in *Escherichia coli* is proposed with twelve states, where three states are the pinned nodes. In Li et al. (2018), a single-input pinning controller design for reachability of Boolean networks is proposed; in addition, different nodes are selected as the pinning ones by solving logical matrix equations and *Drosophila melanogaster* gene regulatory network is used to illustrate the effectiveness and feasibility of the developed method. Finally, Burbano et al. (2019) pinning controllability analysis of multiagent networks subjected to three different types of noise diffusion processes; namely, noise affecting the node dynamics, the communication links, and the pinning control action is done, the effectiveness of the theoretical results is illustrated in the genetic Toggle Switch, originally introduced in *E. coli*. This last publication uses a model based on stochastic differential equations.

Additionally, gene regulatory networks present responses to DNA damage such as cell cycle arrest, DNA repair, senescence, apoptosis among others. Among the main regulators of these responses, tumor suppressor protein p53 (*TP*53) has been recognized as the "guardian of the genome" and is a key component of cellular responses to genotoxic stress (Lane, 1992). p53 regulates the cellular response to genotoxic damage and prevents tumorigenesis by post-translational modifications and gene transactivation (Ashcroft et al., 2000). Without cellular stress, p53 remains inactive and latent due to targeted degradation by the protein E3 ubiquitin-protein ligase Mdm2 (from *MDM*2 proto-oncogene). Mdm2 binds to p53 and marks it for proteasome degradation, preventing p53 accumulation in the nucleus and its transcriptional activity (Momand et al., 1992). In this way, the activity of p53 and its negative regulation by Mdm2 are widely recognized as one of the main regulatory mechanisms in genotoxic stress response (Sionov and Haupt, 1999). The p53-Mdm2 network models have been studied in Wagner et al. (2005), Geva-Zatorsky et al. (2006), Sykes et al. (2006), Wee et al. (2009), Wee et al. (2012), and Hafner et al. (2017), which describe different patterns of response as promotion of transcriptional activities, post-translational modifications, component interactions, and degradation rates. Another important regulator of the p53-Mdm2 network is the tumor suppressor p14ARF (Alternate Reading Frame), one product of the *CDKN*2*A* gene. p14ARF inhibits Mdm2-dependent p53 degradation, through Mdm2-p14ARF complex formation (Zhang et al., 1998). Thus, in response to genotoxic stress induced by gamma-radiation, p14ARF binds directly to Mdm2, leading to an inhibition of Mdm2-mediated p53 ubiquitination and degradation, which increases p53 levels. These events are coupled with downstream signaling pathways, promoting behaviors such as cell cycle arrest, DNA repair, senescence, or apoptosis induction (Zhang et al., 1998; Sionov and Haupt, 1999; Parisi et al., 2002).

The present paper uses a continuous-time approach in a deterministic model of the p53-Mdm2 network regulated by p14ARF under gamma-radiation response (Leenders and Tuszynski, 2013). Pinning control (Li et al., 2004; Chen, 2017) is applied to regulate feedback-loop caused by Mdm2 overexpression stimuli as carcinogenic initial condition (Oliner et al., 1992; Nilbert et al., 1994; Dei Tos et al., 2000; Rayburn et al., 2005), considering no influence of potential mutations. Two desired behaviors are expected; the first one, restoration of an oscillatory pattern, and the second one, the achievement of an increased p53 level expression. p14ARF as pinned node is selected according to the virtual leader methodology (Ren and Beard, 2008; Lewis et al., 2013).

The novelty of the present paper is summarized as follows:

1. The p53 and Mdm2 proteins are controlled by regulating the p14ARF level, based on sensitivity analysis (Dickinson and Gelinas, 1976; Hamby, 1994) as done below.
2. Spanning tree (Ren and Beard, 2008; Lewis et al., 2013) of the p53-Mdm2 network regulated by p14ARF is obtained in order to select the pinned nodes, which are determined based on the virtual leader methodology (Ren and Beard, 2008; Lewis et al., 2013).
3. Pinning control is introduced using a control systems approach where the control dynamics is solved in conjunction with the systems dynamics.
4. The proposed pinning control scheme does not require to apply control inputs to all nodes; in this paper, only one node (the pinned one), ensures oscillatory pattern under gamma-radiation and increased expression of p53 levels for the p53-Mdm2 network.

2. METHODS

2.1. Model Description

The p53-Mdm2 network is key for determining cell behavior in response to cellular stress, such as DNA damage induced by gamma-radiation (Strigari et al., 2014; Hage-Sleiman et al., 2017), which can start a program of p53-dependent consequences such as cell cycle arrest, DNA repair, senescence, or apoptosis induction (Wagner et al., 2005; Geva-Zatorsky et al., 2006; Sykes et al., 2006; Wee et al., 2009, 2012; Hafner et al., 2017). In **Figure 1**, the interaction network between p53, Mdm2, and p14ARF is provided for an individual cell model in two cell compartments: nucleus and cytoplasm. This model includes DNA damage induced by gamma-radiation, which generates the p53 activation and the transactivation of both *TP*53 and *MDM*2 (among other transactivated genes, which are not considered in this model). Initially, the translocation process of mRNAs to the cytoplasm and its subsequent translation into proteins takes place; afterward, proteins are transported back to the nucleus. While p53 remains at high levels, $Mdm2_{nuclear}$ reduces its concentration levels, and vice versa, thus, producing an oscillatory pattern. $Mdm2_{cytoplasmic}$ moves to the nucleus at a constant rate, ignoring all other possible behaviors. The production and degradation rates of p14ARF remain constant (Leenders and Tuszynski, 2013).

Following assumptions for the p53-Mdm2 network regulated by p14ARF are presented:

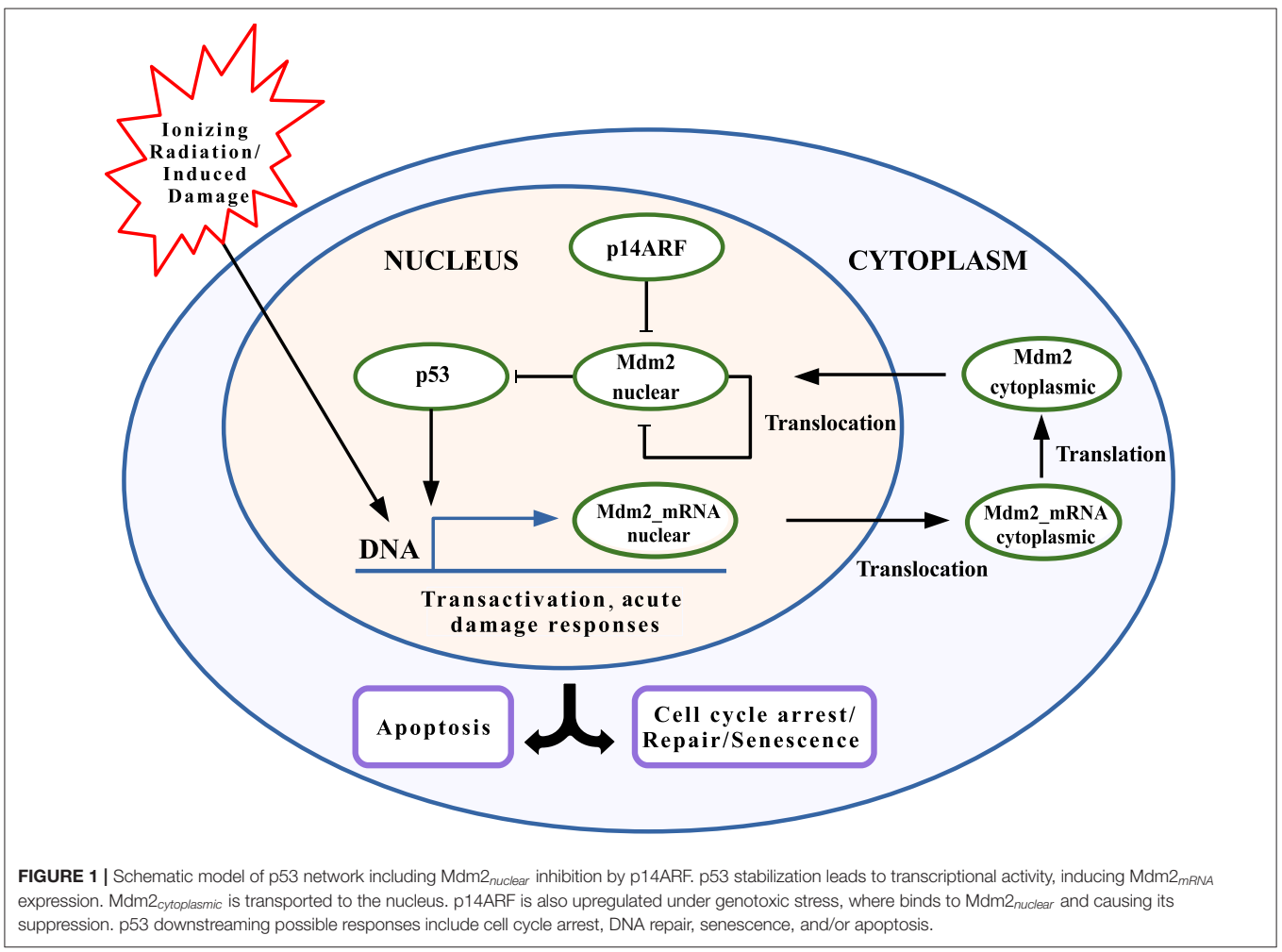

FIGURE 1 | Schematic model of p53 network including $Mdm2_{nuclear}$ inhibition by p14ARF. p53 stabilization leads to transcriptional activity, inducing $Mdm2_{mRNA}$ expression. $Mdm2_{cytoplasmic}$ is transported to the nucleus. p14ARF is also upregulated under genotoxic stress, where binds to $Mdm2_{nuclear}$ and causing its suppression. p53 downstreaming possible responses include cell cycle arrest, DNA repair, senescence, and/or apoptosis.

1. The p53-Mdm2 network is one of the most explored biological mechanisms that exist, which provides adequate information.
2. The biological context requires different scenarios dependent on p53 and failure in the network of p53-Mdm2-p14ARF, which can be represented in equations for simulation output desired reference that have a biological explanation, such as regulation mechanisms in gene expression.
3. The biological responses of p53 that we are interested in exploring are related to the ability to generate cell cycle arrest, DNA repair, senescence, and/or apoptosis, especially the last one since it has a key role in the tumor suppressor response.
4. The model assumes the response to an ionizing radiation stimulus with p53-dependent responses, which presents an oscillatory pattern due to variation between p53 and its inhibitor Mdm2. The main objective is to simulate the p53 suppressed response when Mdm2 is overexpressed, which has been documented in several types of tumors.

2.2. Mathematical Description

Taken from Leenders and Tuszynski (2013), based on the principle of mass-action and the saturable transcription kinetics, the p53-Mdm2 network regulated by p14ARF without control action is mathematically described as follows:

$$\frac{d[p53]}{dt} = k_p - k_1[p53][Mdm2_{nuclear}] - d_p[p53], \quad (1)$$

$$\frac{d[Mdm2\_mRNA_{nuclear}]}{dt} = k_m + k_2\frac{[p53]^{1.8}}{K_D{}^{1.8} + [p53]^{1.8}} - k_0[Mdm2\_mRNA_{nuclear}], \quad (2)$$

$$\frac{d[Mdm2\_mRNA_{cytoplasmic}]}{dt} = k_0[Mdm2\_mRNA_{nuclear}] - d_{rc}[Mdm2\_mRNA_{cytoplasmic}], \quad (3)$$

$$\frac{d[Mdm2_{cytoplasmic}]}{dt} = k_T[Mdm2\_mRNA_{cytoplasmic}] - k_i[Mdm2_{cytoplasmic}], \quad (4)$$

$$\frac{d[Mdm2_{nuclear}]}{dt} = k_i[Mdm2_{cytoplasmic}] - d_{mn}[Mdm2_{nuclear}]^2 - k_3[Mdm2_{nuclear}][p14ARF], \quad (5)$$

$$\frac{d[p14ARF]}{dt} = k_a - d_a[p14ARF] - k_3[Mdm2_{nuclear}][p14ARF]. \quad (6)$$

The deterministic model (1–6) includes: p53 production and degradation (Equation 1), $Mdm2\_mRNA_{nuclear}$ basal transcription (p53-dependent and Mdm2-independent production) in Equation (2); $Mdm2\_mRNA_{cytoplasmic}$ transport from nucleus to cytoplasm (Equation 3), $Mdm2\_mRNA_{cytoplasmic}$ translation rate and $Mdm2_{cytoplasmic}$ protein transport to nucleus (Equation 4), $Mdm2_{cytoplasmic}$ decay through $Mdm2_{nuclear}$-p14ARF complex, which removes Mdm2 and stops $Mdm2_{nuclear}$ ubiquitination rate (Equation 5), p14ARF production and p14ARF decay in nucleus compartment (Equation 6). The parameter values used are presented in **Table 1**.

2.2.1. Model Characteristics of p53-Mdm2 Network Regulated by p14ARF

- The model (1–6) is based on response to gamma radiation in individual MCF-7 cells. This model does not represent all single cell line type responses to gamma radiation.
- It is important to emphasize that the model employed (1–6) describes the core components of the p53 network and are relevant to determine p53 dynamics in response to gamma radiation-induced DNA damage. Mathematical model considers several parameters with non-linear behaviors such as molecule production, degradation, the dissociation constant in the promoter region, translocation of network components, complex formation rate, and translation rate. Other genes/molecules regulations are ignored, gene mutations are not considered, and constant molecule concentrations in the cell are assumed.
- Note that this model includes both experimentally measured and unknown parameters, which are selected manually in order to fit the oscillatory behavior observed for one cell model, as reported in Leenders and Tuszynski (2013).
- Due to the scarcity of parameters used in equations, implications in future researches will involve eliciting responses to stimuli such as gamma radiation, also in

TABLE 1 | Model parameters.

Parameter	Description	Value	References
k_p	p53 production	0.5 proteins/s	Leenders and Tuszynski, 2013
k_1	Mdm2-dependent p53 degradation	9.963×10^{-6}/s	Leenders and Tuszynski, 2013
d_p	p53 decay	1.925×10^{-5}/s	Leenders and Tuszynski, 2013
k_m	p53-independent Mdm2 production	1.5×10^{-3} RNA/s	Leenders and Tuszynski, 2013
k_2	p53-dependent Mdm2 production	1.5×10^{-2}/s	Weinberg et al., 2005
K_D	Dissociation constant in the promoter region	740 proteins	Weinberg et al., 2005
k_0	RNA transport from nucleus to cytoplasm	8.0×10^{-4}/s	Leenders and Tuszynski, 2013
d_{rc}	Mdm2_mRNA decay in cytoplasm	1.444×10^{-4}/s	Hsing et al., 2000
k_T	Translation rate	1.66×10^{-2} proteins/s	Cai and Yuan, 2009
k_i	Protein transport from cytoplasm to nucleus	9.0×10^{-4}/s	Mor et al., 2010
d_{mn}	Mdm2 autoubiquitination	1.66×10^{-7}/s	Leenders and Tuszynski, 2013
k_3	$Mdm2_{nuclear}$-p14ARF complex formation rate	9.963×10^{-6}/s	Northrup and Erickson, 1992
K_a	p14ARF production	0.5 proteins/s	Leenders and Tuszynski, 2013
d_a	p14ARF decay	3.209×10^{-5}/s	Kuo et al., 2004

other stimuli with a p53-dependent and independent responses, in single cells and multiple ones, including detailed characterizations about genome integrity so that a mathematical model reaches excellent precision.

2.3. Pinning Control Methodology

Consider a general network consisting of N nodes with nonlinear couplings, where each node is a scalar nonlinear dynamical system, which represents genes, concentrations of RNAs and proteins, given by

$$\dot{x}_i = f_i(x_i) + h_i(t, x_1, x_2, \ldots, x_N), \quad i = 1, 2, \ldots, N, \tag{7}$$

where $x_i \in \mathbb{R}$ is the state of node i, for $i = 1, 2, \ldots, N$; $f_i : \mathbb{R} \mapsto \mathbb{R}$ represents the self-dynamics of node i related to the degradation process of RNA and proteins, and so on; $h_i : \mathbb{R}^N \mapsto \mathbb{R}$ is the nonlinear coupling function between nodes, associated with the changes of x_i due to transcription, translation, repression, activation, or other interaction processes, and N represents the all network nodes.

The control objective is that (7) tracks a desired output trajectory, given by

$$y = y_r(t).$$

To achieve this objective, local feedback controllers are applied to a reduced number of network nodes, according to the pinning control methodology (Wang and Chen, 2002; Li et al., 2004; Chen, 2017). This methodology is composed of two parts: the first one, pinned nodes l are selected to apply control actions as in (8), where $1 \leq l \leq N$, and l can be as small as one. The second one is the remained network nodes $(N - l)$ without control action as in (9). Thus, the controlled network can be written as

$$\dot{x}_i = f_i(x_i) + h_i(t, x_1, x_2, \ldots, x_N) + g_i(x_i)u_i, \quad i = 1, 2, \ldots, l. \tag{8}$$

$$\dot{x}_i = f_i(x_i) + h_i(t, x_1, x_2, \ldots, x_N), \quad i = l+1, l+2, \ldots, N. \tag{9}$$

where $g_i : \mathbb{R} \mapsto \mathbb{R}$ is a nonlinear function of the node state i, for $i = 1, 2, \ldots, l$ and u_i denotes the control on the node $i \in l$.

In the present paper, u_i in (8), is proposed as a local positive discontinuous feedback control law, described by

$$u_i = \begin{cases} 1 + K_i(1 - e_i), & \text{if} \quad |\varphi_i| < 1, \\ 1 + K_i(1 - sign(e_i)), & \text{if} \quad |\varphi_i| > 1, \end{cases} \tag{10}$$

where K_i is a positive control gain selected by the designer, e_i is the tracking error between the desired output trajectory $[y_r(t)]$ and the controlled state (x_i), is given by

$$e_i = (x_i - y_r(t)), \tag{11}$$

with $\varphi_i = \frac{e_i}{S_i}$ being a proposed auxiliary variable to reject chattering effect caused by $sign(\cdot)$ (signum function extracts the sign of a real number) (), and S_i a signal filter given by

$$\dot{S}_i = -\alpha_i S_i + \omega_i, \quad i = 1, 2, \ldots, l, \tag{12}$$

where α_i and ω_i are positive gains to be selected.

2.4. p14ARF as Pinned Node

To select the pinned nodes, the virtual leader methodology presented in Ren and Beard (2008) and Lewis et al. (2013) is used. The methodology consists on analyzing the interactions between proteins presented in **Figure 1** using the mathematical model (1–6). The nodes that affect directly or indirectly the dynamical behavior everyone else, are candidates as pinned nodes. In this sense, the spanning tree of the p53-Mdm2 network regulated by p14ARF is developed, as in **Figure 2**; based on this analysis, p14ARF is the adequate biological selection as the pinned node.

From Equation (6), the differential equation for p14ARF (pinned node) is defined by three cellular processes, as follows

$$\frac{d[p14ARF]}{dt} = \overbrace{k_a}^{\text{Production}} \overbrace{-d_a[p14ARF]}^{\text{Degradation}} \overbrace{-k_3[Mdm2_{nuclear}][p14ARF]}^{Mdm2_{nuclear}-p14ARF \text{ complex formation}}.$$

In order to control the p53-Mdm2 network, it is necessary to increase the p14ARF concentration levels to regulated $Mdm2_{nuclear}$ production. As can be seen degradation process and $Mdm2_{nuclear} - p14ARF$ complex formation have a negative sign; while, the production process has a positive sign. Due to this fact, we propose to modify the p14ARF production (K_a) process added the control law (10).

Thus, considering Equations (8) and (9), the p53-Mdm2 network regulated by p14ARF with control action is mathematically described as follows:

$$\frac{d[p14ARF]}{dt} = k_a u_1 - d_a[p14ARF] - k_3[Mdm2_{nuclear}][p14ARF], \tag{13}$$

$$\frac{d[p53]}{dt} = k_p - k_1[p53][Mdm2_{nuclear}] - d_p[p53], \tag{14}$$

$$\frac{d[Mdm2\_mRNA_{nuclear}]}{dt} = k_m + k_2\frac{[p53]^{1.8}}{{K_D}^{1.8} + [p53]^{1.8}} - k_0[Mdm2\_mRNA_{nuclear}], \tag{15}$$

$$\frac{d[Mdm2\_mRNA_{cytoplasmic}]}{dt} = k_0[Mdm2\_mRNA_{nuclear}] - d_{rc}[Mdm2\_mRNA_{cytoplasmic}], \tag{16}$$

$$\frac{d[Mdm2_{cytoplasmic}]}{dt} = k_T[Mdm2\_mRNA_{cytoplasmic}] - k_i[Mdm2_{cytoplasmic}], \tag{17}$$

$$\frac{d[Mdm2_{nuclear}]}{dt} = k_i[Mdm2_{cytoplasmic}] - d_{mn}[Mdm2_{nuclear}]^2 - k_3[Mdm2_{nuclear}][p14ARF]. \tag{18}$$

The p53-Mdm2 network regulated by p14ARF without control action (1–6) and with control action (13–18) are simulated using

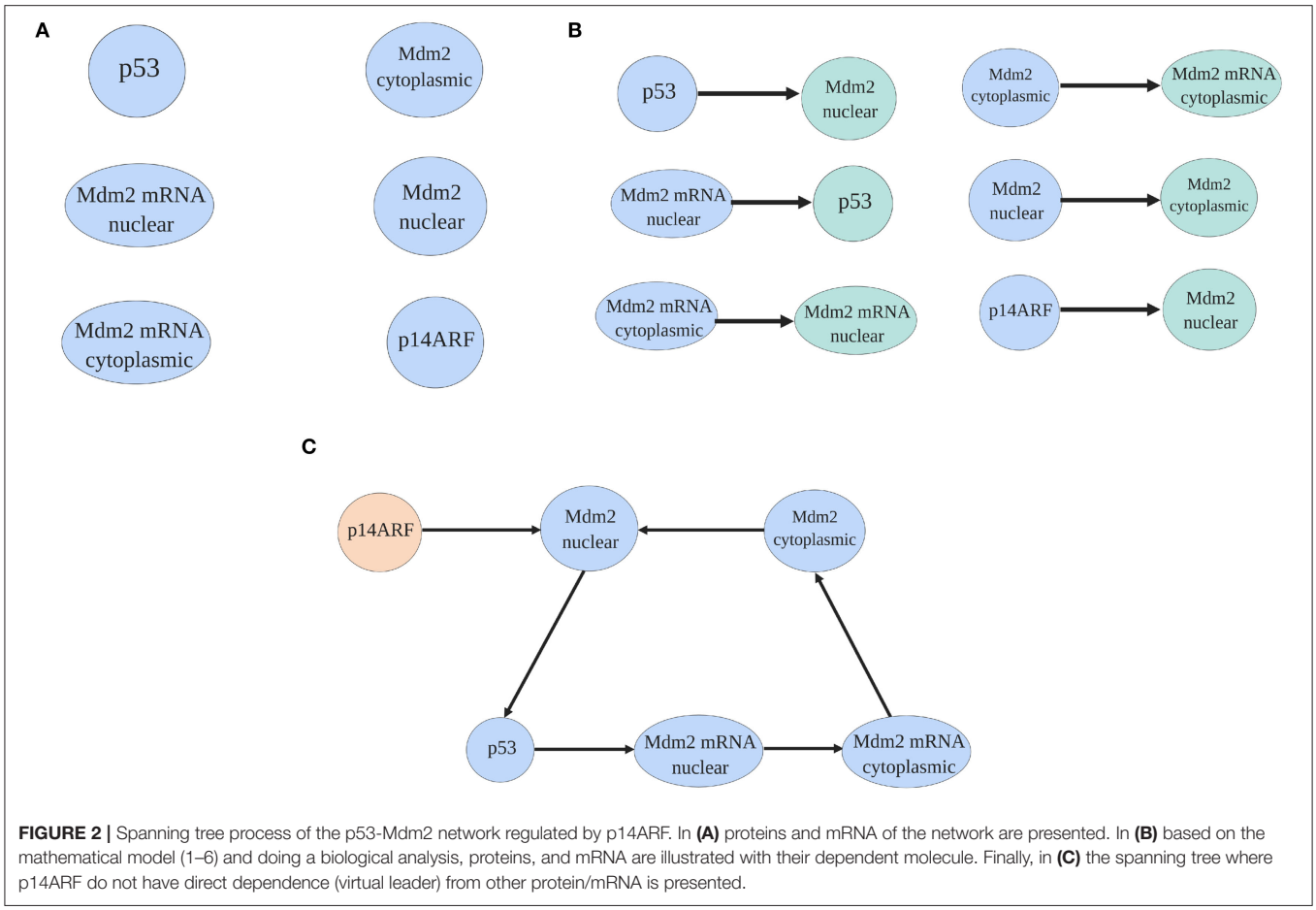

FIGURE 2 | Spanning tree process of the p53-Mdm2 network regulated by p14ARF. In **(A)** proteins and mRNA of the network are presented. In **(B)** based on the mathematical model (1–6) and doing a biological analysis, proteins, and mRNA are illustrated with their dependent molecule. Finally, in **(C)** the spanning tree where p14ARF do not have direct dependence (virtual leader) from other protein/mRNA is presented.

Matlab/Simulink and the fourth-order Runge-Kutta integration method with a fixed step size of 1×10^{-3}.

3. RESULTS

3.1. Behaviors of the p53-Mdm2 Network Regulated by p14ARF Without Control Action

Three different behaviors of the p53-Mdm2 network without control actions are presented in **Figure 3**.

As displayed in **Figure 3A**, the p53-Mdm2 network presents an oscillatory pattern under gamma-radiation, for a lapse of 48 h, using the parameter values shown in **Table 1**. This response is due to post-translational modifications of p53 and the negative interactive loop of Mdm2-mediated ubiquitination, according to Ciliberto et al. (2005) and Geva-Zatorsky et al. (2006); this pattern has not been observed for all cell types and requires wild-type genes (Lahav et al., 2004; Leenders and Tuszynski, 2013).

Figure 3B illustrates that p53-Mdm2 dependent affinity is altered, producing a $Mdm2_{nuclear}$ overexpression when the parameters k_1 and k_2 are set to values five and ten times larger than the original ones, respectively. This behavior is reported in a variety of human soft tissue tumors and in hematological malignancies, as discussed in Oliner et al. (1992), Nilbert et al. (1994), Dei Tos et al. (2000), Bond et al. (2004), and Rayburn et al. (2005). In human tumors, Mdm2 overexpression can inhibit p53 normal regulatory activities and induce a loss of growth-inhibitory signals in cytostatic and apoptotic responses, which may favor a carcinogenic process.

Figure 3C displays an increased expression of p53 levels when the parameters K_a and d_{rc} are set to values ten and one hundred times larger than the original ones, respectively. This p53 dynamical behavior is related to downstream genes involved in signaling pathway, which can produce cell cycle arrest, DNA repair, senescence, and/or apoptotic response (El-Deiry, 1998; Hsing et al., 2000; Khan et al., 2004; Arya et al., 2010; Purvis et al., 2012).

3.2. Behaviors of the p53-Mdm2 Network With Nutlin-3

Several molecular inhibitors of p53-Mdm2 interaction have been proposed, including a small molecule called Nutlin-3 as in Vassilev et al. (2004). Due to Mdm2 deregulation has been reported in various tumor types, Nutlin-3 has been an option to block the p53-binding site of MDM2 competitively and induce the upregulation and activation of p53 pathway (Yee-Lin et al., 2018). These findings motivated us to observe the

possible effect of the external stimulus of Nutlin-3 on the oscillatory pattern behavior in the p53-Mdm2 network regulated by p14ARF. If p53 stabilization is achieved and p53 degradation is avoided, as reported for Nutlin-3 in tumors, p53 can accomplish antiproliferative effects.

In this sense, a simulation to reduce p53-Mdm2 interaction and p53 degradation in the presence of Nutlin-3 is presented in **Figure 4**. To achieve this behavior, k1, which represents Mdm2-dependent p53 degradation, takes values between 25, 50, 75, and 90% less from the original value. It is possible to observe that simulating the effect of Nutlin-3, liberates p53 from the interaction and inhibition by Mdm2 in the network, allowing to desynchronize p53 and MDM2 oscillatory behavior, reaching a stable high concentration when the inhibition by Nutlin-3 is strong.

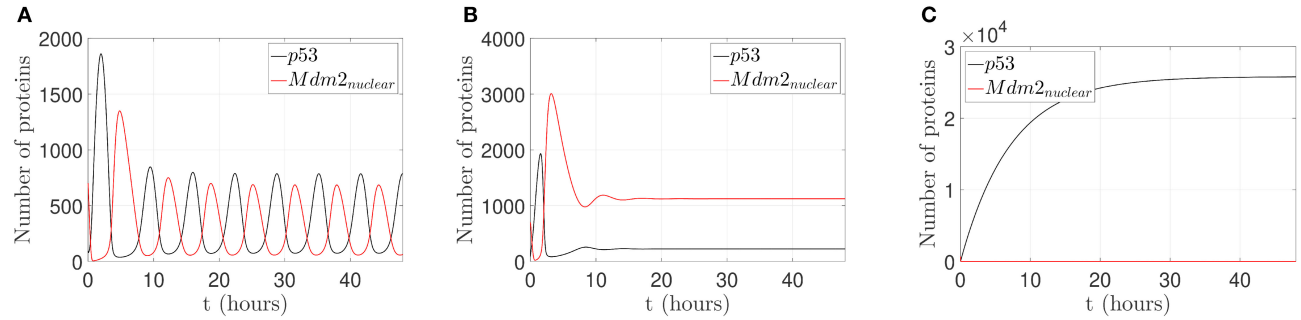

FIGURE 3 | Behaviors of the p53-Mdm2 network regulated by p14ARF without control action. **(A)** p53-$Mdm2_{nuclear}$ oscillatory pattern under gamma-radiation within a lapse of 6.64 h. **(B)** $Mdm2_{nuclear}$ overexpression and p53 downregulation with $k_1 = 1.9926 \times 10^{-6}$/s and $k_2 = 45 \times 10^{-3}$/s; once $Mdm2_{nuclear}$ is overexpressed, p53 is inhibited. It can be conjectured that $Mdm2_{nuclear}$ deregulation will lead to oncogenic behavior through p53 suppression. **(C)** Increased expression of p53 levels response under gamma-radiation induction with $k_a = 5$ proteins/s and $d_{rc} = 1.444 \times 10^{-2}$/s. It can be conjectured that p53 upregulation will lead to cell cycle arrest, DNA repair, senescence, and/or apoptotic response.

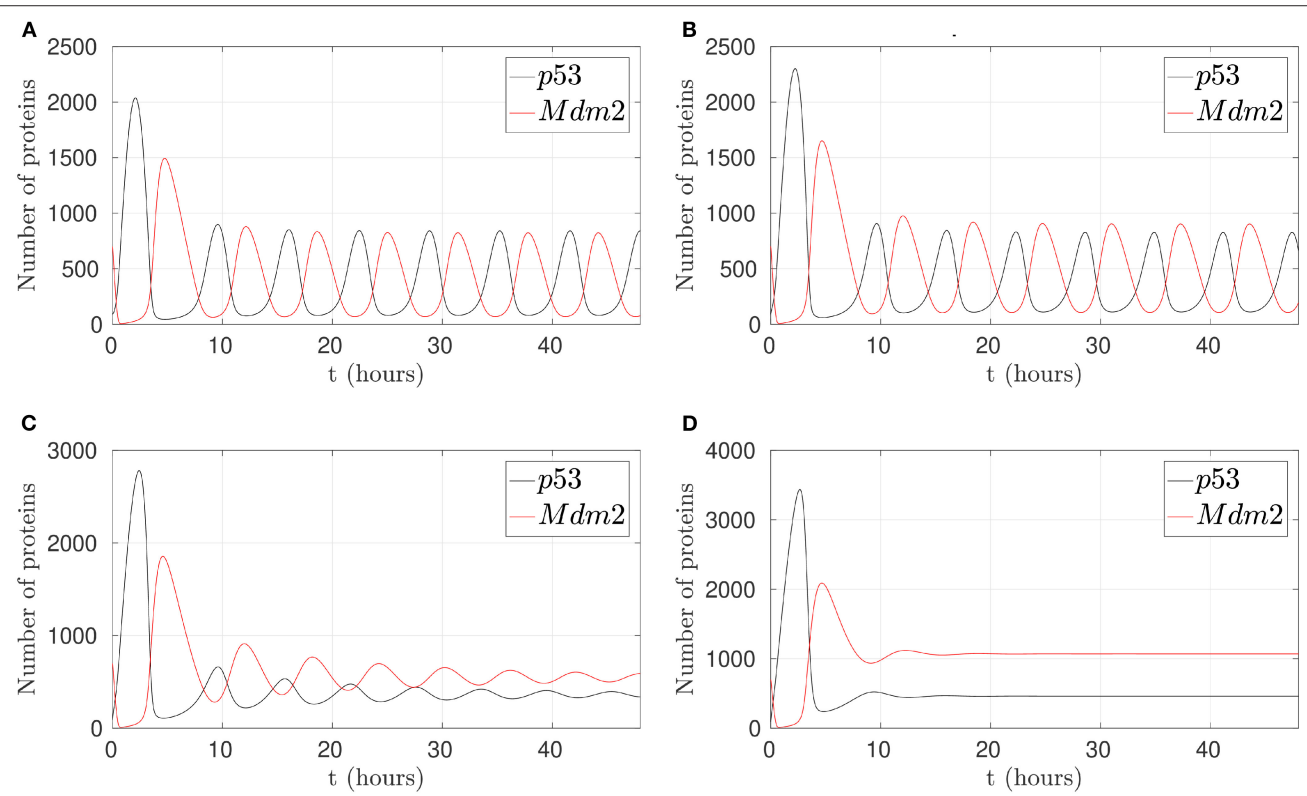

FIGURE 4 | Behaviors of the p53-Mdm2 network regulated by p14ARF under Nutlin-3 effect. **(A)** With 25% activity in k1. **(B)** With 50% activity in k1. **(C)** With 75% activity in k1. **(D)** With 90% activity in k1. **(C,D)** Shows desynchronized p53-Mdm2 interaction.

3.3. Sensitivity Analysis for p14ARF Production (K_a)

This analysis is done on the basis of p14ARF production (K_a) value variation effects on p53 and Mdm2$_{nuclear}$ respectively as can be seen in **Figure 5**. Sensitivity analysis (Dickinson and Gelinas, 1976; Hamby, 1994) determines the K_a values for which the network can not achieve desired behaviors (0–1.5 *proteins*/*s*); the K_a value to reach p53-Mdm2$_{nuclear}$ oscillatory pattern (1.6–9.5 *proteins*/*s*), and the K_a value to generate an increased expression of p53 with Mdm2$_{nuclear}$ downregulation (9.6–50 *proteins*/*s*) are determined from **Figure 5**.

3.4. Behaviors of the p53-Mdm2 Network Regulated by p14ARF With Control Action

To illustrate the p53 and Mdm2$_{nuclear}$ behavior under pinning control actions, two cases are considered: (1) to restore an oscillatory pattern under gamma-radiation, and (2) to achieve an increased expression of p53 level. For a 24 h lapse, the network runs without any control action and presents overexpressed Mdm2-mediated p53 degradation as carcinogenic initial behavior for both cases (Oliner et al., 1992; Nilbert et al., 1994; Dei Tos et al., 2000; Bond et al., 2004; Rayburn et al., 2005), which is displayed in **Figure 3B** above.

3.4.1. Case 1: Restoration of an Oscillatory Pattern Under Gamma-Radiation

For the first case, by considering p14ARF production ($k_a u_1$) in a range between 0 and 1.5 *proteins*/*s*, the proposed controller is turned on after the 24 h initial lapse; however, the network can not achieves the oscillatory pattern as can be seen in **Figure 6A**. Due to the low value of $k_a u_1$, the pinning control technique cannot achieve the desired behavior. Otherwise, with ($k_a u_1$) in a range between 1.6 and 9.5 *proteins*/*s*, the proposed controller

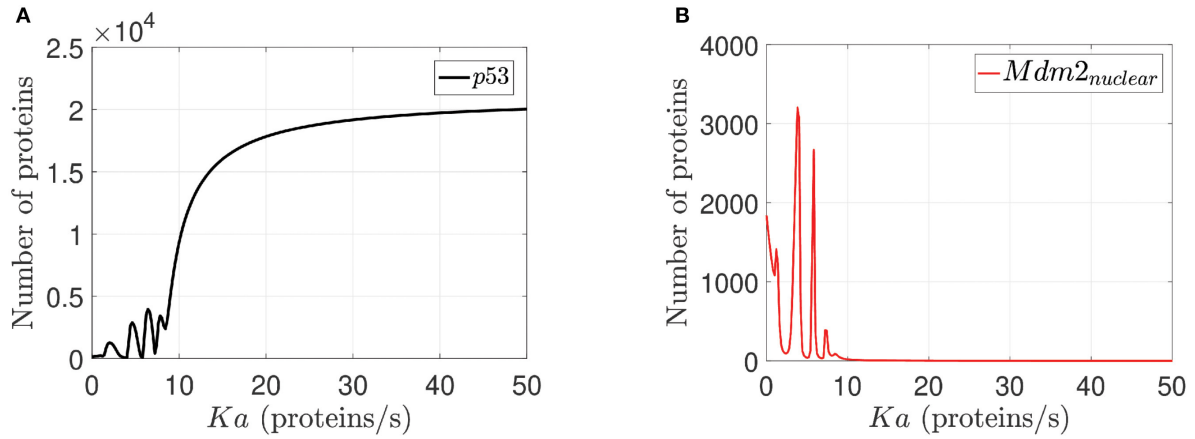

FIGURE 5 | Sensitivity analysis for K_a. In **(A)** p53 response to K_a variations is illustrated. In **(B)** Mdm2$_{nuclear}$ response to K_a variations is presented. For both cases K_a varies between 0 to 50 *proteins*/*s* for all network proteins, but we select p53 and Mdm2$_{nuclear}$ because they are the output desired in the network.

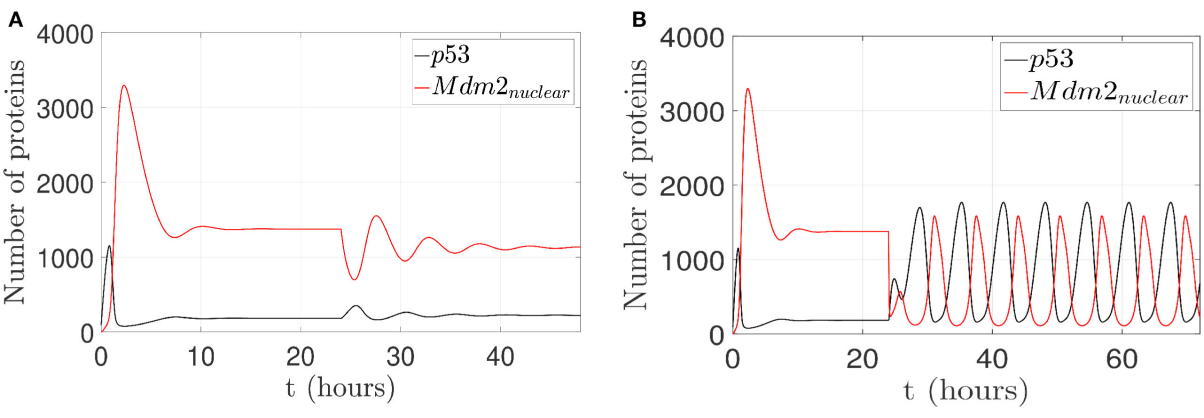

FIGURE 6 | Proposed scenario simulations to achieve restoration of an oscillatory pattern under gamma-radiation in the p53-Mdm2 network regulated by p14ARF. For both tests is possible to observe overexpressed Mdm2-mediated p53 degradation as carcinogenic initial behavior at first 24 h. In **(A)**, control action with ($k_a u_1$) in a range between 0 and 1.5 *proteins*/*s*, the oscillatory pattern under gamma-radiation is not achieved (24–48 h) and likewise in **(B)** oscillatory pattern under gamma-radiation under pinning control (24–72 h) with ($k_a u_1$) in a range between 1.6 and 9.5 *proteins*/*s* is achieved.

forces the network to gradually track the oscillatory pattern as can be seen in **Figure 6B**.

3.4.2. Case 2: Achievement of a p53 Level Increased Expression

For the second case, by considering p14ARF production ($k_a u_1$) in a range between 0 and 1.5 *proteins*/*s*, the proposed controller is turned on; however, the network can not achieves increased expression of p53 as can be seen in **Figure 7A**. Due to the low value of $k_a u_1$, the pinning control technique cannot achieve the desired behavior. Otherwise, with ($k_a u_1$) in a range between 9.6 and 50 *proteins*/*s*, the proposed controller again is turned on, and the system gradually tracks the increased expression of p53 levels as can be seen in **Figure 7B**.

For both cases, the control law ($u_1(t) \in \mathbb{R}$) is applied to p14ARF node as in Equation (13). The respective control actions are displayed in **Figure 8A** for the first case and **Figure 8B** for the second case. From the above results, it can be clearly seen that the pinning controller achieves regulation successfully for the p53-Mdm2 network.

4. DISCUSSION

In biological systems, gene regulatory networks, having complex interactions, present important challenges for the application of control strategies. A mixture of mechanisms such as gene expression patterns, post-translational modifications, translocation, components degradation, and the specific changes of internal and external signals of the cell, generate nonlinear behavior. Hence, to illustrate the applicability of complex network control, we present two cases for a deterministic network

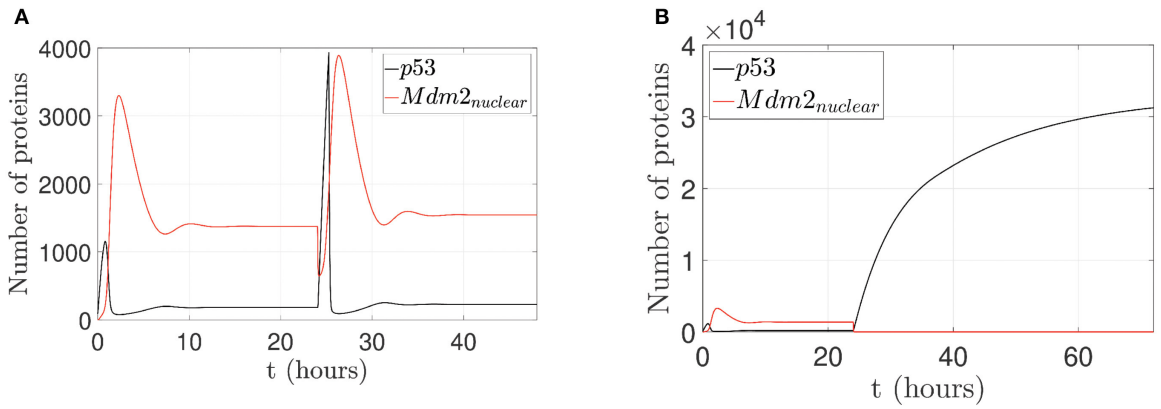

FIGURE 7 | Proposed scenario simulations to achieve p53 level increased expression in the p53-Mdm2 network regulated by p14ARF. For both tests is possible to observe overexpressed Mdm2-mediated p53 degradation as carcinogenic initial behavior at first 24 h. In **(A)**, control action with ($k_a u_1$) in a range between 0 and 1.5 *proteins*/*s*, increased expression of p53 response is not achieved (24–48 h) and likewise in **(B)** increased expression of p53 response under pinning control (24–72 h) with ($k_a u_1$) in a range between 9.6 and 50 *proteins*/*s* is achieved.

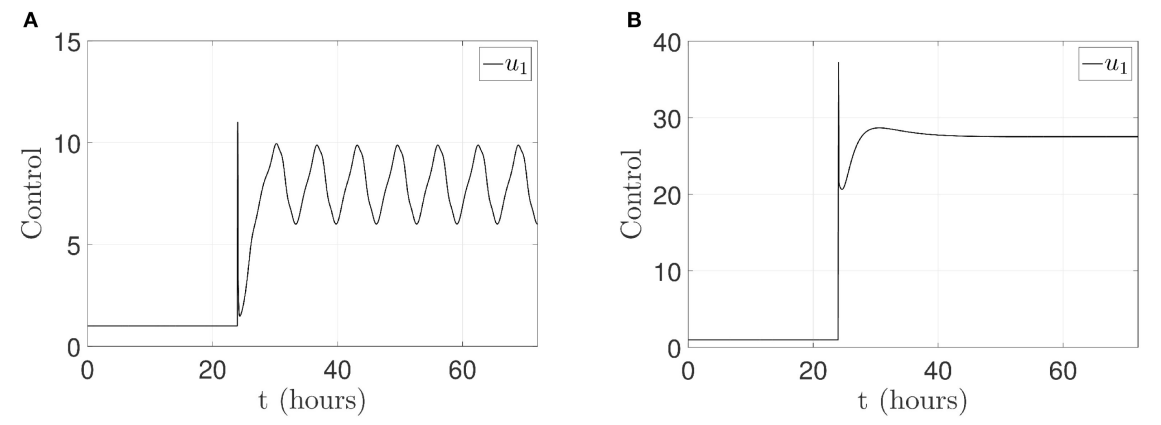

FIGURE 8 | Control signal $u_1(t)$ applied in p14ARF. **(A)** Case 1: control action to achieve an oscillatory pattern under gamma-radiation. **(B)** Case 2: control action to achieve an increased expression of p53 response.

model corresponding to tumor suppressor p53, Mdm2, and p14ARF.

4.1. Behaviors Induced Without Control Action

4.1.1. p53-Mdm2$_{nuclear}$ Oscillatory Pattern

Levels of p53 and Mdm2$_{nuclear}$ proteins present oscillatory behavior, caused by pulses resulting from p53 activation, p53-dependent transactivation, Mdm2 production, Mdm2$_{nuclear}$ sequestration by p14ARF, and Mdm2-mediated ubiquitination. This process allows the cell to respond to double-strand breaks (DSBs) in DNA induced by gamma radiation, which correlates with the number of p53 pulses in individual cells (Lahav et al., 2004; Geva-Zatorsky et al., 2006). For the model used in this paper, there are regulatory factors which have not been considered, as the interaction of other potentially relevant genes transactivated by p53 (nearly 100 genes in pathways such as cell cycle arrest, DNA repair, senescence, and apoptosis; Riley et al., 2008). One of the simpler explanations for the p53 oscillatory pattern is due to repeated activation of ATM (Ataxia Telangiectasia Mutant), which dissociates the p53-Mdm2 complex and stabilizes an increase in p53 levels. These changes are driven by persistent DNA damage induced by radiation (Lahav et al., 2004; Geva-Zatorsky et al., 2006). Therefore, recovering the normal pattern response to DNA damage by the p53-Mdm2 network is fundamental for tumor suppression.

4.1.2. Mdm2$_{nuclear}$ Overexpression and p53 Downregulation

There are gene abnormalities in tumors, carcinogenesis driven by viral infections, and other mechanisms that contribute to inactivate p53 functions and its signaling outcomes (Scheffner et al., 1990; Camus et al., 2003; Rayburn et al., 2005). For example, Mdm2 overexpression leads to p53 downregulation, contributing to losses of tumor suppressor activity. Mdm2 overexpressed is a hallmark in several types of cancer (Nilbert et al., 1994; Dei Tos et al., 2000; Rayburn et al., 2005). As reviewed in Rayburn et al. (2005), Mdm2 is overexpressed in liposarcomas, osteosarcomas, testicular germ cell tumors, embryonic carcinomas, brain tumors (including glioblastomas and astrocytomas), hematological malignancies, bladder cancer, breast cancer, colorectal cancer among others. In this sense, the number of Mdm2 abnormalities is highly variable. Moreover, not all samples from the same type of tumor, show Mdm2 overexpression. Bond et al. (2004) reported other mechanisms that promote Mdm2 overexpression, where a single nucleotide polymorphism (SNP309) in the MDM2 promoter, increases the affinity of the transcriptional activator Sp1, resulting in higher levels of Mdm2 mRNA and Mdm2 translation rate. This behavior illustrates p53 downregulation resulting in a decreased response to DNA damaging agents and acceleration of tumorigenesis.

4.1.3. Increased Expression of p53 Levels

The stabilization and accumulation of p53 levels are part of the DNA damage response to maintain genome integrity. One of the possible response outputs of a fully functional p53 pathway is the induction of apoptosis generated by DNA damage. For apoptosis induction by radiation in certain tissues, a cascade of signaling is generated through pro-apoptotic proteins, such as response mediated by ATM protein (Bakkenist and Kastan, 2003). ATM leads to p53 stabilization, modifying the interaction capabilities with Mdm2 (El-Deiry, 1998). With the activation of p53, its degradation is limited, and p53 levels increase (Oliner et al., 1993). Downstream to ATM/p53 activation, an apoptosis program is carried out by a set of proteins, such as Bim, Puma, Bid, Bmf, Bad, Bik, Noxa, and Hrk, whereby Puma and Noxa can be directly regulated with p53 overexpression by gamma-radiation (Villunger et al., 2003; Chen et al., 2005). This set of proteins can bind and block survival proteins such as Bcl-2, which release death effectors like Bax and Bak. These effectors can lead to a change in the permeability of the outer mitochondria membrane. Furthermore, they can participate in the cell dismantling coupled with caspases (Green and Kroemer, 2004). There are other possible outcomes for the p53 activation pathway and p53 independent responses to DNA damage by gamma-radiation, which can lead the cell to cell cycle arrest, initiate DNA repair, or perform senescence. Experimental evidence indicates that the cellular level of p53 can dictate the response of the cell, such that lower levels of p53 result in arrest, whereas higher level results in apoptosis (Chen et al., 1996; Purvis et al., 2012).

4.1.4. p53-Mdm2 Network With Nutlin-3

Just as p14ARF can stabilize p53 by antagonizing Mdm2 effects (Weber et al., 1999), if we consider molecules or treatments that can function as Mdm2 inhibitors, the small molecule Nutlin-3, is a compound described that binds in the p53-binding pocket within Mdm2, inhibiting its interaction with p53, and preventing p53 tagging for proteasome-mediated degradation (Vassilev et al., 2004; Yee-Lin et al., 2018).

Nutlin-3 is considered cytotoxic in specific wild-type p53 cancer cells (Vassilev et al., 2004; Arya et al., 2010; Yee-Lin et al., 2018), and therefore, p53 executes p53-dependent genes transactivation. Unlike in p53 mutated cell lines, transactivation is defective, and induced Mdm2 expression is impaired (Kamijo et al., 1998). It is essential to consider TP53 mutational status, other p53-dependent responses, and the Nutlin 3 activity in a dose and time-dependent manner (Arya et al., 2010), to predict the effect of Nutlin 3 in Mdm2 binding interactions.

4.2. Behaviors Induced by the Pinning Control Technique

4.2.1. Case 1: Restoration of an Oscillatory Pattern Under Gamma-Radiation

The oscillatory pattern behavior (Geva-Zatorsky et al., 2006; Wee et al., 2009; Batchelor et al., 2011) induced by means of the pinning control technique is illustrated in **Figure 6**, with the purpose of restoring normal network behavior in presence of oncogenic overexpressed Mdm2; this overexpression avoids normal regulatory activities due to a suppressed wild-type p53. The pinning control technique (10) is located at the production of p14ARF node with $K_i = 100$, which allows to regenerate an oscillatory pattern and guarantees that p14ARF production $k_a u_1$ achieves a range between 1.6 and 9.5

proteins/*s*. Enhanced p14ARF production induces a decrease in $Mdm2_{nuclear}$ which in turn increased p53 levels. This approach can help to analyze multiple interaction mechanisms, to induce different cell reprogramming responses. Therefore, the proposed approach motivates future researches on the interdependencies of cellular networks and new ways for treatment designs in tumor suppressor networks.

4.2.2. Case 2: Achievement of an Increased p53 Level Expression

The increased expression of p53 levels by means of the pinning control technique is achieved as shown in **Figure 7**. Such technique generates the desired behavior of p53 progressive accumulation, assuming that this behavior has a post-translational activation mediated by ATM, which could generate the activation of downstream proteins, contributing to the activation of apoptosis or cell cycle arrest (El-Deiry, 1998; Wagner et al., 2005; Geva-Zatorsky et al., 2006; Wee et al., 2009). The pinning control technique (10) is located at the production of p14ARF node with $K_i = 5$, which yields an increased p53 level expression and guarantees that p14ARF production $k_a u_1$ achieves a range between $9.6 - 50\, proteins/s$. p14ARF is moved from the nucleolus to nucleoplasm in response to DNA damage, where $Mdm2_{nuclear}$-p14ARF complex promotes p53 tumor suppressor activity. For this case, it is possible to generate p53 accumulation, which is assumed to be activated by a mechanism linked to post-radiation activation of the ATM protein and p53-dependent induction of downstream proteins (Bakkenist and Kastan, 2003; Villunger et al., 2003; Pauklin et al., 2005). With the proposed pinning control law, it is possible to produce scenarios with different physiological or pathological responses of the p53-Mdm2 network.

For the two cases discussed above, the proposed pinning control strategy for the p53-Mdm2 network dynamics is applied on the p14ARF node based on sensitivity analysis as follows: In the first case, oscillatory pattern activity is achieved; in the second case p53 increased expression and accumulation are obtained. By means of the sensitivity analysis in K_a with respect to $p53$ and $Mdm2_{nuclear}$, it can be established that with K_a low values, the network does not reach the desired behavior. However, for the adequate K_a values mentioned above, the network recovers its oscillatory pattern behavior, or an increase in p14ARF production leads causing a consistently increase p53 level. The proposed pinning control strategy as explained suppresses Mdm2 prooncogenic behavior and allows functional recovery of p53 physiological response.

4.2.3. p14ARF as Pinned Node for p53-Mdm2 Network Control

In this section, we explain the importance of p14ARF as pinned node for the p53-Mdm2 network. Pinned node p14ARF regulates p53 by promoting Mdm2 degradation, preventing the Mdm2-mediated p53 degradation (Kamijo et al., 1998; Zhang et al., 1998; Weber et al., 1999). Experimental evidence indicates that p14ARF can even induce apoptosis through the Bax protein independent of p53 (Suzuki et al., 2003) and activate p53 through phosphorylation in oncogenic damage (Ito et al., 2001). It has also been reported that p14ARF helps to regulate the expression of Rad51, a protein involved in DNA repair; such regulation is explained by the activation of the ATM/ATR pathway, which is activated in response to DNA DSBs damage induced by ionizing radiation (Pauklin et al., 2005). In an *in vitro* approach (Itoshima et al., 2000), esophageal cancer cells were transfected with a plasmid designed to rise ARF expression (exogenous), with the subsequent reduction of endogenous levels of Mdm2 and induced p53 accumulation. Also, Itoshima et al. (2000), observed that mutant form of p53 was also stabilized by ectopic ARF increased expression, consistent with another report where p53 mutated can also be bonded to Mdm2 (Haupt et al., 1997); however, mutant-p53 transcriptional activity is obliterated and the existence of an intact autoregulatory loop between ARF, Mdm2, and p53 is needed to observe full regulatory pathway responses, particularly, a response that leads to apoptosis in cancer cells.

Interestingly, according to a model of overexpression of p14ARF in glioblastoma and astrocytoma cells, p14ARF controls neovascularization, through upregulation of metalloproteinase-3 inhibitor (TIMP3) in a p53-independent signaling pathway, which suppresses angiogenesis (Zerrouqi et al., 2012). Moreover, the upregulation of P14ARF could generate activation or interaction with other factors such as Sp1, c-Jun, AP-1, and JNK that cooperate to prevent tumor-induced microthrombosis through activation of the anticoagulant factor TFPI-2. p14ARF also enhances the transcriptional activity of Sp1, using the interaction of Mdm2 with Sp1. If Mdm2 is inhibited, Sp1 activity can be enhanced and induce tumor suppressor effects independent of p53-mediated signaling, as stated by Zerrouqi et al. (2014).

As future work, it would be essential to generate a new model that includes these and other interactions (disruption of mutant p53 stabilization, p53 post-translational modifications, external inhibitors, or altered proteasome system within the network), which would allow applying these control techniques and other types of analysis to others cell signaling pathways, to computationally model and stir the dynamics of a gene regulatory network to the desired state mainly to reproduce the coordinated fluctuation behavior of core components in p53 network.

Finally, experimental validation for the p53-Mdm2 network regulated by p14ARF as pinned node is not possible to implement due to lack of a biosensor measure the activity in the nucleus and cytoplasm at the cell. Our group is currently developing such a sensor.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

OS, CV, ES, and GC initiated research and provided knowledge about strategies of control in the methods section. AG-S and OR-J

proposed the description model and discussion. EH-V and AA discussed simulations and prepared figures. All authors designed the research, analyzed data, prepared, wrote, and reviewed the paper.

FUNDING

This work was supported by CONACYT, Mexico, Project 257200 and by the Alfons und Gertrud Kassel-Stiftung.

REFERENCES

Ahmad, J., Bernot, G., Comet, J., Lime, D., and Roux, O. (2006). Hybrid modelling and dynamical analysis of gene regulatory networks with delays. *ComPlexUs* 3, 231–251. doi: 10.1159/000110010

Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Sympos. Biocomput.* 4, 17–28. doi: 10.1142/9789814447300_0003

Arya, A., El-Fert, A., Devling, T., Eccles, R., Aslam, M., Rubbi, C., et al. (2010). Nutlin-3, the small-molecule inhibitor of MDM2, promotes senescence and radiosensitises laryngeal carcinoma cells harbouring wild-type p53. *Br. J. Cancer* 103:186. doi: 10.1038/sj.bjc.6605739

Ashcroft, M., Taya, Y., and Vousden, K. (2000). Stress signals utilize multiple pathways to stabilize p53. *Mol. Cell. Biol.* 20, 3224–3233. doi: 10.1128/MCB.20.9.3224-3233.2000

Bakkenist, C., and Kastan, M. (2003). DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* 421:499. doi: 10.1038/nature01368

Batchelor, E., Loewer, A., Mock, C., and Lahav, G. (2011). Stimulus-dependent dynamics of p53 in single cells. *Mol. Syst. Biol.* 7:488. doi: 10.1038/msb.2011.20

Bolouri, H., and Davidson, E. (2002). Modeling transcriptional regulatory networks. *BioEssays* 24, 1118–1129. doi: 10.1002/bies.10189

Bond, G., Hu, W., Bond, E., Robins, H., Lutzker, S., Arva, N., et al. (2004). A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* 119, 591–602. doi: 10.1016/j.cell.2004.11.022

Burbano, D., Russo, G., and Di Bernardo, M. (2019). Pinning controllability of complex network systems with noise. *IEEE Trans. Control Netw. Syst.* 6, 874–883. doi: 10.1109/TCNS.2018.2880300

Cai, L., Friedman, N., and Xie, X. (2006). Stochastic protein expression in individual cells at the single molecule level. *Nature* 440:358. doi: 10.1038/nature04599

Cai, X., and Yuan, Z. (2009). Stochastic modeling and simulation of the p53-MDM2/MDMX loop. *J. Comput. Biol.* 16, 917–933. doi: 10.1089/cmb.2008.0231

Camus, S., Higgins, M., Lane, D., and Lain, S. (2003). Differences in the ubiquitination of p53 by Mdm2 and the HPV protein E6. *FEBS Lett.* 536, 220–224. doi: 10.1016/S0014-5793(03)00054-1

Cao J., Qi X., and Zhao H. (2012). "Modeling gene regulation networks using ordinary differential equations," in *Next Generation Microarray Bioinformatics. Methods in Molecular Biology (Methods and Protocols)*, Vol. 802, eds J. Wang, A. Tan and T. Tian (Humana Press) 185–197. doi: 10.1007/978-1-61779-400-1_12

Chaouiya, C. (2007). Petri net modelling of biological networks. *Brief. Bioinformatics* 8, 210–219. doi: 10.1093/bib/bbm029

Chen, G. (2017). Pinning control and controllability of complex dynamical networks. *Int. J. Autom. Comput.* 14, 1–9. doi: 10.1007/s11633-016-1052-9

Chen, H., Liang, J., and Wang, Z. (2016). Pinning controllability of autonomous Boolean control networks. *Sci. China Inform. Sci.* 59:070107. doi: 10.1007/s11432-016-5579-8

Chen, L., Willis, S., Wei, A., Smith, B., Fletcher, J., Hinds, M., et al. (2005). Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol. Cell* 17, 393–403. doi: 10.1016/j.molcel.2004.12.030

Chen, T., He, H., and Church, G. (1999). "Modeling gene expression with differential equations," in *Biocomputing'99* (Mauna Lani, HI), 29–40. doi: 10.1142/9789814447300_0004

Chen, X., Ko, L., Jayaraman, L., and Prives, C. (1996). p53 levels, functional domains, and DNA damage determine the extent of the apoptotic response of tumor cells. *Genes Dev.* 10, 2438–2451. doi: 10.1101/gad.10.19.2438

Ciliberto, A., Novák, B., and Tyson, J. (2005). Steady states and oscillations in the p53/Mdm2 network. *Cell Cycle* 4, 488–493. doi: 10.4161/cc.4.3.1548

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103. doi: 10.1089/10665270252833208

Dei Tos, A., Doglioni, C., Piccinin, S., Sciot, R., Furlanetto, A., Boiocchi, M., et al. (2000). Coordinated expression and amplification of the MDM2, CDK4, and HMGI-C genes in atypical lipomatous tumours. *J. Pathol.* 190, 531–536. doi: 10.1002/(SICI)1096-9896(200004)190:5<531::AID-PATH579>3.0.CO;2-W

Dickinson, R., and Gelinas, R. (1976). Sensitivity analysis of ordinary differential equation systems–a direct method. *J. Comput. Phys.* 21, 123–143. doi: 10.1016/0021-9991(76)90007-3

El-Deiry, W. (1998). Regulation of p53 downstream genes. *Semin. Cancer Biol.* 8, 345–357. doi: 10.1006/scbi.1998.0097

Elf, J., Li, G., and Xie, X. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* 316, 1191–1194. doi: 10.1126/science.1141967

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Fromentin, J., Eveillard, D., and Roux, O. (2010). Hybrid modeling of biological networks: mixing temporal and qualitative biological properties. *BMC Syst. Biol.* 4:79. doi: 10.1186/1752-0509-4-79

Gao, Z., Chen, X., and Başar, T. (2017). Controllability of conjunctive Boolean networks with application to gene regulation. *IEEE Trans. Control Netw. Syst.* 5, 770–781. doi: 10.1109/TCNS.2017.2746345

Geva-Zatorsky, N., Rosenfeld, N., Itzkovitz, S., Milo, R., Sigal, A., Dekel, E., et al. (2006). Oscillations and variability in the p53 system. *Mol. Syst. Biol.* 2, 1–13. doi: 10.1038/msb4100068

Green, D., and Kroemer, G. (2004). The pathophysiology of mitochondrial cell death. *Science* 305, 626–629. doi: 10.1126/science.1099320

Hafner, A., Stewart-Ornstein, J., Purvis, J., Forrester, W., Bulyk, M., and Lahav, G. (2017). p53 pulses lead to distinct patterns of gene expression albeit similar DNA-binding dynamics. *Nat. Struct. Mol. Biol.* 24:840. doi: 10.1038/nsmb.3452

Hage-Sleiman, R., Bahmad, H., Kobeissy, H., Dakdouk, Z., Kobeissy, F., and Dbaibo, G. (2017). Genomic alterations during p53-dependent apoptosis induced by γ-irradiation of Molt-4 leukemia cells. *PLoS ONE* 12:e0190221. doi: 10.1371/journal.pone.0190221

Hamby, D. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environ. Monitor. Assess.* 32, 135–154. doi: 10.1007/BF00547132

Haupt, Y., Maya, R., Kazaz, A., and Oren, M. (1997). Mdm2 promotes the rapid degradation of p53. *Nature* 387, 296–299. doi: 10.1038/387296a0

Hsing, A., Faller, D., and Vaziri, C. (2000). DNA-damaging aryl hydrocarbons induce Mdm2 expression via p53-independent post-transcriptional mechanisms. *J. Biol. Chem.* 275, 26024–26031. doi: 10.1074/jbc.M002455200

Ito, A., Lai, C., Zhao, X., Saito, S., Hamilton, M., Appella, E., et al. (2001). p300/CBP-mediated p53 acetylation is commonly induced by p53-activating agents and inhibited by MDM2. *EMBO J.* 20, 1331–1340. doi: 10.1093/emboj/20.6.1331

Itoshima, T., Fujiwara, T., Waku, T., Shao, J., Kataoka, M., Yarbrough, W., et al. (2000). Induction of apoptosis in human esophageal cancer cells by sequential transfer of the wild-type p53 and E2F-1 genes: involvement of p53 accumulation via ARF-mediated MDM2 down-regulation. *Clin. Cancer Res.* 6, 285–2859.

Jiao, H., Zhang, L., Shen, Q., Zhu, J., and Shi, P. (2018). Robust gene circuit control design for time-delayed genetic regulatory networks without SUM regulatory logic. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 2086–2093. doi: 10.1109/TCBB.2018.2825445

Kamijo, T., Weber, J., Zambetti, G., Zindy, F., Roussel, M., and Sherr, C. (1998). Functional and physical interactions of the ARF tumor suppressor with p53 and MDM2. *Proc. Natl. Acad. Sci. U.S.A.* 95, 8292–8297. doi: 10.1073/pnas.95.14.8292

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 9:770. doi: 10.1038/nrm2503

Kauffman, S. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437–467. doi: 10.1016/0022-5193(69)90015-0

Khan, S., Guevara, C., Fujii, G., and Parry, D. (2004). p14ARF is a component of the p53 response following ionizing irradiation of normal human fibroblasts. *Oncogene* 23, 6040–6046. doi: 10.1038/sj.onc.1207824

Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* 19, 643–650. doi: 10.1093/bioinformatics/btg027

Kuo, M., den Besten, W., Bertwistle, D., Roussel, M., and Sherr, C. (2004). N-terminal polyubiquitination and degradation of the ARF tumor suppressor. *Genes Dev.* 18, 1862–1874. doi: 10.1101/gad.1213904

Lahav, G., Rosenfeld, N., Sigal, A., Geva-Zatorsky, N., Levine, A., Elowitz, M., et al. (2004). Dynamics of the p53-MDM2 feedback loop in individual cells. *Nat. Genet.* 36:147. doi: 10.1038/ng1293

Lane, D. (1992). p53, guardian of the genome. *Nature* 358, 15–16. doi: 10.1038/358015a0

Leenders, G., and Tuszynski, J. (2013). Stochastic and deterministic models of cellular p53 regulation. *Front. Oncol.* 3:64. doi: 10.3389/fonc.2013.00064

Lewis, F., Zhang, H., Hengster-Movric, K., and Das, A. (2013). *Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches*. London: Springer-Verlag. doi: 10.1007/978-1-4471-5574-4

Li, F., Yan, H., and Karimi, H. (2018). Single-input pinning controller design for reachability of Boolean networks. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 3264–3269. doi: 10.1109/TNNLS.2017.2705109

Li, X., Wang, X., and Chen, G. (2004). Pinning a complex dynamical network to its equilibrium. *IEEE Trans. Circ. Syst. I* 51, 2074–2087. doi: 10.1109/TCSI.2004.835655

Lin, G., Ao, B., Chen, J., Wang, W., and Di, Z. (2014). Modeling and controlling the two-phase dynamics of the p53 network: a Boolean network approach. *New J. Phys.* 16:125010. doi: 10.1088/1367-2630/16/12/125010

Liu, Y., and Barabási, A. (2016). Control principles of complex systems. *Rev. Modern Phys.* 88:035006. doi: 10.1103/RevModPhys.88.035006

Momand, J., Zambetti, G., Olson, D., George, D., and Levine, A. (1992). The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* 69, 1237–1245. doi: 10.1016/0092-8674(92)90644-R

Mor, A., Suliman, S., Ben-Yishay, R., Yunger, S., Brody, Y., and Shav-Tal, Y. (2010). Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nat. Cell Biol.* 12:543. doi: 10.1038/ncb2056

Nilbert, M., Mitelman, F., Mandahl, N., Rydholm, A., and Willén, H. (1994). MDM2 gene amplification correlates with ring chromosomes in soft tissue tumors. *Genes Chromos. Cancer* 9, 261–265. doi: 10.1002/gcc.2870090406

Northrup, S., and Erickson, H. (1992). Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc. Natl. Acad. Sci. U.S.A.* 89, 3338–3342. doi: 10.1073/pnas.89.8.3338

Nowzari, C., Preciado, V., and Pappas, G. (2016). Analysis and control of epidemics: a survey of spreading processes on complex networks. *IEEE Control Syst.* 36, 26–46. doi: 10.1109/MCS.2015.2495000

Oliner, J., Kinzler, K., Meltzer, P., George, D., and Vogelstein, B. (1992). Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* 358:80. doi: 10.1038/358080a0

Oliner, J., Pietenpol, J., Thiagalingam, S., Gyuris, J., Kinzler, K., and Vogelstein, B. (1993). Oncoprotein MDM2 conceals the activation domain of tumour suppressor p53. *Nature* 362:857. doi: 10.1038/362857a0

Papatsenko, D., Waghray, A., and Lemischka, I. (2018). Feedback control of pluripotency in embryonic stem cells: signaling, transcription and epigenetics. *Stem Cell Res.* 29, 180–188. doi: 10.1016/j.scr.2018.02.012

Parisi, T., Pollice, A., Di Cristofano, A., Calabró, V., and La Mantia, G. (2002). Transcriptional regulation of the human tumor suppressor p14arf by E2F1, E2F2, E2F3, and Sp1-like factors. *Biochem. Biophys. Res. Commun.* 291, 1138–1145. doi: 10.1006/bbrc.2002.6591

Pauklin, S., Kristjuhan, A., Maimets, T., and Jaks, V. (2005). ARF and ATM/ATR cooperate in p53-mediated apoptosis upon oncogenic stress. *Biochem. Biophys. Res. Commun.* 334, 386–394. doi: 10.1016/j.bbrc.2005.06.097

Purvis, J., Karhohs, K., Mock, C., Batchelor, E., Loewer, A., and Lahav, G. (2012). p53 dynamics control cell fate. *Science* 336, 1440–1444. doi: 10.1126/science.1218351

Rau, A., Jaffrézic, F., Foulley, J., and Doerge, R. (2010). An empirical Bayesian method for estimating biological networks from temporal microarray data. *Stat. Appl. Genet. Mol. Biol.* 9, 1–26. doi: 10.2202/1544-6115.1513

Rayburn, E., Zhang, R., He, J., and Wang, H. (2005). MDM2 and human malignancies: expression, clinical pathology, prognostic markers, and implications for chemotherapy. *Curr. Cancer Drug Targets* 5, 27–41. doi: 10.2174/1568009053332636

Ren, W., and Beard, R. (2008). *Distributed Consensus in Multi-Vehicle Cooperative Control*. London: Springer-Verlag. doi: 10.1007/978-1-84800-015-5

Riley, T., Sontag, E., Chen, P., and Levine, A. (2008). Transcriptional control of human p53-regulated genes. *Nat. Rev. Mol. Cell Biol.* 9:402. doi: 10.1038/nrm2395

Scheffner, M., Werness, B., Huibregtse, J., Levine, A., and Howley, P. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* 63, 1129–1136. doi: 10.1016/0092-8674(90)90409-8

Selvin, P., and Ha, T. (2008). *Single-Molecule Techniques*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Shmulevich, I., Dougherty, E., Kim, S., and Zhang, W. (2002a). Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 18, 261–274. doi: 10.1093/bioinformatics/18.2.261

Shmulevich, I., Dougherty, E., and Zhang, W. (2002b). From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proc. IEEE* 90, 1778–1792. doi: 10.1109/JPROC.2002.804686

Sionov, R., and Haupt, Y. (1999). The cellular response to p53: the decision between life and death. *Oncogene* 18:6145. doi: 10.1038/sj.onc.1203130

Slotine, J., and Li, W. (1991). *Applied Nonlinear Control*. Englewood Cliffs, NJ: Prentice Hall.

Sorrentino, F., Di Bernardo, M., Garofalo, F., and Chen, G. (2007). Controllability of complex networks via pinning. *Phys. Rev. E* 75:046103. doi: 10.1103/PhysRevE.75.046103

Strigari, L., Mancuso, M., Ubertini, V., Soriani, A., Giardullo, P., Benassi, M., et al. (2014). Abscopal effect of radiation therapy: interplay between radiation dose and p53 status. *Int. J. Radiat. Biol.* 90, 248–255. doi: 10.3109/09553002.2014.874608

Suzuki, H., Kurita, M., Mizumoto, K., Nishimoto, I., Ogata, E., and Matsuoka, M. (2003). p19ARF-induced p53-independent apoptosis largely occurs through BAX. *Biochem. Biophys. Res. Commun.* 312, 1273–1277. doi: 10.1016/j.bbrc.2003.11.071

Sykes, S., Mellert, H., Holbert, M., Li, K., Marmorstein, R., Lane, W., et al. (2006). Acetylation of the p53 DNA-binding domain regulates apoptosis induction. *Mol. Cell* 24, 841–851. doi: 10.1016/j.molcel.2006.11.026

Szallasi, Z., Stelling, J., and Periwal, V. (2006). *System Modeling in Cellular Biology. From Concepts to Nuts and Bolts*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262195485.001.0001

Vassilev, L., Vu, B., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., et al. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science* 303, 844–848. doi: 10.1126/science.1092472

Villunger, A., Michalak, E., Coultas, L., Müllauer, F., Böck, G., Ausserlechner, M., et al. (2003). p53-and drug-induced apoptotic responses mediated by BH3-only proteins puma and noxa. *Science* 302, 1036–1038. doi: 10.1126/science.1090072

Vinayagam, A., Gibson, T., Lee, H., Yilmazel, B., Roesel, C., Hu, Y., et al. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4976–4981. doi: 10.1073/pnas.1603992113

Wagner, J., Ma, L., Rice, J., Hu, W., Levine, A., and Stolovitzky, G. (2005). p53-MDM2 loop controlled by a balance of its feedback strength and effective dampening using ATM and delayed feedback. *IEEE Proc. Syst. Biol.* 152, 109–118. doi: 10.1049/ip-syb:20050025

Wang, H., Qian, L., and Dougherty, E. (2010). Inference of gene regulatory networks using S-system: a unified approach. *IET Syst. Biol.* 4, 145–156. doi: 10.1049/iet-syb.2008.0175

Wang, L., Liu, Y., Wu, Z., Lu, J., and Yu, L. (2019). Stabilization and finite-time stabilization of probabilistic Boolean control networks. *IEEE Trans. Syst. Man Cybernet*. 1–8. doi: 10.1109/TSMC.2019.2898880

Wang, R., Saadatpour, A., and Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol*. 9:055001. doi: 10.1088/1478-3975/9/5/055001

Wang, X., and Chen, G. (2002). Pinning control of scale-free dynamical networks. *Phys. A* 310, 521–531. doi: 10.1016/S0378-4371(02)00772-0

Weber, J., Taylor, L., Roussel, M., Sherr, C., and Bar-Sagi, D. (1999). Nucleolar ARF sequesters MDM2 and activates p53. *Nat. Cell Biol*. 1:20. doi: 10.1038/8991

Wee, K., Surana, U., and Aguda, B. (2009). Oscillations of the p53-AKT network: implications on cell survival and death. *PLoS ONE* 4:e4407. doi: 10.1371/journal.pone.0004407

Wee, K., Yio, W., Surana, U., and Chiam, K. (2012). Transcription factor oscillations induce differential gene expressions. *Biophys. J*. 102, 2413–2423. doi: 10.1016/j.bpj.2012.04.023

Weinberg, R., Veprintsev, D., Bycroft, M., and Fersht, A. (2005). Comparative binding of p53 to its promoter and DNA recognition elements. *J. Mol. Biol*. 348, 589–596. doi: 10.1016/j.jmb.2005.03.014

Yee-Lin, V., Pooi-Fong, W., and Soo-Beng, A. (2018). Nutlin-3, a p53-mdm2 antagonist for nasopharyngeal carcinoma treatment. *Mini Rev. Med. Chem*. 18, 173–183. doi: 10.2174/1389557517666170717125821

Yue, D., Guan, Z., Li, T., Liao, R., Liu, F., and Lai, Q. (2017). Event-based cluster synchronization of coupled genetic regulatory networks. *Phys. A* 482, 649–665. doi: 10.1016/j.physa.2017.04.024

Zerrouqi, A., Pyrzynska, B., Brat, D., and Van Meir, E. (2014). P14arf suppresses tumor-induced thrombosis by regulating the tissue factor pathway. *Cancer Res*. 74, 1371–1378. doi: 10.1158/0008-5472.CAN-13-1951

Zerrouqi, A., Pyrzynska, B., Febbraio, M., Brat, D., and Van Meir, E. (2012). P14arf inhibits human glioblastoma-induced angiogenesis by upregulating the expression of timp3. *J. Clin. Invest*. 122, 1283–1295. doi: 10.1172/JCI38596

Zhang, Y., Xiong, Y., and Yarbrough, W. (1998). ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the RB and p53 tumor suppression pathways. *Cell* 92, 725–734. doi: 10.1016/S0092-8674(00)81401-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 04 September 2020
doi: 10.3389/fphys.2020.01055

Neural Network Deconvolution Method for Resolving Pathway-Level Progression of Tumor Clonal Expression Programs With Application to Breast Cancer Brain Metastases

Yifeng Tao[1,2], *Haoyun Lei*[1,2], *Adrian V. Lee*[3], *Jian Ma*[1] and *Russell Schwartz*[1,4]*

[1] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States, [2] Joint Carnegie Mellon–University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh, PA, United States, [3] Department of Pharmacology and Chemical Biology, UPMC Hillman Cancer Center, Magee-Womens Research Institute, University of Pittsburgh, Pittsburgh, PA, United States, [4] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, United States

OPEN ACCESS

Edited by:
Katharina Jahn,
ETH Zürich, Switzerland

Reviewed by:
Zhihui Wang,
Houston Methodist Research Institute, United States
Yoshitaka Kimura,
Tohoku University, Japan

***Correspondence:**
Russell Schwartz
russells@andrew.cmu.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: 31 January 2020
Accepted: 31 July 2020
Published: 04 September 2020

Citation:
Tao Y, Lei H, Lee AV, Ma J and Schwartz R (2020) Neural Network Deconvolution Method for Resolving Pathway-Level Progression of Tumor Clonal Expression Programs With Application to Breast Cancer Brain Metastases. Front. Physiol. 11:1055. doi: 10.3389/fphys.2020.01055

Metastasis is the primary mechanism by which cancer results in mortality and there are currently no reliable treatment options once it occurs, making the metastatic process a critical target for new diagnostics and therapeutics. Treating metastasis before it appears is challenging, however, in part because metastases may be quite distinct genomically from the primary tumors from which they presumably emerged. Phylogenetic studies of cancer development have suggested that changes in tumor genomics over stages of progression often result from shifts in the abundance of clonal cellular populations, as late stages of progression may derive from or select for clonal populations rare in the primary tumor. The present study develops computational methods to infer clonal heterogeneity and dynamics across progression stages via deconvolution and clonal phylogeny reconstruction of pathway-level expression signatures in order to reconstruct how these processes might influence average changes in genomic signatures over progression. We show, via application to a study of gene expression in a collection of matched breast primary tumor and metastatic samples, that the method can infer coarse-grained substructure and stromal infiltration across the metastatic transition. The results suggest that genomic changes observed in metastasis, such as gain of the *ErbB* signaling pathway, are likely caused by early events in clonal evolution followed by expansion of minor clonal populations in metastasis, a finding that may have translational implications for early detection or prevention of metastasis[1].

Keywords: breast cancer, brain metastases, phylogenetics, deconvolution, pathways, gene modules, transcriptome, matrix factorization

[1]Algorithmic details, parameter settings, and source code are available at https://github.com/CMUSchwartzLab/NND. Additional results and proofs are provided in the **Supplementary Material**.

1. INTRODUCTION

Metastatic disease is the primary mechanism by which cancer results in patient mortality (Chambers et al., 2002; Chaffer and Weinberg, 2011). By the time metastases have appeared, there are generally no viable treatment options (Guan, 2015). Successful treatment thus depends on treating not just the primary tumor but also the seeds of metastasis that may linger after a seemingly successful remission. Identifying successful treatment options for metastasis is problematic, however, since the genomics of primary and metastatic tumors may be quite different even in single patients and metastatic cell populations may be poorly responsive to therapies effective on the primary tumor. Studies of cell-to-cell variation in cancers have revealed often substantial clonal heterogeneity in single tumors, with clonal populations sometimes dramatically shifting across progression stages (Greaves and Maley, 2012). Phylogenetic studies of clonal populations have been inconclusive on the typical evolutionary relationships between primary and metastatic tumors (Schwartz and Schäffer, 2017). It remains a matter of debate whether changes in clonal composition occur primarily through ongoing clonal evolution, which results in novel clones with metastatic potential and resistance to therapy, or from selection on existing clonal heterogeneity already present at the time of first treatment (Ding et al., 2013; de Bruin et al., 2014). The degree to which either answer is true has important implications for prospects for early detection or prophylactic treatment of metastasis.

Brain metastases (BrMs) occur in around 10–30% of metastatic breast cancers cases (Lin et al., 2004). Although recent advances in the treatment of metastatic breast cancer have been able to achieve long-term overall survival, there are limited treatment options for BrMs and clinical prognoses are still disappointing (Witzel et al., 2016). Recent work examining transcriptomic changes between paired primary and BrM samples has demonstrated dramatic changes in expression programs over metastasis, including changes in tumor subtype with important implications for treatment options and prognosis (Priedigkeit et al., 2017; Vareslija et al., 2018). Some past research has sought to infer phylogenetic models to explain the development of brain metastases based on somatic genomic alterations (Brastianos et al., 2015; Körber et al., 2019). Such methods are challenged in drawing robust conclusions about recurrent progression processes, though, by the high heterogeneity within single tumors and across progression stages and patients. While single-cell methods are proving powerful for resolving such problems in other contexts (Qiu et al., 2011; Elyanow et al., 2020), such data is rarely available for studies of metastatic progression, which generally require working with samples archived years before metastases are discovered. Changes in the activity of particular genetic pathways or modules may provide a more robust measure of frequent genomic alterations across cancers.

In the present work, we develop a strategy for tumor phylogenetics to explore how changes in clonal composition, via both novel molecular evolution and shifts in population dynamics of tumor clones and associated stroma, influence changes in expression programs across such progression stages. Our methods make use of multi-site bulk transcriptomic data to profile changes evident in gene expression programs between clones and progression stages. We break from past work in this domain in that we seek to study not clones *per se*, as is typical in tumor phylogenetics (Eaton et al., 2018; Tao et al., 2019b), but what we dub "cell communities": collections of clones or other stromal cell types that persist as a group with similar proportions across samples (section 2.4). We accomplish this via a novel transcriptomic deconvolution approach designed to make use of multiple samples both within and between patients (Schwartz and Shackney, 2010; Zare et al., 2014) while improving robustness to inter- and intra-tumor heterogeneity by integrating deconvolution with pathway-based analyses of expression variation (Park et al., 2009).

2. MATERIALS AND METHODS

2.1. Overview

Cell populations evolve due to genomic perturbations that can result in changes in the activity of various functional pathways between clones. Our overall method for deriving coarse-grained portraits of cell community evolution at the pathway level is illustrated by **Figure 1**. After the preprocessing of transcriptome data (section 2.2), the overall workflow consists of three main steps: First, the bulk expression profiles are mapped into the gene module and pathway space using external knowledge bases to reduce redundancy, noise, and sparsity, and to provide markers of expression variation for the subsequent analysis (section 2.3). Second, a deconvolution step is implemented to resolve cell communities, i.e., coarse-grained mixtures of cell types presumed to represent an associated population of cancer clones and stromal cells, from the compressed pathway representation of samples (section 2.4). Third, phylogenies of these cell communities are built based on the deconvolved communities as well as inferred ancestral (Steiner) communities to reconstruct likely trajectories of evolutionary progression by which cell communities develop—through a combination of genetic mutations, expression changes, and changes in population distributions—as a tumor progresses from healthy tissue to primary and potentially metastatic tumor (section 2.5).

2.2. Transcriptome Data Preprocessing

We applied our methods to raw bulk RNA-Sequencing data of 44 matched primary breast and metastatic brain tumors from 22 patients (each patient gives two samples) (Priedigkeit et al., 2017; Vareslija et al., 2018), where six patients were from the Royal College of Surgeons (RCS) and sixteen patients from the University of Pittsburgh (Pitt). These data profiled the expression levels of ∼60,000 transcripts. These can be represented in the format of a matrix, with rows corresponding to genes and columns to the samples (primary tumors or metastases). We removed the genes that are not expressed in any sample. We also considered only protein-coding genes in the present study. Approximately 20,000 genes remain after the filter. We conducted quantile normalization across samples using the geometric mean to remove possible artifacts (Amaratunga and Cabrera, 2001). The top 2.5% and bottom

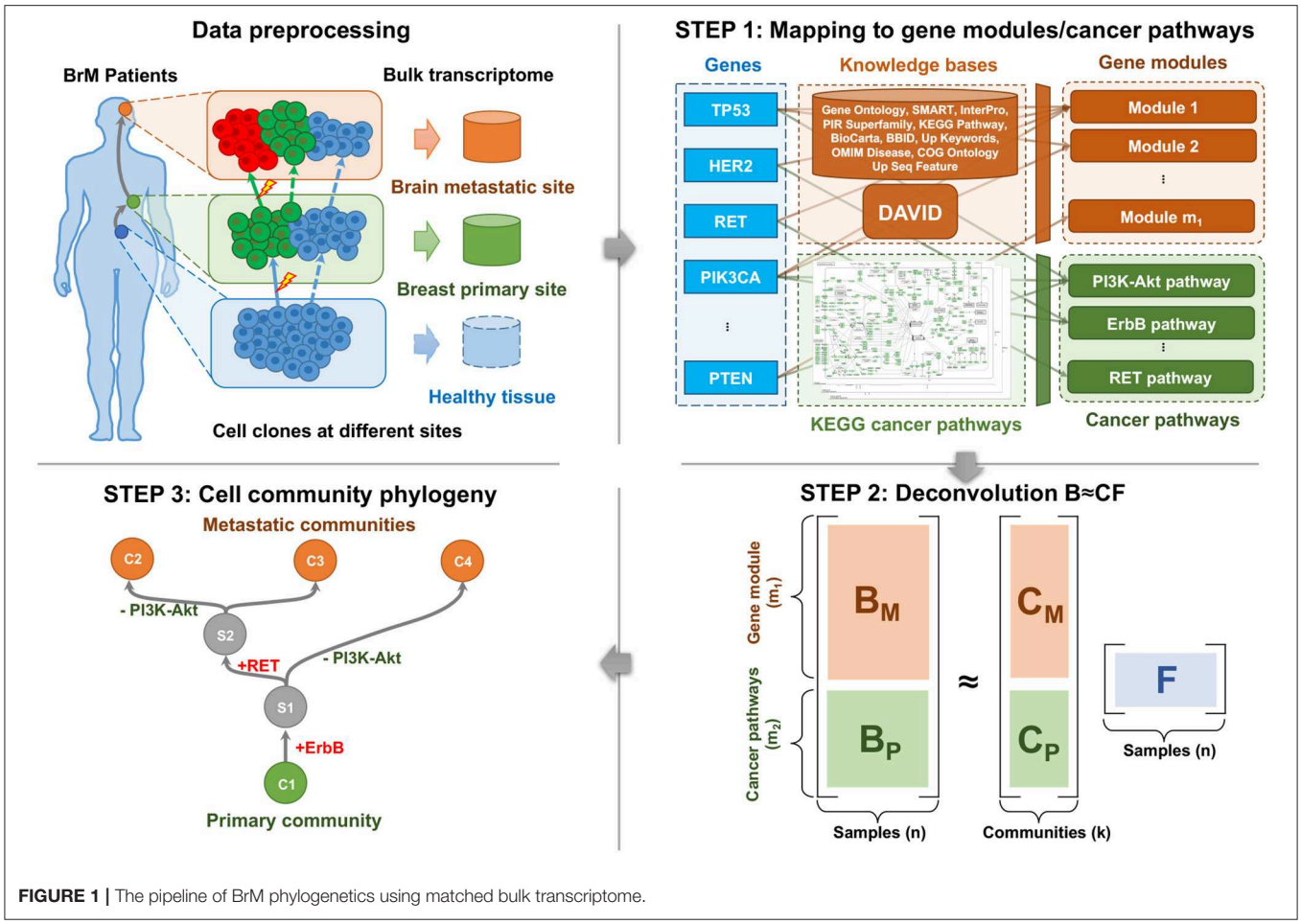

FIGURE 1 | The pipeline of BrM phylogenetics using matched bulk transcriptome.

2.5% of expressions were clipped to further reduce noise. Finally, we transformed the resulting bulk gene expression values into the log space and mapped those for each gene to the interval $[0, 1]$ by a linear transformation. The resulting preprocessed transcriptome data were used as the input of Step 1 (section 2.3).

2.3. Mapping to Gene Modules and Cancer Pathways

The protein-coding gene expressions were mapped into both perturbed gene modules and cancer pathways, using the DAVID tool and external knowledge bases (Huang et al., 2009), as well as the cancer pathways in the KEGG database (Kanehisa and Goto, 2000). This step compresses the high dimensional data and provides markers of cancer-related biological processes (**Figure 1**, Step 1). Note that although both gene module and cancer pathway representations capture recurrent features of metastatic progression, they serve different purposes in our analysis. Gene modules are an essential part of deconvolution in the following steps because they provide the major variance within the data. Cancer pathways serve primarily as probes for *post-hoc* interpretation of the unmixed communities, but are biased relative to the gene module space by the focus only on genes with known relevance to cancer.

2.3.1. Gene Modules

Functionally similar genes are usually affected by a common set of somatic alterations (Park et al., 2009) and therefore are co-expressed in the cells. These genes are believed to belong to the same "gene modules" (Desmedt et al., 2008; Tao et al., 2020). Inspired by the idea of gene modules, we fed a subset of 3,000 most informative genes out of the ∼20,000 genes that have the largest variances into the DAVID tool for functional annotation clustering using several databases (Huang et al., 2009). DAVID maps each gene to one or more modules. We did not force the genes to be mapped into disjunct modules because a gene may be involved in several biological functions and therefore more than one gene module. We removed gene modules that were not enriched (fold enrichment < 1.0) and kept the remaining $m_1 = 109$ modules (and the corresponding annotated functions), where fold enrichment is defined as the EASE score of the current module to the geometric mean of EASE scores in all modules (Hosack et al., 2003). The gene module values of all the $n = 44$ samples were represented as a gene module matrix $\mathbf{B}_M \in \mathbb{R}^{m_1 \times n}$. The i-th gene module value in j-th sample, $(\mathbf{B}_M)_{i,j}$, was calculated by taking the sum of expressions of all the genes in the i-th module. Then $\mathbf{B}_M$ was rescaled row-wise by taking the z-scores across samples to compensate for the effect of variable module sizes.

2.3.2. Cancer Pathways

Although the gene module representation is able to capture the variances across samples and reduce the redundancy of raw gene expressions, it has two disadvantages. The first is a lack of interpretability. Specifically, some annotations assigned by DAVID are not directly related to biological functions, and the annotations of different modules may substantially overlap. The second is that the key perturbed cancer pathways or functions may not always be the ones that vary most across samples. For example, genes in cancer-related KEGG pathways (hsa05200; Kanehisa and Goto, 2000) are not especially enriched in the top 3,000 genes with the largest expression variances. To make better use of prior knowledge on cancer-relevant pathways, we supplemented the generic DAVID pathway sets with a KEGG "cancer pathway" representation of samples $\mathbf{B}_P \in \mathbb{R}^{m_2 \times n}$, where the number of cancer pathways $m_2 = 24$. The cancer-related pathways in the KEGG database are cleaner and easier to explain, more orthogonal to each other, and contain critical signaling pathways to cancer development. We extracted the 23 cancer-related pathways from the following 3 KEGG pathway sets: *Pathways in cancer* (hsa05200), *Breast cancer* (hsa05224), and *Glioma* (hsa05214). An additional cancer pathway *RET pathway* was added, since it was found to be recurrently gained in the prior research (Vareslija et al., 2018). See *y*-axis of **Figure 4D** for the complete list of 24 cancer pathways. We considered all the ∼20,000 protein-coding genes other than top 3,000 genes. The following mapping of cancer pathways and transformation to *z*-scores were similar to that we did to map the gene modules.

Until this step, the raw gene expressions of n samples were transformed into the compressed gene module/pathway representation of samples $\mathbf{B} = \left[\mathbf{B}_M^\intercal, \mathbf{B}_P^\intercal\right]^\intercal \in \mathbb{R}^{m \times n}$, where $m = m_1 + m_2$. The gene module representation $\mathbf{B}_M$ serves for accurately deconvolving and unmixing the cell communities, while the pathway representation $\mathbf{B}_P$ serves as markers/probes and for interpretation purpose.

2.4. Deconvolution of Bulk Data

We applied a type of matrix factorization (MF) with constraints on the pathway-level expression signatures to deconvolve the communities/populations from primary and metastatic tumor samples (**Figure 1**, Step 2) (Koren et al., 2009). Note that common alternatives, such as principal components analysis (PCA) and non-negative matrix factorization (NMF) are not amenable to this case (Lee and Seung, 2000), since PCA does not provide a feasible solution to the constrained problem, and the NMF does not apply to our mixture data, which can be either positive or negative.

2.4.1. Cell Communities

We define a cell community to be a set of clones/clonal subpopulations and other cell types that propagate as a group during the evolution of a tumor. A community may be just a single subpopulation/clone, but is a more general concept in the sense that it usually involves multiple related clones and their associated stroma. For example, a set of immunogenic clones and the immune cells infiltrating them might collectively form a community that has a collective expression signature mixing signatures of the clones and associated immune cells, even if the individual cell types are not distinguishable from bulk expression data alone. While much work in this space has classically aimed to separate individual clones, or perhaps individual cell types more broadly defined, we note that deconvolution may be unable in principle to resolve distinct cell types if they are always co-located in similar proportions. It is particularly true when data is sparse and cell types are fit only approximately, as in the present work, that a model with large complexity to deconvolve the fine-grained populations is prone to overfit. The community concept is intended in part to better describe the results we expect to achieve from the kind of data examined here and in part because identifying these communities is itself of interest in understanding how tumor cells coevolve with their stroma during progression and metastasis. Single-cell methods may provide an alternative, but are not amenable to preserved samples, such as are needed when retrospectively studying primary tumors and metastases that may have been biopsied years apart.

2.4.2. Formulation of Deconvolution

With a matrix of bulk pathway values $\mathbf{B} \in \mathbb{R}^{m \times n}$, the deconvolution problem is to find a component matrix $\mathbf{C} = \left[\mathbf{C}_M^\intercal, \mathbf{C}_P^\intercal\right]^\intercal \in \mathbb{R}^{m \times k}$ that represents the inferred fundamental communities of tumors, and the corresponding set of mixture fractions $\mathbf{F} \in \mathbb{R}_+^{k \times n}$:

$$\min_{\mathbf{C},\mathbf{F}} \quad \|\mathbf{B} - \mathbf{C}\mathbf{F}\|_{\mathrm{Fr}}^2, \tag{1}$$

$$\text{s.t.} \quad \mathbf{F}_{lj} \geq 0, \qquad l = 1, ..., k,\ j = 1, ..., n, \tag{2}$$

$$\sum_{l=1}^{k} \mathbf{F}_{lj} = 1, \qquad j = 1, ..., n, \tag{3}$$

where $||\mathbf{X}||_{\mathrm{Fr}}$ is the Frobenius norm. The column-wise normalization in Equation (3) aims for recovering the biologically meaningful cell communities. In addition, they are equivalent to applying ℓ_1 regularizers and therefore enforce sparsity to the fraction matrix $\mathbf{F}$ (**Supplementary Material**).

2.4.3. Neural Network Deconvolution

Although it is possible to build new algorithms for solving MF by adapting previous work (Lee and Seung, 2000), the additional but necessary constraints of Equations (2) and (3) make the optimization much harder to solve. For the problem of Equations (1)–(3), one can prove that it does not generally guarantee convexity (**Supplementary Material**). A slightly modified version of the algorithm to solve NMF with constraints may guarantee neither good fitting nor convergence (Lei et al., 2019, 2020). Therefore, instead of revising existing MF algorithms, such as ALS-FunkSVD (Funk, 2006; Bell and Koren, 2007; Koren et al., 2009), we developed an algorithm which we call "neural network deconvolution" (NND) to solve the optimization problem using gradient descent. Specifically, the NND was implemented using backpropagation in the form of a neural network (**Figure 2A**) with the PyTorch package (https://pytorch.org/) (Rumelhart et al., 1986; Kingma and Ba,

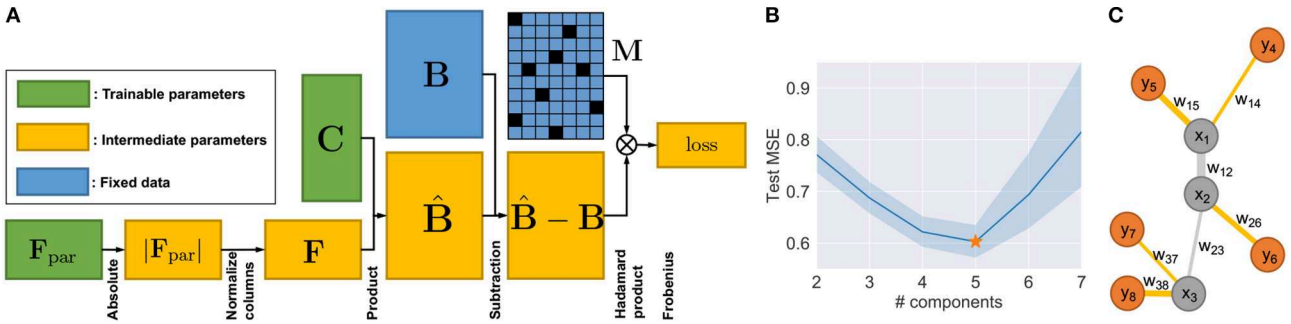

FIGURE 2 | Method details. **(A)** Neural network architecture of NND. **(B)** Test errors of NND using 20-fold CV. Errors in unit of mean square error (MSE). **(C)** Illustration of a phylogeny with five extant nodes and three Steiner nodes.

2014), based on the revised constraints:

$$\min_{\mathbf{C},\mathbf{F}_{par}} \quad \|\mathbf{B} - \mathbf{CF}\|_{Fr}^2, \tag{4}$$

$$\text{s.t.} \quad \mathbf{F} = \text{cwn}\left(|\mathbf{F}_{par}|\right), \tag{5}$$

where $|\mathbf{X}|$ applies element-wise absolute value and $\text{cwn}(\mathbf{X})$ is column-wise normalization, so that each column sums up to 1. The two operations of Equation (5) naturally rephrase and remove the two constraints in Equations (2) and (3), and meanwhile fit the framework of neural networks. An alternative to the absolute value operation $|\mathbf{X}|$ might be rectified linear unit $\text{ReLU}(\mathbf{X}) = \max(\mathbf{0}, \mathbf{X})$. However, this activation function is unstable and leads to inferior performance in our case, since $\mathbf{X}_{lj}$ will be fixed to zero once it becomes negative and will lose the chance to get updated in the following iterations. One may also want to replace the column-wise normalization $\text{cwn}(\mathbf{X})$ with softmax operation $\text{softmax}(\mathbf{X})$. However, the non-linearity introduced by softmax actually changes the original optimization problem (Equations 1–3) and the fitted $\mathbf{F}$ is therefore not sparse.

Based on the revised NND optimization problem (Equations 4 and 5), we built the neural network with the architecture shown in **Figure 2A**. An Adam optimizer other than vanilla gradient descent was used with default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate of 1×10^{-5} (Kingma and Ba, 2014). The mini-batch technique is not required since the data size in our application is small enough not to require it ($\mathbf{B} \in \mathbb{R}^{m \times n}$, $m = 133$, $n = 44$). The training is run until convergence, which is defined as when the relative decrease of training loss is smaller than $\epsilon = 1 \times 10^{-10}$ every 20,000 iterations. This implementation has two main advantages: First, the method can be easily adapted to a wide range of optimization scenarios with various constraints, when existing methods do not or are hard to apply. Second, the NND has the flexibility of allowing for cross-validation, which is important for us in choosing the number of components k and preventing overfitting.

One might be suspicious whether the neural network fits precisely in practice, since it is based on a simple gradient descent optimization. To validate the fitting ability of NND, we plotted the PCA of original samples $\mathbf{B}$ and the fitted samples $\hat{\mathbf{B}}$ (**Supplementary Material**). One can easily see that NND provides a good fit to the data.

2.4.4. Cross-Validation of NND

In order to find the best tradeoff between model complexity and overfitting, we used cross-validation (CV) with the "masking" method to choose the optimal number of components/communities $k = 5$ that has the smallest test error (**Figure 2B**). In each fold of the CV, we used estimated $\hat{\mathbf{B}}$ to only fit some randomly selected elements of $\mathbf{B}$, and then the test error was calculated using the other elements of $\mathbf{B}$. This was implemented by introducing two additional mask matrices $\mathbf{M}_{\text{train}}, \mathbf{M}_{\text{test}} \in \{0,1\}^{m \times n}$, which are in the same shape of $\mathbf{B}$, and $\mathbf{M}_{\text{train}} + \mathbf{M}_{\text{test}} = 1^{m \times n}$. During the training time, with the same constraints in Equation (5), the optimization goal is:

$$\min_{\mathbf{C},\mathbf{F}_{par}} \|\mathbf{M}_{\text{train}} \odot (\mathbf{B} - \mathbf{CF})\|_{Fr}^2, \tag{6}$$

where $\mathbf{X} \odot \mathbf{Y}$ is the Hadamard (element-wise) product. At the time of evaluation, given optimized $\hat{\mathbf{C}}$, $\hat{\mathbf{F}}_{par}$, and therefore optimized $\hat{\mathbf{F}} = \text{cwn}\left(\left|\hat{\mathbf{F}}_{par}\right|\right)$ for the optimization problem during training, the test error was calculated on the test set: $\left\|\mathbf{M}_{\text{test}} \odot \left(\mathbf{B} - \hat{\mathbf{C}}\hat{\mathbf{F}}\right)\right\|_{Fr}^2$. We used 20-fold cross-validation on the NND, so in each fold 95% of positions of $\mathbf{M}_{\text{train}}$ and 5% of positions of $\mathbf{M}_{\text{test}}$ were 1s. Note that the actual number of cell populations is probably considerably larger than 5, and therefore each one of the five communities may contain multiple cell populations. Furthermore, it is likely that with sufficient numbers and precision of measurements, these communities could be more finely resolved into their constituent cell types. However $k = 5$ represents the largest hypothesis space of NND model that can be applied to the current dataset without severe overfitting.

2.5. Phylogeny of Inferred Cell Subcommunities and Pathway Inference of Steiner Nodes

We built "phylogenies" of cell subcommunities and estimated the pathway representation of unobserved (Steiner) nodes (Lu et al., 2003) inferred to be ancestral to them, with the goal of discovering critical communities that appear to be involved in the transition to metastasis and identifying the important changes of functions and expression pathways during this transition

(**Figure 1**, Step 3). Note that we are using the term "phylogeny" loosely here, as these trees are intended to capture evolution of populations of cells not just by accumulation of mutations from a single ancestral clone but also via changes in community structure, for example, due to generating or suppressing an immune response or migrating to a metastatic site. Although an abuse of terminology, we use the term phylogeny here due to the methodological similarity to more proper phylogenetic methods in wide use for analyzing mutational data in cancers (Schwartz and Schäffer, 2017).

2.5.1. Phylogeny of Communities

Given the pathway profiles of the extant communities at the time of collecting tumor samples $\mathbf{C} \in \mathbb{R}^{m\times k}$, a phylogeny of the k extant cell communities was built using the neighbor-joining (NJ) algorithm (Nei and Saitou, 1987), which inferred a tree that contains k extant nodes/leaves, $k-2$ unobserved Steiner nodes, and edges connecting two Steiner nodes or a Steiner node and an extant node. We estimated an evolutionary distance for any pair of two communities u, v as the input of NJ using the Euclidean distance between their pathway vectors $\|\mathbf{C}_{\cdot u} - \mathbf{C}_{\cdot v}\|_2$, similar to that in a prior work (Park et al., 2009).

2.5.2. Inference of Pathways: Setting and Approach

Denote the phylogeny of cell subcommunities as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and $\mathcal{V} = \mathcal{V}_S \cup \mathcal{V}_C$, where the indices of Steiner node $\mathcal{V}_S = \{1, 2, ..., k-2\}$ ($|\mathcal{V}_S| = k-2$), the indices of extant nodes $\mathcal{V}_C = \{k-1, k, ..., 2k-2\}$ ($|\mathcal{V}_C| = k$). For each edge $(u, v) \in \mathcal{E}$, where $1 \le u < v \le 2k-2$, the first node of edge $u \le k-2$ is always a Steiner node. The second node v can be either a Steiner node ($v \le k-2$) or extant node ($v \ge k-1$). Denote the set of weights $\mathcal{W} = \{w_{uv} = 1/d_{uv} \mid (u, v) \in \mathcal{E}\}$ (inverse distance), where the edge length d_{uv} is the output of NJ. For each dimension i of the pathway vectors, we consider them independently and separately, so that each dimension of the Steiner nodes can be solved in the same way. Now let us consider the i-th dimension (and omit the subscript i for brevity) of extant nodes $\mathcal{V}_C$: $\mathbf{y} = \left[y_{k-1}, y_k, ..., y_{2k-2}\right]^\intercal = \mathbf{C}_{i\cdot}^\intercal \in \mathbb{R}^k$ and Steiner nodes $\mathcal{V}_S$: $\mathbf{x} = \left[x_1, x_2, ..., x_{k-2}\right]^\intercal \in \mathbb{R}^{k-2}$. **Figure 2C** illustrates a phylogeny where $k = 5$. The inference of the i-th element in the pathway vector of the Steiner nodes can be formulated as minimizing the following elastic potential energy $U(\mathbf{x}, \mathbf{y}; \mathcal{W})$:

$$\min_{\mathbf{x}} \quad U(\mathbf{x}, \mathbf{y};\ \mathcal{W}) = \sum_{\substack{(u,v)\in\mathcal{E}\\ v\le k-2}} \frac{1}{2} w_{uv}(x_u - x_v)^2 + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \frac{1}{2} w_{uv}(x_u - y_v)^2, \tag{7}$$

which can be rephrased as a quadratic programming problem and solved easily, as we show below.

2.5.3. Inference of Pathways: Derivation of Quadratic Programming, P($\mathcal{W}$), and q($\mathcal{W}$, y)

Theorem 1. *Equation (7) can be further rephrased as a quadratic programming problem:*

$$\min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(\mathcal{W})\mathbf{x} + \mathbf{q}(\mathcal{W}, \mathbf{y})^\intercal \mathbf{x}, \tag{8}$$

where $\mathbf{P}(\mathcal{W})$ *is a function that takes as input edge weights* $\mathcal{W}$ *and outputs a matrix* $\mathbf{P} \in \mathbb{R}^{(k-2)\times(k-2)}$, $\mathbf{q}(\mathcal{W}, \mathbf{y})$ *is a function that takes as input edge weights* $\mathcal{W}$ *and vector* $\mathbf{y}$ *and outputs a vector* $\mathbf{q} \in \mathbb{R}^{k-2}$.

Proof: Based on Equation (7), $U(\mathbf{x}, \mathbf{y}; \mathcal{W}) \ge 0$. Each term inside the first summation ($v \le k-2$) can be written as:

$$\frac{1}{2} w_{uv}(x_u - x_v)^2 = \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x}, \tag{9}$$

where

$$\mathbf{P}(w_{uv}) = \begin{array}{c} \\ u\text{-th row} \\ \\ v\text{-th row} \\ \end{array} \begin{array}{c} \begin{array}{ccccc} u\text{-th col} & & & v\text{-th col} & \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & w_{uv} & \mathbf{0} & -w_{uv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -w_{uv} & \mathbf{0} & w_{uv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array}. \tag{10}$$

Each term ($v \ge k-1$) inside the second summation can be rephrased as:

$$\frac{1}{2} w_{uv}(x_u - y_v)^2 = \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x} + C(w_{uv}, y_v), \tag{11}$$

where

$$\mathbf{P}(w_{uv}) = \begin{array}{c} \\ u\text{-th row} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{c} u\text{-th col} \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0}\ \mathbf{0}\ \mathbf{0} \\ \mathbf{0} & w_{uv} & \mathbf{0}\ \mathbf{0}\ \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}\ \mathbf{0}\ \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}\ \mathbf{0}\ \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}\ \mathbf{0}\ \mathbf{0} \end{bmatrix} \end{array},$$

$$\mathbf{q}(w_{uv}, y_v) = \begin{array}{c} \\ u\text{-th row} \\ \\ \\ \\ \end{array} \begin{bmatrix} \mathbf{0} \\ -w_{uv}y_v \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{12}$$

and $C(w_{uv}, y_v) = \frac{1}{2} w_{uv} y_v^2$ is independent of $\mathbf{x}$. Therefore the optimization in Equation (7) can be calculated and written

as below:

$$\min_{\mathbf{x}} \quad \sum_{\substack{(u,v)\in\mathcal{E}\\ v\le k-2}} \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \left(\frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x}\right), \tag{13}$$

$$\Leftrightarrow \min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^\intercal \left(\sum_{\substack{(u,v)\in\mathcal{E}\\ v\le k-2}} \mathbf{P}(w_{uv}) + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \mathbf{P}(w_{uv})\right)\mathbf{x} + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x}, \tag{14}$$

$$\Leftrightarrow \min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(\mathcal{W})\mathbf{x} + \mathbf{q}(\mathcal{W}, \mathbf{y})^\intercal \mathbf{x}. \quad \square \tag{15}$$

Remark 1. *The optimal* $\mathbf{x}^\star$ *of the Equation (7), or the solution to the quadratic programming problem Equation (8) can be solved by setting the gradient to be* $\mathbf{0}$*:*

$$\mathbf{P}(\mathcal{W})\mathbf{x}^\star + \mathbf{q}(\mathcal{W}, \mathbf{y}) = \mathbf{0}. \tag{16}$$

Therefore,

$$\mathbf{x}^\star = -\mathbf{P}(\mathcal{W})^{-1}\mathbf{q}(\mathcal{W}, \mathbf{y}). \tag{17}$$

Remark 2. *Based on the proof, we can derive how to calculate the matrix* $\mathbf{P}(\mathcal{W})$ *and vector* $\mathbf{q}(\mathcal{W}, \mathbf{y})$*.*

Initialize the matrix and vector with zeros:

$$\mathbf{P} \leftarrow 0^{(k-2)\times(k-2)}, \quad \mathbf{q} \leftarrow 0^{k-2}. \tag{18}$$

For each edge $(u, v) \in \mathcal{E}$ *with weight* w_{uv}*, there are two possibilities of nodes u and v: First, if both of them are Steiner nodes* ($u \le k-2$, $v \le k-2$)*, we update* $\mathbf{P}$ *and keep* $\mathbf{q}$ *the same:*

$$\mathbf{P}_{uu} \leftarrow \mathbf{P}_{uu} + w_{uv}, \mathbf{P}_{vv} \leftarrow \mathbf{P}_{vv} + w_{uv}, \quad \mathbf{P}_{uv} \leftarrow \mathbf{P}_{uv} - w_{uv}, \ \mathbf{P}_{vu} \leftarrow \mathbf{P}_{vu} - w_{uv}. \tag{19}$$

Second, if u is Steiner node and v is an extant node ($u \le k-2$, $v \ge k-1$)*, we update both* $\mathbf{P}$ *and* $\mathbf{q}$*:*

$$\mathbf{P}_{uu} \leftarrow \mathbf{P}_{uu} + w_{uv}, \quad \mathbf{q}_u \leftarrow \mathbf{q}_u - y_v \cdot w_{uv}. \tag{20}$$

We apply the same procedure to all dimension of pathways $i = 1, 2, ..., m$ *to get the full pathway values for each Steiner node.*

3. RESULTS

3.1. NND Deconvolves the Bulk RNA Accurately

Before we applied our deconvolution algorithm NND to the breast cancer brain metastatic samples, we first validated our algorithm on a semi-simulated dataset where the ground truth expressions and fractions of each cell clone in the mixture samples are known.

3.1.1. Semi-simulated GSE11103 Dataset

The semi-simulated dataset is based on the real data of pure clones from the GSE11103 dataset (Abbas et al., 2009; Barrett et al., 2013). Expression profiles of four different cells were measured using microarrays: Raji (B cell), IM-9 (B cell), THP-1 (monocyte), Jurkat (T cell). Each experiment was repeated three times. We took the average of the three replicates to get the expression data of the four pure cell clones. The top 300 genes that varied most across cell types were selected as the ground truth real data of pure cell clones: $\mathbf{C} \in \mathbb{R}_+^{300\times 4}$. We then created 100 mixture samples of the four pure clones *in silico* $\mathbf{B} \in \mathbb{R}_+^{300\times 100}$ by randomly generating the fraction matrix $\mathbf{F} \in \mathbb{R}_+^{4\times 100}$. The fraction matrix was generated in the following way:

$$\mathbf{F}_{lj} \leftarrow U(0,1), \qquad l = 1, ..., 4, \ j = 1, ..., 100, \tag{21}$$

$$\mathbf{F}_{lj} \leftarrow \frac{\mathbf{F}_{lj}}{\sum_{l'=1}^{4} \mathbf{F}_{l'j}}, \qquad j = 1, ..., 100, \tag{22}$$

where $U(0, 1)$ is a uniform distribution in the interval $[0, 1]$. The semi-simulated bulk expression matrix $\mathbf{B}$ was then generated from $\mathbf{C}$, $\mathbf{F}$, with a log-normal noise:

$$(\mathbf{B})_{ij} = (\mathbf{CF})_{ij} + 2^{\mathcal{N}(0,(s\sigma)^2)}, \quad i = 1, ..., 300, \ j = 1, ..., 100, \tag{23}$$

where $\mathcal{N}\left(0, (s\sigma)^2\right)$ is a Gaussian distribution; s controls the noise level, which we set to 0, 0.4, 0.9, and 1.3 for test; σ is the standard deviation of $\log_2$-transformed original GSE11103 data.

3.1.2. Performance Evaluation

Given the bulk matrix $\mathbf{B}$, we applied NND and other two algorithms to infer the estimated $\hat{\mathbf{C}}$, $\hat{\mathbf{F}}$ and $\hat{\mathbf{B}} = \hat{\mathbf{C}}\hat{\mathbf{F}}$, and compared the accuracy between estimated and actual values using the following metrics. For $\mathbf{C}$, we used L_1 loss (Zhu et al., 2018):

$$L_1 \text{ loss}(\mathbf{C}) = \frac{\|\hat{\mathbf{C}} - \mathbf{C}\|_1}{\|\mathbf{C}\|_1}. \tag{24}$$

For $\mathbf{F}$ and $\mathbf{B}$, we used root mean square error (RMSE):

$$\text{RMSE}(\mathbf{F}) = \sqrt{\|\hat{\mathbf{F}} - \mathbf{F}\|_{\text{Fr}}^2}, \tag{25}$$

$$\text{RMSE}(\mathbf{B}) = \sqrt{\frac{\|\hat{\mathbf{B}} - \mathbf{B}\|_{\text{Fr}}^2}{\|\mathbf{B}\|_{\text{Fr}}^2}} \tag{26}$$

Different levels of noise s were added to test the robustness of models and the performance of different models under different conditions. We repeated all the experiments for 10 times to get the boxplot.

3.1.3. Competing Algorithms

There are two competing algorithms for the deconvolution problem. Geometric unmixing is an algorithm that borrows the intuition from computational geometry (Schwartz and Shackney, 2010), which first identifies the corners of a simplex containing all the mixture sample points, and then infers the fraction matrix.

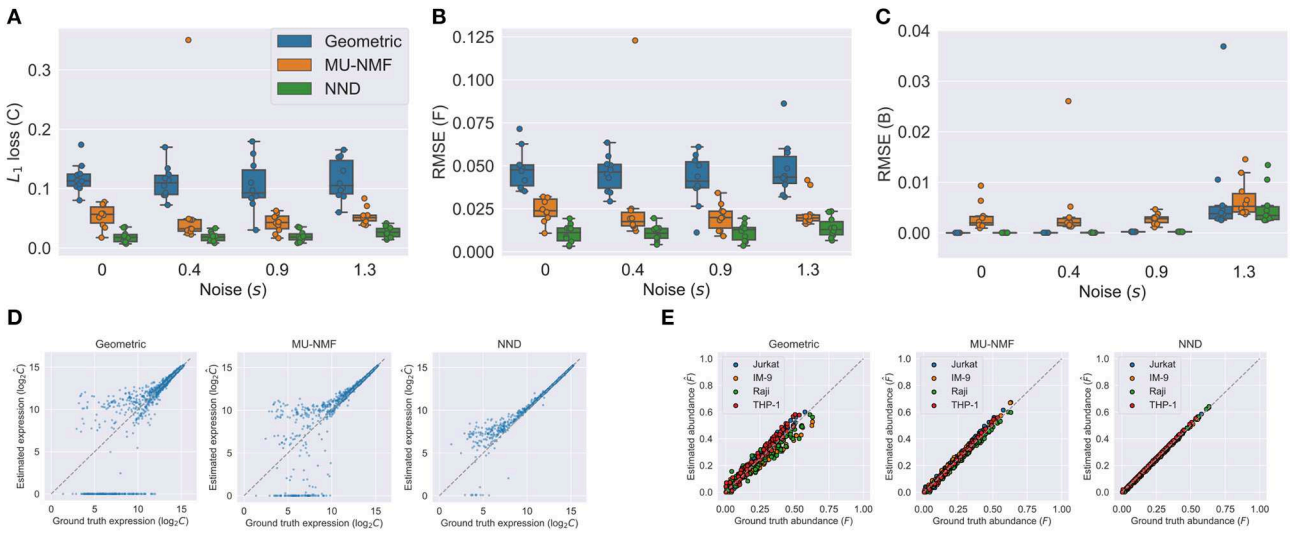

FIGURE 3 | Comparison of NND and other algorithms on the semi-simulated GSE11103 dataset shows the better accuracy and robustness of NND algorithm. **(A–C)** Accuracies in estimated **C**, **F**, and **B** with three different algorithms. We tested four noise levels and repeated the experiments for ten replicates. **(D)** Estimated $\hat{\mathbf{C}}$ and ground truth **C** using three different algorithms with noise level *s* of 1.3. **(E)** Estimated $\hat{\mathbf{F}}$ and ground truth **F** using three algorithms with noise level *s* of 1.3.

However, the algorithm does not directly optimize the problem (Equations 1–3). Another intuitive algorithm is based on the popular multiplicative update (MU) rule that solves general NMF problem (Lee and Seung, 2000): an additional update step of $\mathbf{F}_{lj} \leftarrow \frac{\mathbf{F}_{lj}}{\sum_{l'=1}^{k} \mathbf{F}_{l'j}}, j = 1, ..., n$ can be added to the loop. Although the original MU rule guarantees the non-increasing of the objective function, this additional update step can lead to an increasing objective and we need to stop the iteration once this happened. Since the two competing algorithms work on non-negative space, we adapted the NND by adding an element-wise absolute value operator after the **C** in the network (**Figure 2A**).

3.1.4. Superiority of NND

We show the results in **Figure 3**. **Figures 3A–C** show the accuracies of both **C**, **F**, and **B** using the three algorithms under various noise levels. One can easily see that NND achieves lower L_1 loss of **C**, RMSE of **F**, and RMSE of **B**. What is more, it is also much more robust than the geometric and MU-NMF algorithms, as there are fewer outliers that have huge errors. MU-NMF has a reasonable estimation accuracy of **C** and **F**. However, its overall fitting ability is limited due to its non-convergence-guaranteed MU optimization algorithm. We can also visualize the estimation accuracy by plotting the estimated values and ground truth values at a specific noise level, as is shown in **Figures 3D,E**. One can see the superiority of NND qualitatively over the other two algorithms in estimating expression profiles and fractions of individual pure clones.

3.2. Gene Modules/Pathways Provide an Effective Representation

Gene expressions of samples were mapped into the gene module and pathway space in order to reduce the noise of raw transcriptome data and reduce redundancy (section 2.3). We verified that the gene module/pathway representation is effective in the sense that it captures distinguishing features of primary/metastatic sites and individual samples well and is able to identify recurrently gained or lost pathways.

3.2.1. Feature Space of the Gene Module and Pathway Representation

As one can see in **Figure 4A**, the first principal component analysis (PCA) dimension of the gene module and pathway representation accounts for the difference between primary and metastatic samples, while the second and third PCA dimensions mainly capture variability between patients. This observation suggests the feasibility of using the gene module/pathway representation to distinguish recurrent features of metastatic progression across patients despite heterogeneity between patients. To make a direct comparison of the noise and redundancy between the gene module/pathway and raw gene expression representations, we applied hierarchical clustering to the 44 samples using Ward's minimum variance method (Ward, 1963). Two hierarchical trees were built based on the two different representations (**Figure 4B**). The gene module/pathway features more effectively separate the primary and metastatic samples into distinct clusters (**Figure 4B**, right panel) than do the raw gene expression values (**Figure 4B**, left panel). This is consistent with the PCA results that the largest mode of variance in the pathway representation distinguishes primary from metastatic samples. We do notice that in a few cases, matched primary and metastatic samples from the same patient are neighbors with pathway-based clustering. For example, 29P_Pitt:29M_Pitt and 51P_Pitt:51M_Pitt are grouped in the same clades using the pathway representation, showing that in a minority of cases, features of individual patients dominate

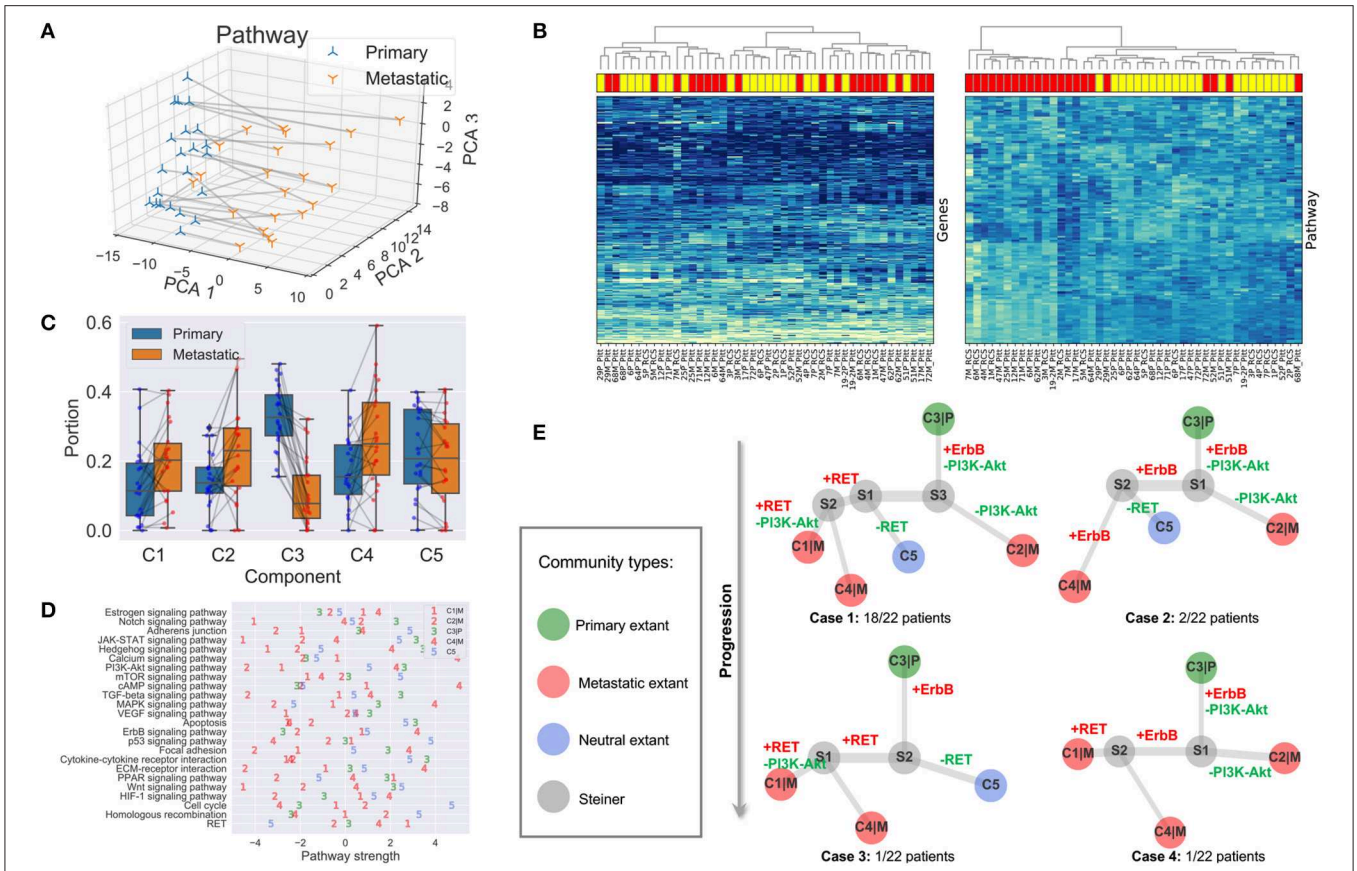

FIGURE 4 | Results and analysis. **(A)** First three gene module/pathway representation PCA dimensions of matched primary and metastatic samples. Matched samples are connected. **(B)** Hierarchical clustering of tumor samples based on raw gene expressions (left panel) and compressed gene module/pathway representation (right panel). Metastatic samples are shown in red rectangles and primary ones in yellow. **(C)** Portions and changes of the five communities in primary and metastatic sites. Each gray line connects the portions of a community in the primary site (blue node) and metastatic site (red node) from the same patient. **(D)** Pathway strengths across cell communities. **(E)** Phylogeny of cell subcommunities.

over primary vs. metastatic features. Following previous work (Park et al., 2009), we quantified the ability of the hierarchical tree to group the samples of the same labels using four metrics. (1) MSD: Mean square distance of edges that connect nodes of the same label (primary vs. metastatic). (2) z_{MSD}: The labels of all nodes were shuffled and the MSD is recalculated for 1,000 times to get the mean μ_{MSD} and standard deviation σ_{MSD}, which were used to get the z-score of the current assignment $z_{\text{MSD}} = (\text{MSD} - \mu_{\text{MSD}})/\sigma_{\text{MSD}}$. (3) rMSD: The ratio of MSD of edges that connect same label nodes and MSD of edges that connect distinct label nodes. (4) z_{rMSD}: as with MSD, a z-score of rMSD was calculated by shuffling labels for 1,000 times. Intuitively, the smaller values the MSD, z_{MSD}, rMSD, and z_{rMSD} are, the better is the feature representation at grouping same label samples together. The shortest paths and distances between all pairs of nodes were calculated using the Floyd-Warshall algorithm (Floyd, 1962; Warshall, 1962). All the edge lengths were considered as 1.0 to account for the different scales of pathway and gene representations. The pathway representation has significantly lower values for all four metrics (**Table 1**), indicating its strong grouping ability.

TABLE 1 | Quantitative performance of hierarchical clustering.

Feature representation	MSD	rMSD	z_{MSD}	z_{rMSD}
Gene expression	99.62	0.93	−2.60	−2.57
Gene module/pathway	86.23	0.66	−13.37	−11.42

3.2.2. Recurrently Perturbed Cancer Pathways

We next identified differentially expressed pathways in the primary and metastatic tumors using bulk data $\mathbf{B}_P \in \mathbb{R}^{24\times44}$, prior to deconvolving cellular subcommunities. We conducted the Student's t-test followed by FDR correction on each of the 24 pathways. Eleven pathways are significantly different between the two sites (FDR < 0.05; **Table 2**). The signaling pathways related to neurotransmitter and calcium homeostasis, including *cAMP* and *Calcium* (Hofer and Lefkimmiatis, 2007), are enriched in metastatic samples, which we can suggest may reflect stromal contamination by neural cells in the brain metastatic samples. We also observed recurrent gains in *ErbB* pathway, as indicated by the primary studies (Priedigkeit et al., 2017; Vareslija et al.,

TABLE 2 | Differentially expressed cancer pathways between primary and metastatic samples (FDR < 0.05).

Gain/Loss after metastasis	Differentially expressed pathways	FDR
Relative gain	cAMP signaling pathway	6.88e-03
Relative gain	ErbB signaling pathway	2.09e-02
Relative gain	Calcium signaling pathway	4.39e-02
Relative loss	Cytokine-cytokine receptor interaction	4.37e-06
Relative loss	Apoptosis	8.53e-04
Relative loss	JAK-STAT signaling pathway	8.53e-04
Relative loss	Wnt signaling pathway	3.97e-03
Relative loss	Hedgehog signaling pathway	4.50e-03
Relative loss	PI3K-Akt signaling pathway	1.35e-02
Relative loss	TGF-beta signaling pathway	4.56e-02
Relative loss	Notch signaling pathway	4.56e-02

2018). Three pathways related to immune activity are under-expressed in metastatic samples, including *Cytokine-cytokine receptor interaction* (Lee and Margolin, 2011), *JAK-STAT* (Lee and Margolin, 2011), and *Notch* (Aster et al., 2017), consistent with the previous inference of reduced immune cell expression in metastases in general and brain metastasis most prominently (Zhu et al., 2019). We can suggest that this result similarly may reflect expression changes in infiltrating immune cells, due to the immunologically privileged environment of the brain, rather than expression changes in tumor cell populations. Five other signaling pathways, including *Apoptosis* (Wong, 2011), *Wnt* (Zhan et al., 2016), *Hedgehog* (Gupta et al., 2010), *PI3K-Akt* (Brastianos et al., 2015), and *TGF-beta* (Massagué, 2008), show reduction in metastatic samples and in each case, their loss or dysregulation has been reported to promote the tumor growth and brain metastasis. Note that the primary references for these data define pathways using the co-expression pattern of genes (Priedigkeit et al., 2017; Vareslija et al., 2018), while our work uses external knowledge bases. Previous research also used somatic mutations or copy number variation to analyze perturbed genes (Brastianos et al., 2015; Priedigkeit et al., 2017), while we focus exclusively on the transcriptome. Despite large differences in data types and pathway definitions, our observations are consistent with the prior analysis, especially with respect to variation in the *HER2/ErbB2* and *PI3K-Akt* pathways.

3.3. Landscape of Deconvolved Cell Communities in Tumors

We unmixed the bulk data **B** into five components using NND (section 2.4). The deconvolution enables us to produce at least a coarse-grained landscape of major cell communities **C** and their distributions in primary and metastatic tumors **F**. The number of components ($k = 5$) was chosen through 20-fold cross-validation (section 2.4; **Figure 2B**). Although the true heterogeneity of the samples may be much larger, we fit k to provide a balance between excessively coarse-grained communities if k is too small vs. excessively high variance and thus unstable deconvolution if k is too large.

3.3.1. Community Distributions Across Samples F

The portions of the five components in all the 44 samples are represented as the mixture fraction matrix $\mathbf{F} \in \mathbb{R}^{5\times 44}$ (**Figure 4C**). A primary or metastatic community is one inferred to change proportions substantially (magnitude > 0.05) in the tumor samples after metastasis, or perhaps to be entirely novel to or extinct in the metastatic sample (denoted by a $|P$ or $|M$ suffix). Otherwise, the component is classified as a neutral community. Three components ($C1|M$, $C2|M$, $C4|M$) are classified as metastatic communities; one ($C3|P$) as primary; and one ($C5$) as neutral (**Figure 4C**). Some components may be missing in both samples of some patients, e.g., $C1|M$, $C2|M$, $C5|M$ are absent in two, one, and one patient. We note that these five communities represent rough consensus clusters of cell populations inferred to occur frequently, but not universally, among the samples. Based on this rule, we can define four basic cases of patients in total. Twelve subcases can be found using a more detailed classification method based on the existence of communities in both primary and metastatic samples (**Supplementary Material**).

3.3.2. Pathway Values of Communities C

We are especially interested in the pathway part $\mathbf{C}_P$ of the cell community inferences, since it serves as the marker and provides results easier to interpret. The pathway values of five subcommunities using $\mathbf{C}_P$ provides a much more fine-grained description of samples (**Figure 4D**), compared with that in section 3.2, which is only able to distinguish the differentially expressed pathways in bulk samples. As noted in section 2.4, it is likely that true cellular heterogeneity is greater than the methods are able to discriminate and that communities inferred by our model may each conflate one or more distinct cell types and clones. We observe that the metastatic community $C4|M$ most prominently contributes to the enrichment for functions related to neurotransmitter and ion transport, since its strongest pathways (*cAMP*, *Calcium*) are greatly enriched relative to those of the other four communities. We might interpret this community as reflecting at least in part stromal contamination from neural cells specific to the metastatic site. $C4|M$ also contributes most to the gains of *ErbB* in brain samples. The metastatic subcommunity $C1|M$ is probably most closely related to the loss of immune response in metastatic samples as it has the lowest pathway values of *Notch*, *JAK-STAT*, and *Cytokine-cytokine receptor interaction*. This component might thus in part reflect the effect of relatively greater immune infiltration in the primary vs. the metastatic site. $C1|M$ also has the lowest pathway values of *Apoptosis*, *Wnt*, and *Hedgehog*. The metastatic community $C2|M$ is most responsible for the loss of *PI3K-Akt* and *TGF-beta* pathways. We also note that although *RET* does not show up in the list of **Table 2**, it seems to be quite over-expressed in the metastatic communities $C1|M$ and $C4|M$ but not in the metastatic community $C2|M$.

3.4. Phylogenies of BrM Communities Reveal Common Order of Perturbed Pathways

We built phylogenies of cell communities and calculated the pathway representations of their Steiner nodes (section 2.5). The phylogenies' topologies provide a way to infer a likely evolutionary history of cancer cell communities and thus their constitutive cell types. At the same time, the perturbed pathways along their edges suggest the order of genomic alterations or changes in community composition.

3.4.1. Topologically Similar BrM Phylogenies

All five cell components do not appear in each BrM patient. We analyze the distribution of communities in each patient based on whether the community is inferred to be present in the patient (**Supplementary Material**). There are four different cases in general (**Figure 4E**). Case 1: all five communities are found in the patient (majority; 18/22 patients). Case 2: only $C1|M$ missing (minority; 2/22). Case 3: only $C2|M$ missing (minority; 1/22). Case 4: only $C5$ missing (minority; 1/22). Although not all communities exist in Cases 2–4, the topologies are similar to that of Case 1 and can be seen as special cases of Case 1, representing some inferred common mechanisms of progression across all the BrM patients.

3.4.2. Common Order of Altered Cancer Pathways

After inferring the pathway values for Steiner nodes, the most perturbed pathways can also be found by subtracting the pathway vectors of nodes that share an edge. We focus on the top five gained or lost pathways along the evolutionary trajectories and the changes of magnitude larger than 1.0 (**Supplementary Material**). We further examine those perturbed cancer pathways that were specifically proposed in the study that generated the data examined here, as well as others that are clinically actionable (Brastianos et al., 2015; Priedigkeit et al., 2017; Vareslija et al., 2018), i.e., *ErbB*, *PI3K-Akt*, and *RET* (**Figure 4E**). As one may see from Case 1, the primary community $C3|P$ first evolves to community $S3$ by gaining expressions in *ErbB* and losing functions in *PI3K-Akt*. Then, if it continues to lose *PI3K-Akt* activity, it will evolve into the metastatic community $C2|M$. If it gains in *RET* activity, it will instead evolve into metastatic communities $C1|M$ and $C4|M$. The perturbed pathways along the trajectories of Cases 2–4 are similar to those of Case 1, with minor differences. We therefore draw to the conclusion that the evolution of BrMs follows a specific and common order of pathway perturbations. Specifically, the gain of *ErbB* reproducibly happens before the loss of *PI3K-Akt* and the gain of *RET*. Different subsequently perturbed pathways lead to different metastatic tumor cell communities. These inferences are consistent with the hypothesis that at least some major changes in expression programs between primary and metastatic communities occur by selecting for heterogeneity present early in tumor development rather than solely deriving from novel functional changes immediately prior to or after metastasis.

4. DISCUSSION

Cancer metastasis is usually a precursor to mortality with no successful treatment options. Better understanding mechanisms of metastasis provides a potential pathway to identify new diagnostics or therapeutic targets that might catch metastasis before it ensues, treat it prophylactically, or provide more effective treatment options once it occurs. The present work developed a computational approach intended to better reconstruct mechanisms of functional adaption from multisite RNA-Seq data to help us understand at the level of cancer pathways the mechanisms by which progression frequently proceeds across a patient cohort. Our method compresses expression data into a gene module/pathway representation using external knowledge bases, deconvolves the bulk data into putative cell communities where each community contains a set of associated cell types or subclones, and builds evolutionary trees of inferred communities with the goal of reconstructing how these communities evolve, adapt, and reconfigure their compositions across metastatic progression. Results on semi-simulated data show the method to yield improved accuracy in mixture deconvolution relative to prior deconvolution algorithms. We applied the pipeline to matched transcriptome data from 22 BrM patients and found that although there are slight differences of tumor communities across the cohort, most patients share a similar mechanism of tumor evolution at the pathway level. Specifically, the methods infer a fairly conserved mechanism of early gain of *ErbB* prior to metastasis, followed post-metastasis gain of *RET* or loss of *PI3K-Akt* resulting in intertumor heterogeneity between samples. Our methods provide a novel way of viewing the development of BrM with implications for basic research into metastatic processes and potential translational applications in finding markers or drug targets of metastasis-producing clones prior to the metastatic transition.

The results suggest several possible avenues for future development. In part, they suggest a need for better separating phylogenetically-related mixture components (i.e., distinct tumor cell clones) from unrelated infiltrating cell types (e.g., healthy stroma from the primary or metastatic site or infiltrating immune cells). The methods are likely finding only a small fraction of the true clonal heterogeneity of the tumors and stroma, and might benefit from algorithms capable of better resolution or from integration of multi-omics data (e.g., RNA-Seq, DNA-Seq, methylation) that might have complementary value in finer discrimination of cell types. The present methods are also using only a limited form of temporal constraint in considering a two-stage progression process and without use of quantitative time measurements. Models might be extended in future work to consider true time-series data, such as is becoming available through "liquid biopsy" technologies. In addition, we know of no data with known ground truth that models the kind of progression process studied here nor of other tools designed for modeling similar progression processes from expression data, leaving us reliant on validating based on consistency with prior research on brain metastasis (Brastianos et al., 2015; Priedigkeit et al., 2017; Vareslija et al., 2018). Future

work might compare to prior approaches for reconstruction of clonal evolution from expression data more generically (Desper et al., 2004; Riester et al., 2010; Schwartz and Shackney, 2010) and seek replication on additional real or simulated expression data or artificial mixtures of different cell types (Qiu et al., 2011) designed to mimic metastasis-like progression. The general approach might also have broader application than studying metastasis, for example in reconstructing mechanisms of other progression processes, such as pre-cancerous to cancerous, as well as to other tumor types or independent data sets. Finally, much remains to be done to exploit the translational potential of the method in better identifying diagnostic signatures and therapeutic targets, and what type of effective and safe clinical strategies can be taken to prevent metastasis at an early stage.

DATA AVAILABILITY STATEMENT

The breast cancer brain metastases dataset analyzed for this study can be found on Github: https://github.com/lizhu06/TILsComparison_PBTvsMET. The simulated dataset generated, and data for analysis for this study can be found on Github: https://github.com/CMUSchwartzLab/NND.

AUTHOR CONTRIBUTIONS

RS, JM, and AL contributed to the conceptualization. RS, JM, AL, YT, and HL contributed to the methodology. YT contributed to the software. RS, JM, AL, and YT contributed to the formal analysis, writing-review and editing, and funding acquisition. AL contributed to the resources. YT and HL contributed to the writing of the original draft. RS supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by a grant from the Mario Lemieux Foundation, U.S. N.I.H. awards R21CA216452 and R01HG010589, Pennsylvania Department of Health award 4100070287, Breast Cancer Alliance, Susan G. Komen for the Cure, and by a fellowship to YT from the Center for Machine Learning and Healthcare at Carnegie Mellon University. It was also supported in part by the AWS Machine Learning Research Awards. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

ACKNOWLEDGMENTS

An earlier version of this work was published in the International Symposium on Mathematical and Computational Oncology 2019 (Tao et al., 2019a). We would like to thank to the reviewers for their helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.01055/full#supplementary-material

REFERENCES

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* 4:e6098. doi: 10.1371/journal.pone.0006098

Amaratunga, D., and Cabrera, J. (2001). Analysis of data from viral DNA microchips. *J. Am. Stat. Assoc.* 96, 1161–1170. doi: 10.1198/016214501753381814

Aster, J. C., Pear, W. S., and Blacklow, S. C. (2017). The varied roles of Notch in cancer. *Annu. Rev. Pathol.* 12, 245–275. doi: 10.1146/annurev-pathol-052016-100127

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bell, R. M., and Koren, Y. (2007). "Scalable collaborative filtering with jointly derived neighborhood interpolation weights," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (Omaha, NE), 43–52. doi: 10.1109/ICDM.2007.90

Brastianos, P. K., Carter, S. L., Santagata, S., Cahill, D. P., Taylor-Weiner, A., Jones, R. T., et al. (2015). Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* 5, 1164–1177. doi: 10.1158/2159-8290.CD-15-0369

Chaffer, C. L., and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science* 331, 1559–1564. doi: 10.1126/science.1203543

Chambers, A. F., Groom, A. C., and MacDonald, I. C. (2002). Dissemination and growth of cancer cells in metastatic sites. *Nat. Rev. Cancer* 2, 563–572. doi: 10.1038/nrc865

de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., et al. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* 346, 251–256. doi: 10.1126/science.1253462

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., et al. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.* 14, 5158–5165. doi: 10.1158/1078-0432.CCR-07-4756

Desper, R., Khan, J., and Schäffer, A. A. (2004). Tumor classification using phylogenetic methods on expression data. *J. Theor. Biol.* 228, 477–496. doi: 10.1016/j.jtbi.2004.02.021

Ding, L., Raphael, B. J., Chen, F., and Wendl, M. C. (2013). Advances for studying clonal evolution in cancer. *Cancer Lett.* 340, 212–219. doi: 10.1016/j.canlet.2012.12.028

Eaton, J., Wang, J., and Schwartz, R. (2018). Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics* 34, i357–i365. doi: 10.1093/bioinformatics/bty270

Elyanow, R., Dumitrascu, B., Engelhardt, B. E., and Raphael, B. J. (2020). netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res.* 30, 195–204. doi: 10.1101/gr.251603.119

Floyd, R. W. (1962). Algorithm 97: shortest path. *Commun. ACM* 5, 344–348. doi: 10.1145/367766.368166

Funk, S. (2006). *Netflix Update: Try This at Home*. Technical report.

Greaves, M., and Maley, C. C. (2012). Clonal evolution in cancer. *Nature* 481, 306–313. doi: 10.1038/nature10762

Guan, X. (2015). Cancer metastases: challenges and opportunities. *Acta Pharma. Sin. B* 5, 402–418. doi: 10.1016/j.apsb.2015.07.005

Gupta, S., Takebe, N., and Lorusso, P. (2010). Targeting the Hedgehog pathway in cancer. *Ther. Adv. Med. Oncol.* 2, 237–250. doi: 10.1177/1758834010366430

Hofer, A. M., and Lefkimmiatis, K. (2007). Extracellular calcium and cAMP: second messengers as "Third Messengers?" *Physiology* 22, 320–327. doi: 10.1152/physiol.00019.2007

Hosack, D. A., Dennis Jr, G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4:R70. doi: 10.1186/gb-2003-4-10-r70

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kingma, D., and Ba, J. (2014). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations* (San Diego, CA).

Körber, V., Yang, J., Barah, P., Wu, Y., Stichel, D., Gu, Z., et al. (2019). Evolutionary trajectories of IDHWT glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell* 35, 692–704.e12. doi: 10.1016/j.ccell.2019.02.007

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* 42, 30–37. doi: 10.1109/MC.2009.263

Lee, D. D., and Seung, H. S. (2000). "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00* (Cambridge, MA: MIT Press), 535–541.

Lee, S., and Margolin, K. (2011). Cytokines in cancer immunotherapy. *Cancers* 3, 3856–3893. doi: 10.3390/cancers3043856

Lei, H., Gertz, E. M., Schäffer, A. A., Fu, X., Tao, Y., Heselmeyer-Haddad, K., et al. (2020). Tumor heterogeneity assessed by sequencing and fluorescence *in situ* hybridization (fish) data. *bioRxiv*. doi: 10.1101/2020.02.29.970392

Lei, H., Lyu, B., Gertz, E. M., Schäffer, A. A., Shi, X., Wu, K., et al. (2019). "Tumor copy number deconvolution integrating bulk and single-cell sequencing data," in *Research in Computational Molecular Biology*, ed L. J. Cowen (Cham: Springer International Publishing), 174–189. doi: 10.1007/978-3-030-17083-7_11

Lin, N. U., Bellon, J. R., and Winer, E. P. (2004). CNS metastases in breast cancer. *J. Clin. Oncol.* 22, 3608–3617. doi: 10.1200/JCO.2004.01.175

Lu, C. L., Tang, C. Y., and Lee, R. C.-T. (2003). The full Steiner tree problem. *Theor. Comput. Sci.* 306, 55–67. doi: 10.1016/S0304-3975(03)00209-3

Massagué, J. (2008). TGFβ in cancer. *Cell* 134, 215–230. doi: 10.1016/j.cell.2008.07.001

Nei, M., and Saitou, N. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Park, Y., Shackney, S., and Schwartz, R. (2009). Network-based inference of cancer progression from microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6, 200–212. doi: 10.1109/TCBB.2008.126

Priedigkeit, N., Hartmaier, R. J., Chen, Y., Vareslija, D., Basudan, A., Watters, R. J., et al. (2017). Intrinsic subtype switching and acquired ERBB2/HER2 amplifications and mutations in breast cancer brain metastases. *JAMA Oncol.* 3, 666–671. doi: 10.1001/jamaoncol.2016.5630

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., et al. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* 29, 886–891. doi: 10.1038/nbt.1991

Riester, M., Stephan-Otto Attolini, C., Downey, R. J., Singer, S., and Michor, F. (2010). A differentiation-based phylogeny of cancer subtypes. *PLoS Comput. Biol.* 6:e1000777. doi: 10.1371/journal.pcbi.1000777

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323:533. doi: 10.1038/323533a0

Schwartz, R., and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* 18:213. doi: 10.1038/nrg.2016.170

Schwartz, R., and Shackney, S. E. (2010). Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics* 11:42. doi: 10.1186/1471-2105-11-42

Tao, Y., Cai, C., Cohen, W. W., and Lu, X. (2020). "From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer," in *Pacific Symposium on Biocomputing* (Hawaii).

Tao, Y., Lei, H., Lee, A. V., Ma, J., and Schwartz, R. (2019a). "Phylogenies derived from matched transcriptome reveal the evolution of cell populations and temporal order of perturbed pathways in breast cancer brain metastases," in *Mathematical and Computational Oncology*, eds G. Bebis, T. Benos, K. Chen, K. Jahn, and E. Lima (Cham: Springer International Publishing), 3–28. doi: 10.1007/978-3-030-35210-3_1

Tao, Y., Rajaraman, A., Cui, X., Cui, Z., Eaton, J., Kim, H., et al. (2019b). Improving personalized prediction of cancer prognoses with clonal evolution models. *bioRxiv*. doi: 10.1101/761510

Vareslija, D., Priedigkeit, N., Fagan, A., Purcell, S., Cosgrove, N., O'Halloran, P. J., et al. (2018). Transcriptome characterization of matched primary breast and brain metastatic tumors to detect novel actionable targets. *J. Natl. Cancer Inst.* 111, 388–398. doi: 10.1093/jnci/djy110

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Warshall, S. (1962). A theorem on boolean matrices. *J. ACM* 9, 11–12. doi: 10.1145/321105.321107

Witzel, I., Oliveira-Ferrer, L., Pantel, K., Muller, V., and Wikman, H. (2016). Breast cancer brain metastases: biology and new clinical perspectives. *Breast Cancer Res.* 18:8. doi: 10.1186/s13058-015-0665-1

Wong, R. S. Y. (2011). Apoptosis in cancer: from pathogenesis to treatment. *J. Exp. Clin. Cancer Res.* 30:87. doi: 10.1186/1756-9966-30-87

Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., et al. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* 10:e1003703. doi: 10.1371/journal.pcbi.1003703

Zhan, T., Rindtorff, N., and Boutros, M. (2016). Wnt signaling in cancer. *Oncogene* 36:1461. doi: 10.1038/onc.2016.304

Zhu, L., Lei, J., Devlin, B., and Roeder, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* 12, 609–632. doi: 10.1214/17-AOAS1110

Zhu, L., Narloch, J. L., Onkar, S., Joy, M., Broadwater, G., Luedke, C., et al. (2019). Metastatic breast cancers have reduced immune cell recruitment but harbor increased macrophages relative to their matched primary tumors. *J. Immunother. Cancer* 7:265. doi: 10.1186/s40425-019-0755-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 08 September 2020
doi: 10.3389/fonc.2020.553342

NFATc Acts as a Non-Canonical Phenotypic Stability Factor for a Hybrid Epithelial/Mesenchymal Phenotype

Ayalur Raghu Subbalakshmi[1], Deepali Kundnani[2], Kuheli Biswas[3], Anandamohan Ghosh[3], Samir M. Hanash[2], Satyendra C. Tripathi[2,4] and Mohit Kumar Jolly[1]**

[1] Centre for BioSystems Science and Engineering, Indian Institute of Science, Bengaluru, India, [2] Department of Clinical Cancer Prevention, UT MD Anderson Cancer Center, Houston, TX, United States, [3] Department of Physical Sciences, Indian Institute of Science Education and Research, Kolkata, India, [4] Department of Biochemistry, All India Institute of Medical Sciences, Nagpur, India

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Jean Clairambault,
Institut National de Recherche en Informatique et en Automatique (INRIA), France
Gabor Balazsi,
Stony Brook University, United States

***Correspondence:**
Satyendra C. Tripathi
sctripathi@aiimsnagpur.edu.in
Mohit Kumar Jolly
mkjolly@iisc.ac.in;
mkjolly.15@gmail.com

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 18 April 2020
Accepted: 13 August 2020
Published: 08 September 2020

Citation:
Subbalakshmi AR, Kundnani D, Biswas K, Ghosh A, Hanash SM, Tripathi SC and Jolly MK (2020) NFATc Acts as a Non-Canonical Phenotypic Stability Factor for a Hybrid Epithelial/Mesenchymal Phenotype. Front. Oncol. 10:553342. doi: 10.3389/fonc.2020.553342

Metastasis remains the cause of over 90% of cancer-related deaths. Cells undergoing metastasis use phenotypic plasticity to adapt to their changing environmental conditions and avoid therapy and immune response. Reversible transitions between epithelial and mesenchymal phenotypes – epithelial–mesenchymal transition (EMT) and its reverse mesenchymal–epithelial transition (MET) – form a key axis of phenotypic plasticity during metastasis and therapy resistance. Recent studies have shown that the cells undergoing EMT/MET can attain one or more hybrid epithelial/mesenchymal (E/M) phenotypes, the process of which is termed as partial EMT/MET. Cells in hybrid E/M phenotype(s) can be more aggressive than those in either epithelial or mesenchymal state. Thus, it is crucial to identify the factors and regulatory networks enabling such hybrid E/M phenotypes. Here, employing an integrated computational-experimental approach, we show that the transcription factor nuclear factor of activated T-cell (NFATc) can inhibit the process of complete EMT, thus stabilizing the hybrid E/M phenotype. It increases the range of parameters enabling the existence of a hybrid E/M phenotype, thus behaving as a phenotypic stability factor (PSF). However, unlike previously identified PSFs, it does not increase the mean residence time of the cells in hybrid E/M phenotypes, as shown by stochastic simulations; rather it enables the co-existence of epithelial, mesenchymal and hybrid E/M phenotypes and transitions among them. Clinical data suggests the effect of NFATc on patient survival in a tissue-specific or context-dependent manner. Together, our results indicate that NFATc behaves as a non-canonical PSF for a hybrid E/M phenotype.

Keywords: hybrid epithelial/mesenchymal, NFATc, cancer systems biology, epithelial–mesenchymal transition, mathematical modeling, phenotypic stability factor

INTRODUCTION

Metastasis remains clinically insuperable and causes over 90% of cancer related deaths (1). A hallmark of metastasizing cells is phenotypic plasticity, which empowers them to adapt to their ever-changing microenvironment, while evading therapy and immune response (2). Cells displaying phenotypic plasticity can have profound consequences: an identical genetic background can give rise to varying phenotypes under different environmental conditions, enabling non-genetic heterogeneity (3, 4), due to stochasticity in cell-fate decision making (5). A crucial axis of phenotypic plasticity during metastasis is epithelial-mesenchymal plasticity, which allows bidirectional switching of cells among an epithelial phenotype, a mesenchymal phenotype, and one or more hybrid epithelial/mesenchymal (E/M) phenotypes (6). These hybrid E/M cells can be more metastatic than cells in epithelial or mesenchymal states (7, 8) and can exhibit collective cell migration as clusters of circulating tumor cells (CTCs) (9–11) – the major drivers of metastasis (12). Thus, understanding the molecular mechanisms enabling one or more hybrid E/M phenotype(s) is key to decoding and eventually restricting metastasis.

Epithelial–mesenchymal transition (EMT) is influenced by various pathways such as transforming growth factor β (TGF-β), Wnt–β-catenin, bone morphogenetic protein (BMP), Notch, Hedgehog, and receptor tyrosine kinases (13). These EMT signals alter the levels of one or more EMT-inducing transcription factors (EMT-TFs) such as ZEB and SNAIL which can directly repress various epithelial molecules such as E-cadherin and/or induce the expression of various mesenchymal ones (6). ZEB and SNAIL form mutually inhibitory feedback loops with two microRNA families miR-200 and miR-34 where the transcription factors and the micro-RNAs mutually inhibit each other (14–17). Overexpression of ZEB promotes EMT and silences the micro-RNAs which act as a safeguard for maintaining an epithelial phenotype (14). Recent studies have indicated the involvement of phenotypic stability factors (PSF) such as GRHL2, OVOL2, NUMB, and NRF2 that can maintain the cells in a hybrid E/M phenotype(s) and prevent the cells from undergoing a complete EMT (18–23). Knockdown of these PSFs usually drove hybrid E/M cells toward a completely mesenchymal phenotype, as observed in H1975 non-small cell lung cancer (NSCLC) cells which can maintain a hybrid E/M phenotype stably over multiple passages *in vitro* (21). Higher levels of these PSFs also increased the mean residence times (MRTs) of cells in a hybrid E/M phenotype (24) and associated with poor patient survival, thus highlighting the clinical significance of hybrid E/M phenotypes (25).

Here, we investigate the role of the nuclear factor of activated T-cell (NFATc) in mediating EMT. NFATc is a family of five transcription factors (NFAT1–5), four of which (NFATc1–4) are regulated by calcium Ca^{2+} signaling (26). Initially identified as functionally important for T lymphocytes, the NFAT family regulates cell cycle progression, gene expression and apoptosis (26). Abnormalities in NFATc signaling have been reported in many carcinomas as well as lymphoma and leukemia (27). Recent evidence has suggested the interconnections of NFATc with EMT circuitry. On one hand, overexpression of NFATc increased the levels of TWIST, ZEB1, SNAI1; its downregulation decreased the levels of these EMT-TFs as well as mesenchymal markers such as N-cadherin and Vimentin (28). On the other hand, NFATc transcriptional activity was also shown to be crucial for maintaining E-cadherin levels (29), which can inhibit ZEB1 indirectly through controlling the membranous localization of β-catenin (30, 31) and restrict EMT. Moreover, NFATc can activate SOX2 (28, 32) which can upregulate levels of miR-200 (33); overexpression of miR-200 can drive mesenchymal–epithelial transition (MET) (34). These opposing interactions of NFATc with EMT circuitry lead to the question of whether NFATc promotes EMT or inhibits it. Here, we have developed a mechanism-based mathematical model that captures the interconnections between NFATc signaling and EMT circuitry. Our analysis predicts that NFATc can stabilize a hybrid E/M phenotype, facilitating cellular plasticity. Knockdown of NFATc in H1975 cells pushed the hybrid E/M cells into a mesenchymal state, validating our prediction that NFATc can function as a PSF.

RESULTS

NFATc Inhibits the Progression of Complete EMT

To determine the role of NFATc in EMT, we first investigated the dynamics of crosstalk between NFATc and an EMT regulatory circuit (shown in the dotted rectangle) that includes miR-200, ZEB, and SNAIL (**Figure 1A**). This crosstalk was captured through a set of coupled ordinary differential equations (ODEs).

First, we examined the temporal dynamics of a cell in response to SNAIL levels. SNAIL represents the effect of an exogenous EMT-inducing signal such as TGF-β signaling. In the absence of NFATc, a cell that started in an epithelial state (high miR-200, low ZEB) first transitioned to a hybrid E/M phenotype and later to a mesenchymal state (low miR-200, high ZEB). The presence of NFATc, however, delayed this transition to a mesenchymal state (**Figure 1B** and **Supplementary Figure S1A**). Interestingly, the steady state value of ZEB mRNA levels was higher in case of NFATc-EMT coupled network as compared to the control case (circuit bounded by the dotted rectangle in **Figure 1A**); this difference can be ascribed to the activation of ZEB by NFATc (**Figure 1B**). Indeed, strengthening the activation of miR-200 by NFATc and/or weakening the activation of ZEB by NFATc (represented by dotted horizontal lines) prevented upregulation of ZEB levels and consequent attainment of the mesenchymal phenotype (**Supplementary Figures S1B,C**).

Next, we calculated a bifurcation diagram of cellular EMT phenotypes in response to increasing levels of SNAIL, an external EMT-inducing signal (**Figure 1B**). We observe that the system switches from epithelial to hybrid E/M and then to mesenchymal state with increasing SNAIL signal, as indicated by increasing values of ZEB mRNA and decreasing values of miR-200 (**Figure 1C** and **Supplementary Figure S1D**). Next, we compared the bifurcation diagram of the NFATc coupled network with that of the control case (i.e., without NFATc –the circuit bounded by the dotted rectangles in **Figure 1A**) to determine the

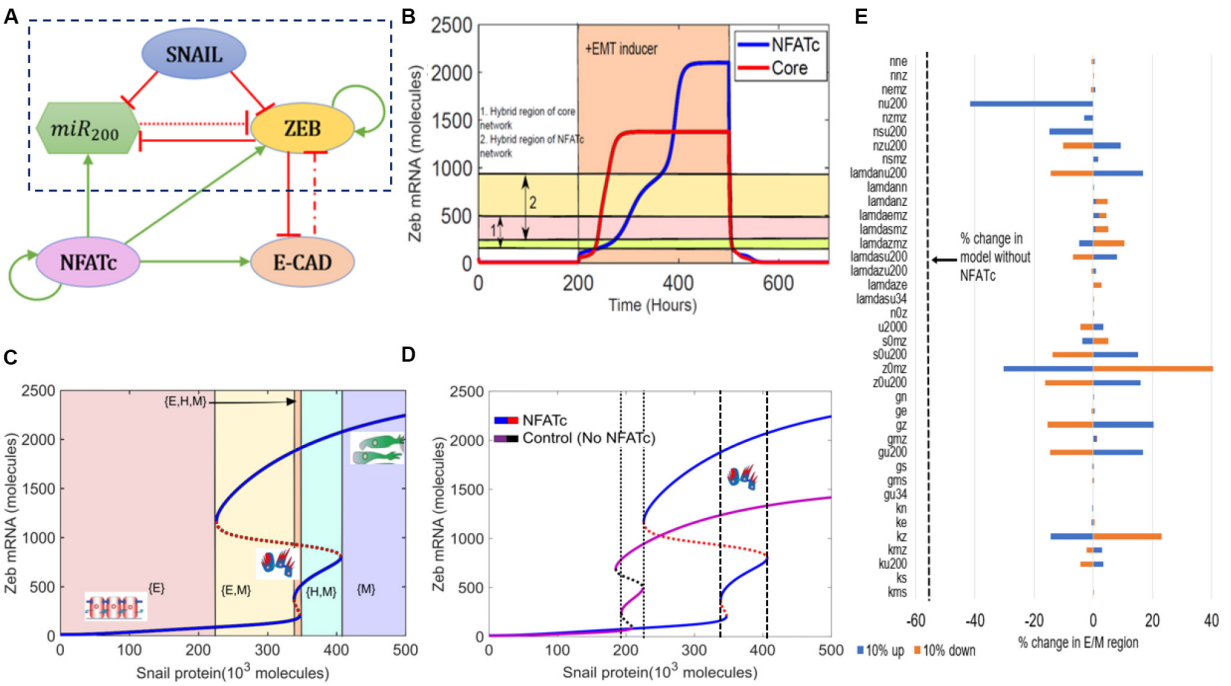

FIGURE 1 | NFATc inhibits the progression of complete EMT. **(A)** Schematic representation of the EMT network coupled with NFATc. Red bars denote inhibition, green arrows indicate activation. Solid arrows represent transcriptional regulation, dash-dotted line represents indirect regulation and dotted line represent micro-RNA mediated regulation. **(B)** Temporal dynamics of ZEB mRNA levels in a cell starting in an epithelial phenotype, when exposed to a high level of S = 330,000 molecules (orange-shaded region) for circuit shown in **(A)** (NFATc) and that shown in the dotted rectangle in **(A)** (core). Horizontal shaded regions denote the levels of ZEB mRNA corresponding to hybrid E/M region of both NFATc (cream and pink shaded region) and the core (yellow and pink shaded region) networks. **(C)** Bifurcation diagram of ZEB mRNA levels as driven by SNAIL signal for the circuit shown in **(A)**. Curves denote the value of ZEB mRNA upon equilibration, where continuous curves stand for stable steady states and dotted lines represent unstable solutions. Different colored shaded regions show the existence/co-existence of different phenotypes; cartoons depict epithelial (E), hybrid E/M (H), and mesenchymal (M) states. **(D)** Bifurcation diagrams indicating the ZEB mRNA levels for increasing SNAIL levels for the coupled EMT-NFATc circuit (solid blue and dotted red curve) and the core EMT circuit (solid purple and dotted black curve). **(E)** Sensitivity analysis indicating percent change in the interval of SNAIL levels for stable hybrid E/M region, when corresponding parameter values are varied by ±10%. The black dotted line indicates the percent change in the stable hybrid region in the absence of NFATc (core network) when compared to the coupled network with NFATc.

changes in the system behavior conferred by NFATc (**Figure 1D**). In the presence of NFATc, a higher value of SNAIL, i.e., a stronger external signal, was required for the cells to exit the epithelial phenotype. Moreover, in the presence of NFATc, the cell maintained a hybrid E/M phenotype over a broader range of SNAIL levels (compare the range of SNAIL levels bounded by dotted lines vs. dashed lines in **Figure 1D**), thus requiring a much stronger stimulus to undergo a complete EMT. Stochastic simulations revealed possible cellular transitions among different phenotypes, depending on the level of SNAIL. At lower SNAIL levels, cells could possibly directly transition to a mesenchymal state from the epithelial state (**Supplementary Figure S2A**); however, at intermediate levels, we saw the emergence of the hybrid E/M state (**Supplementary Figure S2B**). At higher levels of SNAIL, the epithelial state disappears and cells can transition between the hybrid E/M and mesenchymal states (**Supplementary Figure S2C**).

Finally, to ascertain the robustness of the effect of NFATc in associating a larger range of SNAIL values for the existence of hybrid E/M phenotype, a sensitivity analysis was performed where each parameter of the model was varied – one at a time – by ±10%, and the corresponding change in the range of values of SNAIL enabling a hybrid E/M phenotype (i.e., the interval of x-axis between dashed lines) was measured. For most of the model parameters, the relative change in this range of values was quite small, suggesting the robustness of the model predictions (**Figure 1E**). A change in few selected parameters such as the interaction between NFATc and miR200, and self-activation of ZEB exhibited stronger sensitivity; nonetheless, even in these few cases, the decrease in range of SNAIL levels enabling a hybrid E/M phenotype is smaller when compared to the case in absence of NFATc (dotted line in **Figure 1E**). Put together, these observations suggest that NFATc may inhibit the progression to a complete EMT and can behave as a "phenotypic stability factor" for hybrid E/M phenotype.

Knockdown of NFATc in H1975 Cells Promotes Complete EMT

To validate our model prediction that NFATc functions as a PSF for the hybrid E/M phenotype, we knocked down NFATc1 using siRNAs in NSCLC H1975 cells with a stable hybrid E/M phenotype.

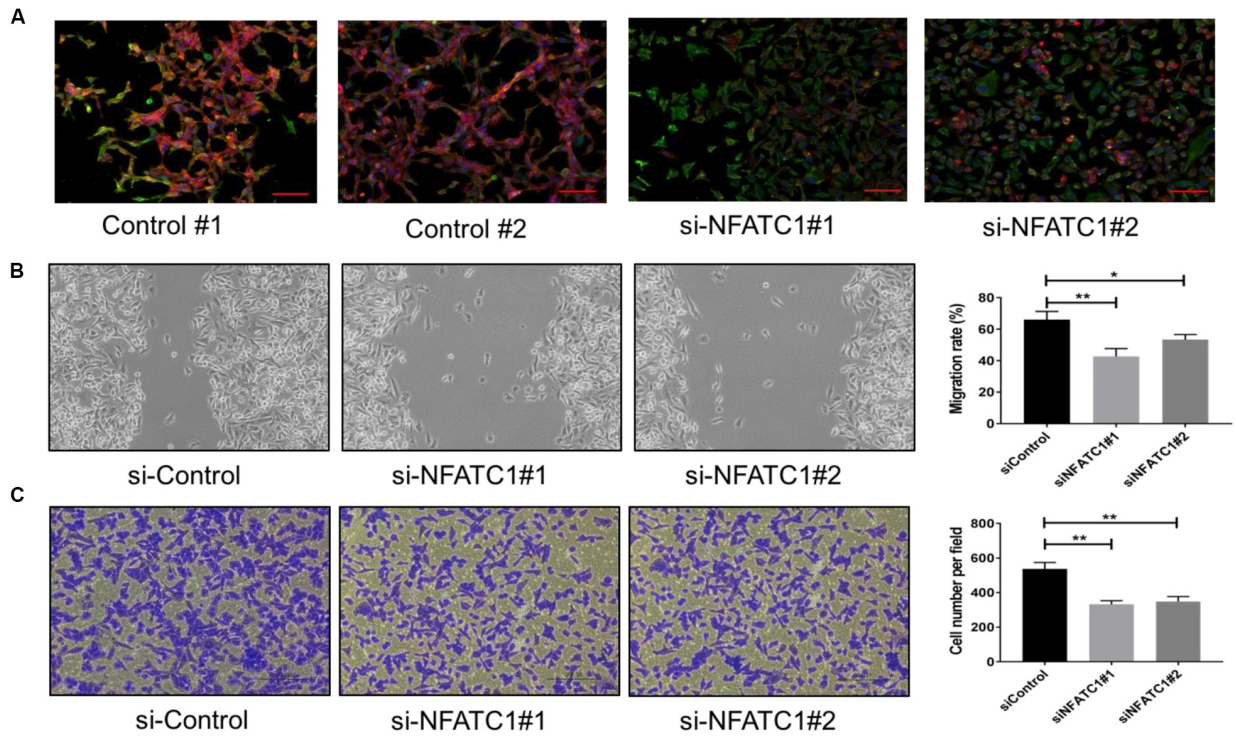

FIGURE 2 | NFATc knockdown in H1975 cells promotes progression toward complete EMT. **(A)** Expression of CDH1 (E-cadherin, Red) and VIM (Vimentin, Green) examined by immuno-fluorescence staining in H1975 cells for control and NFATc1 knockdown case. Scale bar 100 μm. **(B)** Scratch assay for control H1975 cells and those treated with siRNAs against NFATc. Magnification: 100× (quantification in last column). **(C)** Same as panel **(B)** but for trans-well migration assay. $^{*}p < 0.05$, $^{**}p < 0.005$ using Student's *t*-test; $n = 3$.

Individual H1975 cells can co-express E-cadherin and Vimentin (21). We observed that cells treated with NFATc1 siRNAs mostly lost E-cadherin staining (**Figure 2A**). NFATc knockdown decreased the E-cadherin levels and increased the levels of ZEB, SNAIL, and Vimentin, both at protein and mRNA levels (**Supplementary Figures S3A–C**). Thus, NFATc1 knockdown can push stable hybrid E/M cells into a mesenchymal state. Additionally, NFATc1 knockdown reduced the migration potential of H1975 cells as observed in scratch and trans-well migration assays. These findings indicate that the hybrid E/M cells may exhibit greater migratory and invasive potential when compared to mesenchymal cells (**Figures 2B,C**), reminiscent of our previous observations comparing hybrid E/M cells with mesenchymal ones (18). Overall, these experimental results provide a proof-of-principle validation of our model predictions that NFATc can stabilize a hybrid E/M phenotype.

NFATc Does Not Increase the Mean Residence Time of the Hybrid E/M Phenotype

In addition to extending the range of SNAIL levels enabling a hybrid E/M phenotype, the previously identified PSFs – GRHL2, OVOL1/2, and ΔNP63α – had another trait: their presence increased the mean residence time (MRT) of cells in hybrid E/M phenotype. MRT is the average time spent by the cells in a particular phenotype (basin of attraction) – E, M, and hybrid E/M – calculated via stochastic simulations (24). Thus, the phenotype with a larger MRT implies a relatively higher stability of the same, as compared to other co-existing phenotypes/states. Hence, beyond enabling a larger range of values of SNAIL (or any other EMT inducing signal) for the existence of a hybrid E/M phenotype (as shown via bifurcation diagrams), increased MRT can be considered as another hallmark trait of a PSF. We next investigated whether NFATc increased the MRT of cells in a hybrid E/M phenotype.

Even though NFATc extended the range of SNAIL values enabling a hybrid E/M state, the hybrid E/M state always co-existed with epithelial and/or mesenchymal states ({E, H, M} and {H, M} phases in **Figure 1B**); no monostable regime ({H}) for a hybrid E/M state was seen in the case of NFATc, as observed with GRHL2, OVOL1/2, ΔNP63α, NUMB, and NRF2 (18–22, 35). Similarly, compared to the other PSFs, the presence of NFATc does not increase the absolute value of MRT for a hybrid E/M phenotype as compared to the case without NFATc. In the case of control circuit, the MRT of the epithelial state is higher than that of the mesenchymal state in the {E, M} bi-stable phase. In the {H, M} phase, the MRT of the mesenchymal state dominates that of the hybrid E/M state as SNAIL values are increased (**Figure 3A**). Similar trends are seen in the case of NFATc-EMT coupled network; however, in the case of {E, H, M} phase in presence of NFATc, the MRT of hybrid E/M state is not higher as compared to that of epithelial or mesenchymal states (compare the red curve

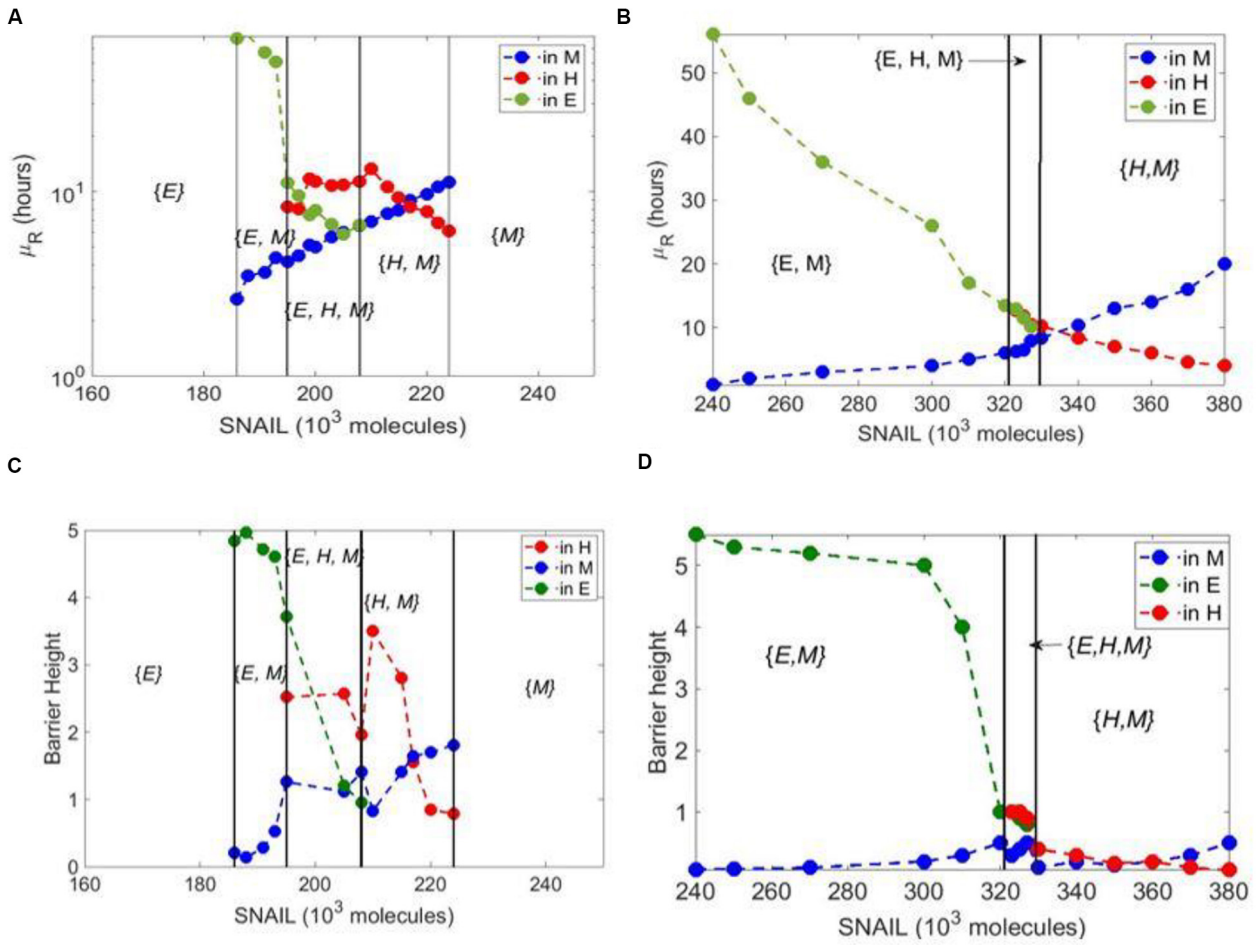

FIGURE 3 | Mean residence time analysis for core EMT circuit vs. coupled EMT-NFATc circuit. **(A)** Variation of mean residence time (μ_R) with varying SNAIL levels for the core circuit. **(B)** Same as panel **(A)** but for coupled EMT-NFATc circuit. **(C)** Barrier height for core EMT circuit at variable SNAIL levels. **(D)** Same as panel **(C)** but for coupled EMT-NFATc circuit.

in **Figure 3B** vs. that in **Figure 3A**). This trend is also seen in the barrier height calculated from the potential difference between the local minimum and saddle points corresponding to these states (**Figures 3C,D**).

We also plotted the potential landscapes for the NFATc-EMT coupled network at varying SNAIL levels (**Supplementary Figure S4**), which were consistent with the trend of barrier heights seen; for instance, at $S = 323 \times 10^3$ or $S = 330 \times 10^3$ molecules, the barrier height of hybrid E/M state was more than that of mesenchymal state, but at $S = 380 \times 10^3$, that of mesenchymal state was higher. Put together, these results suggest NFATc does not increase the MRT of hybrid E/M state.

RACIPE Analysis of NFATc Network Reveals Its Non-canonical Behavior as a PSF

To analyze the underlying design principles of the NFATc-EMT coupled network, we employed a recently developed computational method – random circuit perturbation (RACIPE) (36). RACIPE takes as input the topology of a regulatory network and generates an ensemble of mathematical models corresponding to the network topology, each with a randomly chosen set of kinetic parameters. Then, for each mathematical model, various possible steady states (phenotypes) are identified. Finally, statistical tools are used to identify the robust dynamical properties emerging from the network topology. Here, each mathematical model is a set of five coupled ODEs, where each ODE tracks the temporal dynamics of the five species constituting the regulatory network (SNAIL, ZEB, miR200, E-CAD, and NFATc).

Among the 10,000 parameter sets generated via RACIPE, we found cases where the network topology can give rise to the existence of phases with one steady state (mono-stable) or more – two (bi-stable), and three (tri-stable) steady states (**Figure 4A**). We performed RACIPE on the core EMT network (miR-200/ZEB/SNAIL), its coupling to other PSFs (OVOL, GRHL2, OCT4, NRF2), and the coupled EMT-NFATc network. For a given parameter set, one or more steady states were obtained depending on the initial conditions chosen; each steady state solution was

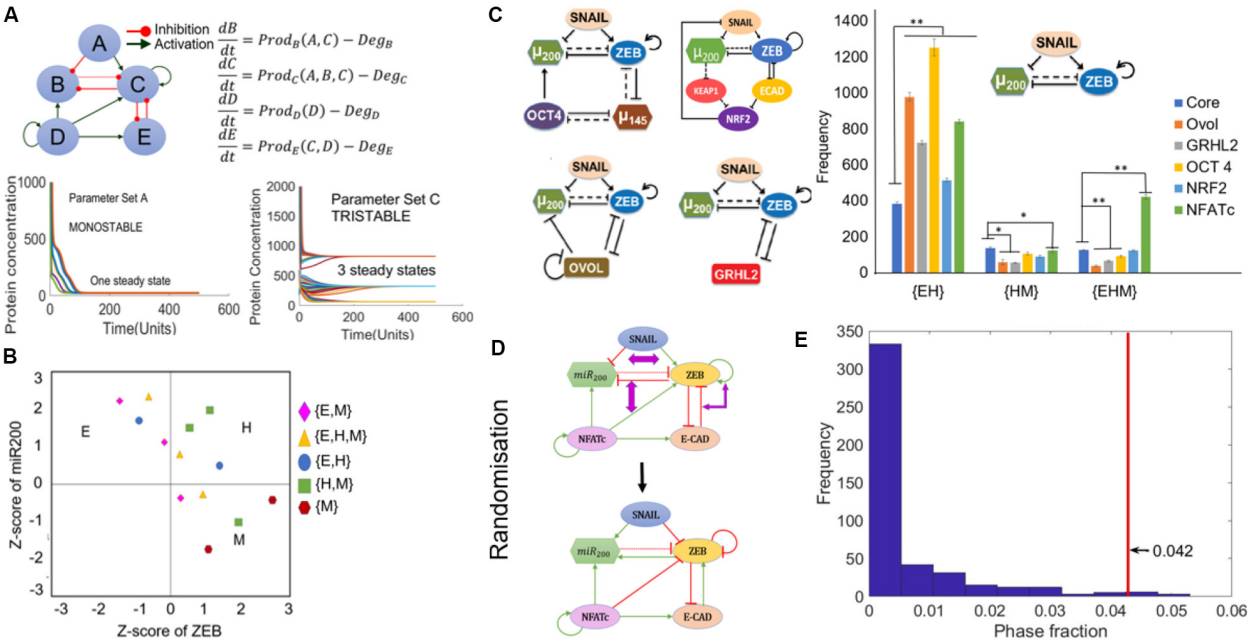

FIGURE 4 | RACIPE analysis of NFATc network. **(A)** Schematic of coupled EMT-NFATc network simulated for various parametric combinations through RACIPE. For different parameters, dynamics of this regulatory network, as modeled by a set of ordinary differential equations representing the interactions in the network, can end up in a mono- or a multi-stable regime. **(B)** Phases are calculated after *z*-scores of ZEB and miR200 levels calculated from RACIPE are used to classify steady states obtained from each parameter set into Epithelial (E) – (high miR-200, low ZEB)/Hybrid (H) – (high miR200, high ZEB)/Mesenchymal (M) – (low miR-200, high ZEB) phenotypes. Different symbols denote different phases. **(C)** Frequencies of different multi-stable phases for core EMT network (SNAIL/miR-200/ZEB), its coupling to other PSFs (shown to the left of the graph) and coupled EMT-NFATc coupled network. Error bars denote standard deviation; $n = 3$. **(D)** Depiction of network randomization methodology. For every node, the degree of incoming and outgoing edges are conserved but not necessarily the number of inhibitory or excitatory edges arriving at or emanating from a node. **(E)** Frequency distribution of the {E, H, M} phase fraction for randomized networks. The red line denotes the phase fraction of {E, H, M} phase in the wild-type (WT), i.e., NFATc network. $^*p < 0.05$, $^{**}p < 0.005$ using two tailed Student's *t*-test.

binned as epithelial, hybrid E/M or mesenchymal, based on the *z*-scores of miR-200 and ZEB for that case. Thus, each parameter was categorized into a given monostable or multi-stable phase; for instance, a parameter set that enabled both epithelial and hybrid E/M phenotypes for different initial conditions was classified as {E, H} (**Figure 4B**). Compared to the core network, each of these networks enabled a higher number of parameter sets enabling the co-existence of E and hybrid E/M states ({E, H}) (**Figure 4C**). Conversely, in most cases, the frequency of {H, M} (co-existence of M and hybrid E/M states) and {E, H, M} (co-existence of E, hybrid E/M, and M states) was decreased. However, in the case of NFATc, there was a significant increase in the frequency of {E, H, M} phase (**Figure 4C**) unlike other PSFs, suggesting that the presence of NFATc may enhance cellular plasticity among epithelial, hybrid E/M and mesenchymal states.

To test whether these results for NFATc are specific to the network topology of coupled NFATc-EMT network, we generated many randomized networks by swapping the edges between nodes in the network, such that the number of incoming and outgoing edges for every node was maintained the same (an example shown in **Figure 4D**). RACIPE analysis was performed on each of these randomized networks ($n = 461$; see section "Materials and Methods" for details), and the output obtained was separated into different phases i.e., {E}, {H}, {M}, {E, M}, {E, H}, {H, M}, and {E, H, M} as mentioned above. We quantified the frequency for multi-stable phases containing the hybrid state i.e., {E, H}, {H, M}, and {E, H, M} for all randomized networks, and calculated the frequencies of these phases, i.e., number of parameter sets out of 10,000 that enabled a given phase. The frequency distribution revealed that most of the randomized circuits gave rise to a lower fraction of {E, H, M} phase as compared to that for the wild-type NFATc-EMT coupled circuit (the value denoted by the red vertical line) (**Figure 4E**). However, such stark differences were not observed for the {E, H} (**Supplementary Figure S5A**) and {H, M} phases (**Supplementary Figure S5B**), suggesting that the NFATc-EMT coupled circuit topology is enriched for enabling co-existence and consequent possible switching among the epithelial, mesenchymal and hybrid E/M phenotypes.

NFATc Confers Stability to the Hybrid E/M Phenotype in a Multi-Stable Phase

We observed that the presence of NFATc in the network increased the frequency of multi-stable phases containing the hybrid state, particularly the {E, H, M} phase. This led us to investigate the relative stability of the different states in a given multi-stable phase. To quantify relative stability, every parameter set giving rise to either {E, H}, {H, M}, or {E, H, M} phases was simulated using 1000 random initial conditions and each time we tabulated

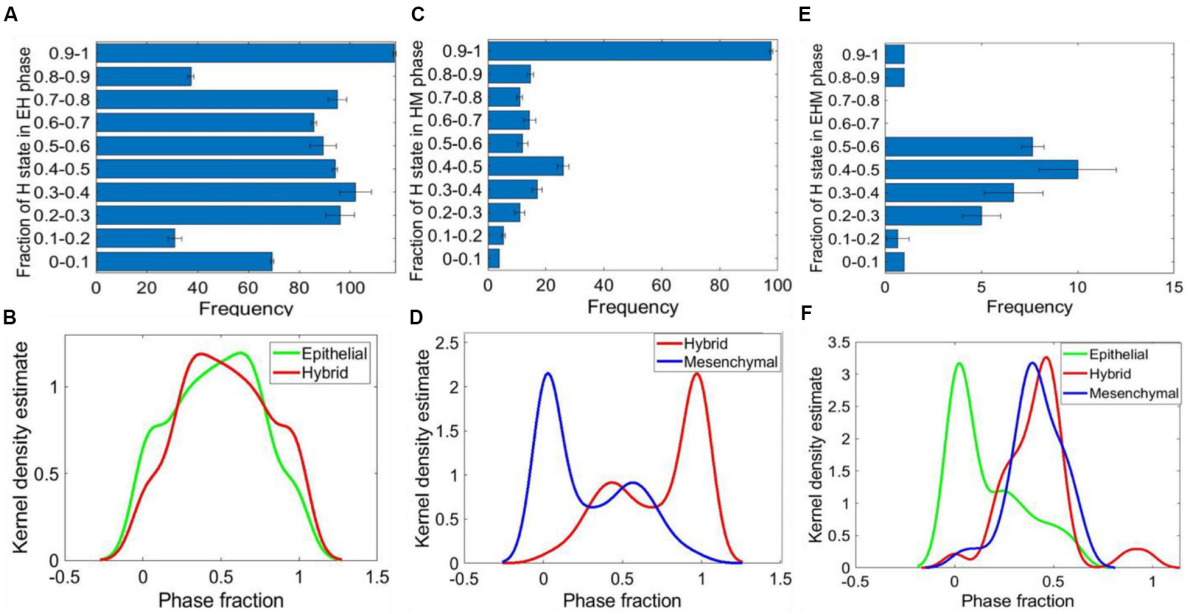

FIGURE 5 | Relative stability analysis. **(A)** Frequency distribution of H state in {E, H} phase. **(B)** Kernel density plot showing the frequency distribution of E and H states in the {E, H} phase. **(C)** Frequency distribution of H state in the {H, M} phase. **(D)** Kernel density plot showing the distribution of M and H states in the {H, M} phase. **(E)** Frequency distribution of H state in the {E, H, M} phase. **(F)** Kernel density plot showing the distribution of E, H, and M states in the {E, H, M} phase. For panels **(A,C,E)** the error bars represent the mean ± standard deviation for three sets of independently chosen initial conditions for a given parameter set obtained from one RACIPE run.

how many initial conditions led to which state – E, H, or M. For individual parameter sets, we observed heterogeneity in terms of relative stability of H states in the {E, H} phase, i.e., some parameter sets seemed to have a deeper "basin of attraction" for the epithelial attractor as compared to the hybrid E/M one and *vice versa* (**Figure 5A** and **Supplementary Figure S6A**). Nonetheless, there were similarities in the frequencies of E and the H state obtained across parameter sets obtained from independent RACIPE replicates, as represented by their similar and overlapping kernel density estimates (**Figure 5B** and **Supplementary Figures S6C,D**).

This analysis for the {H, M} phase revealed that the H state was more stable; i.e., the number of parameter cases for which the relative stability of H state was higher as compared to M state was more than the number of cases when the M state was relatively more stable (**Figure 5C** and **Supplementary Figure S5B**). This trend was maintained for parameter sets obtained across three RACIPE replicates (**Figure 5D** and **Supplementary Figures S5E,F**). Similar analysis on the {E, H, M} phase suggested that the E state was relatively less stable than the H and M states (**Figures 5E,F** and **Supplementary Figures S7A–D**). Together, these results indicate that the presence of NFATc can confer high stability to the hybrid E/M state for parametric combinations enabling the co-existence of multiple phenotypes.

NFATc Affects Clinical Outcome in a Tissue Specific Manner

The hybrid E/M phenotype is often attributed to drive tumor aggressiveness (7, 8). This trend is further supported by clinical data where PSFs such as GRHL2 and NRF2 correlate with poor patient survival (18, 21). We investigated whether the levels of NFATc can correlate with clinical response, and observed that the association of NFATc with clinical outcomes is context dependent. Higher levels of NFATc correlated with better relapse free survival among breast cancer patients (**Figures 6A–C**) but with poor relapse free survival among lung cancer patients (**Figure 6D**). We observed similar context-dependent behavior of NFATc in terms of overall survival, even within the same tissue (**Supplementary Figures S8A–C**). Also, NFATc was positively correlated with better metastasis free survival among breast cancer patients (**Supplementary Figure S8D**). Thus, the correlation of NFATc with patient survival is highly likely to be context dependent.

DISCUSSION

Recent *in vitro*, *in vivo,* and *in silico* investigations have emphasized the existence and significance of hybrid E/M phenotype(s) in various cancer types (37). These hybrid E/M phenotypes can exhibit maximum plasticity (38), possess traits of cancer stem cell-like traits, evade drug resistance, and thus be the "fittest" for metastasis (8). These preclinical experimental observations are supported by clinical analysis of carcinoma samples suggesting that the presence of hybrid E/M cells in a patient at the time of diagnosis associates with poor patient outcomes. Interestingly, even a very small percentage of hybrid E/M cells (score >2%) was found to be sufficient to confer poor

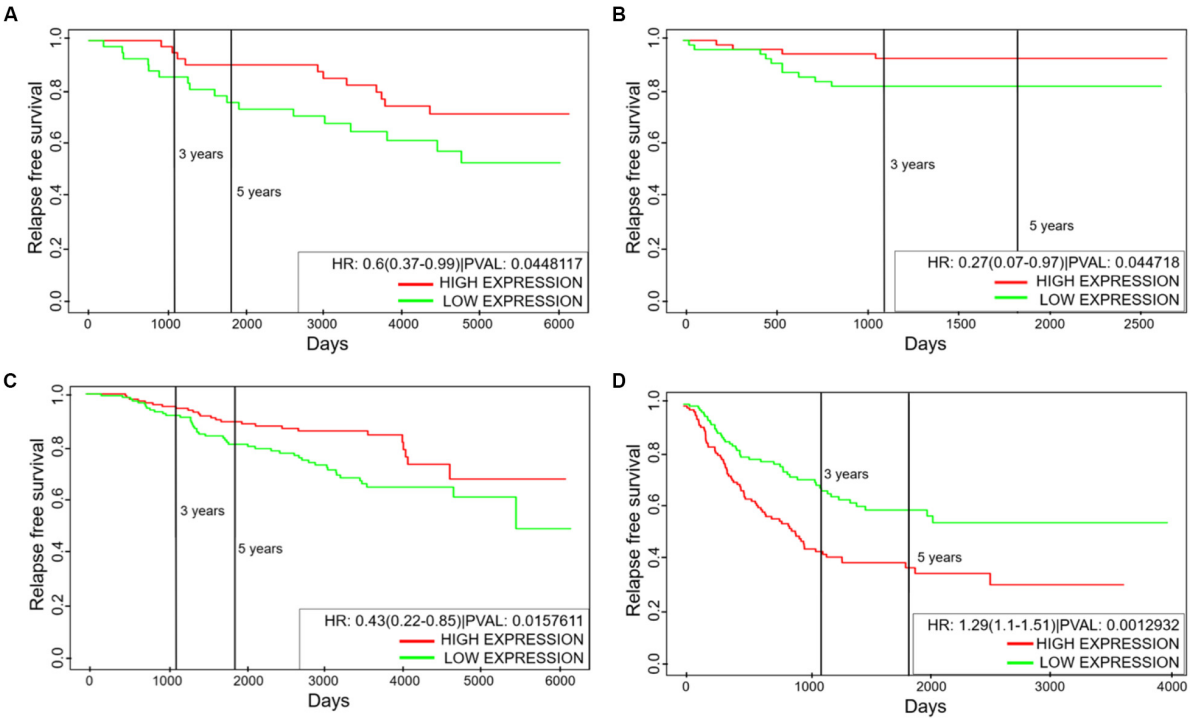

FIGURE 6 | Kaplan–Meier analysis for NFATc1 levels. **(A–C)** Relapse free survival for GSE6532, GSE19615, and GSE17705 (breast cancer tissue samples). **(D)** Relapse free survival for GSE41271 (non-small cell lung cancer tissue samples). Green curve shows the group of patients with low NFATc1; red curve shows the group with high NFATc1 levels. All cohorts are divided at median values of NFATc1; HR denotes hazard ratio; PVAL denotes *p*-value.

prognosis (39). Thus, identifying mechanisms that can maintain cells in hybrid E/M phenotypes is of crucial importance in our efforts to curb metastatic load.

Here, we developed a computational modeling framework to identify the transcription factor NFATc as a potential PSF for hybrid E/M phenotypes. In presence of NFATc, cells undergo a delayed or stalled EMT; thus maintaining cells in hybrid E/M phenotypes; knockdown of NFATc in H1975 NSCLC cells drove the progression toward a complete EMT phenotype, reminiscent of observations made for other PSFs – GRHL2, OVOL2, NUMB, and NRF2 (18–23). Similar effects of NFATc knockdown were also seen in MCF10A and DLD1 cells where treatment with VIVIT, a soluble inhibitor of NFATc transcriptional activity, significantly reduced E-cadherin expression and protein level, and increased Slug and Vimentin levels (29), thus driving EMT. NFATc transcriptional activity was shown to be capable of maintaining E-cadherin levels even in the presence of TGFβ induced EMT (29), suggesting that NFATc acted as a "molecular brake" or "guardian" of epithelial traits, preventing a complete EMT (40). Consistently, NFATc1 was identified to be a master regulator of chromatin remodeling to regulate hybrid E/M phenotypes in skin cancer *in vivo*; the proportion of hybrid E/M phenotypes was also shown to be increased by GRHL2, OVOL1/2, and ΔNP63α at the expense of complete EMT cells, thus lending further credence to our results indicating a functional equivalence between NFATc1 and previously identified PSFs such as GRHL2, OVOL1/2, ΔNP63α, and NRF2 (7). In developmental EMT scenario, NFATc1 is implicated in a key role during heart valve development; NFATc1-null embryos exhibit excessive EMT and impaired valve formation. Transcriptional repression of Snail1 and Snail 2 by NFATc1 can inhibit EMT and help maintain vascular E-cadherin levels required for cellular adhesiveness (41, 42). These observations across multiple contexts highlight that NFATc may maintain cell-cell contacts in a hybrid E/M phenotype.

Hybrid E/M phenotype(s) are also often associated with higher stem-like behavior and enhanced metastasis across cancer types *in vitro* and *in vivo* (25). Consistently, NFAT transcriptional activity contributes to metastasis in colon cancer; inhibition of NFATc1 reduced metastatic growth in an immunocompetent mouse model. Further, genes upregulated by NFATc1 significantly correlated with worse clinical outcomes for Stage II and III colorectal cancer patients (43). Similarly, NFATc2 was overexpressed in lung adenocarcinoma tumor-initiating cells; it supported tumorigenesis *in vivo* and its knockdown *in vitro* reduced 60–70% tumor-spheres and restricted the renewability of tumor-spheres (32). NFAT/calcineurin signaling pathway is also activated in breast cancer and aggravates tumorigenic and metastatic potential of mammary tumor cells *in vitro* and *in vivo* (44, 45). Furthermore, NFATc1 levels were found to be significantly upregulated in spheroid-forming cells in pancreatic cancer, where NFATc1 promotes SOX2 transcriptionally (28). One of the targets of NFAT/SOX2 signaling pathway is ALDH1A1 (32) – a *bona fide* marker

for hybrid E/M cells behaving as cancer stem cells in breast cancer (46). Therefore, these observations underscore the connection between NFAT signaling, stemness, and metastatic aggressiveness.

Our results show that while NFATc increased the parametric range of SNAIL levels enabling a hybrid E/M phenotype, it did not increase the MRT of hybrid E/M cells, suggesting that the role of NFATc may be non-canonical in terms of behaving as a PSF. This non-canonical behavior is further elucidated by RACIPE analysis, where, unlike other PSFs such as GRHL2, NFATc increased the frequency of parametric combinations containing co-existing epithelial, hybrid E/M and M phenotypes, and possible interconversions among them. Thus, NFATc may be thought of as a driver of phenotypic plasticity, and targeting NFAT signaling may curb cancer cell adaptation (47) – a distinctive property of metastasis-initiating cells (48). Given that at least a full-blown EMT by itself need not be necessary for metastasis, the emergent dynamics of metastatic networks (49, 50) can also hold clues for identifying other perturbations to curb metastatic load.

MATERIALS AND METHODS

Mathematical Modeling

As per the schematic shown in **Figure 1A**, the dynamics of all five molecular species (miR-200, SNAIL, ZEB, E-cadherin, and NFATc) was described by a system of coupled ODEs. The level of a protein, mRNA or micro-RNA (X) is described via a chemical rate equation that assumes the generic form:

$$\frac{dX}{dt} = g_X H^S(A, A_0, n, \lambda) - k_X X$$

Where the first term of the equation signifies the basal rate of production (g_X); the terms multiplied to g_X represent the transcriptional/translational/post-translational regulations due to interactions among the species in the system, as defined by the Hills function $[H^S(A, A_0, n, \lambda)]$. The term $k_X X$ accounts for the rate of degradation of the species (X) based on first order kinetics. The complete set of equations and parameters are presented in the **Supplementary Material**.

Cell Culture and siRNA Treatments

H1975 cells were cultured in RPMI 1640 medium containing 10% fatal bovine serum and 1% penicillin/streptomycin cocktail (Thermo Fisher Scientific). Cells were transfected at a final concentration of 50 nM siRNA using Lipofectamine RNAiMAX (Thermo Fisher Scientific) according to the manufacturer's instructions using following siRNAs: siControl (Thermo Fisher Scientific), siNFATC1 #1 (Invitrogen), siNFATC1 #2 (Invitrogen). Regular mycoplasma testing was also carried out to exclude any possible cell culture contamination.

RT-PCR

Total RNA was isolated following manufacturer's instructions using RNAeasy kit (Qiagen). cDNA was prepared using iScript gDNA clear cDNA synthesis kit (Bio-Rad). A TaqMan PCR assay was performed with a 7500 Fast Real-Time PCR System using TaqMan PCR master mix, commercially available primers, and FAM^{TM}-labeled probes for CDH1, VIM, ZEB1, NFATC1, SNAIL, and VIC^{TM}-labeled probes for 18 S, as per manufacturer's instructions (Life Technologies). Each sample was run in biological and technical triplicates. Ct values for each gene was calculated and normalized to Ct values for 18 S (ΔCt). The $\Delta\Delta$Ct values were then calculated by normalization to the ΔCt value for control.

Western Blotting Analysis

H1975 cells were lysed in RIPA lysis assay buffer (Pierce) supplemented with enzyme inhibitor cocktail (Roche). The samples were separated on a 4–20% SDS-polyacrylamide gel (Bio-Rad). After transfer to PVDF membrane, incubation was carried out with primary antibodies anti-CDH1 (1:1000; Cell Signaling Technology), anti-vimentin (1:1000; Cell Signaling Technology), anti-Zeb1 (1:1000; Cell Signaling Technology), anti-SNAIL (1:1000; Cell Signaling Technology), and anti-beta actin (1:10 000; Abcam) and subsequent secondary antibodies. Membranes were exposed using the ECL method (GE Healthcare) as per manufacturer's instructions.

Immunofluorescence

H1975 Cells were fixed in 3.4% paraformaldehyde, permeabilized with 0.2% Triton X-100, and then stained with primary antibodies against CDH1 (1:100; Abcam) and vimentin (1:100; Cell Signaling Technology). Alexa conjugated secondary antibodies (Life Technologies) were used to detect the expression of respective proteins. DAPI was used to counterstain the nuclei.

Wound-Healing Assay

Scratch wound-healing assay was performed to determine cell migration using confluent cultures (80-90% confluence). Briefly, H1975 cells (1 × 105 cells/ml) were seeded in 6-well tissue culture plate. After cells attain expected confluency, they were starved for 24 h using 0.2% serum in growth media. Next day, a sterile p200 pipet tip was used to create a wound on the confluent monolayer and media was replenished. Images were acquired at 0 and 16 h; the experiments were repeated three times. Images of the scratch wounds were taken and measured by ImageJ software to calculate the mean and standard deviation. Each group was compared with the control group. Cell migration was expressed as the migration rate: (original scratch width – new scratch width)/original scratch width × 100%.

Trans-Well Migration Assay

H1975 cells were grown in 6-well plates and treated with siNFATC1 for 24 h. After 48 h of NFATC1 knockdown, cell monolayers were harvested, and 2 × 104 viable cells/200 μl cell concentration was prepared in serum free medium. The cell suspension was transferred on top of a 0.8 μm pore diameter Transwell insert (Millipore) and placed on a 24 well cell culture plate. A 10% fetal bovine serum solution was added as chemo-attractant at the bottom of the insert and plate was

incubated at 37°C for 18 h. Non-migrated cells were removed by gently swabbing inside each insert. Cells were fixed and stained with a 0.5% crystal violet solution for 10 min. The inserts were thoroughly washed with and air dried completely before visualizing under a microscope. Cell numbers were counted at ×200 magnification. The experiment was repeated three times and statistically analyzed with five fields of view, and the mean values were taken as the migratory cell number.

Mean Residence Time Analysis

The MRT was calculated as follows: As the degradation rate of ZEB mRNA is much greater than that of ZEB protein and miR-200 and also the production rate of E-cad and NFATc is much more larger than that of ZEB protein and miR-200, we assumed that ZEB mRNA, E-cad, and NFATc reach to the equilibrium much faster relatively, that is, $\frac{dm_Z}{dt} = 0$, $\frac{dE}{dt} = 0$, and $\frac{dN}{dt} = 0$. This assumption reduces the equations given in **Supplementary Material:** to two coupled ODEs of ZEB and miR-200. Then we simulated the dynamical system in presence of external noise and obtained the time evolution of ZEB protein and miR-200 using Euler–Maruyama simulation. From the time evolution of ZEB and miR-200 the dynamical states of the system were coarse grained as an itinerary of basins visited. Then the MRT was calculated by multiplying the total number of successive states with Δt. Detailed methods are outlined in the publication by Biswas et al. (24).

RACIPE

Random circuit perturbation (36) algorithm was run on the coupled EMT-NFATc network and its randomized counterparts. Continuous steady state levels were obtained as output for the five variables, for ensembles of mathematical models; each model has a randomly chosen parameter set corresponding to intrinsic production/degradation of all species as well as those representing the regulatory links. The algorithm was used to generate 10,000 mathematical models, each with a different set of parameters. Hundred initial conditions were chosen for each model, and all steady state solutions obtained were compiled together. With this consolidated data, the *z*-scores of steady state levels of all the biomolecules in the individual networks were calculated. Based on the *z*-score of ZEB and miR-200, the phenotype for a given steady state solution is decided, i.e., if ($z_{zeb} > 0$) and ($z_{mir-200} < 0$), it is counted as mesenchymal state, ($z_{zeb} > 0$) and ($z_{mir-200} > 0$) is counted as a hybrid state, and ($z_{zeb} < 0$) and ($z_{mir-200} > 0$) is counted as an epithelial state. Similarly, states are determined for all solutions and based on the state of each steady state for a given set of parameters, the phases are determined.

Network Randomization

The following rules were employed to generate an ensemble of randomized networks. For each node, in each instance of a randomization of the wild-type network (**Figure 1A**), the number of incoming and outgoing edges were kept constant. The number of activation edges and the number of inhibitory edges were also kept fixed at 6 and 5, respectively [The same number as that in the wild-type network (**Figure 1A**)]. Furthermore, the source node and the target node for each of the edges were kept fixed but the identity of the edge in terms of it being an activation or inhibition link was allowed to change. Hence, 461 ($\binom{11}{5}C - 1 = \frac{11!}{6!5!} - 1 = 461$) such randomized networks were constructed excluding the wild-type case.

Relative Stability Analysis

The relative stability analysis was performed in MATLAB. RACIPE generated parameter sets that give rise to multi-stable phases were first determined. These parameter sets were simulated in MATLAB using 1000 random initial conditions and the steady state each time was determined. Then, the total number of times each of the possible steady states was reached was calculated.

Kaplan–Meier Analysis

ProgGene (51) was used for conducting Kaplan–Meier analysis for respective datasets. The number of samples in NFATc1-high vs. NFATc1-low categories are given below:

GSE6532 (breast cancer): *n* (High) = 44, *n* (Low) = 43
GSE19615 (breast cancer): *n* (High) = 58, *n* (Low) = 57
GSE17705 (lung cancer): *n* (High) = 149, *n* (Low) = 149
GSE41271 (lung cancer): *n* (High) = 138, *n* (Low) = 137
GSE30219 (lung cancer): *n* (High) = 141 *n* (Low) = 141
GSE19536 (breast cancer): *n* (High) = 56, *n* (Low) = 44
GSE14814 (lung cancer): *n* (High) = 44, *n* (Low) = 44.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

MKJ and ST designed the research. MKJ, ST, AG, and SH supervised the research. ARS, DK, and KB performed the research. All authors contributed to analyzing the data and writing and editing of the manuscript.

FUNDING

This work was supported by Ramanujan Fellowship awarded by SERB, DST, Government of India to MKJ (SB/S2/RJN-049/2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.553342/full#supplementary-material

REFERENCES

1. Gupta GP, Massagué J. Cancer metastasis: building a framework. *Cell.* (2006) 127:679–95.
2. Jolly MK, Tripathi SC, Somarelli JA, Hanash SM, Levine H. Epithelial-mesenchymal plasticity: how have quantitative mathematical models helped improve our understanding? *Mol Oncol.* (2017) 11:739–54. doi: 10.1002/1878-0261.12084
3. Jolly MK, Kulkarni P, Weninger K, Orban J, Levine H. Phenotypic plasticity, bet-hedging, and androgen independence in prostate cancer: role of non-genetic heterogeneity. *Front Oncol.* (2018) 8:50. doi: 10.3389/fonc.2018.00050
4. Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell.* (2011) 146:633–44. doi: 10.1016/j.cell.2011.07.026
5. Balázsi G, van Oudenaarden A, Collins JJ. Cellular decision making and biological noise: from microbes to mammals. *Cell.* (2011) 144:910–25. doi: 10.1016/j.cell.2011.01.030
6. Nieto MA, Huang RY, Jackson RA, Thiery JP. EMT: 2016. *Cell.* (2016) 166:21–45.
7. Pastushenko I, Blanpain C. EMT transition states during tumor progression and metastasis. *Trends Cell Biol.* (2019) 29:212–26. doi: 10.1016/j.tcb.2018.12.001
8. Jolly MK, Mani SA, Levine H. Hybrid epithelial/mesenchymal phenotype(s): the 'fittest' for metastasis? *Biochim Biophys Acta Rev Cancer.* (2018) 1870:151–7. doi: 10.1016/j.bbcan.2018.07.001
9. Campbell K, Rossi F, Adams J, Pitsidianaki I, Barriga FM, Garcia-Gerique L, et al. Collective cell migration and metastases induced by an epithelial-to-mesenchymal transition in *Drosophila* intestinal tumors. *Nat Commun.* (2019) 10:2311. doi: 10.1038/s41467-019-10269-y
10. Liao T-T, Yang M-H. Hybrid epithelial/mesenchymal state in cancer metastasis: clinical significance and regulatory mechanisms. *Cells.* (2020) 9:623. doi: 10.3390/cells9030623
11. Bocci F, Jolly MK, Onuchic JN. A biophysical model uncovers the size distribution of migrating cell clusters across cancer types. *Cancer Res.* (2019) 79:5527–35. doi: 10.1158/0008-5472.CAN-19-1726
12. Giuliano M, Shaikh A, Lo HC, Arpino G, De Placido S, Zhang XH, et al. Perspective on circulating tumor cell clusters: why it takes a village to metastasize. *Cancer Res.* (2018) 78:845–52. doi: 10.1158/0008-5472.CAN-17-2748
13. Gonzalez DM, Medici D. Signaling mechanisms of the epithelial-mesenchymal transition. *Sci Signal.* (2014) 7:re8. doi: 10.1126/scisignal.2005189
14. Burk U, Schubert J, Wellner U, Schmalhofer O, Vincan E, Spaderna S, et al. A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *EMBO Rep.* (2008) 9:582–9. doi: 10.1038/embor.2008.74
15. Lu M, Jolly MK, Levine H, Onuchic JN, Ben-Jacob E. MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc Natl Acad Sci USA.* (2013) 110:18174–9. doi: 10.1073/pnas.1318192110
16. Siemens H, Jackstadt R, Hünten S, Kaller M, Menssen A, Götz U, et al. miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions. *Cell Cycle.* (2011) 10:4256–71. doi: 10.4161/cc.10.24.18552
17. Park S-M, Gaur AB, Lengyel E, Peter ME. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev.* (2008) 22:894–907. doi: 10.1101/gad.1640608
18. Bocci F, Tripathi SC, Vilchez MSA, George JT, Casabar J, Wong P, et al. NRF2 activates a partial epithelial-mesenchymal transition and is maximally present in a hybrid epithelial/mesenchymal phenotype. *Integr Biol.* (2019) 11:251–63. doi: 10.1101/390237
19. Watanabe K, Villarreal-Ponce A, Sun P, Salmans ML, Fallahi M, Andersen B, et al. Mammary morphogenesis and regeneration require the inhibition of EMT at terminal end buds by Ovol2 transcriptional repressor. *Dev Cell.* (2014) 29:59–74. doi: 10.1016/j.devcel.2014.03.006
20. Jia D, Jolly MK, Boareto M, Parsana P, Mooney SM, Pienta KJ, et al. OVOL guides the epithelial-hybrid-mesenchymal transition. *Oncotarget.* (2015) 6:15436–48. doi: 10.18632/oncotarget.3623
21. Jolly MK, Tripathi SC, Jia D, Mooney SM, Celiktas M, Hanash SM, et al. Stability of the hybrid epithelial/mesenchymal phentoype. *Oncotarget.* (2016) 7:27067–84.
22. Bocci F, Jolly MK, Tripathi SC, Aguilar M, Hanash SM, Levine H, et al. Numb prevents a complete epithelial-mesenchymal transition by modulating Notch signaling. *J R Soc Interface.* (2017) 14:20170512. doi: 10.1098/rsif.2017.0512
23. Hong T, Watanabe K, Ta CH, Villarreal-Ponce A, Nie Q, Dai X. An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput Biol.* (2015) 11:e1004569. doi: 10.1371/journal.pcbi.1004569
24. Biswas K, Jolly M, Ghosh A. Stability and mean residence times for hybrid epithelial/mesenchymal phenotype. *Phys Biol.* (2019) 16:025003. doi: 10.1088/1478-3975/aaf7b7
25. Jolly MK, Somarelli JA, Sheth M, Biddle A, Tripathi SCC, Armstrong AJJ, et al. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol Ther.* (2019) 194:161–84. doi: 10.1016/j.pharmathera.2018.09.007
26. Mognol GP, Carneiro FRG, Robbs BK, Faget DV, Viola JPB. Cell cycle and apoptosis regulation by NFAT transcription factors: new roles for an old player. *Cell Death Dis.* (2016) 7:e2199. doi: 10.1038/cddis.2016.97
27. Medyouf H, Ghysdael J. The calcineurin/NFAT signaling pathway: a novel therapeutic target in leukemia and solid tumors. *Cell Cycle.* (2008) 7:297–303. doi: 10.4161/cc.7.3.5357
28. Singh SK, Chen N, Hessmann E, Siveke J, Lahmann M, Singh G, et al. Antithetical NFAT c1–Sox2 and p53–miR200 signaling networks govern pancreatic cancer cell plasticity. *EMBO J.* (2015) 34:517–30. doi: 10.15252/embj.201489574
29. Gould R, Bassen DM, Chakrabarti A, Varner JD, Butcher J. Population heterogeneity in the epithelial to mesenchymal transition is controlled by NFAT and phosphorylated Sp1. *PLoS Comput Biol.* (2016) 12:e005251. doi: 10.1371/journal.pcbi.1005251
30. Schmalhofer O, Brabletz S, Brabletz T. E-cadherin, beta-catenin, and ZEB1 in malignant progression of cancer. *Cancer Metastasis Rev.* (2009) 28:151–66. doi: 10.1007/s10555-008-9179-y
31. Mooney SM, Jolly MK, Levine H, Kulkarni P. Phenotypic plasticity in prostate cancer: role of intrinsically disordered proteins. *Asian J Androl.* (2016) 18:704–10. doi: 10.4103/1008-682X.183570
32. Xiao ZJ, Liu J, Wang SQ, Zhu Y, Gao XY, Tin VPC, et al. NFATc2 enhances tumor-initiating phenotypes through the NFATc2/SOX2/ALDH axis in lung adenocarcinoma. *eLife.* (2017) 6:e26733. doi: 10.7554/eLife.26733
33. Wang G, Guo X, Hong W, Liu Q, Wei T, Lu C, et al. Critical regulation of miR-200 / ZEB2 pathway in Oct4 / Sox2-induced mesenchymal-to-epithelial transition and induced pluripotent stem cell generation. *Proc Natl Acad Sci USA.* (2013) 110:2858–63. doi: 10.1073/pnas.1212769110
34. Somarelli JA, Shelter S, Jolly MK, Wang X, Bartholf Dewitt S, Hish AJ, et al. Mesenchymal-epithelial transition in sarcomas is controlled by the combinatorial expression of miR-200s and GRHL2. *Mol Cell Biol.* (2016) 36:2503–13. doi: 10.1128/MCB.00373-16
35. Jolly MK, Boareto M, Debeb BG, Aceto N, Farach-Carson MC, Woodward WA, et al. Inflammatory breast cancer: a model for investigating cluster-based dissemination. *NPJ Breast Cancer.* (2017) 3:21. doi: 10.1101/119479
36. Huang B, Lu M, Jia D, Ben-Jacob E, Levine H, Onuchic JN. Interrogating the topological robustness of gene regulatory circuits by randomization. *PLoS Comput Biol.* (2017) 13:e1005456. doi: 10.137/journal.pcbi.1005456
37. Jolly MK, Celia-Terrassa T. Dynamics of phenotypic heterogeneity associated with EMT and stemness during cancer progression. *J Clin Med.* (2019) 8:1542. doi: 10.3390/jcm8101542
38. Tripathi S, Chakraborty P, Levine H, Jolly MK. A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. *PLoS Comput Biol.* (2020) 16:e1007619. doi: 10.1101/592691
39. Godin L, Balsat C, van Eycke Y, Allard J, Royer C, Remmelink M, et al. A novel approach for quantifying cancer cells showing hybrid epithelial / mesenchymal states in large series of tissue samples: towards a new prognostic marker. *Cancers (Basel).* (2020) 12:906. doi: 10.3390/cancers12040906
40. Li S, Yang J. Ovol proteins: guardians against EMT during epithelial differentiation. *Dev Cell.* (2014) 29:1–2. doi: 10.1016/j.devcel.2014.04.002

41. Wu B, Baldwin HS, Zhou B. Nfatc1 directs the endocardial progenitor cells to make heart valve primordium. *Trends Cardiovasc Med.* (2013) 23:294–300. doi: 10.1016/j.tcm.2013.04.003
42. Wu B, Wang Y, Lui W, Langworthy M, Tompkins KL, Hatzopoulos AK, et al. Nfatc1 coordinates valve endocardial cell lineage development required for heart valve formation. *Circ Res.* (2011) 109:183–92. doi: 10.1161/CIRCRESAHA.111.245035
43. Tripathi MK, Deane NG, Zhu J, An H, Mima S, Wang X, et al. Nuclear factor of activated T-cell activity is associated with metastatic capacity in colon cancer. *Cancer Res.* (2014) 74:6947–57. doi: 10.1158/0008-5472.CAN-14-1592
44. Quang CT, Lebouche S, Passaro D, Furhmann L, Nourieh M, Vincent-Salomon A, et al. The calcineurin/NFAT pathway is activated in diagnostic breast cancer cases and is essential to survival and metastasis of mammary cancer cells. *Cell Death Dis.* (2015) 6:e1658. doi: 10.1038/cddis.2015.14
45. Yiu GK, Toker A. NFAT induces breast cancer cell invasion by promoting the induction of cyclooxygenase-2. *J Biol Chem.* (2006) 281:12210–7. doi: 10.1074/jbc.M600184200
46. Colacino JA, Azizi E, Brooks MD, Harouaka R, Fouladdel S, McDermott SP, et al. Heterogeneity of human breast stem and progenitor cells as revelaed by transcriptional profiling. *Stem Cell Rep*. (2018) 10:1596–609. doi: 10.1016/j.stemcr.2016.05.008
47. Qin J-J, Nag S, Wang W, Zhou J, Zhang W-D, Wang H, et al. NFAT as cancer target: mission possible? *Biochim Biophys Acta.* (2014) 1846:297–311. doi: 10.1016/j.bbcan.2014.07.009
48. Celià-Terrassa T, Kang Y. Distinctive properties of metastasis-initiating cells. *Genes Dev*. (2016) 30:892–908. doi: 10.1101/gad.277681.116
49. Li C, Balazsi G. A landscape view on the interplay between EMT and cancer metastasis. *npj Syst Biol Appl.* (2018) 4:34. doi: 10.1038/s41540-018-0068-x
50. Lee J, Lee J, Farquhar KS, Yun J, Frankenberger CA, Bevilacqua E, et al. Network of mutually repressive metastasis regulators can promote cell heterogeneity and metastatic transitions. *Proc Natl Acad Sci USA*. (2014) 111:E364–73. doi: 10.1073/pnas.1304840111
51. Goswami CP, Nakshatri H. PROGgeneV2?: enhancements on the existing database. *BMC Cancer*. (2014) 14:970. doi: 10.1186/1471-2407-14-970

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 17 September 2020
doi: 10.3389/fphys.2020.00806

Application of the Moran Model in Estimating Selection Coefficient of Mutated CSF3R Clones in the Evolution of Severe Congenital Neutropenia to Myeloid Neoplasia

Khanh N. Dinh[1], *Seth J. Corey*[2] *and Marek Kimmel*[3,4]*

[1] *Irving Institute for Cancer Dynamics and Department of Statistics, Columbia University, New York, NY, United States,* [2] *Departments of Pediatric and Cancer Biology, Cleveland Clinic, Cleveland, OH, United States,* [3] *Departments of Statistics and Bioengineering, Rice University, Houston, TX, United States,* [4] *Department of Systems Biology and Engineering, Gliwice, Poland*

Bone marrow failure (BMF) syndromes, such as severe congenital neutropenia (SCN) are leukemia predisposition syndromes. We focus here on the transition from SCN to pre-leukemic myelodysplastic syndrome (MDS). Stochastic mathematical models have been conceived that attempt to explain the transition of SCN to MDS, in the most parsimonious way, using extensions of standard processes of population genetics and population dynamics, such as the branching and the Moran processes. We previously presented a hypothesis of the SCN to MDS transition, which involves directional selection and recurrent mutation, to explain the distribution of ages at onset of MDS or AML. Based on experimental and clinical data and a model of human hematopoiesis, a range of probable values of the selection coefficient s and mutation rate μ have been determined. These estimates lead to predictions of the age at onset of MDS or AML, which are consistent with the clinical data. In the current paper, based on data extracted from published literature, we seek to provide an independent validation of these estimates. We proceed with two purposes in mind: (i) to determine the ballpark estimates of the selection coefficients and verify their consistency with those previously obtained and (ii) to provide possible insight into the role of recurrent mutations of the G-CSF receptor in the SCN to MDS transition.

Keywords: clinical data, G-CSF receptor (G-CSFR), recurrent mutation, myeloid neoplasia, Moran model, selective advantage

OPEN ACCESS

Edited by:
Ernesto Augusto Bueno Da Fonseca Lima,
University of Texas at Austin, United States

Reviewed by:
Diego Samuel Rodrigues,
Campinas State University, Brazil
Alfredo Rodríguez,
National Institute of Pediatrics, Mexico

***Correspondence:**
Marek Kimmel
kimmel@rice.edu

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *17 March 2020*
Accepted: *17 June 2020*
Published: *17 September 2020*

Citation:
Dinh KN, Corey SJ and Kimmel M (2020) Application of the Moran Model in Estimating Selection Coefficient of Mutated CSF3R Clones in the Evolution of Severe Congenital Neutropenia to Myeloid Neoplasia. Front. Physiol. 11:806. doi: 10.3389/fphys.2020.00806

1. INTRODUCTION

Bone marrow failure (BMF) syndromes, such as severe congenital neutropenia (SCN) are leukemia predisposition syndromes. In addition to SCN, these heterogeneous groups of disorders include Fanconi anemia, dyskeratosis congenita, Diamond-Blackfan anemia, Shwachman-Diamond syndrome, and *GATA2* deficiency (West and Churpek, 2017; Kennedy and Shimamura, 2019). Each of these clinically defined disorders are monogenic with mutations in one or more genes in a pathway. For example, Fanconi anemia results from germline mutations in genes involved in DNA repair and Diamond-Blackfan anemia in ribosome structure (Oyarbide et al., 2019). What is less well-understood are the somatic mutations that arise during transformation of a BMF syndrome to myeloid neoplasia (Rafei and DiNardo, 2019).

Focusing on the SCN to MDS to AML transition, individuals with a germline mutation in the *ELANE* gene develop at early age a severe neutropenia (Touw, 2015). This profound neutropenia makes them susceptible to recurrent infections, which can be only partly managed by antibiotics. Treatment introduced in the 1990s involves administration of large doses of recombinant human granulocyte colony stimulating factor (G-CSF), which boosts neutrophil production (Bonilla et al., 1989). Unfortunately, in about 30% of patients, either myelodysplatic syndrome (MDS), a preleukemic disorder, or acute myeloid leukemia (AML) emerges. In 70% MDS or AML cases arising from SCN, somatic mutations of the G-CSF Receptor (*CSF3R*) occur (Link, 2019). These are almost always nonsense mutations. The truncated *CSF3R* affects altered signaling, gene expression, and phenotype within the neutrophil lineage. There is enhanced proliferation and impaired neutrophilic differentiation to G-CSF.

Stochastic mathematical models have been conceived, which attempt to explain the transition of SCN to MDS and then to AML, in the most parsimonious way, using suitable extensions of standard processes of population genetics and population dynamics, such as the branching (Kimmel and Corey, 2013) and the Moran processes (Wojdyla et al., 2019). Specifically, the latter paper presented a hypothesis of the SCN → MDS transition, which involves the Moran process with directional selection (Durrett, 2008) and recurrent mutation, to explain the distribution of ages at onset of MDS or AML. As argued in Wojdyla et al. (2019), starting in the fetal life, *CSF3R* mutations arise as a random process and are selected for when G-CSF is administered to boost neutrophil production. Based on experimental and clinical data and a model of human hematopoiesis, a range of probable values of the selection coefficient s and mutation rate μ have been determined. These estimates lead to predictions of the age at onset of MDS or AML, which are consistent with the clinical data.

In the current paper, based on data extracted from published literature, we seek to provide an independent validation of these estimates. We will use the model of evolution of the mutant receptors in the hematopoietic stem cells (HSC) in the bone marrow in the form of a Moran process with selection and recurrent mutation. This is the same process we used in Wojdyla et al. (2019), except that here, to simplify computations, we assume constant HSC population size and develop an analytical approximation of the expected values of the mutant receptor occurrence among HSC under the assumption that initial count of mutants is already substantial (Methods and Data). We proceed with two purposes in mind. Our first purpose is to determine the ballpark estimates of the selection coefficients and verify their consistency with those obtained in Wojdyla et al. (2019). Our second purpose is to provide insight into the relative role of recurrent mutations of the G-CSF receptor in the SCN to MDS transition.

2. METHODS AND DATA

2.1. Moran Process

In the monograph by Durrett (2008), the Moran process with selection is defined as follows

- Constant population of N individuals.
- At each discrete time moment, a randomly chosen individual dies, and, at the same moment, another randomly chosen individual proliferates (for mathematical completeness, it can be the same individual).
- In the model with directional selection, there are individuals of two types: wildtype (WT) and mutant (M) and the choice of individual that proliferates is biased. The wildtype is chosen with weight $1-s$, $s \in (0,1)$.

It is instructive to consider the discrete-time case first. Let us denote the number of mutants by i. There are four possibilities

- WT dies, with probability $(N-i)/N$
  - WT proliferates, with probability $(1-s)(N-i)/[(1-s)(N-i)+i]$
  - M proliferates, with probability $i/[(1-s)(N-i)+i]$
- M dies, with probability i/N
  - WT proliferates, with probability $(1-s)(N-i)/[(1-s)(N-i)+i]$
  - M proliferates, with probability $i/[(1-s)(N-i)+i]$

Only the WT→M and M→WT options lead to change in the number of mutants

$$p_{i,i+1} = \frac{N-i}{N}\frac{i}{[(1-s)(N-i)+i]},$$

$$p_{i,i-1} = \frac{i}{N}\frac{(1-s)(N-i)}{[(1-s)(N-i)+i]},$$

the M→M and WT→WT options jointly contribute to $p_{i,i}$. States $\{0\}$ and $\{N\}$ are absorbing. The probability of being eventually absorbed in $\{N\}$, if at time 0 there are i mutants, is equal to

$$P[T_N < T_0] = \frac{1-(1-s)^i}{1-(1-s)^N}$$

in the case with selection, which leads to

$$P[T_N < T_0] = i/N$$

in the neutral case.

The continuous-time version is defined by transition intensities

$$q_{i,i+1} = (N-i)\frac{i}{N},\ \ q_{i,i-1} = i\frac{(1-s)(N-i)}{N},$$

which have different denominators than the transition probabilities. However, they lead to the same absorption formula. The expected time to absorption in $\{N\}$ (fixation of the mutant) has a commonly used asymptotics

$$E_1(T_N) \sim \frac{2}{s}\ln(N)$$

as $N \to \infty$ in the case with selection, which however is not very accurate.

2.2. Time-Continuous Moran Process With Directional Selection and Recurrent Mutations

A time-continuous Moran process with directional selection may be supplemented with recurrent mutation by adding a term of the form $\mu(N-i)$ to the $q_{i,i+1}$ transition intensity. This can be interpreted as an equal and independent chance $\mu\Delta t + o(\Delta t)$, for each of the $N-i$ WT cells, of becoming a mutant in a short time interval $(t, t+\Delta t)$. The complete set of transition rules for the chain $\{X(t),\ t \geq 0\}$ assumes the form

$$q_{i,i+1} = (N-i)\frac{i}{N} + \mu(N-i),\quad q_{i,i-1} = i\frac{(1-s)(N-i)}{N},\ i = 0, \ldots N. \tag{1}$$

Because the state space is finite, the chain is eventually absorbed in the state N at random time T_N; cf. **Figure 1** for a heuristic illustration.

2.3. Simulation of Trajectories of the Moran Process With Recurrent Mutation

Simulation of a time-continuous Markov Chain is based directly on application of the transition intensities as expressed in Equation (1). Briefly, if the mutant cell chain $X(t)$ is in state i at time t, then the time to the next jump is a random variable τ distributed exponentially with parameter $q_{i,i-1} + q_{i,i+1}$. The direction of the jump at time $t+\tau$ is then decided by a random choice, $i \to i-1$ with probability $\frac{q_{i,i-1}}{q_{i,i-1}+q_{i,i+1}}$ and $i \to i+1$ with probability $\frac{q_{i,i+1}}{q_{i,i-1}+q_{i,i+1}}$, respectively. This algorithm (know also popularly as the Gillespie algorithm) is based on the properties of holding times and jumps of time-continuous Markov Chain, as explained for example in the book (Grimmett and Stirzaker, 2001). Simulations depicted in **Figure 2** were executed using this algorithm.

2.4. Approximation of the Moran Process With Recurrent Mutation and Estimation of Selection Coefficient and Mutation Rate

In the current study, we are not as much concerned with a mathematically rigorous theory of the Moran process with selection and recurrent mutation, as with obtaining computable expressions that lead to ballpark estimates of the selection coefficient and mutation rate. Based on transitions spelled out in Equation (1), we obtain the following expression for the conditional expectations:

$$\mathrm{E}[X(t+\Delta t)|X(t) = x] = x + \left(\frac{(N-x)xs}{N} + \mu(N-x)\right)\Delta t + o(\Delta t) \tag{2}$$

Corresponding expression for variance is more involved. From Equation (2), assuming that $x = X(t)$ can be replaced by its expectation and denoting the latter by $x(t)$ we formally obtain the following ordinary differential equation (ODE) for $x(t)$.

$$\dot{x}(t) = \left(\frac{(N-x(t))x(t)s}{N} + \mu(N-x(t))\right),\quad t \in [t_0, t_1] \tag{3}$$

Following a change of variables $y(t) = x(t)/N \in [0,1]$, this leads to

$$\dot{y}(t) = (1-y(t))y(t)s + \mu(1-y(t)),\quad t \in [t_0, t_1] \tag{4}$$

This latter equation has an explicit solution

$$y(t) = \frac{1-(\mu/s)\alpha_0\exp(-(\mu+s)(t-t_0))}{1+\alpha_0\exp(-(\mu+s)(t-t_0))},\quad t \in [t_0, t_1] \tag{5}$$

where

$$\alpha_i = (1-y_i)/(y_i+u),\quad y_i = y(t_i),\quad i = 0, 1, \text{ and } u = \mu/s \tag{6}$$

This curve is very similar to that derived in the initial part of the well-known study by Gerrish and Lenski (1998), under branching process hypotheses. Let us also notice that population size N does not play a role in the expression for $y(t)$. However, as evidenced by a comparison between the simulations in **Figures 2A,F**, larger N reduces process variance and the slight bias of $y(t)$ as the estimate of $X(t)/N$.

Let us note that Equation (5) yields

$$\frac{1-y(t)}{y(t)+\mu/s} = \alpha_0\exp(-(\mu+s)(t-t_0)), \tag{7}$$

which after substitution $t = t_1$ yields

$$\alpha_1 = \frac{1-y_1}{y_1+u} = \alpha_0\exp(-(\mu+s)(t_1-t_0)), \tag{8}$$

which yields

$$\mu + s = \frac{\ln(\alpha_1/\alpha_0)}{t_1-t_0} \tag{9}$$

The latter can be written alternatively as

$$s = \frac{1}{(1+u)}\frac{\ln(\alpha_1/\alpha_0)}{(t_1-t_0)} \tag{10}$$

Knowing y_0 and y_1 (and therefore also knowing α_0 and α_1), we can thus now find the set of values (s, μ) such that $y_i = y(t_i)\ \ i = 0, 1$.

The latter expression embodies the trade-off between selection and mutation. To understand it, let us notice that the RHS of Equation (10) is equal to $C = \frac{\ln(\alpha_1/\alpha_0)}{(t_1-t_0)}$ if $u = 0$, and it changes very little if u is small. The magnitude of C (which is the estimate of s when $\mu = 0$) as computed from data varies between 0.002 and 0.05 if we disregard the sole negative value -0.059. To significantly influence (i.e., by say 10%) the lowest estimate 0.002, the mutation per cell per nucleotide rate should be equal to at least 0.0002, which is five orders of magnitude higher than the standard human rate. We use per nucleotide rates since we are discussing specific mutation sites in each case.

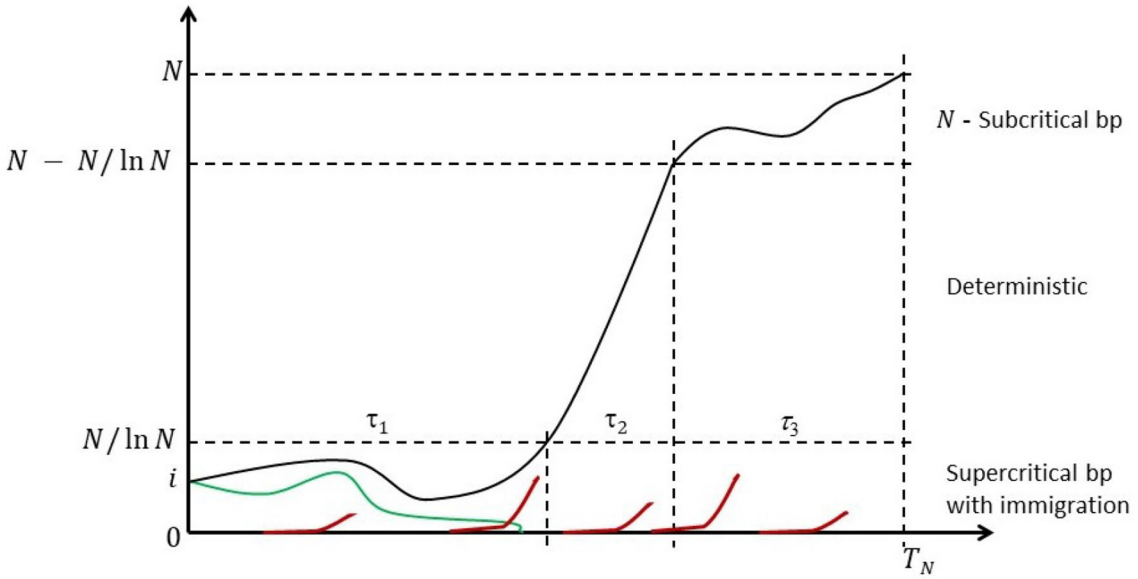

FIGURE 1 | Anatomy of the Moran process with recurrent mutations. Some mutant count trajectories may become transiently extinct (green line), but they will be resurrected by one of the recurring mutations events (red lines). Eventually the mutants are fixed.

In summary, estimates depicted in **Table 1** and **Figure 3** were obtained using the method of the preceding paragraph under assumption $\mu = u = 0$.

2.5. Empirical Observations of Increase in Mutant Receptor Frequency Over Time

Several papers documented the process of increase of the frequency of mutant receptor, based on genome sequencing of bone marrow cells of the SCN patients, in two or more time points. We focus our attention on three of these papers (Beekman et al., 2012; Skokowa et al., 2014; Klimiankou et al., 2019). In these papers, patient data were recorded with changing frequency of mutant receptors over time. Of these cases, we selected only the ones that displayed unambiguous monotonous trend. Theses case are listed in **Table 1**. The estimates that are obtained using expression (10) are ambiguous since the expression provides only a relationship between s and μ. However, if s estimated under the hypothesis that μ is small, which is plausible unless the mutation rate is orders of magnitude higher than normal, then the estimates of s differ only slightly from those obtained under $\mu = 0$, as explained in the preceding subsection. This effect is very similar to that observed in simulations in Wojdyla et al. (2019).

3. RESULTS

3.1. Approximate Mean Expression vs. Simulations

We address here the accuracy of the agreement of the approximate expression (5) for expected value of the Moran process with directional selection and recurrent mutations, with direct simulations. **Figure 2**, presents a comparison of 1,000 simulated trajectories and their mean and standard deviation to the $y(t)$ function. We observe an almost complete agreement of the approximate mean and simulation average, which become indistinguishable with cell count N increasing from 1,000 to 10,000 [compare panels (A–F)]. Additionally, the simulation variance decreases almost inversely proportionally to N.

It is very important to compare the influence of selection coefficient s and mutation rate μ on the expectation of the process. If μ does not exceed 0.001 per generation, its influence can be disregarded, while the influence of s is decisive [panels (A–C)]. Only when μ reaches 0.01, its influence becomes important. In the estimates of the selection coefficient s, based on expression (10), this effect is represented by the coefficient $u = \mu/s$ (also present in expressions for α_0 and α_1), the magnitude of which determines the departure from the case $\mu = 0$.

3.2. Estimates of Selection Coefficients

Table 1 depicts the estimates of the selection coefficient s obtained using Equation (10) with $\mu = 0$. All details of the data used are included in the **Table 1**. The estimates depend on the assumed average interdivision time of the HSC (including some self-renewing CMP). It is assumed to be equal to 1/24 of 1 year (15 days). Changing this assumption leads to different estimates, as it can be tested by modifying the parameter λ in the spreadsheet (λ equals the inverse of the interdivision time). Overall, the estimates span a range from 0 to 0.05, with the exception of patient 13 of publication (Klimiankou et al., 2019) who has a negatively estimated selection coefficient.

Figure 3 depicts estimated selection coefficients $\hat{s}$ from **Table 1**, plotted against the time interval $t_1 - t_0$ between the first and the second instance of sequencing. There seems to exist a negative association between $\hat{s}$ and $t_1 - t_0$. In addition, the MDS cases (red circles), seem to have lower values of $\hat{s}$ than the

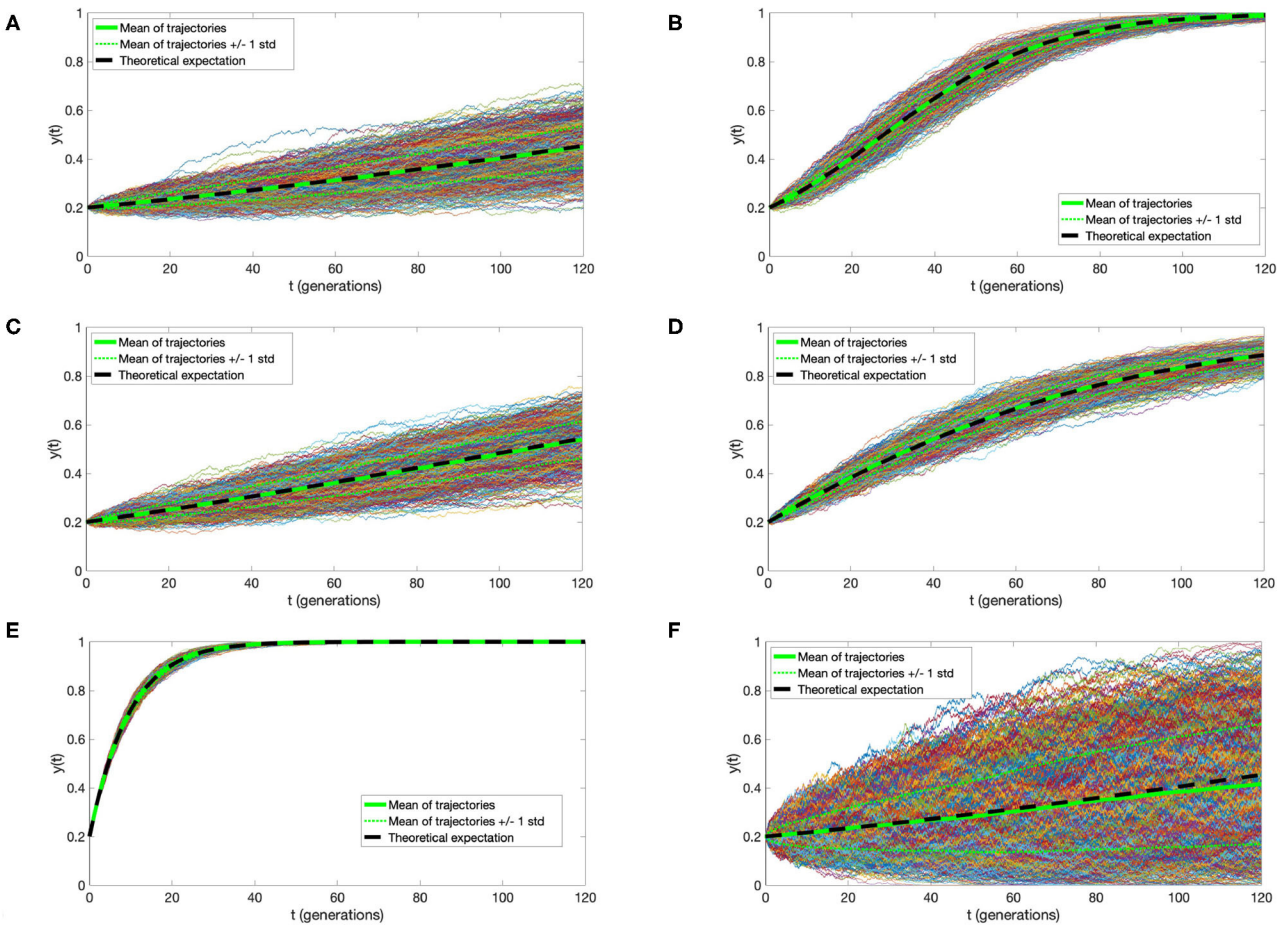

FIGURE 2 | Trajectories of the Moran process with directional selection and recurrent mutation. One thousand simulations were plotted, with thicker green line denoting their mean, thinner green lines mean ± standard deviation, and black dashed line the solution of Equation (4). Parameter values **(A)** $s = 0.01$, $\mu = 0.000001$, **(B)** $s = 0.05$, $\mu = 0.000001$, **(C)** $s = 0.01$, $\mu = 0.001$, **(D)** $s = 0.01$, $\mu = 0.01$, **(E)** $s = 0.01$, $\mu = 0.1$; and $N = 10000$, $t_0 = 0$, $y_0 = 0.2$ in **(A–E)**. **(F)** For comparison, trajectories under $s = 0.01$, $\mu = 0.1$ as in **(A)**, but with $N = 1,000$ are depicted, illustrating the impact of the cell count on variance of the process. In all panels, time is expressed in average cell generation units, μ in inverse time units, and s is dimensionless.

AML cases (green circles). Cases labeled as "CN-MDS/AML" in Klimiankou et al. (2019) are denoted by black circles.

4. DISCUSSION

The results of the present paper provide estimates of the selection coefficients that may underlie the fixation of the mutant G-CSF receptor in the SCN to MDS transition, which are consistent with the range deduced in Wojdyla et al. (2019) based on epidemiological evidence. Let us emphasize that our initial and final fractions of mutant receptor data come from sequencing of samples from patients with SCN. Availability of these sequencing data in papers (Beekman et al., 2012; Skokowa et al., 2014; Klimiankou et al., 2019) is at the stem of our results.

Severe congenital neutropenia is not the only inherited BMF syndrome with predisposition to MDS and AML; however, we believe that SCN provides the most robust and accurate disease to model because acquisition of *CSF3R* mutation is so common (70-80%) as a secondary hit (see discussion of sources in Wojdyla et al., 2019). On the other hand, *TP53* mutations in Shwachman-Diamond are controversial in that the mutations do not augur for transformation (Xia et al., 2018). Furthermore, the prevalence of mutations in *TP53* and in other genes such *RUNX1* in Fanconi anemia or dyskeratosis congenita appears to be much less than that of *CSF3R* in SCN (Chao et al., 2017; Lane, 2017; Kirschner et al., 2018). Despite this, modified Moran model might be applicable to other bone marrow failure syndromes that are associated with leukemia transformation. Since the variant allele frequencies of *CSF3R* is not reported for most of these rare patients, but our model provides an accurate prediction, it is conceivable that uncommon secondary mutations, such as *TP53* or *PPM1D* or *RUNX1*, could be used in our modified Moran model.

We do not use the cases with non-monotonic change in variant receptor frequency. The reason is that, at this range of frequency, Moran model is very unlikely to exhibit persistent reversals. Therefore, it is more likely that factors that cannot be included in the Moran model play a role.

TABLE 1 | Tabular summary of truncated receptor data from three publications, and resulting estimates of selection coefficient for the Moran model without recurrent mutation.

Identifier	References	Figure	Case	Phase	y_0	y_1	$t_1 - t_0$	λ	α_1	α_2	$\hat{s}$	Mutation
1	Klimiankou et al. (2019)	Figure 4E	CN pt. 21		0.13	0.14	0.45	24	0.15	0.16	0.002	Q749
2			CN pt. 11		0.03	0.10	1.15	24	0.03	0.12	0.047	Q754
3			CN pt. 27		0.01	0.06	1.75	24	0.01	0.06	0.053	Q741
4			CN pt. 19		0.13	0.15	2.90	24	0.15	0.18	0.002	Y752
5			CN pt. 13		0.13	0.07	0.45	24	0.14	0.08	-0.059	Q741
6	Beekman et al. (2012)	Figure S4	pt. ph. 1	MDS	0.06	0.11	6.00	24	0.07	0.12	0.004	D715
7			pt. ph. 2	AML	0.11	0.49	9.00	24	0.12	0.97	0.010	D715
8	Skokowa et al. (2014)	Figure S3	pt. 6 ph. 1	MDS	0.24	0.56	13.00	24	0.32	1.27	0.004	Q726X
9			pt. 6 ph. 2	AML	0.56	0.83	3.00	24	1.27	4.88	0.019	Q726X
10			pt. 10	MDS	0.01	0.02	3.00	24	0.01	0.02	0.007	Q726P
11			pt. 16 ph. 1	MDS	0.10	0.30	4.50	24	0.11	0.43	0.012	Q720X
12			pt. 16 ph. 2	AML	0.30	0.33	0.33	24	0.43	0.49	0.017	Q720X
13			pt. 19 ph. 1	MDS	0.28	0.43	2.25	24	0.39	0.75	0.012	Y729X
14			pt. 19 ph. 2	AML	0.43	0.65	0.75	24	0.75	1.86	0.050	Y729X

Figure numbers are these in original publications. MDS/AML column based on the disease history of the patient in Beekman et al. (2012) and ***Figure 3*** *in the Supplement to Skokowa et al. (2014); Klimiankou et al. (2019) is listing "CN-MDS/AML" as a single category. y_0, initial fraction of mutant receptor, y_1, final fraction of mutant receptor, $t_1 - t_0$, duration of time interval (years), λ, average count of cell divisions (year^{-1}), also equal to the inverse of the expected interdivision time (years), α_0, α_1, as defined in Equation (8), and $\hat{s}$, estimate of the selection coefficient. "pt." is patient, "ph." is phase.*

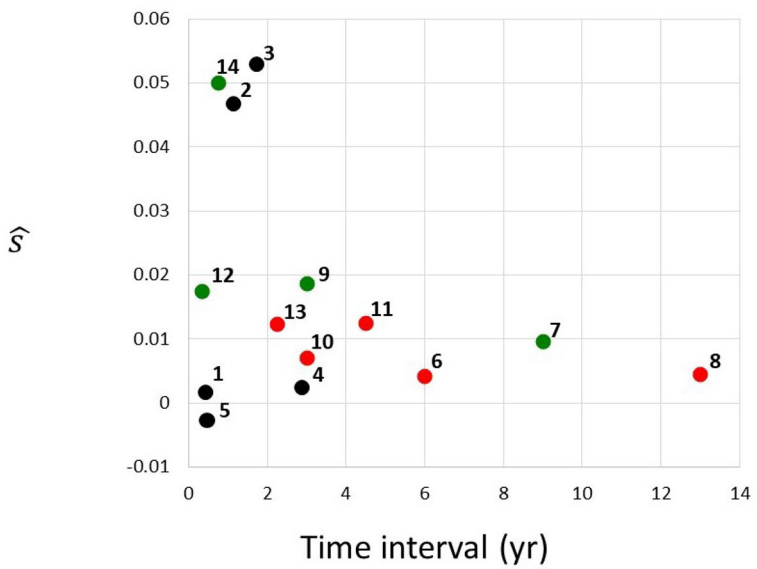

FIGURE 3 | Estimated selection coefficients from **Table 1**, plotted against the time interval between the first and the second instance of sequencing. Red, MDS; green, AML; black, unclassified. Numbering of cases follows the identifiers in **Table 1**.

One of the important questions in understanding cancer evolution is the balance among different genetic forces, such as mutation and selection. The problem has been studied for solid cancers, e.g., by Ling et al. (2015). In essence, mutant frequency in the cell population can increase in a similar way with different (negatively associated) values of s and μ. In particular, if a fit to the mutant frequency increase observed over a time interval is obtained under $\mu = 0$, as in **Table 1**, then under $\mu > 0$, the estimate of s will only be smaller. The decrease will depend on the value of $u = \mu/s$. However, as discussed in the Results, unless the mutation rate in cells is five orders of magnitude higher than in normal cells, i.e., $\mu \approx 10^{-4}$ per cell generation per nucleotide, mutation does not make much difference for the estimates. Therefore, recurrent mutation is an important factor only if the *CSF3R* mutation sites are extremely strong mutational hot-spots. We also examined the magnitude of the correlation coefficient, depending on whether the observed transition was from SCN to MDA or to AML (**Figure 3**), wherever the data have been available. The selection coefficients in the transitions to AML seem to be greater.

An additional effect may be due to the fact that not one, but several types of mutant receptors are observed in MDS. The most frequent is the truncated D715 variant, but there are a number of other, as documented in Beekman et al. (2012), Skokowa et al. (2014), and Klimiankou et al. (2019). Therefore, the basic mutation rate should be multiplied by the number of alternative mutants. Assuming that there are no more than 10 of these mutants, the effect does not seem to play a major role.

It is interesting to observe the apparent effect of ascertainment bias on the data from papers (Beekman et al., 2012; Skokowa et al., 2014; Klimiankou et al., 2019) which we use in our study. **Figure 3** in the Results depicts estimated selection coefficients $\hat{s}$ from **Table 1**, plotted against the time interval $\Delta t = t_1 - t_0$ between the first and the second instance of sequencing. The negative association seen in **Figure 3** may be explained by the fact that the second time at which the frequency of the mutant receptor is observed, arrives sooner if the progression of the disease is faster, i.e., when the coefficient s is higher. This trend may also lead to our estimated s being in general an overestimate, under the assumption that cases with very low s are never diagnosed. Impact of ascertainment bias on estimates of progression in solid cancers has been studied (see, e.g., Kimmel and Flehinger, 1991), and similar methods may be used. However, such study exceeds the framework of the current paper.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. Specifically, MK conceived the study and derived mathematical expressions, SC searched for relevant data and interpreted the model in the context of data, and KD designed and carried out Moran model computations and simulations.

FUNDING

KD acknowledged the support from the Herbert and Florence Irving Institute for Cancer Dynamics and Department of Statistics at Columbia University. MK and SC were supported by NIH grant 5R01HL128173-04.

REFERENCES

Beekman, R., Valkhof, M. G., Sanders, M. A., Van Strien, P. M., Haanstra, J. R., Broeders, L., et al. (2012). Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia. *Blood* 119, 5071–5077. doi: 10.1182/blood-2012-01-406116

Bonilla, M. A., Gillio, A. P., Ruggeiro, M., Kernan, N. A., Brochstein, J. A., Abboud, M., et al. (1989). Effects of recombinant human granulocyte colony-stimulating factor on neutropenia in patients with congenital agranulocytosis. *N. Engl. J. Med.* 320, 1574–1580. doi: 10.1056/NEJM198906153202402

Chao, M. M., Thomay, K., Goehring, G., Wlodarski, M., Pastor, V., Schlegelberger, B., et al. (2017). Mutational spectrum of fanconi anemia associated myeloid neoplasms. *Klin. P¨adiatr.* 229, 329–334. doi: 10.1055/s-0043-117046

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. New York, NY: Springer Science & Business Media.

Gerrish, P. J., and Lenski, R. E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica* 102:127. doi: 10.1023/A:1017067816551

Grimmett, D. and Stirzaker, D. (2001). *Probability and Random Processes*. 3rd Ed. New York, NY: Clarendon Press.

Kennedy, A. L., and Shimamura, A. (2019). Genetic predisposition to mds: clinical features and clonal evolution. *Blood* 133, 1071–1085. doi: 10.1182/blood-2018-10-844662

Kimmel, M., and Corey, S. (2013). Stochastic hypothesis of transition from inborn neutropenia to AML: interactions of cell population dynamics and population genetics. *Front. Oncol.* 3:89. doi: 10.3389/fonc.2013.00089

Kimmel, M., and Flehinger, B. J. (1991). Nonparametric estimation of the size-metastasis relationship in solid cancers. *Biometrics* 47, 987–1004. doi: 10.2307/2532654

Kirschner, M., Maurer, A., Wlodarski, M. W., Ferreira, M. S. V., Bouillon, A.-S., Halfmeyer, I., et al. (2018). Recurrent somatic mutations are rare in patients with cryptic dyskeratosis congenita. *Leukemia* 32, 1762–1767. doi: 10.1038/s41375-018-0125-x

Klimiankou, M., Uenalan, M., Kandabarau, S., Nustede, R., Zeidler, C., Welte, K., et al. (2019). Ultra-sensitive CSF3R deep sequencing in patients with severe congenital neutropenia. *Front. Immunol.* 10:116. doi: 10.3389/fimmu.2019.00116

Lane, D. A. (2017). Correcting the hemophilic imbalance. *Blood* 129, 10–11. doi: 10.1182/blood-2016-11-748822

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6496–E6505. doi: 10.1073/pnas.1519556112

Link, D. C. (2019). Mechanisms of leukemic transformation in congenital neutropenia. *Curr. Opin. Hematol.* 26, 34. doi: 10.1097/MOH.0000000000000479

Oyarbide, U., Topczewski, J., and Corey, S. J. (2019). Peering through zebrafish to understand inherited bone marrow failure syndromes. *Haematologica* 104, 13–24. doi: 10.3324/haematol.2018.196105

Rafei, H., and DiNardo, C. D. (2019). Hereditary myeloid malignancies. *Best Pract. Res. Clin. Haematol.* 32, 163–176. doi: 10.1016/j.beha.2019.05.001

Skokowa, J., Steinemann, D., Katsman-Kuipers, J. E., Zeidler, C., Klimenkova, O., Klimiankou, M., et al. (2014). Cooperativity of RUNX1 and CSF3R mutations in severe congenital neutropenia: a unique pathway in myeloid leukemogenesis. *Blood* 123, 2229–2237. doi: 10.1182/blood-2013-11-538025

Touw, I. P. (2015). Game of clones: the genomic evolution of severe congenital neutropenia. *Hematology Am. Soc. Hematol. Educ. Program* 2015, 1–7. doi: 10.1182/asheducation-2015.1.1

West, A. H., and Churpek, J. E. (2017). Old and new tools in the clinical diagnosis of inherited bone marrow failure syndromes. *Hematology Am. Soc. Hematol. Educ. Program* 2017, 79–87. doi: 10.1182/asheducation-2017.1.79

Wojdyla, T., Mehta, H., Glaubach, T., Bertolusso, R., Iwanaszko, M., Braun, R., et al. (2019). Mutation, drift and selection in single-driver hematologic malignancy: example of secondary myelodysplastic syndrome following treatment of inherited neutropenia. *PLoS Comput. Biol.* 15:e1006664. doi: 10.1371/journal.pcbi.1006664

Xia, J., Miller, C. A., Baty, J., Ramesh, A., Jotte, M. R., Fulton, R. S., et al. (2018). Somatic mutations and clonal hematopoiesis in congenital neutropenia. *Blood* 131, 408–416. doi: 10.1182/blood-2017-08-801985

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 23 September 2020
doi: 10.3389/fphys.2020.511071

Predicting Relapse in Patients With Triple Negative Breast Cancer (TNBC) Using a Deep-Learning Approach

Guangyuan Yu[1,2†], Xuefei Li[2†‡], Ting-Fang He[3], Tina Gruosso[4,5], Dongmei Zuo[4], Margarita Souleimanova[4], Valentina Muñoz Ramos[4], Atilla Omeroglu[6], Sarkis Meterissian[5,7], Marie-Christine Guiot[6,8], Li Yang[1], Yuan Yuan[9], Morag Park[4,5,10], Peter P. Lee[3*] and Herbert Levine[11,12*]*

[1] Department of Physics and Astronomy, Rice University, Houston, TX, United States, [2] Center for Theoretical Biological Physics, Rice University, Houston, TX, United States, [3] Department of Immuno-Oncology, City of Hope Comprehensive Cancer Center, Duarte, CA, United States, [4] Goodman Cancer Research Centre, McGill University, Montreal, QC, Canada, [5] Department of Oncology, McGill University, Montreal, QC, Canada, [6] Department of Pathology, McGill University Health Centre, Montreal, QC, Canada, [7] Department of Surgery, McGill University Health Centre, Montreal, QC, Canada, [8] Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada, [9] Department of Medical Oncology and Therapeutics Research, City of Hope Comprehensive Cancer Center, Duarte, CA, United States, [10] Department of Biochemistry, McGill University, Montreal, QC, Canada, [11] Department of Bioengineering, Northeastern University, Boston, MA, United States, [12] Department of Physics, Northeastern University, Boston, MA, United States

The abundance and/or location of tumor infiltrating lymphocytes (TILs), especially $CD8^+$ T cells, in solid tumors can serve as a prognostic indicator in various types of cancer. However, it is often difficult to select an appropriate threshold value in order to stratify patients into well-defined risk groups. It is also important to select appropriate tumor regions to quantify the abundance of TILs. On the other hand, machine-learning approaches can stratify patients in an unbiased and automatic fashion. Based on immunofluorescence (IF) images of $CD8^+$ T lymphocytes and cancer cells, we develop a machine-learning approach which can predict the risk of relapse for patients with Triple Negative Breast Cancer (TNBC). Tumor-section images from 9 patients with poor outcome and 15 patients with good outcome were used as a training set. Tumor-section images of 29 patients in an independent cohort were used to test the predictive power of our algorithm. In the test cohort, 6 (out of 29) patients who belong to the poor-outcome group were all correctly identified by our algorithm; for the 23 (out of 29) patients who belong to the good-outcome group, 17 were correctly predicted with some evidence that improvement is possible if other measures, such as the grade of tumors, are factored in. Our approach does not involve arbitrarily defined metrics and can be applied to other types of cancer in which the abundance/location of $CD8^+$ T lymphocytes/other types of cells is an indicator of prognosis.

Keywords: triple negative breast cancer (TNBC), relapse prediction, immunofluorescence images, tumor-infiltrating T cells, machine-learning

OPEN ACCESS

Edited by:
George Bebis, University of Nevada, Reno, United States

Reviewed by:
Hermann Frieboes, University of Louisville, United States
Anthony Elias, University of Colorado, United States

***Correspondence:**
Morag Park
morag.park@mcgill.ca
Peter P. Lee
plee@coh.org
Herbert Levine
h.levine@northeastern.edu

[†] These authors have contributed equally to this work

[†] Present address:
Xuefei Li, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *09 November 2019*
Accepted: *31 August 2020*
Published: *23 September 2020*

Citation:
Yu G, Li X, He T-F, Gruosso T, Zuo D, Souleimanova M, Ramos VM, Omeroglu A, Meterissian S, Guiot M-C, Yang L, Yuan Y, Park M, Lee PP and Levine H (2020) Predicting Relapse in Patients With Triple Negative Breast Cancer (TNBC) Using a Deep-Learning Approach. Front. Physiol. 11:511071. doi: 10.3389/fphys.2020.511071

INTRODUCTION

In nearly all cancer types, it has been demonstrated that patients with higher numbers of tumor infiltrating lymphocytes (TILs) in their solid tumors usually have better prognosis in term of the overall survival as well as the disease-free survival (Gooden et al., 2011). Most studies focused on $CD8^+$ T lymphocytes (Sato et al., 2005; Galon et al., 2006; Sharma et al., 2007; Mahmoud et al., 2011; Rahbar et al., 2015; Carstens et al., 2017), which can recognize and kill cancer cells with specific antigens (Martínez-Lostao et al., 2015). For example, in colorectal cancer and melanoma (Pagès et al., 2009; Galon et al., 2016), the ratio of T-cell density in the core of a tumor (CT) to that at the invasive margin (IM), i.e., the Immunoscore, has demonstrated its power to indicate prognosis.

However, due to the heterogeneity of the abundance of TILs within tumors, selection of the threshold-value for defining patient categories can be ambiguous. Furthermore, the exact threshold-value as well as the choice of a suitable metric (such as the Immunoscore defined in colorectal cancer) can vary from one type of cancer to another. In order to reduce such ambiguities, a machine-learning approach can be helpful due to its parameter-free formulation. Indeed, there have recently been a few successful applications of machine-learning approaches in cancer research: Agarap (2017) compared six machine-learning (ML) algorithms on the Wisconsin Diagnostic Dataset for a binary prediction problem of benign vs. malignant tumor; Heidari et al. (2018) developed a machine-learning approach to predict short-term cancer risk by comparing asymmetry of the left vs. right breasts; Saltz et al. (2018) trained a convolutional neural network (CNN) to recognize TILs in the H&E histological images from the TCGA database and generated TIL maps of TCGA samples; here the authors showed that TIL densities and spatial structure can be associated with features such as tumor types, immune subtypes, and tumor molecular subtypes.

In this work, using immunofluorescence (IF) images of $CD8^+$ T lymphocytes and cancer cells, we developed a machine-learning approach to predict the risk of relapse for patients with Triple Negative Breast Cancer (TNBC). We first used tumor-section images of 24 patients with either poor or good outcome to train a specific convolutional neural network (CNN) called MXNet. Subsequently, the trained CNN was applied to predict whether a patient is expected to have a good or poor outcome in an independent test set. This test set is a distinct cohort of TNBC patients (29 of them) from a different medical center.

An overall workflow of our approach is shown in **Figure 1**. Our results, to be detailed below, show that the 6 patients (out of 29) who belong to the poor outcome group are all correctly predicted by our procedure; for the 23 patients (out of 29) who belong to the good-outcome group, 17 of them are correctly predicted. This number might increase if additional factors such as tumor grade or nodal involvement are taken into account.

Compared to other metrics, such as the overall $CD8^+$ T-cell density or the infiltration level into tumor islets of $CD8^+$ T cells, we show that our machine-learning approach has better predictive power. Due to the automatic nature of our procedure, we believe this approach could be readily applied to other types of cancer where the abundance/location of $CD8^+$ T lymphocytes (or any type of non-cancer cells) is likely to be an indicator of prognosis. Furthermore, our algorithm does not rely on clinical training or experience, which means this method could be widely adopted.

MATERIALS AND METHODS

Patients and Specimens

There are two independent cohorts in our study: one from the City of Hope (CH, 24 patients in total) and the other from McGill University (MG, 29 patients in total). All these patients had TNBC and underwent surgery. All 55 patients were treatment-naive before the surgery. Details of the sample collection for the two cohorts are described in the following.

For the City of Hope cohort, samples from patients diagnosed with triple-native breast cancer, invasive ductal carcinoma (IDC) type, and treated at COH from 1994 to 2015 were retrieved. At the time of surgery, none of the patients had prior treatment. Archived formalin-fixed paraffin-embedded (FFPE) tumor tissues were sectioned into 5 μm thick slides and baked onto glass microscope slides and labeled with anti-pan cytokeratin (AE1/AE3, Dako) and anti-CD8 (SP16, Biocare) using the Opal TSA system (Akoya Bioscience). Stained samples were further counterstained with DAPI, cover-slipped with ProLong® Gold Antifade mounting media, and imaged by Vectra automated imaging system.

For the McGill cohort, it is a subset of the cohort published in Gruosso et al. (2019). Samples were collected from patients undergoing breast surgeries at the McGill University Health Centre (MUHC) between 1999 and 2012. All tissues were snap-frozen in O.C.T. Tissue-Teck Compound within 30 min of removal. For the purposes of this study, samples were selected according to the following criteria: therapy-naive at time of surgical excision, clinically documented lack of expression/amplification of ER, PR and HER2, a histological subtype assignment of invasive ductal carcinoma [not otherwise specified) (IDC (NOS)] and availability of matched formalin-fixed paraffin-embedded (FFPE) tumor blocks. Information regarding clinical variables and disease course (follow-up) was obtained through review of Medical Records at the MUHC. Five micro meter sections from frozen tissue were prepared for each sample, subjected to routine hematoxylin and eosin (H&E) staining, and evaluated by an attending clinical pathologist with expertise in breast tissue to identify invasive, *in situ* and normal components. Cancer cells and $CD8^+$ T cells were labeled by pan-cytokeratin (PanCK) and CD8 immune-fluorescence (IF) antibodies, respectively. Primary antibodies for immunofluorescence (IF) as well as the IF protocol were described and detail in Li et al. (2019).

Patients were divided into two outcome-groups: any patient who had a relapse within 3 years of the surgery belongs to the poor-outcome group; conversely, any patient who survived and did not have a relapse within 5 years belongs to the good-outcome

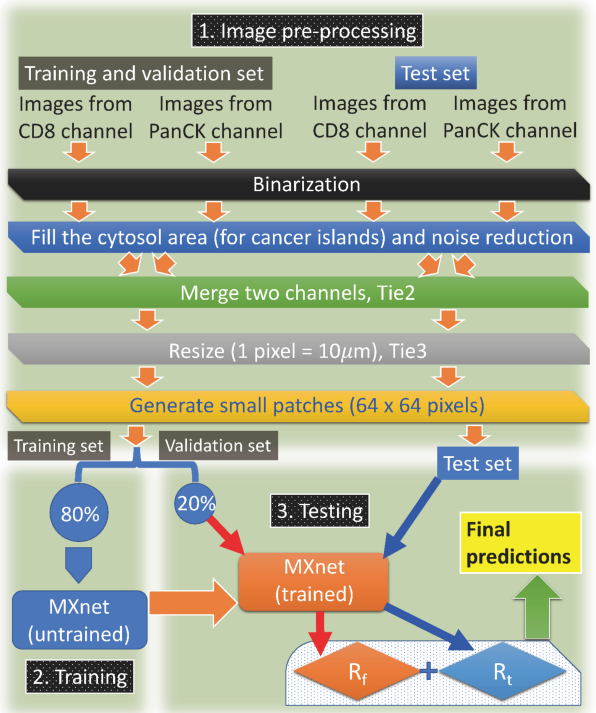

FIGURE 1 | The flowchart. Overall description of our machine-learning approach.

group. Apparently, there is 1 year gap between the good and poor outcome groups and rare individuals that fall in-between are dropped. We chose this particular standard because we want to ensure that patients in the two groups are well-separated.

Pre-processing of the Immunofluorescent Images

The original resolution of our images is 0.5 μm (CH) and 0.975 μm (MG), respectively. We first binarize the two IF channels for all images: pixels with an IF intensity in the top 90% (among all pixels within an individual image) are assigned as 1 and others are assigned as 0. Then the binary images of cancer cells are labeled in white (1) and black (0) and the binary images of $CD8^+$ T cells are labeled in red (1) and black (0). In addition, we remove the isolated connected-areas with an area smaller than 200 μm^2, which might be due to the noise in the IF signals. Furthermore, for the binary images of cancer cells, since PanCK is a cell-surface marker, the cytosol of cancer cells might be black. In order to faithfully represent the area of cancer islets, we automatically fill holes with an area smaller than 200 μm^2. The two corresponding binary images for cancer cells and $CD8^+$ T cells are subsequently merged together to generate images (Tie2) for further processing.

For deep-learning, the suitable image-size is usually 32 to a few hundred pixels in one dimension, while our original images can be around 20,000 pixels in on dimension. Therefore, we resize all Tie2 images to the same scale for the two independent cohorts so that 1 pixel in each image (Tie3) corresponds to 10 μm. Next, each image (Tie3) is divided into smaller (adjacent) patches (64 × 64 pixels). If the patch has a number of white ($PanCK^+$) pixels that is less than a quarter of the total number of pixels in this patch, it will be discarded. If the patch has no $CD8^+$ pixels in it, it will also be discarded. After this step, we now have final images (Tie4) for the machine-learning procedure later. Note that according to our standard, some of the areas at the invasive margin of a tumor might be discarded because of the lack of cancer cells, though some parts of the invasive margins are kept in the analysis. An example illustrating the areas kept in the analysis is shown in **Supplementary Figure S1**. Furthermore, in the Supplementary Material, we tested the effects of changing the spatial resolution of patches in detail (**Supplementary Table S1**). The results indicate that our baseline procedure is optimal for the current dataset. Finally, we also demonstrated that discarding patches without T cells does not substantially change our original results (**Supplementary Table S2**).

Training

For the training set, we use patches from patients in the CH cohort. A detailed table of the total number of small patches for each patient in the training set can be found in **Table 1**. Briefly,

TABLE 1 | Number of patches derived from the images of the training set and clinical information.

ID	Patch #	Outcome	Rtn	Grade	Nodal-status
P1	44	Good	0.17–0.41	III	No
P2	811	Good	0.92–0.97	III	No
P3	810	Good	0.87–0.99	III	No
P4	220	Good	0.98–1	III	No
P5	834	Good	0.99–1	III	NA
P6	226	Good	0.93–1	III	No
P7	171	Good	0.72–0.84	II	Yes
P8	92	Good	0.77–1	III	No
P9	30	Good	0.80–0.87	III	No
P10	387	Good	0.68–0.77	III	No
P11	471	Good	0.75–0.94	III	Yes
P12	228	Good	1	III	No
P13	260	Good	0.29–0.58	II	No
P14	30	Good	0.33–0.59	II	NA
P15	243	Good	0.31–0.45	III	Yes
P16	218	Poor	0	III	No
P17	84	Poor	0–0.22	III	No
P18	129	Poor	0.04–0.31	NA	Yes
P19	82	Poor	0.06–0.11	III	No
P20	290	Poor	0–0.02	III	No
P21	113	Poor	0–0.06	III	No
P22	144	Poor	0.08–0.24	III	Yes
P23	256	Poor	0.13–0.22	III	No
P24	235	Poor	0–0.08	III	No

Column 4 is range of R_{tn} derived from the validation set (20% of patches from each patient in the CH cohort) after 5 rounds of testing using randomly selected patches. For each patient, R_{tn} is the percentage of the patches that is predicted to be from patients with the good outcome.

there are 9 patients with the poor outcome and 15 patients with the good outcome.

For the CH cohort, 80% of the small patches from each patient are used for training. The other 20% are used for the threshold-selection procedure (validation) described later. There are more patches from patients with the good outcome in the training set (4857 vs. 1551); hence, in order to make the training balanced between samples from poor and good outcomes, we generated 3 additional copies of each small patch from patients with the poor outcome and added them to the training set. Examples of patches are shown in **Figures 2A, B**. In addition, to test whether generating additional copies would bias the model prediction, we investigated other methods of balancing the number patches from the two prognostic groups in the training set (**Supplementary Table S3**). The details can be found in the (**Supplementary Table S4**), and there was no substantial difference between the two balancing methods, thus overfitting was less of a concern.

For the deep-learning network, we use "deepflow" from MXNet (Chen et al., 2015; Krizhevsky et al., 2017) for this project. The code that we developed can be found at https://github.com/xun6000/deepflow. Note that the procedure to feed these images into MXNet is a bit complex, and the github file contains the command line instruction to do so properly. In the following, we will describe the procedure of the training algorithm.

With one input small patch (64 × 64 pixels), a probability can be computed by MXNet to determine whether this patch is from a patient with the poor outcome. Since we know *a priori* where this patch comes from, based on the difference between this probability and its known value (0 for the good outcome or 1 for the poor outcome), the internal parameters of MXNet are updated automatically using the optimization algorithm called RMSProp (Ruder, 2016). We choose the input parameters for RMSProp as learning_rate = 0.0005, weight_decay = 0.01, factor_epoch = 10, lr_factor = 0.25. In addition, the mini-batch size for RMSProp is related to the performance of network, i.e., larger mini-batch size will make the net harder to find the global minimum (Keskar et al., 2016). The mini-batch size is the number of images that are fed together to MXNet for one round of update for the internal parameters in MXNet. Specifically, we use 20 images as our mini-batch size.

One epoch is defined as the process in which all patches were served as the input to train the MXNet based on the defined outcome (the other input information). We run 100 epochs to train the MXNet after which the accuracy should have been stabilize (**Figure 2C**). If we select the cut-off probability between a poor-outcome patch and a good-outcome one to be 0.5, i.e.,

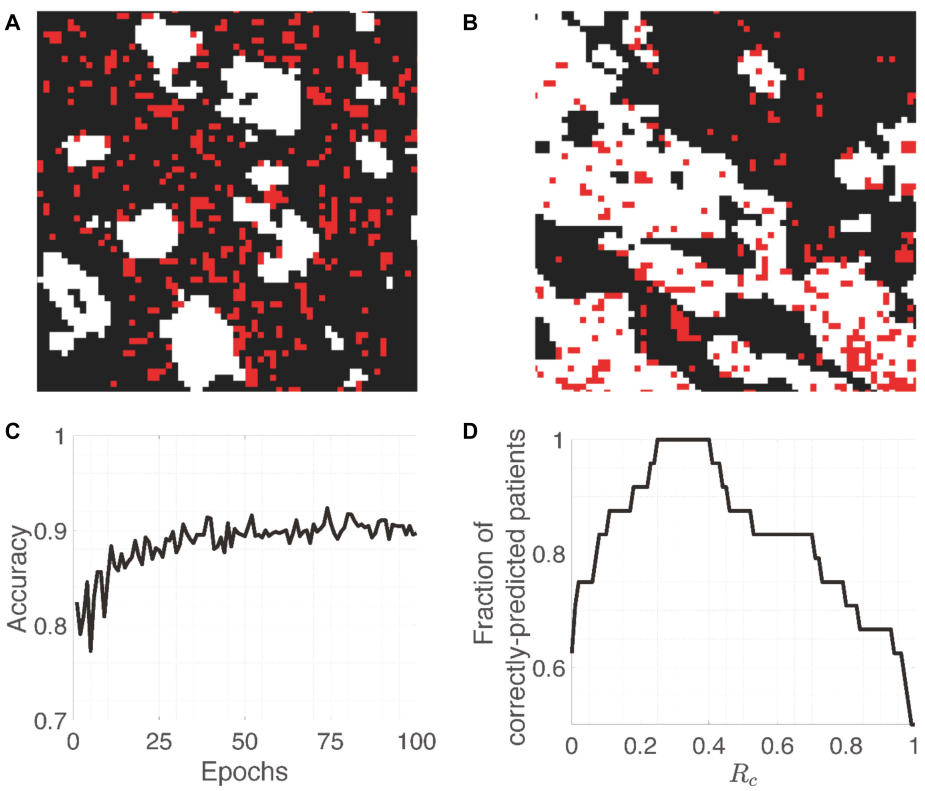

FIGURE 2 | Training images and accuracy. **(A,B)** Representative patches (64 × 64 pixels) from patients with the poor and good outcome, respectively. While and red pixels represent PanCK-positive (cancer cells) and CD8-positive ($CD8^+$ T cells) areas, respectively. **(C)** Evolution of the accuracy on training patches as a function of Epochs. **(D)** The fraction of correctly-predicted patients as a function of the cut-off percentage (R_c) of patches that are classified as arising from a patient with the good outcome.

probability >0.5 means that the patch is from a patient with the poor outcome, then the training and validation accuracy are around 0.9 and 0.85, respectively. This is understandable because individual patches from patients with the good outcome may resemble those from patients with the poor outcome, and vice versa. The whole training process takes 3–4 h on a Tesla K80 NVIDIA GPU.

RESULTS

Patient Stratification Criteria Based on Deep-Learning Predictions

After training, the network can predict whether a small patch is from a patient with the good or poor outcome, which is named as a "good" or "poor" patch. We then used the trained CNN to predict the remaining small patches (the 20% mentioned before) so that we can determine the percentage (Rtn) of the "good" patches in each patient. For a given cut-off percentage Rc, we discover whether Rtn of a patient is higher or lower than Rc. If we assume that any patient whose Rtn < Rc is predicted to have the poor outcome (and vice versa), we will achieve some degree of accuracy of the prediction (Ac) by the trained MXNet. We then change Rc until Ac reaches the maximum, selecting the percentage (Ropt) that best-separates the two groups of the 24 patients in the training cohort.

Since the 20% of small patches from the CH cohort are randomly selected and the CNN can also have some randomness, the Rtn for each patient can vary for different realizations (column 4, **Table 1**). Furthermore, for each realization, there is a range of Rc that gives the same accuracy. After going through 5 realizations, we find the Ropt should be between 0.14 and 0.40. Most of the times (4 out 5), we can find a Ropt that makes a perfect separation (**Figure 2D**). Specifically, we select the average of the values that can give a perfect separation in those 5 realizations, which is Rc =0.30.

Predicted Prognosis for the Independent Cohort

Next, the percentage (Rt) of good patches can be determined for each patient in the test set (MG cohort). Thus, these patients will be predicted to have the poor (Rt < Rc) or good (Rt > Rc) outcome. Our results show that for the 6 patients (out of 29) who belong to the poor-outcome group, they are all correctly predicted by our approach; for the 23 patients (out of 29) who belong to the good-outcome group, 17 of them are correctly predicted (**Table 2**).

Furthermore, the imperfect prediction could be due to factors other than the $CD8^+$ T cells. For example, if we also integrate other clinical information such as the nodal status and the tumor grade, for the 6 patients that are not correctly predicted by our approach, there are 2 of them (Patients 18 and 19) whose tumor grade is lower (grade II). Note that all other 27 patients in the MG cohort have grade III tumors (**Table 2**); and all 3 patients with grade II tumors in the CH cohort belong the good-outcome group (**Table 1**). In addition, it is valuable to notice that the Nodal status of the other 4 incorrect predictions (rows highlighted in red in **Table 2**) is either No or NA, whereas the Nodal status of correctly-predicted poor-outcome patients is mostly Yes or NA with only one exception (P5) out of 6 patients (P1–P5 and P22). However, in the poor-outcome group of the training cohort (CH), only 2 out of 9 patients have a positive Nodal status. Therefore, combining the prediction using our approach with the information on the tumor grade and the nodal status, the accuracy might be improved significantly. This needs to be tested in the future for a data-set with more complete annotation regarding Nodal status.

In addition, we tested the degree to which the accuracy of our prediction is diminished if the size of the section samples decreases to half or a quarter of the original samples (columns 4 and 5 in **Table 2**). For 3 of the patients (out of 29), because of the inhomogeneity of the tumors, the prediction is not perfectly robust to the region of selection.

TABLE 2 | Predicted vs. actual outcome for the test set (MG cohort).

ID	Predicted-outcome	Ground-truth	Half	Quarter	Grade	Nodal-status
P1	Poor	Poor	Poor	Poor	III	Yes
P2	Poor	Poor	Poor	Poor	III	Yes
P3	Poor	Poor	Poor	Poor	III	NA
P4	Poor	Poor	Poor	Poor	III	NA
P5	Poor	Poor	Poor	Poor	III	No
P6	Good	Good	Good	Good	III	No
P7	Good	Good	Good	Good	III	No
P8	Good	Good	Good	Good	III	Yes
P9	**Poor**	Good	Poor	Poor	III	No
P10	**Poor**	Good	Poor	Good	III	NA
P11	**Poor**	Good	Good	Good	III	No
P12	Good	Good	Good	Good	III	NA
P13	Good	Good	Poor	Good	III	No
P14	Good	Good	Good	Good	III	No
P15	Good	Good	Good	Good	III	No
P16	Good	Good	Good	Good	III	NA
P17	Good	Good	Good	Good	III	NA
P18	Poor	Good	Poor	Poor	II	No
P19	Poor	Good	Poor	Poor	II	No
P20	Good	Good	Good	Good	III	No
P21	Good	Good	Good	Good	III	No
P22	Poor	Poor	Poor	Poor	III	Yes
P23	Good	Good	Good	Good	III	No
P24	Good	Good	Good	Good	III	Yes
P25	Good	Good	Good	Good	III	No
P26	Good	Good	Good	Good	III	No
P27	Good	Good	Good	Good	III	Yes
P28	Good	Good	Good	Good	III	Yes
P29	**Poor**	Good	Poor	Good	III	NA

The actual outcome of individual patient is shown in column 3. Prediction of our machine-learning approach using the full-, half-, and quarter-size section samples are shown in columns 2, 4, and 5. Information on the Grade and Nodal status of the tumors is shown in columns 6 and 7. Rows that are labeled in blue are for patients with Grade II tumors. Rows that are labeled in red are for patients with a poor-outcome prediction (column 2) but a good-outcome in reality (column 3). NA stands for not applicable.

When defining poor outcome as positive, the confusion matrix equals as follows:

	ground truth 1	ground truth 0
predict 1	6	6 (Type II error)
predict 0	0 (Type I error)	17

The recall = 6/6 = 1, precision = 6/12 = 0.5

Comparison of the Prediction-Accuracy Between the Deep-Learning Method and CD8$^+$ T-Cell Number or Infiltration Level

The machine-learning approach gives a reasonably good prediction of the 3 year relapse likelihood. We tried to compare the accuracy of this prediction with other possible metrics, such as the density of CD8$^+$ T cells inside cancer-cell islands, the absolute numbers of CD8$^+$ T cells and cancer cells, etc. In **Figure 3**, there exists an apparent overlap between the poor- and good-outcome group using the density of CD8$^+$ T cells inside cancer-cell islands (**Figure 3A**) or the absolute numbers of CD8$^+$ T cells and cancer cells (**Figure 3B**). Note that for the CH cohort, using our deep-learning approach (**Figure 2D**), we can have a perfect separation between the two groups of patients. Nevertheless, if we manually select the "perfect" cut-offs according to the data, as demonstrated by the dash lines in **Figure 3**, the maximum stratification accuracy considering the density of CD8$^+$ T cells inside cancer-cell islands or the absolute numbers of CD8$^+$ T cells and cancer cells will be 85 and 87%, respectively. Even though the accuracy using these methods is comparable to our deep-learning approach, the selection of the cut-off is not statistically justified.

To further test whether other clinical data could better predict the outcome, we performed hierarchical clustering and principal component analysis based on the clinical characteristics collected for the CH cohort (see **Supplementary Tables S5, S6**). In short, these analyses did not give an adequate separation of the two prognostic groups, whereas our current baseline procedure was successful. More details are provided in section 5 of the **Supplementary Material**.

Information Extracted by Our Machine-Learning Approach in Determining the Outcome

Finally, we describe the information extracted by our machine-learning approach in determining the outcome. The results suggest that the absolute density or number of CD8$^+$ T cells might not be the most important factor but instead the (relative) infiltration of CD8$^+$ T cells is more crucial. As demonstrated in **Figure 4**, we generally observe that patches from patients with the good outcome have more red pixels (CD8$^+$) as compared to their counterparts. However, for patches from patients with poor outcome, we can still observe patch samples with many CD8$^+$ T cells (red pixels) but these pixels are outside of the cancer islands (white areas); and we observe patches with fewer red pixels (CD8$^+$ T cell) but most of them are inside white areas (cancer islands), from patients with the good outcome. In summary, our results indicate that the relative infiltration level of CD8$^+$ T cells into cancer-cell islands is the most important factor to determine whether a patch would be predicted to arise from a patient with the good outcome.

DISCUSSION

In this work, we developed a machine-learning approach to predict the 3 year relapse likelihood based on IF images of cancer cells and CD8$^+$ T cells. While the approach is effective with an accuracy 86% or higher, there is still room to further improve the accuracy of our approach by including additional features that can be measured in addition to CD8 makers. In the following, we will further discuss possible candidates.

First, the molecular states of CD8$^+$ T cells can be diverse (Guo et al., 2018), including different levels of exhaustion (Wherry and Kurachi, 2015). Therefore, it would be more informative

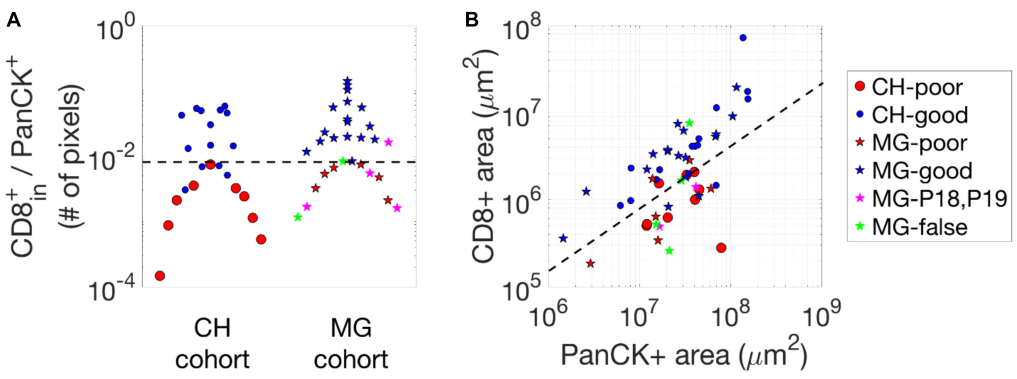

FIGURE 3 | Statistics of CD8$^+$ pixels in images. **(A)** The density of CD8$^+$ pixels inside cancer-cell islands, i.e., the number of CD8 + pixels divided by the number of PanCK$^+$ pixels. **(B)** CD8-positive and PanCK-positive areas (in μm^2) of the two cohorts. The black dash lines in each figure are selected manually to separate the groups (poor- vs. good-outcome) of the patients which give rise to the highest accuracy when comparing to the actual outcome.

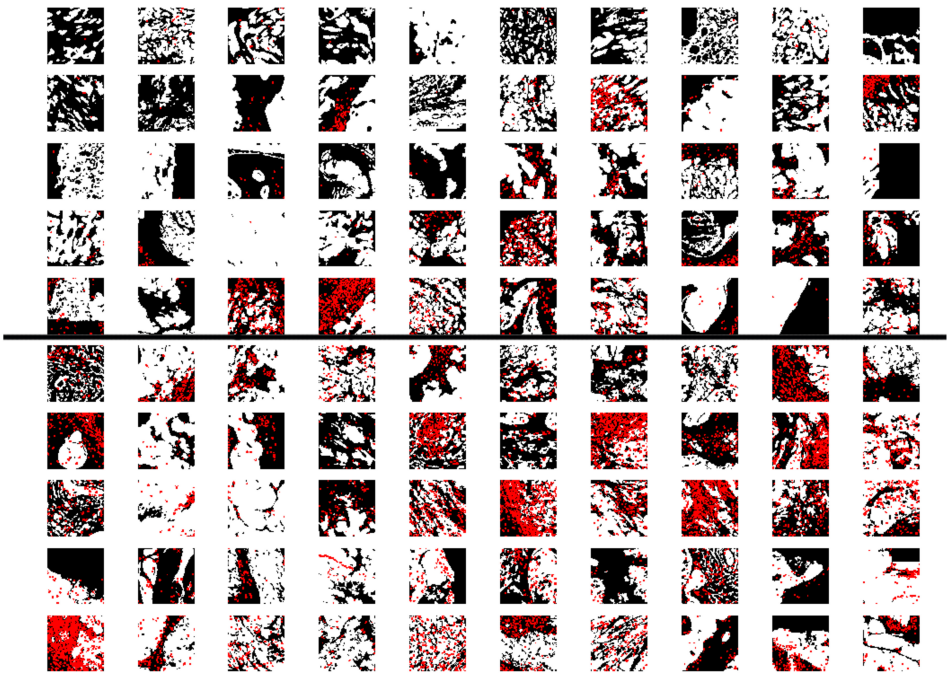

FIGURE 4 | Examples of patches. Patches that are predicted to be from patients with the poor outcome (upper 50) or good outcome (lower 50). The size of each patch is 64 pixels (64 pixels, with 1 pixel = 10 μm White and red pixels represent PanCK-positive (cancer cells) and CD8-positive ($CD8^+$T cells) areas, respectively. For the patches in the upper 50, we still observe many samples have a lot of red pixels (CD8-positive areas), however, compared to patches in the lower 50, most of the red dots are in the tumor stroma (black areas) instead of cancer-cell islands (white areas).

to also assess the functional states of individual $CD8^+$ T cells via additional markers, such as Granzyme B, EOMES, T-bet, PD-1, and so on. By incorporating these additional features of $CD8^+$ T cells, the prediction-accuracy of our approach could be further improved.

Secondly, there are other types of immune cells beyond $CD8^+$ T cells that have been demonstrated to have predictive power in patient prognosis, such as tumor-associated macrophages (Zhang et al., 2012) as well as Tregs (Shang et al., 2015). Having an image with the information regarding several types of these immune cells again might improve the accuracy of our approach.

Thirdly, properties of cancer cells also matter in predicting outcome, in addition to the spatial information of cancer cells and $CD8^+$ T cells. For example, in our test cohort, we found that two good-outcome patients who are predicted to have poor outcome actually have Grade II tumors, where the proliferation rate of tumor cells is low. In fact, all patients with Grade II tumors belong to the good-outcome group. Another related possibility still to be investigated is the EMT status of cancer cells. This is motivated by the fact that markers for epithelial-to-mesenchymal transition (EMT) of cancer cells are usually indicative for progression of disease (Tsoukalas et al., 2017; Luo et al., 2018).

The current formulation of our algorithm is a binary class problem in nature. We might imagine changing this binary class problem to a triple class problem, where the 3rd class is the patches that we currently discarded. This can make the application of the algorithm much simpler. The details can be found in the **Supplementary Material**.

Our final remark concerns one weak point of our approach. Due to the heterogeneity of tumors, it is not currently possible to accurately predict the outcome of a patient based on a small sub-sample of one part of a tumor. For example, for P13 shown in **Table 2**, using half of the patches from the tumor gives the opposite prediction, compared to using a quarter or the whole section. Changing the location of the selection of the half can also change the prediction; again this is due to the heterogeneity of the tumor itself. Therefore, to predict prognosis based on a limited number of patches, it is important to sample multiple sites of a tumor instead of from only one part.

In summary, we developed a machine-learning approach that can predict the 3 year relapse risk of TNBC based on the IF images of cancer cells and $CD8^+$ T cells, with an accuracy 86% or higher. The advantage of this approach is that the standards to determine outcome are relatively objective. Therefore, it can readily be applied to other types of samples. With more training samples and more features measured, this approach should reach even higher prediction accuracy and become useful for rapid clinical prognosis.

DATA AVAILABILITY STATEMENT

The original images used and analyzed in this study can be accessed via the link provided in the 'data.txt' file on Github (https://github.com/xun6000/deepflow/blob/master/data.txt).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the City of Hope Institutional Review Board (IRB #14346) via the City of Hope Biospecimen Repository McGill University Health Centre (MUHC) review board (study approval SUR-2000-966 and SUR-99-780). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GY, XL, and HL designed the study. AO, SM, M-CG, YY, and PL included and followed patients. T-FH, MS, and VR provided clinical data. T-FH, DZ, and TG performed the IHC analyses. GY and XL developed the algorithm. LY developed the triple-class algorithm. MP, PL, and HL supervised the study. GY, XL, T-FH, TG, MP, PL, and HL wrote the manuscript. All coauthors read and approved the final manuscript.

FUNDING

This work was supported by the National Science Foundation Center for Theoretical Biological Physics (NSF PHY-1427654), NSF DMS-1361411, Stand Up to Cancer, Breast Cancer Research Foundation, and The V Foundation. This study was also supported by funding from CQDM (Consortium québécois sur la découverte du medicament/Québec Consortium for Drug Discovery) (52811) and the National Institutes of Health (NIH 2P01CA097189-06) (to MP) and Merck, Sharpe & Dohme Corp./McGill Faculty of Medicine Grants for Translational Research (238371) (to MP). The breast tissue and data bank at McGill University is supported by funding from the Database and Tissue Bank Axis of the Réseau de Recherche en Cancer of the Fonds de Recherche du Québec-Santé and the Québec Breast Cancer Foundation (to MP). TG has been supported by the Charlotte and Leo Karassik Oncology fellowship.

ACKNOWLEDGMENTS

We are grateful to B. Clavieri (Microscopy Imaging Lab, University of Toronto) for scanning IF slides. We thank Jo-Ann Bader and the staff at the Histology Core Facility at the Goodman Cancer Research Centre for assistance with sample preparation. We thank R. Deagle and E. Tse-Luen (Advanced BioImaging Facility, McGill University) for help with sample imaging. We thank members of the Departments of Surgery, Pathology and Anaesthesia at the McGill University Health Centre for their assistance with sample collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.511071/full#supplementary-material

REFERENCES

Agarap, A. F. (2017). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. *arxiv* [Preprint], Avaliable at: https://arxiv.org/abs/1711.07831 (accessed December 12, 2018).

Carstens, J. L., Correa de Sampaio, P., Yang, D., Barua, S., Wang, H., Rao, A., et al. (2017). Spatial computation of intratumoral T cells correlates with survival of patients with pancreatic cancer. *Nat. Commun.* 8:15095. doi: /10.1038/ncomms15095

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., et al. (2015). MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. *arxiv* [Preprint]. Avaliable at: https://arxiv.org/abs/1512.01274 (accessed October 25, 2017).

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-PageÌs, C., et al. (2006). Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 313, 1960–1964. doi: 10.1126/science.1129139

Galon, J., Fox, B. A., Bifulco, C. B., Masucci, G., Rau, T., Botti, G., et al. (2016). Immunoscore and immuno-profiling in cancer: an update from the melanoma and immunotherapy bridge 2015. *J. Transl. Med.* 14:273. doi: 10.1186/s12967-016-1029-z

Gooden, M. J., de Bock, G. H., Leffers, N., Daemen, T., and Nijman, H. W. (2011). The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. *Br. J. Cancer* 105, 93–103. doi: 10.1038/bjc.2011.189

Gruosso, T., Gigoux, M., Manem, V. S. K., Bertos, N., Zuo, D., Perlitch, I., et al. (2019). Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. *J. Clin. Invest.* 129, 1785–1800. doi: 10.1172/jci96313

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., et al. (2018). Global characterization of T cells in nonsmall-cell lung cancer by single-cell sequencing. *Nat. Med.* 24, 978–985. doi: 10.1038/s41591-018-0045-3

Heidari, M., Khuzani, A. Z., Hollingsworth, A. B., Danala, G., Mirniaharikandehei, S., Qiu, Y., et al. (2018). Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. *Phys. Med. Biol.* 63:035020. doi: 10.1088/1361-6560/aaa1ca

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: generalization gap and sharp minima. *arxiv* [Preprint]. Avaliable at: https://arxiv.org/abs/1609.04836 (accessed March 6, 2018).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Li, X., Gruosso, T., Zuo, D., Omeroglu, A., Meterissian, S., Guiot, M. C., et al. (2019). Infiltration of CD8+ T cells into tumor cell clusters in triple-negative breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 116, 3678–3687. doi: 10.1073/pnas.1817652116

Luo, Y., Yu, T., Zhang, Q., Fu, Q., Hu, Y., Xiang, M., et al. (2018). Upregulated N-cadherin expression is associated with poor prognosis in epithelial-derived solid tumours: a meta-analysis. *Eur. J. Clin. Invest.* 48:e12903. doi: 10.1111/eci.12903

Mahmoud, S. M., Paish, E. C., Powe, D. G., Macmillan, R. D., Grainge, M. J., Lee, A. H. G., et al. (2011). Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. *J. Clin. Oncol.* 29, 1949–1955. doi: 10.1200/jco.2010.30.5037

Martínez-Lostao, L., Anel, A., and Pardo, J. (2015). How do cytotoxic lymphocytes kill cancer cells? *Clin. Cancer Res.* 21, 5047–5056. doi: 10.1158/1078-0432.ccr-15-0685

Pagès, F., Kirilovsky, A., Mlecnik, B., Asslaber, M., Tosolini, M., Bindea, G., et al. (2009). In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.* 27, 5944–5951. doi: 10.1200/jco.2008.19.6147

Rahbar, M., Naraghi, Z. S., Mardanpour, M., and Mardanpour, N. (2015). Tumor-infiltrating CD8+ lymphocytes effect on clinical outcome of mucocutaneous melanoma. *Indian J. Dermatol.* 60:212.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arxiv* [Preprint]. Avaliable at https://arxiv.org/abs/1609.04747 (accessed September 20, 2018).

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., et al. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* 23, 181.e7–193.e7. doi: 10.1016/j.celrep.2018.03.086

Sato, E., Olson, S. H., Ahn, J., Bundy, B., Nishikawa, H., Qian, F., et al. (2005). Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18538–18543. doi: 10.1073/pnas.0509182102

Shang, B., Liu, Y., Jiang, S. J., and Liu, Y. (2015). Prognostic value of tumor-infiltrating FoxP3+ regulatory T cells in cancers: a systematic review and meta-analysis. *Sci. Rep.* 5:15179. doi: 10.1038/srep15179

Sharma, P., Shen, Y., Wen, S., Yamada, S., Jungbluth, A. A., Gnjatic, S., et al. (2007). CD8 tumor-infiltrating lymphocytes are predictive of survival in muscle-invasive urothelial carcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 104, 3967–3972. doi: 10.1073/pnas.0611618104

Tsoukalas, N., Aravantinou-Fatorou, E., Tolia, M., Giaginis, C., Galanopoulos, M., Kiakou, M., et al. (2017). Epithelial-mesenchymal transition in non small-cell lung cancer. *Anticancer. Res.* 37, 1773–1778. doi: 10.21873/anticanres.11510

Wherry, E. J., and Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nat. Rev. Immunol.* 15, 486–499. doi: 10.1038/nri3862

Zhang, Q. W., Liu, L., Gong, C. Y., Shi, H. S., Zeng, Y. H., Wang, X. Z., et al. (2012). Prognostic significance of tumor-associated macrophages in solid tumor: a meta-analysis of the literature. *PLoS One* 7:e50946. doi: 10.1371/journal.pone.0050946

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 19 October 2020
doi: 10.3389/fphys.2020.583333

Digital Pathology Analysis Quantifies Spatial Heterogeneity of CD3, CD4, CD8, CD20, and FoxP3 Immune Markers in Triple-Negative Breast Cancer

Haoyang Mi[1], Chang Gong[1], Jeremias Sulam[1,2], Elana J. Fertig[1,3], Alexander S. Szalay[4,5], Elizabeth M. Jaffee[3,6], Vered Stearns[3], Leisha A. Emens[7], Ashley M. Cimino-Mathews[3,8] and Aleksander S. Popel[1,3]*

[1] Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [2] Johns Hopkins Mathematical Institute for Data Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, United States, [3] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, United States, [4] Henry A. Rowland Department of Physics and Astronomy, Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD, United States, [5] Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, United States, [6] The Bloomberg~Kimmel Institute for Cancer Immunotherapy, Johns Hopkins School of Medicine, Baltimore, MD, United States, [7] Department of Medicine/Hematology-Oncology, Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, PA, United States, [8] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

OPEN ACCESS

Edited by:
Russell C. Rockne, City of Hope National Medical Center, United States

Reviewed by:
Hermann Frieboes, University of Louisville, United States
Krithika Kodumudi, Moffitt Cancer Center, United States

***Correspondence:**
Haoyang Mi
hmi1@jhmi.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *14 July 2020*
Accepted: *24 September 2020*
Published: *19 October 2020*

Citation:
Mi H, Gong C, Sulam J, Fertig EJ, Szalay AS, Jaffee EM, Stearns V, Emens LA, Cimino-Mathews AM and Popel AS (2020) Digital Pathology Analysis Quantifies Spatial Heterogeneity of CD3, CD4, CD8, CD20, and FoxP3 Immune Markers in Triple-Negative Breast Cancer. Front. Physiol. 11:583333. doi: 10.3389/fphys.2020.583333

Overwhelming evidence has shown the significant role of the tumor microenvironment (TME) in governing the triple-negative breast cancer (TNBC) progression. Digital pathology can provide key information about the spatial heterogeneity within the TME using image analysis and spatial statistics. These analyses have been applied to CD8+ T cells, but quantitative analyses of other important markers and their correlations are limited. In this study, a digital pathology computational workflow is formulated for characterizing the spatial distributions of five immune markers (CD3, CD4, CD8, CD20, and FoxP3) and then the functionality is tested on whole slide images from patients with TNBC. The workflow is initiated by digital image processing to extract and colocalize immune marker-labeled cells and then convert this information to point patterns. Afterward invasive front (IF), central tumor (CT), and normal tissue (N) are characterized. For each region, we examine the intra-tumoral heterogeneity. The workflow is then repeated for all specimens to capture inter-tumoral heterogeneity. In this study, both intra- and inter-tumoral heterogeneities are observed for all five markers across all specimens. Among all regions, IF tends to have higher densities of immune cells and overall larger variations in spatial model fitting parameters and higher density in cell clusters and hotspots compared to CT and N. Results suggest a distinct role of IF in the tumor immuno-architecture. Though the sample size is limited in the study, the computational workflow could be readily reproduced and scaled due to its automatic

nature. Importantly, the value of the workflow also lies in its potential to be linked to treatment outcomes and identification of predictive biomarkers for responders/non-responders, and its application to parameterization and validation of computational immuno-oncology models.

Keywords: digital pathology, image informatics, spatial patterns, breast cancer, tumor heterogeneity, immuno-architecture, QuPath

INTRODUCTION

Triple negative breast carcinoma (TNBC) is an aggressive form of breast cancer that is negative for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER-2). Treatments for TNBC have historically been confined to surgery, radiation and chemotherapies due to the lack of biologic targets which enable endocrine therapy (ER and PR) and targeted therapy (HER2) in other subgroups of breast cancer. Recent studies deciphering the role of the immune system in cancer revealed a significant effect of the tumor microenvironment (TME) in modulating the tumor progression, especially how the tumor hijacks the anti-inflammatory mechanism of inhibitory immune checkpoint molecules to develop its immune resistance and evasion capability. These studies inspired the emergence of anti-cancer immunotherapy by promoting the host anti-tumor immunity. This idea has led to an array of successful treatments against cancers. Recently, the Impassion130 study demonstrated that first-line treatment with atezolizumab and nab-paclitaxel resulted in an overall survival benefit in patients with advanced programmed death-ligand 1 (PD-L1) positive TNBC, and this is a new standard of care (Schmid et al., 2020). Meanwhile, immune checkpoint blockade also advances the treatment outcomes for other cancer types including melanoma (Long et al., 2016), non-small cell lung cancer (Melosky et al., 2020), and renal-cell carcinoma (Motzer et al., 2019). However, not all patients experience therapeutic benefit from immunotherapy. Heterogeneity within the TME may account for some of the variability in patient response to immunotherapies (Klemm and Joyce, 2015). Therefore, characterization of the tumor immuno-architecture is a critical step toward understanding the complex interplay between pro-and anti-tumor immunity.

Previous research indicates that the existence of tumor-infiltrating lymphocytes (TILs) possesses a unique predictive or prognostic value in different types of cancer (Haanen et al., 2006; Kawai et al., 2008; Suzuki et al., 2010; Ladányi, 2015; Lianyuan et al., 2018), including TNBC (Denkert et al., 2018) as the infiltrating profiles are associated with favorable patient outcomes. Specifically, high overall survival scores are often accompanied by high levels of cytotoxic CD8+ T cells, whereas forkhead box protein 3+ (FoxP3+) regulatory T cells and Type 2/Type 17 CD4+ helper T cells (Th_2 and Th_{17} T cells) diminish this effect (Fridman et al., 2011). Therefore, monitoring the distribution of different TILs and their associations should yield insights into how cancer progresses. Hence, further studies are needed to elucidate the underlying mechanisms from the spatio-temporal perspective.

Digital pathology is an emerging discipline that allows quantitative analysis of digital images of histological specimens using computational approaches. Digitized images provide easy access for pathologists to high-resolution histological data with typically gigapixel content (Al-Janabi et al., 2012). Therefore, tissue contexts are well preserved and amenable to computer-assisted techniques for quantitative analysis of the spatial immuno-architecture. We started the development of a workflow for spatial statistical analysis in a previous study with a single immune marker for CD8+ T cells (Gong et al., 2018). The process started with image processing, during which tumor specimen images were segmented to map CD8+ T cells into a Cartesian coordinate system, then the density of point pattern within each subregion (first-order property in spatial statistics terminology) were gathered to reveal spatial variations. For each subregion, cell coordinates were converted into spatial point patterns, then Thomas cluster process was fitted to clustered patterns (assessed by complete spatial randomness test). For the entire point pattern, cell clustering morphometrics were performed. Collectively, fitted clustering parameters and morphometric measurements were harvested to characterize the immuno-architecture (Gong et al., 2018). Other investigators examined the CD8+ T cells infiltration profile by constructing profiles of cell pixel density vs. distance from the tumor boundary and then used a computational model to interpret the data (Li et al., 2019). Alternative approaches introduced the pattern of tumor cells as a reference to measure the infiltration of lymphocytes. CD8+ T cells and the tumor cells were colocalized and then a series of metrics were introduced to measure the spatial interactions such as quantifying the nearest neighbor distribution function for two different cell types using spatial G-function (Barua et al., 2018) and evaluating the spatial clustering using Morisita index and Getis-Ord hotspots analysis (Yuan, 2016). These studies adopt different metrics to interpret spatial distributions of various entities extracted from pathology data; spatial heterogeneity is a universally observed hallmark of cancer whether it is gauged in terms of density, model fitting parameters, clustering size, or infiltration level. Such variations can also be linked to treatment outcomes to better understand the effects of the TME and to assist clinicians in making more accurate diagnoses.

The integration of image processing, statistical analysis, and computational biology has already shown to be powerful in characterizing and interpreting the spatial heterogeneity in multiple tumor types including breast cancer (Brown et al., 2014; Mani et al., 2016; Altan et al., 2018; Du et al., 2019; Wong et al., 2019). Nevertheless, high dimensional quantitative measurements of the interaction between immune markers

and assessment of their spatial correlations have not been conducted. Such metrics will not only provide additional layers of tumor spatial heterogeneity information valuable for patient stratification, but also allow us to better understand the mechanisms behind the formation of the patterns we observe in the TME. In this study, we propose a multi-module workflow to quantify the spatial patterns of five immune markers that control the functional status of T cells, on consecutive pathology sections: CD3, CD4, CD8, CD20, and FoxP3, and from a small patient population (n = 5) with TNBC to obtain statistically and pathologically meaningful results. For each patient, our workflow starts with image processing, evaluation of point patterns from three perspectives, and implementation of region-based characterization. To the best of our knowledge, our analysis evaluates the heterogeneity of TNBC in a broad immune context for the first time, therefore paving the way to the identification of reliable predictive biomarkers and the design of innovative therapies when properly correlated with clinical outcomes.

In addition, the cell densities derived from the workflow can be converted to 3D numerical densities to facilitate development and calibration of spatially resolved computational immuno-oncology models. For example, 3D densities of different cell types calculated from point patterns can be utilized to populate *in silico* computational agent-based models (ABMs) that have the potential of predicting spatio-temporal TME (Gong et al., 2017; Norton et al., 2017, 2018, 2019). Such three-dimensional ABM could be combined with Quantitative Systems Pharmacology (QSP) models for whole patient to enable mechanistic systems biology modeling of different drugs or combinations (Cheng et al., 2017; Rieger et al., 2018; Bai et al., 2019; Jafarnejad et al., 2019; Milberg et al., 2019; Wang et al., 2019, 2020; Sové et al., 2020).

MATERIALS AND METHODS

Pathology Specimen Materials and Methods

This study was approved by the Institutional Review Board of the Johns Hopkins Medical Institutions. Digitally scanned slides from a subset of previously described primary breast tumors were evaluated (Cimino-Mathews et al., 2016). Briefly, formalin fixed, paraffin embedded blocks from surgically resected primary breast carcinomas with no prior neoadjuvant chemotherapy were randomly selected from the pathology archives at Johns Hopkins Medical Institutions (associated response data to treatment not available). TNBC was defined as negative for ER, PR, and the HER-2. Consecutive sections (approximately 5 μm each) from whole tumor were individually stained for CD3 (mouse monoclonal, clone PS1, catalog no. ORG-8982; Leica Microsystems, Bannockburn, IL, United States), CD4 (rabbit monoclonal, clone Sp35, catalog no. 790-4423; Ventana Medical Systems), CD8 (mouse monoclonal, clone C8/C8144B, catalog no. 760-4250; Cell Marque, Rocklin, CA, United States), CD20 (monoclonal, clone MS/L26, catalog no. 760-2531; Ventana Medical Systems, Tucson, AZ, United States), and FoxP3 (mouse monoclonal, clone 236A/E7, catalog no. 14-4777-80, dilution 1:50; eBioscience; San Diego, CA, United States). Immunohistochemically labeled slides were scanned at 20× objective (0.49 microns/pixel) using the Aperio Scanscope AT (Aperio/Leica Biosystems, Vista, CA, United States). Five (5) cases of TNBC were selected for this current study based on intact tissue integrity on the scanned images (i.e., complete cross sections and lack of tissue folds). To simplify the analysis, two tumor islands in Case 1 are split into Case 1A (upper left island) and Case 1B (lower right island). **Supplementary Figure S1** shows a representative biomarker panels across Cases 1–5.

Computational Methods

The overall workflow includes a central module and four submodules (**Figure 1**). First, the stained (positive) cell nuclei are detected and the tissue annotation module is launched to identify regions of normal tissue (N), central tumor (CT), and invasive front (IF). In this module, image registration is performed for each case on the five slides with different labels, and raw coordinates obtained from image segmentation are mapped to the same reference coordinate system using the transformation matrix. Cell densities are the output of this step, and therefore can be directly visualized using 3D and waterfall graphical representations to visualize intra- and inter-tumor heterogeneity. Also, cell density vs. distance profiles can be constructed. In the spatial point model-fitting module, for each slide, the full point pattern is divided into smaller patches with overlaps, and the Thomas point process model is then fitted based on subregion data if complete spatial randomness (CSR) hypothesis is rejected for this subregion (Baddeley et al., 2015). In the clustering and morphometrics module, for the full point pattern of each slide, the cell clusters are detected using a hierarchical clustering algorithm (Tang et al., 2016). For each detected cluster, morphometrics including convexity, circularity, and eccentricity are calculated and recorded. In the correlation analysis module, a clustering and degree of colocalization-based method is applied to quantify the correlations between immune marker pairs. All spatial statistical measurements used in this study have shown promising application values in the context of digital pathology. Results from these metrics are classified based on tissue type regions and then used for statistical comparison to reveal intra-tumoral heterogeneity. Such workflow is repeated for all cases, thus capturing inter-tumoral heterogeneity.

Image Processing

Cell nucleus segmentation and coordinate extraction from IHC slides

Stained (positive) nucleus segmentation is performed using software platform QuPath (v0.2.0-m10) (Bankhead et al., 2017). QuPath is selected for this study because it is a flexible open-source software with well-managed version control and technical support, and it is capable of a wide range of digital pathology analyses. As the IHC staining may vary both between and within each case, the image processing workflow is initiated by manual correction to the stain estimates for each whole slide image using the 'estimate stain vectors' function. Nucleus detection is then carried out using an unsupervised watershed algorithm with custom parameters tuned on a validation set of WSIs from Case

1, 2, and 3. This built-in algorithm has been implemented by a wide range of peer-reviewed studies (Bankhead et al., 2018; Zhang et al., 2018; Acs et al., 2019a,b; Ferré et al., 2019; Habets et al., 2019; Kather et al., 2019; Santiago et al., 2019; Berben et al., 2020; Blagih et al., 2020; Tsakiroglou et al., 2020). Importantly, the performance was found equivalent to commercial software and pathologists' manual annotations (Bankhead et al., 2018; Acs et al., 2019b; Berben et al., 2020). While the nuclei are identified, centroid coordinates are recorded to represent the cells' location. Afterward, pseudo cell objects are formed by expanding the nuclei boundaries for 7.5 μm and then a list of features is calculated based on intensity and morphometry measurements. For each IHC biomarker, 25 regions of interests (ROIs) are randomly selected for Cases 1–3 (75 in total) and a Random Tree classifier (Breiman, 2001) is trained using the aforementioned features by annotating regions in a subset of subregions. The classification results are updated in the form of color marks when each annotation is drawn. The classifier is then validated on the remaining WSIs to ensure the robustness. A low-resolution image in the testing set is shown in **Figure 2A**. Exemplar segmentation and classification results are shown in **Figures 2B,C**. Segmentation and classification settings are shown in **Supplementary Material** and **Supplementary Tables S1, S2**.

To evaluate the performance of the image segmentation algorithm, 20 subregions are sampled from each slide for each case using the random sampling method. **Figure 2D** shows four exemplar subregions for performance evaluation, where red outlines indicate QuPath segmentation results and green dots indicate manual approach. For each subregion, we also manually detect labeled cells, and then measure the sensitivity (recall) and precision of our algorithm. Results indicate that there is a strong correlation between manual and automatic approaches (Spearman's correlation coefficient $\rho = 0.978$). Details of the evaluation of QuPath can be found in the **Supplementary Material**.

Registration and coordinate transformation of IHC slides

The pathology images available for this study were single label IHC slides produced with consecutive sections from each tumor excision. In this process, z-axis difference for each section, location and rotation when placed onto slides, as well as possible folding of tissue during preparation, all contributed to discrepancies between coordinate systems of each slide from the same patient. These discrepancies are minimized by image registration. As the cutting sequence of these immune marker slides was unknown, all slides are treated equally and the CD4+ slide is selected as the reference for all cases. Global automatic registration by Matlab application "Registration Estimator" is first performed for all pairs (**Figure 3A**). The registration accuracies are manually assessed based on tissue overlap level: among 20 registration pairs, we find that global registration produces high accuracy for five pairs; for the remaining, the local registration is required, which is performed using software Icy (De Chaumont et al., 2012) (**Figure 3B**). Both global and local registrations generate transformation matrices for the corresponding regions. These matrices can be used to estimate registration accuracy. First, global registrations are performed for all pairs as the baseline. Next, transformation matrices generated from global and local registrations are used to register tissue contours, separately. Dice Similarity Coefficients (DSCs) (Guy et al., 2019) are calculated, respectively, and cumulative results are collected (**Figure 3C**). Finally, we compared the registration accuracy by performing the Wilcoxon rank-sum test between global and local DSC groups (**Figure 3D**) and observed a significant improvement ($\boldsymbol{p}$ = **4.90e-3**). Results also show that the average global registration DSC scores for those five pairs (0.916) are very similar to the average local registration DSC scores (0.917). Technical details on global and local registrations and performance evaluation can be found in the **Supplementary Material**.

Measuring Intra- and Inter-tumoral Heterogeneity

Region characterization based on pathologist's annotations

For each case, the breast cancer pathologist (AMC-M) annotated (outlined) the tumor boundary, which was considered the 'ground truth' for the present analysis. Green contours in **Figure 4A** and **Supplementary Figure S2** indicate the annotations for Case 1 and Cases 2–5. Annotations are converted into coordinate sets and then registered to the reference slide (CD4+ slide) using the transformation matrices. Each annotation is a closed curve so that the corresponding coordinate set can form a polygon. Next, we create a score map by overlaying the five coregistered polygons (from five slides, one for each immune marker) and recording the number of polygons each pixel of the WSI resides within. Now that each pixel is assigned a score, we apply a smoothing filter to the score map and threshold all pixels with scores exceeding 2.5 to determine the consensus tumor boundary. Finally, we obtain a dense region, the contour of which functions as the averaged boundary between normal tissue [sometimes referred to as stroma (Tyekucheva et al., 2017)] and CT. Next, we buffered the boundary with 0.5 mm inward and outward to create the IF, which separates the normal tissue (N) and CT with a band of 1 mm (Pages et al., 2009; Halama et al., 2011; Hendry et al., 2017). Note that in some studies the region is considered 0.5 mm wide (Gong et al., 2018; Li et al., 2019); we will show below the quantitative implications of either assumption. Afterward, we remove those pixels that fall within the IF from the tumor mask, thus the contour of remaining points gives the outline of CT. Similarly, to extract the normal tissue, we computed the mean value of the RGB channels for each pixel. To exclude background and noise, we set a customized threshold to rule out high-intensity pixels. The remaining pixels contain complete WSI foreground information, and the normal region can be easily extracted when IF and CT pixels are removed. For all pixels associated with each region, we obtain the outline to form a polygon to represent the region. In this step, specimen images and annotations from pathologist are the only inputs. Pixel coordinate maps are generated using python. All subsequent computations are performed using R: the point-in-polygon test is performed using 'point.in.polygon' function from R package 'sp' (Pebesma and Bivand, 2005); the outlines are generated using 'concaveman' function from R package 'concavemann' (Gombin et al., 2017).

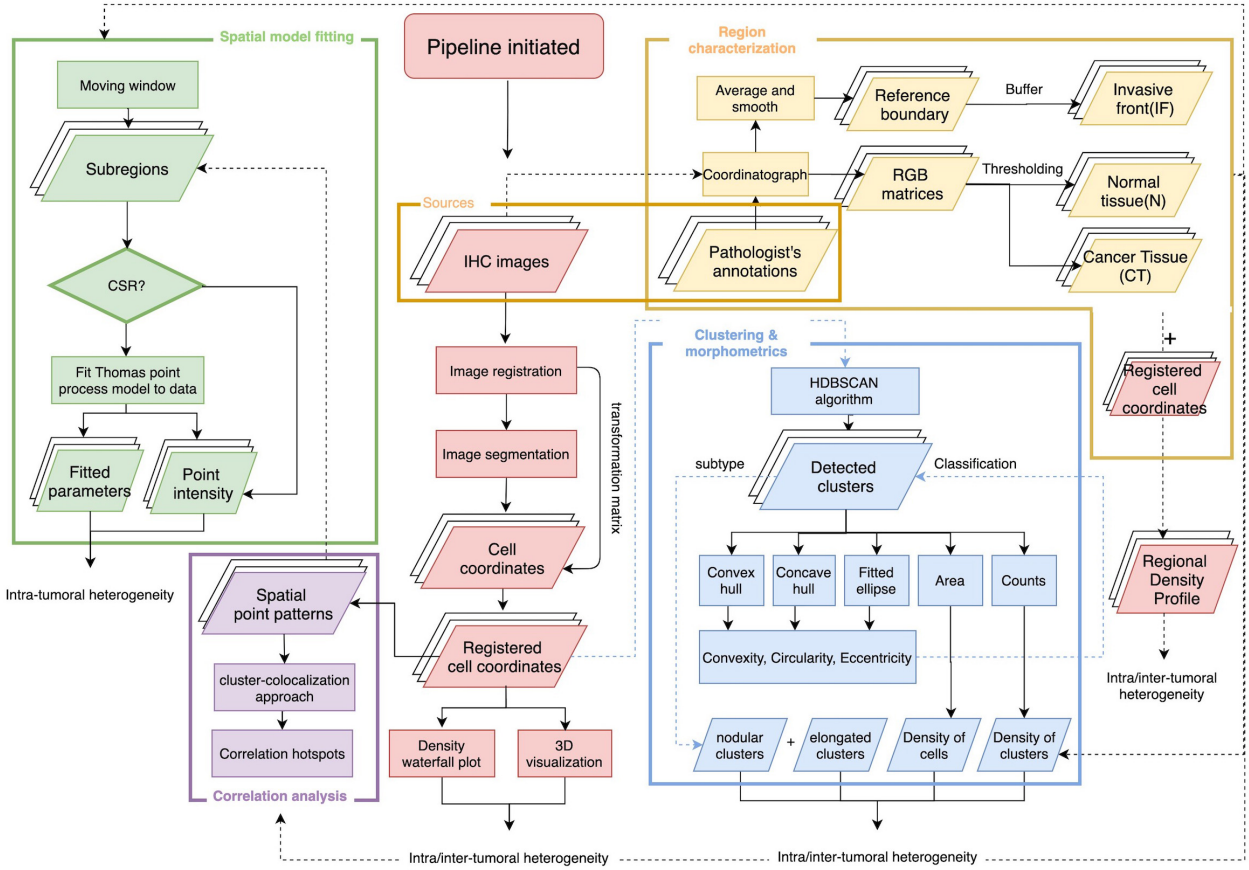

FIGURE 1 | Overall workflow of spatial pattern quantification for immune markers. The workflow is initiated by two steps, first the image processing for IHC slides to extract coordinates of immune marker-labeled cells; second, the tissue type regions are characterized based on pathologist's annotations of the tumor boundary and original IHC images. The results from these two steps construct regional density profiles. The point patterns are fed to all remaining submodules to quantify intra- and inter-tumoral heterogeneity. In the spatial model-fitting submodule, point patterns within subregions are tested for CSR, and the Thomas model is fitted to data if test is rejected. In the clustering and morphometric submodule, point clusters are detected and multiple shape descriptors are calculated for each cluster. In the correlation analysis submodule, a cluster-colocalization based method gauges the spatial distributions near each point to identify correlation hotspots, where highly correlated immune marker pairs are engaged in. For each slide, the collective results from the aforementioned metrics capture intra-tumoral heterogeneity and the analyses repeated across all cases capture inter-tumoral heterogeneity.

Whole slide image partitioning and extraction of first-order properties

We partition the WSI into subregions for local spatial analysis using a moving rectangular window with edge lengths of x_{window} and y_{window}, which traverses the WSI with step size of x_{step} and y_{step}. The window size should be large enough to capture local density variations, and sufficiently small to have multiple subregions and stationary underlying point pattern processes. Based on these considerations, we performed fractal analysis (see **Supplementary Material** and **Supplementary Figure S3**) and determined the window lengths and step size as $x_{window} = y_{window}$ =0.4 mm and $x_{step} = y_{step}$ =0.2 mm (complete discussion is in **Supplementary Material**). In this study, we define all individual rectangular areas that the moving window has scanned as subregions. As the window is moving through the whole slides, first-order properties such as number and density of points are recorded for subsequent visualization and local statistical analysis.

Measuring the heterogeneity with spatial entropy measurement

A form of Shannon's entropy (Claramunt, 2005) uses the entropy as a measure of diversity of density of multiple point species in space. This modified version incorporates the factor of distance. Assuming that the increase of distance between same type of points and the decrease of distance between different types of points will result in the increase of entropy, this spatial entropy is defined as:

$$H_{SC} = -\sum_{i=1}^{n} \frac{d_i^{int}}{d_i^{ext}} p_i \log_2 p_i \qquad (1)$$

where d_i^{int} is the average Euclidean distance between all points of type i; d_i^{ext} is the average Euclidean distance between all points of type i and the points of other types; p_i is the percentage of type i within the subregion.

For each case, we first map registered and reference point patterns into a Cartesian coordinate system. Next, we calculate

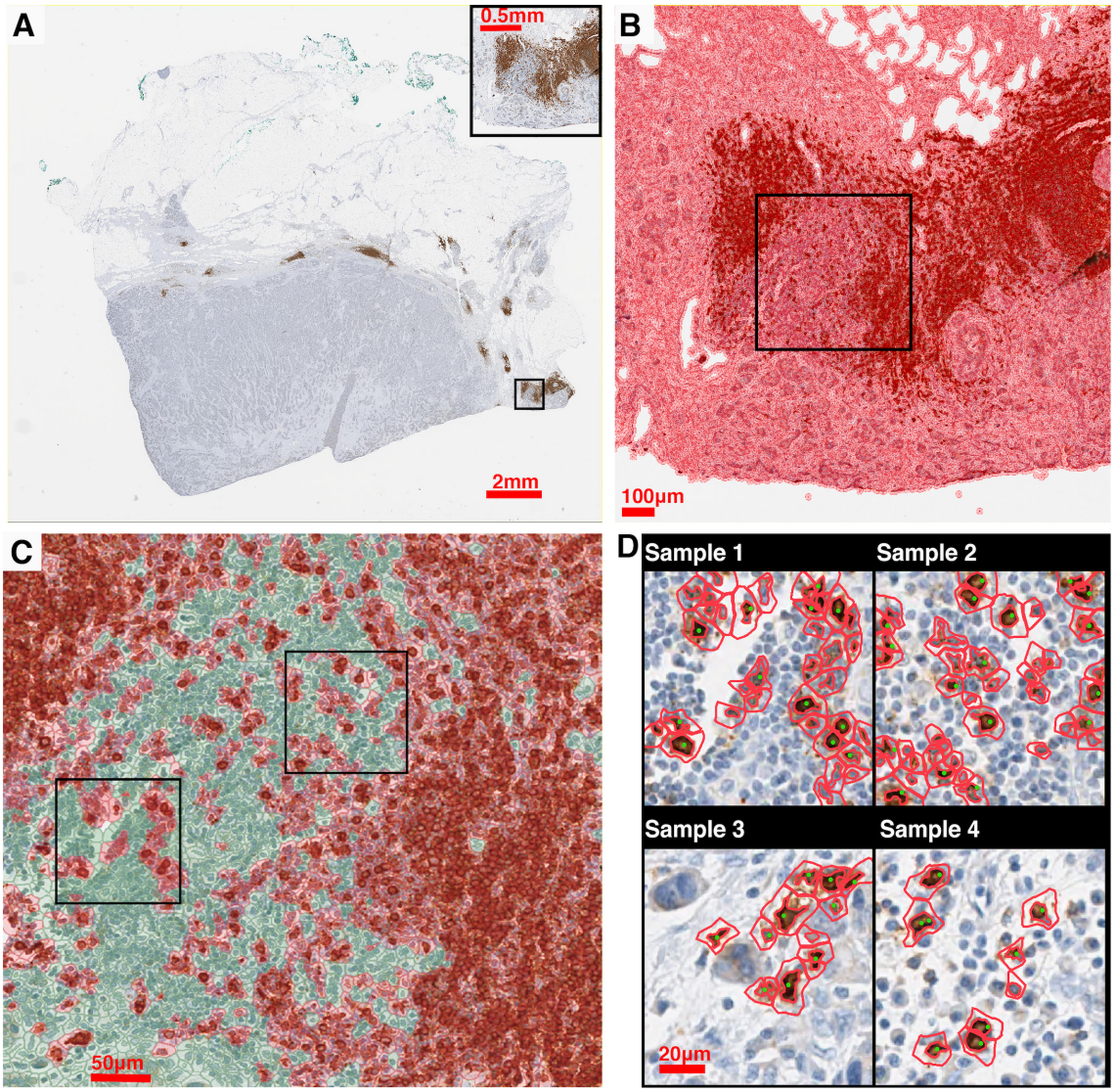

FIGURE 2 | IHC image segmentation, classification using QuPath and performance validation. **(A)** Original IHC image (Case 5, CD20). **(B)** Nucleus detection results of the subregion marked in panel **(A)**. In this process, nucleus boundaries are detected (inner red contour) and then expanded outward for 7.5 μm (outer red contour) to form a pseudo cell. Then morphology and intensity features of the cell objects are fed to the classifier to identify positive cells. **(C)** Classification results of the subregion marked in panel **(C)** using the corresponding classifier. Green: negative (non-stained) cells. Red: positive (stained) cells. **(D)** Results of manual detection (green dots) and algorithmic detection (red outlines) are mapped together to evaluate the performance. 20 subregions are randomly sampled across the slide. Two subregions marked in panel **(C)** are shown as examples.

the H_{sc} for the multi-type point pattern within subregions. Collective results are then classified based on tissue type associated with their locations. Then we examine intra- and inter-tumoral heterogeneity of spatial entropy in each tissue type by showing their distributions with a series of probability density functions.

Constructing cell density versus distance profile for whole slide image

To compute cell density-distance profiles, infiltration intensities are quantified as the immune-marker labeled cell densities at the corresponding distance from consensus tumor boundary obtained in the previous step. To generate the density-distance profile, we utilized pixel-distance based algorithm to segment the whole tissue into multiple equal-width band sections: all foreground pixels are classified according to their locations, denoted as CT-, IF-, and N-pixels; for both CT- and IF-pixels, the distance toward the IF inner (adjacent to CT) boundary are computed; for each N-pixel, the distance toward the IF outer (adjacent to normal tissue) boundary are computed; next we group the pixels into multiple intervals with the interval length of 150 μm (first interval: 0–150 μm; second interval: 150–300 μm; ...) based on their distance value; then we extract the shape outline for each group by computing the

concave hull; consequently each group comprises a polygon with 150 μm width. Minor adjustments are needed to ensure all points within each shape outline are properly arranged so that they can be connected in clockwise or counterclockwise direction, if necessary.

To ensure the density-distance profile proceeds along the direction of immune infiltration, first we assign index 0 to the central band polygon of the IF as the reference polygon. Then we render negative indexes, decreasing from 0 until we reach the most distant band within the normal tissue, and render positive indexes, increasing from 0 until we reach the most distant polygon within the tumor. With all groups are arranged in a consecutive numerical order that expands from the edge of normal tissue toward the CT, the cell density is generated by calculating the area and counting the cells inside. The described methods are summarized in **Figures 4A–C**.

For all aforementioned calculations, the concave hull is computed using a function from R package 'concaveman'; the pixel distance and polygon areas are computed using the function 'gDistance' and 'gArea' from R package 'rgeos' (Bivand and Rundel, 2017).

Constructing cell density versus distance profile within invasive front

The IF is segmented into sections along its own direction: the length of central reference line is first calculated; then multiple equispaced points along the line are sampled based on a given interval. A Voronoi tessellation was created within the IF region based on Euclidian distance to each chosen point. The area of each resulting polygon and the number of cells it encloses are computed to determine cell density, and all polygons are indexed in a numerical order starting from 1 with the left-most point, in a clockwise direction. The methods are summarized in **Figure 4D**. The length of the reference line is calculated using the function 'lineLength' from R package 'SDraw' (McDonald, 2016); the length of interval is set to be 0.2 mm; the points are sampled using the function "spsample" from R package 'sp.'

Constructing 95% confidence interval (CI) along with the cell density profile

Depending on the shape of whole slide images, polygons at a certain distance may be small in size, which could result in inaccurate estimation of the average cell density of that distance. Therefore, to test the reliability of all average cell density point estimates along the 2D projection, we construct a 95% CI along the profile. For better accuracy, we assume the variance of density between each window within one distance polygon band is equal to the variance of the entire region with the same tissue type as the polygon of interest, which is denoted as σ. Hence the confidence interval can be computed according to the formula (Efron, 1981):

$$D \pm T_c \cdot \frac{\sigma}{\sqrt{n}} \tag{2}$$

where D is the density of cells within a polygon, T_c is the critical t-value. In this study T_c is defined as 1.96 (95% confidence level); n is the number of samples, and is calculated according to the equation:

$$n = \frac{A}{s_l \times s_w} \tag{3}$$

where A is the area of the polygon; s_l and s_w are the length and width of the window. In this study, window size is defined as $s_l = s_w = 0.4$ mm.

Confidence intervals are calculated along with the density profile; however, when visualizing the data, we truncated the portion below zero, and use 80% of the density as the threshold to filter out locations at which the density estimates are not reliable as the mean of the region (labeled as red dots).

Constructing the three-dimensional immune landscape

For each slide, the recorded density mapped to original locations to construct the landscape. The entire landscape is then characterized to reflect region-specific information (N, IF, and CT). The landscape data visualization is implemented using software Blender 2.80 (Hess, 2007).

Measuring the heterogeneity from spatial point pattern process model fitting results

In our study, the local point pattern is defined as the point pattern of the immune marker within a given subregion. For each captured local point pattern, Complete Spatial Randomness (CSR) is tested using the Clark-Evans test with the null hypothesis being a uniform Poisson process (one-tailed, H_A: clustered distribution, significance level $\alpha = 0.05$) (Baddeley et al., 2015). If the pattern failed to pass the CSR test, we fit a Thomas point process model to the local point pattern and record fitted parameters. The model assumes cluster patterns are generated in two steps: in the first step, a pattern of parent points within the window is generated according to a homogeneous Poisson process given the intensity κ; in the second step, a random number of offspring points is generated, so that the number of offspring points that belong to any parent point also follows Poisson distribution with intensity μ, at the same time the location follows isotropic Gaussian distribution with standard deviation σ. The theoretical Ripley's K function of the Thomas process is:

$$K(r) = \pi r^2 + \frac{1}{\kappa}(1 - e^{-\frac{r^2}{4\sigma^2}}) \tag{4}$$

where r is the distance of a sample random point of the point pattern within which the function is evaluated. For each sub-region, the fitted parameters κ, μ, and σ are biologically interpreted as features of the clustering pattern of immune marker-labeled cells. κ stands for the number of labeled cell clusters per unit area; μ is the number of labeled cells per cluster. We further use the distance of the cell toward the cluster center to quantify the internal cell distribution of each cluster (Thomas, 1949; Waagepetersen, 2007; Tanaka et al., 2008). We observe that the components of distance vector are normally distributed and independent since each point in the generated clustered pattern is produced from an isotropic Gaussian process so that the collective distance profile within each cluster follows a Rayleigh distribution with a probability density function:

$$G(r\,;\sigma) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \tag{5}$$

The average moment can be calculated as

$$\mu(r) = \sigma\sqrt{\pi/2} \tag{6}$$

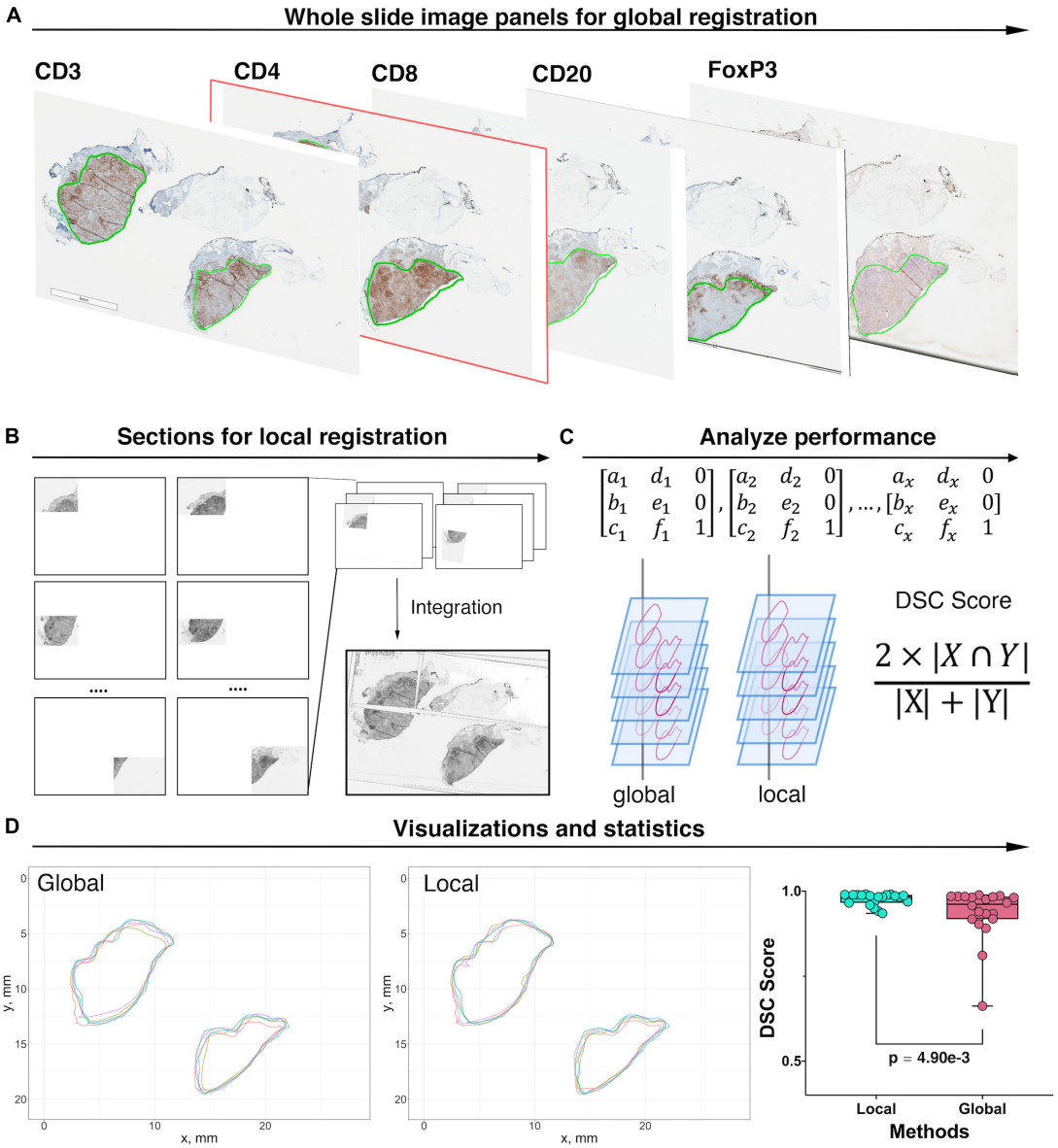

FIGURE 3 | Image registration and performance evaluation workflow, Case 1 is taken as an example for illustration. **(A)** Whole slide image panels with pathologist's annotations (green outlines) for Case 1. In this study, CD4+ slide is selected as the reference. Global registration is first applied to all registration pairs. The performance for each pair is then evaluated manually and prepared for local registration if necessary. **(B)** Poorly registered slides are subject to local registration. Slides and references are segmented into multiple subregions and using software Icy to perform local pairwise registration. **(C)** Transformation matrices obtained from both local and global registrations are applied to tissue contours. For each method, the DSC is then computed between registered contours and the contours of the reference slide. **(D)** Registered contours from two methods and reference contours are mapped to the same coordinates. DSC is computed by calculating their respective and intersection areas. The Wilcoxon rank-sum test is performed when DSCs for all 20 registrations pairs are collected. The result showed that the local registration performs significantly better than global registration (Wilcoxon rank-sum p = **4.90e-3**).

and the radius of the circle where 95% of the cells would fall in is calculated as:

$$Q(F, \sigma) = \sigma\sqrt{-2ln(1 - F/100)} \quad (7)$$

where F = 95. The CSR testing is performed using functions "clarkevans.test" and "kppm" from R package "spatstat," with parameters "clustered" and "Thomas," respectively (Baddeley et al., 2015).

For each slide, we measure and compare spatial statistics between different regions (intra-tumoral) and between different cases (inter-tumoral). We use the quartile coefficient of dispersion (QCoD) and the coefficient of variation (CoV) to assess the variability of the metrics (density and spatial model fitting parameters) within one slide. Furthermore, the median value of each metric is used to represent the corresponding case for the case-wise comparison. QCoD

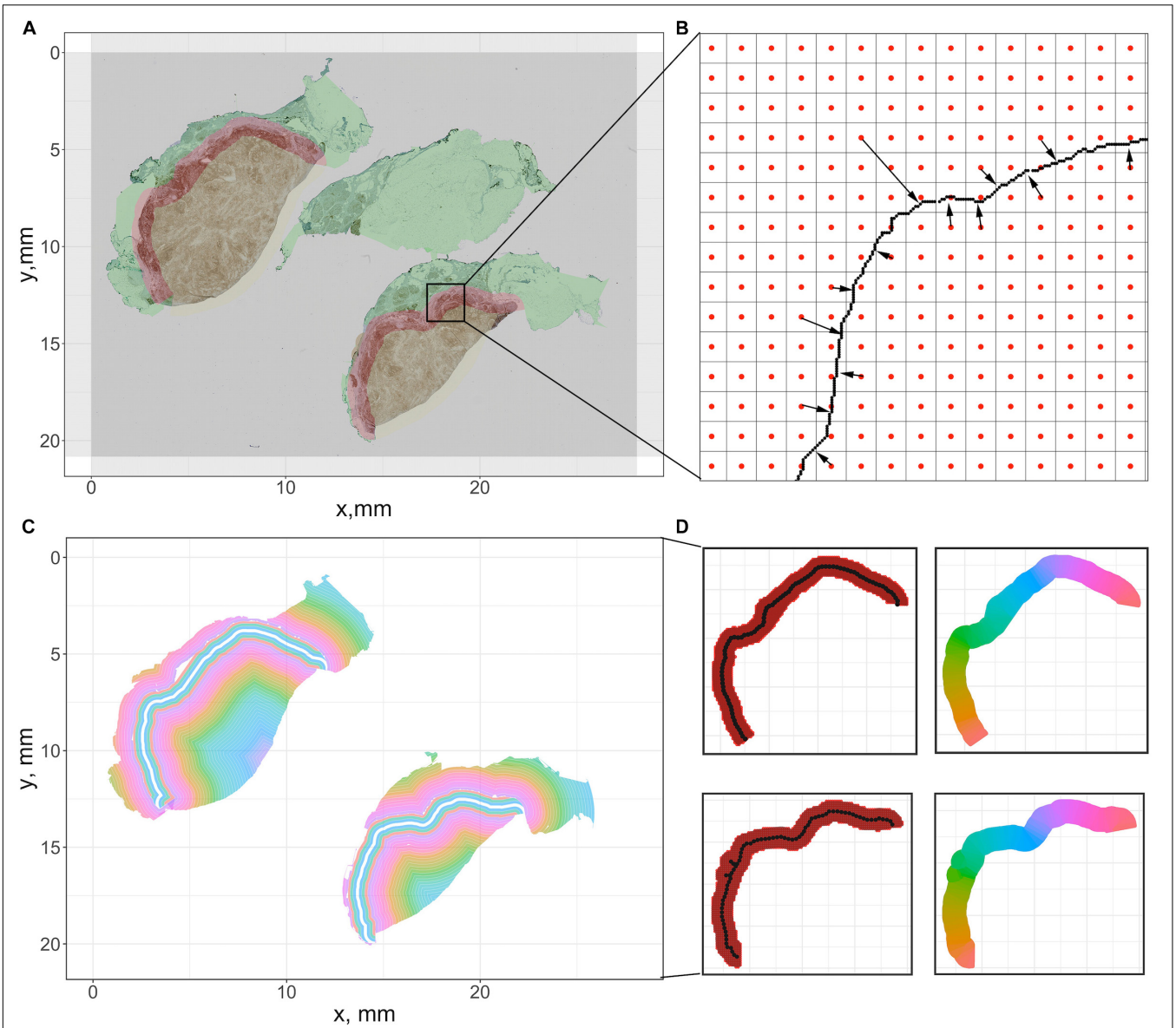

FIGURE 4 | Summary of the pixel-distance based method to generate sections for cell density-distance profiles. We use Cases 1A and 1B here for illustration. **(A)** For the downsampled image, we generate a grid whose size is equal to the number of pixels. Then the center coordinate of each square is recorded to represent the pixel. **(B)** For all pixels within tissues, we calculated the distance toward the tumor boundary of the corresponding case. **(C)** All pixels are grouped based on their distances to generate equidistant sections. **(D)** To segment the invasive front, multiple equidistant points are sampled along the boundary, and each pixel within the IF is assigned to their nearest sample point. Then all pixels associated with each sample points form a polygon to calculate density. The color codes in the right panels indicate different sections.

and CoV are computed using the following formulas:

$$QCoD = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{8}$$

$$CoV = \frac{\sigma}{\mu} \tag{9}$$

where Q_1, Q_3, μ, and σ are the first quartile, third quartile, standard deviation, and mean of each metric.

Measuring the heterogeneity from clustering and morphometric analysis

Immune contexture heterogeneity in the TME holds a significant value to the study of the anti-tumor immune response (Beck et al., 2011; Schwen et al., 2018). Therefore, we performed immune cell cluster analysis to assess the intra- and inter-tumoral heterogeneity. We first identify clusters from the global point patterns using an adjusted version of the clustering algorithm Hierarchical DBSCAN (HDBSCAN) from R package 'largeVis' (Tang et al., 2016). This method

generates the cluster hierarchy based on density-adjusted distance connectivity, and parent and child cluster stabilities are compared to extract clusters. The algorithm arguments are defined as minPts (minimum cells per cluster) = 30, and K (the number of cells in the core neighborhood) = 4. Next, morphological analysis is performed on each previously identified cluster by calculating shape descriptors. To describe the structure of clusters, we introduced α-shape, which envelops a set of points by point pairwise segments that could be regarded as a chord of a circle with a given radius α. To identify the α-shape which exactly harbors all points within the given region, the value of α is increased from 10 μm until the polygon reaches the ideal size. α-shapes are computed using the function "ashape" from R package 'alphahull' (Rodríguez Casal and Pateiro López, 2010). Then the morphometrics for each cluster are calculated for the following measurements:

Convexity: Measures the degree of the object. Convexity is mathematically defined as:

$$f_{convex} = \frac{A_\alpha}{A_{convex}} \tag{10}$$

where $A_\alpha$is the area of the α-shape and $A_{convex}$is the area of the convex hull, generated upon the same dataset.

Circularity: Measures the roundness of the object. Circularity is mathematically defined as:

$$f_{circularity} = \frac{4 \cdot \pi \cdot A_\alpha}{P_\alpha^2} \tag{11}$$

where P_α is the perimeter of the α-shape.

Eccentricity: Measures the degree of deviation of the object from being circular. To generate the ellipse, we first assume the points within each cluster follow chi-squared distribution $Q \sim x^2(k)$. Then the eigenvectors can be calculated from the covariance matrix, which indicates orientations. Now the semi-major and -minor axis lengths can be computed as:

$$a = \sqrt{\lambda_1 X_2^2(0.95)}\,, b = \sqrt{\lambda_2 X_2^2(0.95)} \tag{12}$$

where λ_1 and λ_2 are eigenvalues of the covariance matrix. By this definition, the ellipse is represented as the contour where 95% of points were covered. Thus, the eccentricity is computed as:

$$e = \sqrt{1 - \frac{b^2}{a^2}} = \sqrt{1 - \frac{\lambda_2}{\lambda_1}} \tag{13}$$

Correlating the Spatial Patterns of Different Immune Markers

The metrics above all referred to spatial distributions of a single marker. However, it is of interest to know if the distributions of cells with different markers are correlated. For example, CD8+ cells generally inhibit tumor growth whereas FoxP3+ cells generally promote tumor growth; if they are colocalized their effects might cancel each other. Such assumption simplifies the definition of T cell lineages due to the limitations in materials and the scale of biomarker panel. In this study, we implement a variation of the Clus-DoC (clustering-degree of colocalization) approach to analyze spatial correlation between different immune cell labels (Pageon et al., 2016). We focused on the correlation between three pairs of the spatial patterns: CD3+/CD8+, CD4+/FoxP3+, and CD8+/FoxP3+ as representations for anti-tumor immunity regions, immunosuppression regions, and immune-crosstalk regions. With each pair of full point patterns, for both channels, a DoC score is assigned to each point. This step requires the comparison of the spatial distribution of all the neighboring points from both channels for every single point. Centered at each point of type A, circles with increasing radius are formed to compute the associated density gradients of points from both channels. Then for each point of type A, the correlations between the density gradients between both channels are measured by Spearman's rank coefficient ρ_{AB}. Next, each coefficient ρ_{AB} is converted to a DoC score by normalization using the equation:

$$DoC_A = \rho_{AB} \cdot e^{-\left(\frac{N_{AB}}{R_{max}}\right)} \tag{14}$$

where N_{AB} is the distance of the current point of type A to the nearest neighbor of type B, R_{max} is the maximum search radius. Thus, the DoC score is bounded within [−1, 1], where 1 indicates a strong correlation (colocalization) and −1 indicates anti-correlation (segregation). These calculations are performed for both channels and DoC scores are then used to identify correlated points. To select a proper DoC threshold, we create a synthetic point pattern by shifting the full point patterns of a given slide (we use CD8+ slide from Case 2 in our study, but it can be any WSI within the study cohort) to a given direction by a minor distance to simulate a well-localized pattern pair. For simplicity, we unify the shift directions for all points to left and with a distance to 10 μm plus uncertainty caused by the registration error. The averaged DoC score is then selected as the threshold. Spatially, points with high DoC scores (correlated) are close to other points of both channels, whereas points with low DoC scores (non-correlated) are not close to points of at least one channel.

In the second step, the threshold is used to select highly correlated points for each channel. Points from both channels are then mapped to the same coordinate system. Next, we use the density-based clustering algorithm described in the morphometric module to detect clusters that contain points of both types A and B. Such clusters highlight regions with strong mutual interactions of immune markers. The search of neighbors for each point is calculated and accelerated using the C++ implemented k-dimensional tree search algorithm in Python library 'SciPy' (Virtanen et al., 2020); the distance of a point to its nearest neighbor is calculated using function 'nncross' in R package 'spatstat' (Baddeley et al., 2015); the density-based clustering is performed using function HDBSCAN in R package 'largeVis,' with arguments K = 4, and minPts = 30.

RESULTS

Assessing Intra- and Inter-tumoral Heterogeneity With Multiple Metrics

Immune Cell Density Distribution and Infiltration Profiles

A summary of first-order properties analysis is shown in **Figure 5**. Cases 1A and 1B are used here for illustration; both parts of the tumor are present on the same slide but are separated. For each slide, the region annotations (**Figure 5A**) and coordinates extraction (**Figure 5B**) are performed first to characterize the spatial distribution. **Figure 5C** shows profiles of cell density vs. distance from the boundary in mm for each case for the five labels. These are the immune-infiltration profiles for different cell types. Two definitions of the IF are introduced here, with a width equals 0.5 mm (blue vertical dashed lines) and 1 mm (red vertical dashed lines) in accordance with pathologists' convention. 95% confidence intervals are depicted in gray bands. Most of the sections consist of sufficient 0.15 × 0.15 mm windows to estimate CI; however, near the edges of the slide, in normal tissue (N) and CT, the areas may be small resulting in wide confidence intervals; in these cases we first exclude all the small regions (area < 1.1 mm^2) from analysis and then we manually set a threshold with ±80% of real density to label the remaining unreliable sections (red dots). Importantly, we observe that the immune-infiltration profiles are unbiased regardless of how IF is defined. The maxima for the different immune cells are shifted from the boundary toward the CT, but still are within the IF. In other words, the cell densities peak around the 0.2–0.35 mm band and then drop gradually toward the innermost of the CT. To further corroborate that such infiltration profiles are caused by tumor heterogeneity, we compare the actual cell distribution pattern to a binomial distribution pattern by quadrat test. Theoretically, if a point set is randomly generated over a region which consists of multiple sections, then the expected number of points each section harbored can be calculated as the total number of points multiplied by the probability a point happened to be in this section. Therefore, for each WSI, we harvest the actual cell counts (frequencies) and theoretical frequencies in each section and performed chi-square independent test. Results show that the null hypothesis (two types of observation are independent) is rejected by all trials, suggesting that the actual cell distribution pattern is not a realization of randomness, rather it is driven by heterogeneity (**Supplementary Figure S4**).

Waterfall plots are frequently used in other studies to present results of clinical trials, when patients' responses are ranked from best to worse, using tumor size as a metric, each patient is represented as a bar in the plot. In this case we use waterfall plot to rank cell densities from largest to smallest, from 0.4 × 0.4 mm windows throughout the tissue, for each label with colors corresponding to N (green), IF (red) and CT (yellow). The results are shown in **Figure 5D**. Waterfall plots indicate that CT and IF tend to have higher cell densities whereas fewer cells tend to accumulate in N, as the left-hand side of the chart contains more red (IF) bars and right-hand side have more green bars; the plots illustrate a high degree of heterogeneity as bars of different color are interspersed throughout the tissue. As our study focuses on heterogeneity of tumor characteristics, it is important to assess the level of heterogeneity within the IF. **Figure 5E** depicts the cell density distribution of CD4+ T cell plotted as a function of the distance along the middle of the IF; the densities are averaged over the width of the IF of 1 mm. Clearly, the spatial heterogeneity is present not only between different tumor regions, but also within the IF. Again, quadrat test is performed to assess the distribution pattern across all tessellations within each IF. Similarly, tests reject the null hypothesis so that the tumor heterogeneity is also the key factor that dominate the infiltration profiles in IF (**Supplementary Figure S5**). We then depict the immune landscape by visualizing the cell distribution in 3D (**Figure 5F**). For each slide, we map recorded subregion densities to their corresponding locations and generate surface plot with density represented by magnitude and region categories represented by different colors. 3D landscape representation directly depicts the regional density variations. We repeat the analysis for Cases 2–5 for all five labels and the results are presented in **Supplementary Figure S6** (infiltration profiles), **Supplementary Figure S7** (waterfall plots), **Supplementary Figure S8** (infiltration profiles in IF), and **Supplementary Figures S9, S10** (3D plots). Note that the cell density level in Cases 1A and 1B are significantly higher compared to other cases, which may reflect an efficient immune infiltration.

Spatial Entropy of Multitype Point Patterns

The results above visualize the heterogeneity of cell density distributions within and between the different regions of the specimens. The coefficient of variation CoV is one metric that characterizes the level of heterogeneity. We will also use a spatially adjusted Shannon's entropy as a formal metric of spatial heterogeneity. For each WSI, the point pattern for each subregion is mapped to a Cartesian coordinate system to form a series of multitype point patterns. For each multitype point pattern, the modified Shannon's entropy is calculated, and collective statistics are presented using probability density functions (PDFs, **Figure 6**). The results show clear clustering patterns of the entropy scores around 1.5 over H_{SC} measurement spaces in IF across all cases. The results indicate that regions with higher entropies are more likely to associate with IF. Biologically, such 'chaos' is possibly driven by the engagement of various components within the TME, namely the spatial intra-tumoral heterogeneity. We also note that while the H_{SC} scores in CT also appear to cluster in Cases 1A, 1B, 2, 4, and 5, the distribution in Case 3 is comparatively flatten and is similar to N; whereas the H_{SC} scores in N are normally flatten with lower magnitude, the distribution in Case 5 is intense and sharp. This phenomenon is possibly caused by the different infiltration level of lymphocytes. An efficient immune response can facilitate the recruitment of infiltrating T cells into the battlefield to either fuel the immunosuppression or promote immunoactivation, depending on the recruited T cell subtypes. Once the infiltration barriers (as seen in **Figure 5C** and **Supplementary Figure S6**) are broken, the

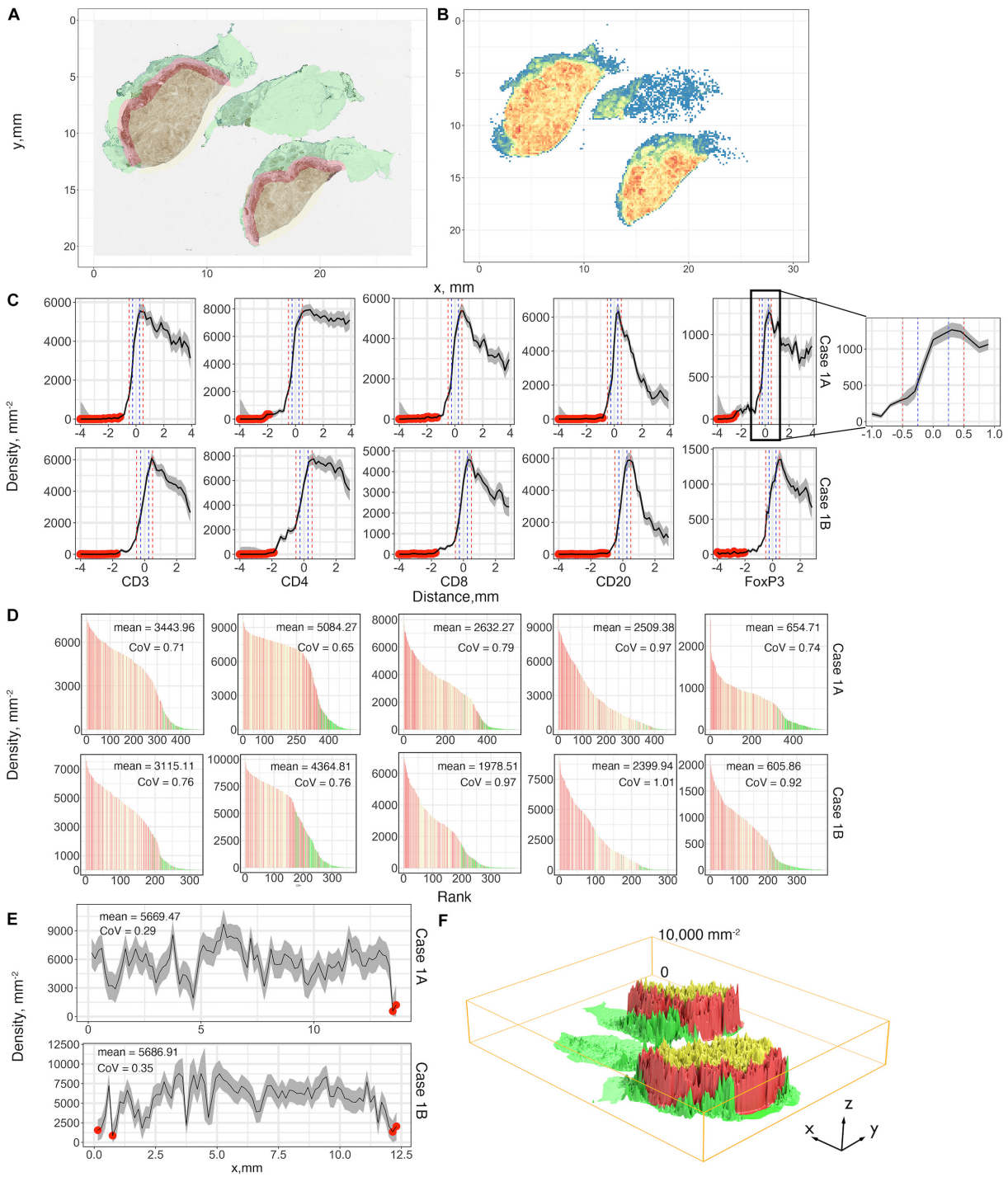

FIGURE 5 | First-order variables statistics summary, Case 1, AB, is taken as an example for illustration. **(A)** Region characterization for WSI. Green: normal tissue (N); Red: invasive front (IF); Yellow: central tumor (CT). **(B)** CD4+ cell density visualized using heatmap with bin size = 0.15 mm. Color code: blue to red corresponds to low to high. **(C)** Cell density-distance profiles with a pop-up window for Cases 1A and 1B. Whole tissues are segmented into equidistant sections. Densities of different immune markers are calculated for each section and mapped with their distances to the invasive boundary, respectively. 95% confidence intervals are calculated upon the profile, and we use 80% of the density as the threshold to label those unreliable locations (red dots). Two definitions of IF are introduced here and are indicated as vertical lines, blue: width of 0.5 mm; red: width of 1 mm. **(D)** Densities of subregions are visualized using waterfall plots. For each slide, the densities are shown as bar heights, which are ranked from highest to lowest with colors corresponding to their locations. Color codes are consistent with **(A)**. **(E)** The invasive front with thickness of 1 mm is sectioned along its horizontal direction, and the same process is repeated to construct the cell density-distance profile. **(F)** 3D visualization for the density of each subregion with location labels. Color codes are consistent with **(A)**. See also **Supplementary Material** (**Supplementary Figures S6, S7**) for additional visualizations for Cases 2–5.

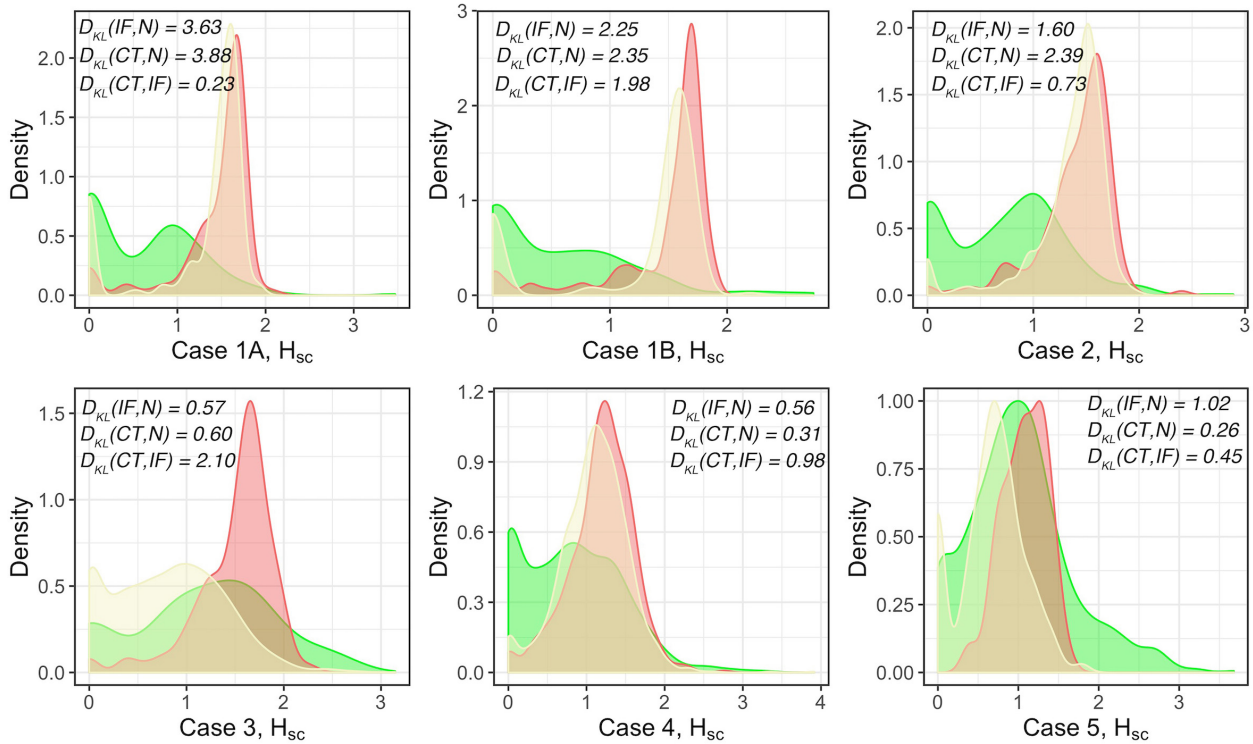

FIGURE 6 | Modified Shannon's method to quantify the spatial entropy of multitype point patterns, denoted as H_{SC}. Green: normal tissue (N); Red: invasive front (IF); Yellow: central front (CT). For each case, full point patterns for each label are mapped to the same coordinate system and the H_{SC} scores are measured and presented as the PDFs. In general, the higher the H_{SC} score is, the more disorder/heterogeneity the subregion contains. Strong heterogeneity is observed in IF as indicated by the clustered H_{SC} scores across cases.

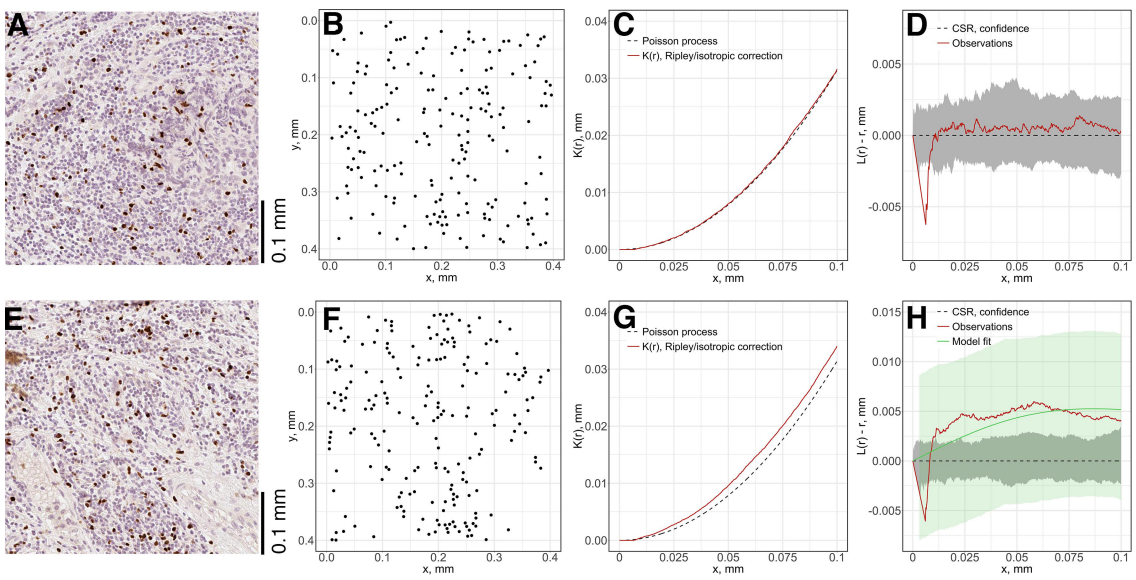

FIGURE 7 | Local spatial point pattern analysis for subregions. We used a moving window to gauge local characteristics across each slide. **(A,E)** Original exemplar IHC subregions. **(B,F)** Associated point patterns obtained from image segmentation and coordinate extraction. **(C,G)** K-estimation using Ripley's border correction for pattern **(B)** and **(F)**. Clark-Evans method is performed for CSR test, pattern (F) failed to pass the test and clustering model-fitting is performed. **(D,H)** L-transformation of *K* function and 95% confidence interval. For pattern **(B)**, the modified Thomas clustering process model is fit to pattern **(F)**. Results are evaluated using Dao-Genton goodness-of-fit test (green envelope).

immuno-architecture may tend to uniform and mitigate the spatial heterogeneity within TME.

Spatial Point Pattern Model Fitting

For each subregion (**Figures 7A,E**), we perform the Complete Spatial Randomness (CSR) to check whether the associated point pattern (**Figures 7B,F**) follows homogeneous Poisson distribution (**Figures 7C,G**); if that window failed to pass the CSR test, we further fit the Thomas process model to the point pattern to quantify the clustering (**Figures 7D,H**). In this study, we either directly use these parameters, e.g., μ as the number of immune marker-labeled cells per cluster, or perform transformation for intuitive interpretation, e.g., σ^2 is used to calculate the average distances of points to cluster areas and center.

As the window is moved through the slide, local features of spatial point patterns are quantified. Collective results are further classified based on their regions; therefore, the variance captures the intra-tumoral heterogeneity. Statistics for each region are also comparable among cases, which can reveal inter-tumoral heterogeneity. In this study, we use the following metrics to characterize each region for each case: mean cell density (counts/mm^2), average number of cells per cluster, mean distance to cluster core, and mean cluster area. QCoD of the mean values for each parameter across cases are summarized in **Table 1**. They collectively reflect intra- and inter-tumoral heterogeneity.

Cell Cluster Distributions and Morphometrics

We also extend our heterogeneity analysis to a global scale by quantifying cell clustering patterns. For each slide, the whole point pattern is clustered using a hierarchical clustering algorithm (**Supplementary Figures S11A,B**). For each detected cell cluster, we describe and characterize the shape by morphometrics, including convexity, circularity, and eccentricity. We first generate the alpha-hull (**Supplementary Figure S11C**) and convex hull from the point set that forms the cluster, upon which we derive the convexity and circularity; then we obtain the minor and major axes of the fitted ellipse (**Supplementary Figure S11D**) that covers 95% of the points of the cluster, from which we derive eccentricity.

Similar to model-fitting analysis, the regional variations of morphometrics capture intra-tumoral heterogeneity, and variations among cases capture inter-tumoral heterogeneity. We use the following metrics to characterize regions, respectively: average density of cells within clusters, average density of clusters, nodular cluster density (**Supplementary Figures S11E,G**) and elongated cluster density (**Supplementary Figures S11F,H**). For each label, we use boxplots to show variations between cases, and we perform the Wilcoxon rank-sum test between groups. **Figure 8** shows that even though IF is not always distinguished across labels and metrics, the nodular cluster densities and elongated cluster densities in IF are consistently higher than N. We also observe that CD3+ and CD8+ elongated cluster densities in IF are higher than CT. Such unique immuno-architecture may suggest that the elongated cytotoxic T cell cluster density in IF is a potential biomarker for further exploration.

Correlation Analysis of the Spatial Patterns of Different Immune Markers

The threshold for classification of the degree of colocalization (DoC) scores is determined by comparison of synthesized colocalization pairs. Considering the overall accuracy of our local registration algorithm is 0.976, we shift the point patterns of CD8+ left by 0.01/0.976 mm. This results in the majority ($\geq$90%) of DoC for each individual point are larger than 0.84, hereby the threshold is established. In this study, two types of correlations are considered: positive correlation means co-occurrence of points of both channels are likely to be observed in subjects' neighborhood. High DoC scores of points from both channels account for this type; negative correlation means co-occurrence of points of only one channel are likely to be observed in subject's neighborhood. High DoC scores of one channel whereas low DoC scores of the other account for this type. Based on such criteria, we analyze the correlations between CD3+/CD8+, CD4+/FoxP3+, and CD8+/FoxP3+. Results are shown in **Figure 9** with Cases 1A, B as an example, and clusters are represented by outlines in different colors.

Positive Correlations of CD3+ and CD8+ Immune Markers Identify Possible Anti-tumor Hotspots

For each case, the DoC scores, defined by Eq. 14, are assigned to each point of CD3+ and CD8+ markers. Next we use the threshold established above to select points with high DoC scores as candidates for clustering analysis using the same algorithm in cell clustering and morphometrics module (**Figures 9A,B**), and finally, the detected clusters that contain correlated cells from two channels are defined as hotspots of correlation and the cells within such hotspots are defined as correlated cells (**Figure 9C**).

We first examined the correlations between CD3+ and CD8+ marker pairs. As **Table 2** shows, the numbers of correlated cells differ drastically from case to case, however, the ratios are relatively consistent across cases. The density of hotspots in the IF and CT are significantly higher than normal tissue (N) (Wilcoxon rank-sum $\boldsymbol{p}$ = **2.20e-3** and **8.70e-3**), but there is no difference between IF and CT (Wilcoxon rank-sum $\boldsymbol{p}$ = **0.3095**). Such distribution pattern of correlation clusters of CD3+ and CD8+ cells indicates possible sites of tumor infiltrate conferring anti-tumor immunity (**Figure 9D**).

Positive Correlations of CD4+ and FoxP3+ Immune Markers Identify Possible Immunosuppression Hotspots

The same workflow is repeated for the CD4+ and FoxP3+ pair. A summary of the results is presented in **Table 3**. Similar patterns are observed when compared to CD3+ and CD8+ statistics. However, the ratios of correlated cells are generally higher than CD3+ and CD8+ pairs. The density of hotspots within IF and CT are significantly higher than N (Wilcoxon rank-sum $\boldsymbol{p}$ = **2.20e-3** and **8.70e-3**), plus the density at IF is also significantly higher than CT (Wilcoxon rank-sum $\boldsymbol{p}$ = **4.11e-2**).

Collectively, our analysis identifies several hotspots in which CD3+/CD8+ and CD4+/FoxP3+ pairs are

TABLE 1 | Quartile coefficient of dispersion (QCoD), Eq. 8, for spatial model fitting parameters (range and mean).

Marker	Region	Density	Cell/cluster	Mean distance	Cluster area
CD3	N	0.61–0.82 (0.67)	0.50–0.73 (0.64)	0.12–0.22 (0.16)	0.24–0.41 (0.31)
	IF	0.55–0.71 (0.64)	0.67–0.87 (0.77)	0.15–0.20 (0.17)	0.28–0.38 (0.32)
	CT	0.42–0.68 (0.57)	0.58–0.79 (0.71)	0.16–0.20 (0.18)	0.30–0.39 (0.34)
CD4	N	0.48–0.72 (0.62)	0.58–0.73 (0.64)	0.13–0.19 (0.17)	0.26–0.37 (0.32)
	IF	0.49–0.68 (0.57)	0.64–0.90 (0.76)	0.13–0.18 (0.15)	0.26–0.35 (0.30)
	CT	0.26–0.68 (0.53)	0.46–0.78 (0.65)	0.11–0.21 (0.16)	0.22–0.41 (0.32)
CD8	N	0.51–0.82 (0.61)	0.60–0.77 (0.67)	0.16–0.21 (0.18)	0.31–0.40 (0.35)
	IF	0.48–0.69 (0.59)	0.59–0.83 (0.71)	0.14–0.20 (0.17)	0.28–0.38 (0.33)
	CT	0.53–0.78 (0.64)	0.54–0.75 (0.66)	0.14–0.21 (0.18)	0.28–0.40 (0.35)
CD20	N	0.48–0.68 (0.59)	0.55–0.67 (0.60)	0.14–0.17 (0.15)	0.27–0.32 (0.29)
	IF	0.67–0.76 (0.71)	0.65–0.84 (0.78)	0.14–0.20 (0.17)	0.28–0.38 (0.33)
	CT	0.36–0.70 (0.50)	0.51–0.65 (0.61)	0.13–0.20 (0.16)	0.26–0.38 (0.32)
FoxP3	N	0.42–0.62 (0.48)	0.44–0.88 (0.60)	0.13–0.55 (0.23)	0.25–0.88 (0.42)
	IF	0.52–0.63 (0.58)	0.60–0.85 (0.73)	0.16–0.24 (0.19)	0.31–0.45 (0.36)
	CT	0.36–0.64 (0.50)	0.56–0.76 (0.67)	0.11–0.24 (0.18)	0.21–0.45 (0.35)

co-localized. The majority of identified hotspots are located within the IF. Such hotspots are characterized by strong correlations of CD3+/CD8+ and CD4/FoxP3+. These findings further reveal a strong heterogeneity within IF, as these two T cell subpopulations carry distinct immune characteristics.

Negative Correlations of CD8+ and FoxP3+ Immune Markers May Identify an Immune Response Landscape

We analyzed the negative correlation of CD8+ and FoxP3+ immune markers as a potential indicator of immune response landscape involving cytotoxic T cells (CTLs) and regulatory T cells (Tregs). In this section, two types of hotspots are defined. First, the hotspot with correlated CD8+ cells and non-correlated FoxP3+ cells; such clusters include FoxP3 cells that are surrounded by CD8+ cells, namely CD8-dominant hotspot. Biologically such hotspots may indicate places where CTLs may efficiently inhibit the strong immunosuppression of Tregs. Second is the opposite type, namely FoxP3-dominate hotspot, where the anti-tumor immunity is possibly impaired by Tregs.

The identification of such landscape can contribute to evaluating the immunotherapy treatment outcomes. Interactions are visualized by identifying negative correlation hotspots of two channels using previous workflow. Unlike the positive hotspots, negative hotspots are widespread throughout the entire tumor tissue for both types. For both types, the density of hotspots at CT and IF (**Supplementary Table S3**) are significantly higher than N (Wilcoxon rank-sum p = **1.52e-2** and **2.20e-3**) but no significant difference is observed between IF and CT (Wilcoxon rank-sum p = **0.0649**). For FoxP3-dominate hotspots, the CT and IF (**Supplementary Table S4**) are significantly higher than N (Wilcoxon rank-sum p = **1.52e-2** and **2.20e-3**) and the density in IF is also significantly higher than CT (Wilcoxon rank-sum p = **2.20e-3**). Results are shown in **Supplementary Figure S12** with Cases 1 A, B as an example, and clusters are represented by outlines in different colors.

DISCUSSION

In this study, we proposed a digital pathology computational workflow to systematically analyze whole slide images (WSI) and quantify the intra- and inter-tumoral heterogeneity through multiple metrics. We analyzed immunohistochemistry (IHC) slides of tumor resections from five patients with TNBC. The sample size is limited to make inferences at the population level, but it is sufficient to look in-depth at each sample with multiple immune markers and to develop a methodology for spatial statistical characterization and build a platform that could be extended to large number of samples, including multiplex IHC and immunofluorescence microscopy (mIF). It should also be noted that each WSI contains enormous amount of information and large numbers of cells of different type to fulfill our need to obtain statistically and biologically meaningful results. In principle, this approach is consistent with personalized medicine where inferences could be made at the level of individual patient. For each patient, five biomarkers: CD3, CD4, CD8, CD20, and FoxP3, were labeled using IHC staining methods. The whole computational workflow starts with image processing: we use cell nucleus segmentation to obtain location information of labeled cells in their original slides and perform image registration using a multimodal protocol to calculate transformation matrices which map all slides to the reference CD4+ slide. We further proposed a pixel-distance based method, and with its application, we can characterize the whole slide into normal tissue (N), IF, and CT. From this point, all subsequent analysis results can be classified into tissue type/region categories to reveal intra-tumoral heterogeneity; for each region, we compare the results between cases for inter-tumoral heterogeneity.

By visualizing the density distributions for each slide and spatial entropy analysis, we identified significant spatial variations of cell densities within and across slides, which qualitatively characterized the intra- and inter-tumoral heterogeneity. In addition, we are particularly interested in the spatial profiles along the direction from N through the IF to the innermost of CT. We observe that for each slide, the cell densities increase sharply within IF and then drop, in some

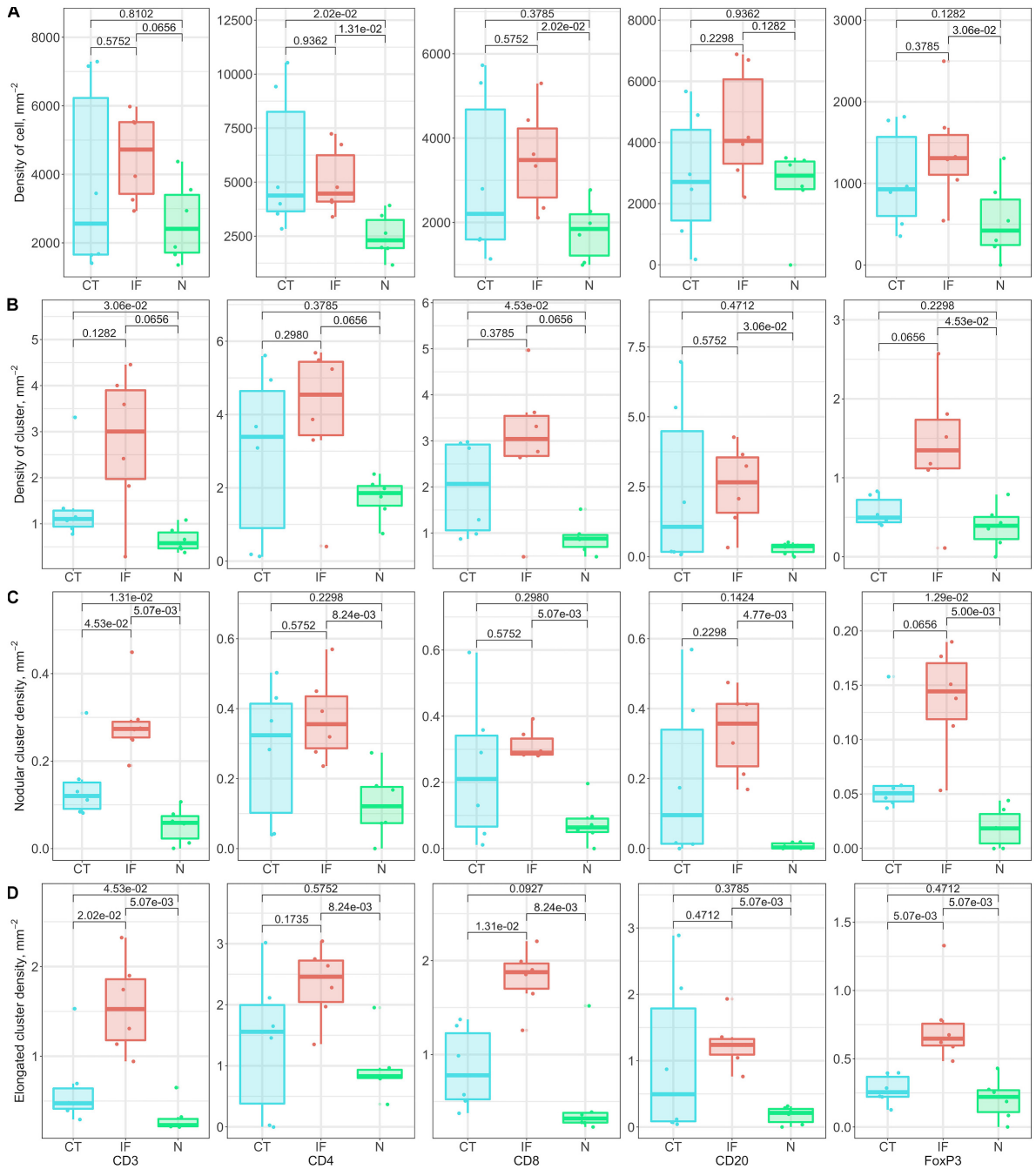

FIGURE 8 | Relationship of cluster morphometrics between regions for specific markers. **(A)** Average density of corresponding immune cells within a cluster. **(B)** Density of clusters within specific regions. **(C)** Number of nodular clusters; standard: convexity > 0.8, circularity > 0.5, and eccentricity < 0.8. **(D)** Number of elongated clusters; standard: convexity < 0.3, or circularity < 0.3, or eccentricity > 0.9.

cases to a plateau and in some precipitously and exhibiting fluctuations; hence corroborating the distinct role of IF in the immuno-architecture. This allows us to propose a hypothesis that the source of the immune cells in the IF is not in the normal tissue into which the tumor grows, but rather the cells extravasate from the tumor vasculature whose density is known to be higher at the rim of the tumor (Stamatelos et al., 2019); this hypothesis needs to be tested in future studies. We then fit a spatial point process model to data within subregions to capture local variabilities. Statistical results indicate that strong intra- and inter-tumoral heterogeneities co-exist across our study cohort. For each slide, we also evaluated the cell clustering using a hierarchical based algorithm against full point patterns. We then gauged the first-order properties and morphometrics of each

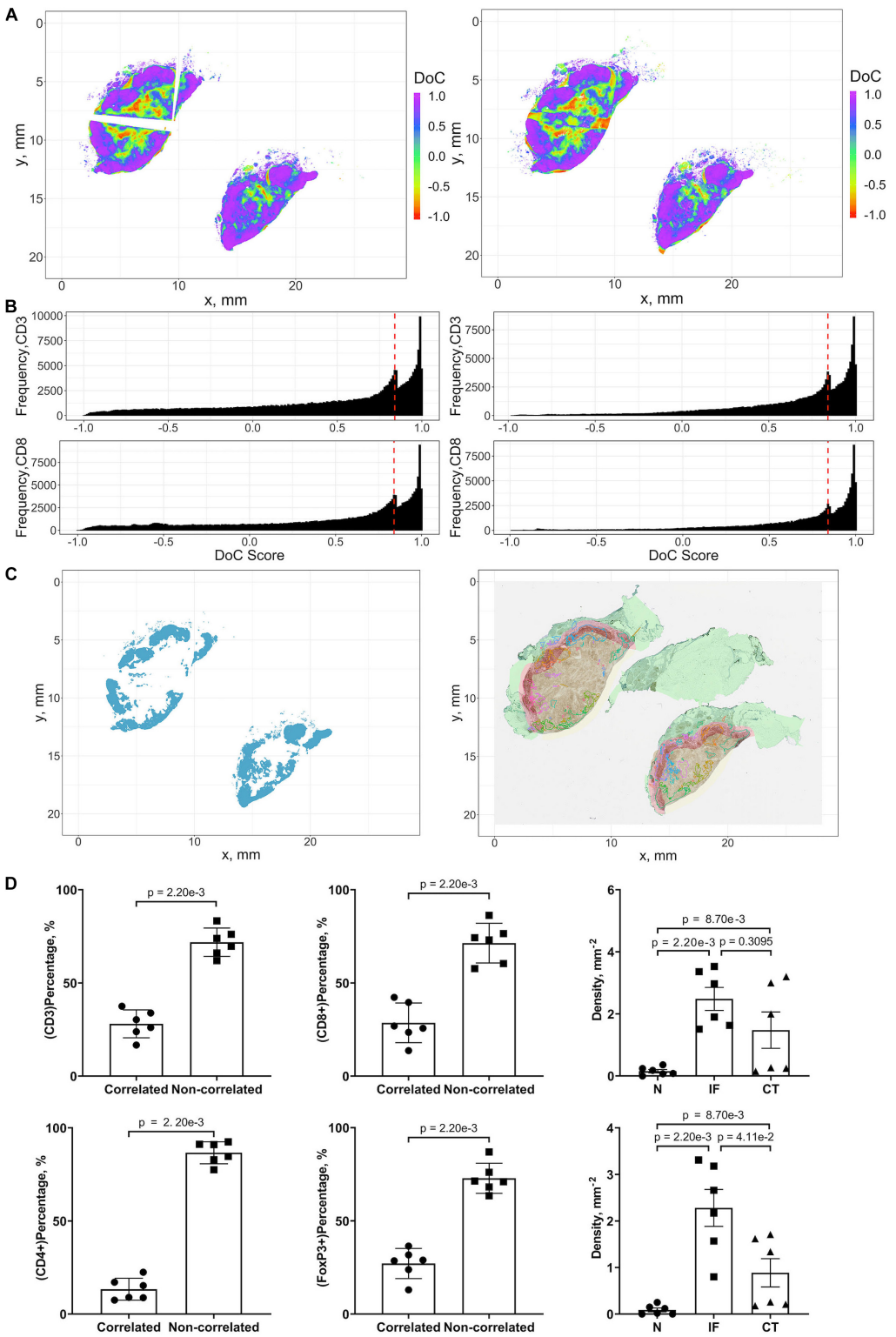

FIGURE 9 | Results of positive correlation analysis. In this study, we analyze two pairs of markers: CD3+, CD8+ and CD4+, FoxP3+. Each marker within the correlation pair is defined as a "channel". **(A)** DoC scores distributions for Cases 1A (left island) and 1B (right island). **(B)** Histograms of CD3+ and CD8+ DoC scores for Case 1A (left) and 1B (right). Red line is the predetermined threshold to select potential correlated cells. **(C)** Selected cells for both channels are mapped for clustering analysis (left) using HDBSCAN algorithm. We then define those clusters that both channels are involved as correlation hotspots (right, each cluster is represented by a colored outline). Green: normal tissue (N); Red: invasive front (IF); Yellow: central tumor (CT). **(D)** Statistical analysis. Top row: proportions of correlated CD3+ and CD8+ cells and density of correlated hotspots in different regions; bottom row: same metrics for CD4+ and FoxP3+.

cluster. Results revealed that variations are more likely to occur in CT and IF but less likely in N. We also identified a unique distribution pattern of nodular cytotoxic T cell clusters. As our recent study has shown (Gong et al., 2018), the distribution and shape of clusters have certain relations to the treatment outcomes, thus our findings may lead to predictive biomarkers that could eventually be used clinically when tested on a large number of specimens. Finally, we performed correlation analysis and discovered that the IF is multifaceted and may bear pro- and anti-tumor functions simultaneously, e.g., with higher expressions of CD8+ and FoxP3+ cells.

In addition to characterizing intra- and inter-tumoral heterogeneity, the characteristics obtained from tissue samples, such as spatial cell density profiles of different immune cells, the magnitude of spatial cell density fluctuations, and the spatial correlations between the densities of different immune cell types, can facilitate development and parameterization of spatially resolved computational immuno-oncology models. Recently, QSP models have been applied to immuno-oncology research as a platform for conducting virtual clinical trials (Cheng et al., 2017; Bai et al., 2019; Jafarnejad et al., 2019; Milberg et al., 2019; Ma et al., 2020). These models capture system scale behavior in cancer patients and are capable of population level predictions of disease trajectories in response to intervention. On tissue-cellular scale, ABMs have been employed and used for spatially explicit simulations to investigate emergent behavior arising from interactions between cancer and immune cells, such as spatial and spatio-temporal variations in tumor morphology and immuno-architecture (Kim et al., 2009; Shi et al., 2014; Wells et al., 2015; Gong et al., 2017; Norton et al., 2017, 2019; Pourhasanzade et al., 2017; Hoehme et al., 2018; Ji et al., 2019). When combining QSP models with ABM, cancer models can be further enhanced by taking advantage of both model types: while the QSP module captures whole-body temporal dynamics including lymph nodes, blood, peripheral compartment, and tumor, ABM simulation accounts for crucial aspects of high-granularity features such as cancer cell clonal evolution and TME heterogeneity. The resulting hybrid model will be able to closely track and predict the course of cancer development, both primary tumors and metastases, and potentially during treatment in individual patients by incorporating patient-specific TME characteristics, which can be quantified using our digital pathology platform. Such synergy would enable a better understanding of impact of spatial heterogeneities in the CT and IF on the pathophysiological parameters and variables. Strictly speaking, cell densities calculated directly from digitally segmented pathology images as described in this study represent the number of cell profiles per unit area in the tissue slide typically with a 4–5 micron thickness, which is a common metric in pathology, where a cell signature is a section of cell with an area larger than the detectable threshold set in our segmentation algorithm. However, in computational models the cell concentrations are usually represented as the number of cells per unit volume rather than unit area. Using methods from the field of stereology (Weibel et al., 1966), 3D numerical densities (N_V) can be estimated from 2D density (N_A) using

TABLE 2 | Statistical summary for CD3+ and CD8+ immune markers correlation analysis.

Case	QDoC (DoC Score)		Correlated cell counts		Percentage, %		Cluster density, mm^{-2}		
	CD3	CD8	CD3	CD8	CD3	CD8	N	IF	CT
1A	0.95	0.94	71,630	66,097	23.9	25.7	0.36	3.36	2.00
1B	0.35	0.29	61,342	53,715	33.9	39.6	0.24	2.97	3.20
2	0.15	0.15	91,147	70,859	37.5	42.3	0.08	3.53	3.00
3	0.29	0.39	25,773	17,804	30.3	26.9	0	1.90	0.25
4	0.24	0.30	26,340	22,180	26.1	23.5	0.07	1.63	0.27
5	0.61	0.91	11,672	7,214	16.7	13.7	0.18	1.51	0.15

N, normal tissue; IF, invasive front; CT, central tumor.

TABLE 3 | Statistical summary for CD4+ and FoxP3+ immune markers correlation analysis.

Case	QDoC (DoC Score)		Correlated cell counts		Percentage, %		Cluster density, mm^{-2}		
	CD4	FoxP3	CD4	FoxP3	CD4	FoxP3	N	IF	CT
1A	0.68	0.60	83,363	17,902	17.1	29.0	0.25	3.18	1.71
1B	0.40	0.36	65,267	14,719	22.5	36.5	0.12	3.31	1.62
2	0.52	0.53	59,281	15,630	15.3	23.9	0.02	2.17	1.34
3	1.30	0.39	25,172	12,085	9.0	31.8	0	2.66	0.26
4	0.85	0.90	19,385	6,547	8.8	13.0	0	0.80	0.21
5	0.87	0.49	18,596	7,586	7.5	28.6	0.15	1.57	0.18

N, normal tissue; IF, invasive front; CT, central tumor.

the following equation: $N_V = N_A/(t + D - 2h)$, where t is the thickness of the section, D is the diameter of stained cells (which are lymphocytes in the scope of this study), and h is the minimum height of detectable spherical cap (which can be derived from cellular segmentation algorithm parameters) (Royet, 1991). In this equation, $2h$ in the denominator accounts for loss of undetected parts of the cell. N_V indicates number of cells per unit volume and can directly be used to inform 3D spatial models of tumor-immune interactions. Using this equation, one could convert the 2D densities (in mm^{-2}) to 3D densities (in mm^{-3}); in the conversion the slide thickness is typically t = 4.5–5 μm, $h = D/2 - \sqrt{(D/2)^2 - A_{crit}/\pi}$, where A_{crit} is the minimum area detectable during the segmentation, typically ∼10 μm^2; diameter values reported for T cell (5–7.1 μm) and for B cell (5.5–9 μm) are also necessary for the conversion (Chapman et al., 1981; Turgeon, 2005; Tsourkas et al., 2007; Strokotov et al., 2009; El Hentati et al., 2010; Mrozek-Gorska et al., 2019; Renner et al., 2020).

Depending on the purpose of each computer simulation, one can either derive overall 3D density and use it to populate the *in silico* TME; or if spatial heterogeneity is of interest, the variability of cell density can be taken into account by sampling multiple N_A from different regions of the digital pathology analysis output to initiate the simulated TME with a range of N_V values in space. After simulation, the same methods employed in this study to analyze spatial correlations between different cell types can be applied to virtual sections of model-generated three-dimensional tumor, which would enable quantitative comparisons between model-generated spatial patterns of cancer and immune cells and patient pathology images. QSP and ABM have been used to model the tumor growth and invasion of several cancer types, such as melanoma (Wang et al., 2013; Milberg et al., 2019), breast cancer (Bates et al., 2006; Bianca and Pennisi, 2012), colorectal (Kather et al., 2017), and non-small cell lung cancer (Jafarnejad et al., 2019).

Future work should focus on increasing the scale of the current workflow. In this study, cells expressing CD8/FoxP3 are considered as cytotoxic/regulatory T cells. Such loose criterion serves the need to test the functionality of the workflow using preliminary computational results from pathology images. However, a comprehensive biomarker panel is required to account for the complexity in cell lineage definition and further to characterize the components in TME. Such materials will be obtained in subsequent studies applied to multiplex labeled specimens. Improving the performance of image processing is another critical issue. We recognize the power of artificial intelligence in digital pathology and such techniques could be incorporated in an extension of the workflow. For example, traditional segmentation algorithms may not adequately distinguish cell boundaries due to staining issues. A possibility is to introduce convolutional neural network (CNN) trained on well-defined ground-truth (Khosravi et al., 2018). In this study, immune markers are stained on consecutive slices of tumor resections, therefore artifacts may be introduced such as distortions. Registration reduces uncertainty introduced by these artifacts, but cannot fully compensate for the location mismatch, and errors may be introduced in derived point patterns and subsequent analysis. Such problems can be alleviated by harvesting data from multiplex images in the first place by labeling different cells and molecules on the same slide; in this case artificial location shifts, sample folds and z-axis differences are essentially eliminated. In the tissue type characterization step, we identified the IF by averaging annotations provided by expert pathologist. To pinpoint IF, deep learning methods can be applied for automated tissue segmentation. In point pattern analysis stage, we extract copious intra- and inter-tumoral heterogeneity information from collective slides; when correlated with treatment outcomes, these results can provide more useful information for pathologists and immuno-oncologists.

DATA AVAILABILITY STATEMENT

The datasets generated or analyzed during this work are available from the corresponding author on reasonable request. The codes for computational methods are made available at https://github.com/popellab/SpatHeterogeneity-TNBC. The open source software QuPath may be downloaded at https://qupath.github.io/. The open source software Blender may be downloaded at https://www.blender.org/download/. The open source software Icy may be downloaded at http://icy.bioimageanalysis.org/download/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of the Johns Hopkins Medical Institutions. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

HM, CG, and AP designed the workflow. HM processed the images, implemented the statistical tests, and produced the results. AC-M provided the specimens, assisted in the interpretation of results, and edited the manuscript. HM, CG, JS, EF, AS, EJ, VS, LE, AC-M, and AP critically edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The authors gratefully acknowledge financial support from National Institutes of Health (Grant Nos. R01CA138264 and U01CA212007).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.583333/full#supplementary-material

REFERENCES

Acs, B., Ahmed, F. S., Gupta, S., Wong, P. F., Gartrell, R. D., Pradhan, J. S., et al. (2019a). An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma. *Nat. Commun.* 10, 1–7. doi: 10.1038/s41467-019-13043-2

Acs, B., Pelekanou, V., Bai, Y., Martinez-Morilla, S., Toki, M., Leung, S. C., et al. (2019b). Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab. Invest.* 99, 107–117. doi: 10.1038/s41374-018-0123-7

Al-Janabi, S., Huisman, A., and Van Diest, P. J. (2012). Digital pathology: current status and future perspectives. *Histopathology* 61, 1–9. doi: 10.1111/j.1365-2559.2011.03814.x

Altan, M., Kidwell, K. M., Pelekanou, V., Carvajal-Hausdorf, D. E., Schalper, K. A., Toki, M. I., et al. (2018). Association of B7-H4, PD-L1, and tumor infiltrating lymphocytes with outcomes in breast cancer. *NPJ Breast Cancer* 4:40. doi: 10.1038/s41523-018-0095-1

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R. R Package Version 1.64-1*. Available online at: https://cran.r-project.org/web/packages/spatstat/index.html (accessed September 30, 2020).

Bai, J. P., Earp, J. C., and Pillai, V. C. (2019). Translational quantitative systems pharmacology in drug development: from current landscape to good practices. *AAPS J.* 21:72. doi: 10.1208/s12248-019-0339-5

Bankhead, P., Fernández, J. A., Mcart, D. G., Boyle, D. P., Li, G., Loughrey, M. B., et al. (2018). Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer. *Lab. Invest.* 98, 15–26. doi: 10.1038/labinvest.2017.131

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., Mcart, D. G., Dunne, P. D., et al. (2017). QuPath: open source software for digital pathology image analysis. *Sci. Rep.* 7, 1–7. doi: 10.1038/s41598-017-17204-5

Barua, S., Solis, L., Parra, E. R., Uraoka, N., Jiang, M., Wang, H., et al. (2018). A functional spatial analysis platform for discovery of immunological interactions predictive of low-grade to high-grade transition of pancreatic intraductal papillary mucinous neoplasms. *Cancer Inform.* 17:1176935118782880. doi: 10.1177/1176935118782880

Bates, G. J., Fox, S. B., Han, C., Leek, R. D., Garcia, J. F., Harris, A. L., et al. (2006). Quantification of regulatory T cells enables the identification of high-risk breast cancer patients and those at risk of late relapse. *J. Clin. Oncol.* 24, 5373–5380. doi: 10.1200/JCO.2006.05.9584

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., Van De Vijver, M. J., et al. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* 3:108ra113. doi: 10.1126/scitranslmed.3002564

Berben, L., Wildiers, H., Marcelis, L., Martinez, A. A., Bosisio, F., Hatse, S., et al. (2020). Computerized scoring protocol for identification and quantification of different immune cell populations in breast tumor regions using QuPath software. *Histopathology* 77, 79–91. doi: 10.1111/his.14108

Bianca, C., and Pennisi, M. (2012). The triplex vaccine effects in mammary carcinoma: a nonlinear model in tune with SimTriplex. *Nonl. Anal.* 13, 1913–1940. doi: 10.1016/j.nonrwa.2011.12.019

Bivand, R., and Rundel, C. (2017). *rgeos: Interface to Geometry Engine-Open Source (GEOS). R Package Version 0.3-23*. Available online at: https://cran.r-project.org/web/packages/rgeos/index.html (accessed September 30, 2020).

Blagih, J., Zani, F., Chakravarty, P., Hennequart, M., Pilley, S., Hobor, S., et al. (2020). Cancer-specific loss of p53 leads to a modulation of myeloid and T cell responses. *Cell Rep.* 30, 481–4966. doi: 10.1016/j.celrep.2019.12.028

Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brown, J. R., Wimberly, H., Lannin, D. R., Nixon, C., Rimm, D. L., and Bossuyt, V. (2014). Multiplexed quantitative analysis of CD3, CD8, and CD20 predicts response to neoadjuvant chemotherapy in breast cancer. *Clin. Cancer Res.* 20, 5995–6005. doi: 10.1158/1078-0432.CCR-14-1622

Chapman, E. H., Kurec, A. S., and Davey, F. (1981). Cell volumes of normal and malignant mononuclear cells. *J. Clin. Pathol.* 34, 1083–1090. doi: 10.1136/jcp.34.10.1083

Cheng, Y., Thalhauser, C. J., Smithline, S., Pagidala, J., Miladinov, M., Vezina, H. E., et al. (2017). QSP toolbox: computational implementation of integrated workflow components for deploying multi-scale mechanistic models. *AAPS J.* 19, 1002–1016. doi: 10.1208/s12248-017-0100-x

Cimino-Mathews, A., Thompson, E., Taube, J. M., Ye, X., Lu, Y., Meeker, A., et al. (2016). PD-L1 (B7-H1) expression and the immune tumor microenvironment in primary and metastatic breast carcinomas. *Hum. Pathol.* 47, 52–63. doi: 10.1016/j.humpath.2015.09.003

Claramunt, C. (2005). "A spatial form of diversity," in *Lecture Notes in Computer Science*, eds A. G. Cohn, and D. M. Mark, (Berlin: Springer), 218–231. doi: 10.1007/11556114_14

De Chaumont, F., Dallongeville, S., Chenouard, N., Hervé, N., Pop, S., Provoost, T., et al. (2012). Icy: an open bioimage informatics platform for extended reproducible research. *Nat. methods* 9:690. doi: 10.1038/nmeth.2075

Denkert, C., Von Minckwitz, G., Darb-Esfahani, S., Lederer, B., Heppner, B. I., Weber, K. E., et al. (2018). Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* 19, 40–50. doi: 10.1016/S1470-2045(17)30904-X

Du, Z., Lin, J.-R., Rashid, R., Maliga, Z., Wang, S., Aster, J. C., et al. (2019). Qualifying antibodies for image-based immune profiling and multiplexed tissue imaging. *Nat. Protoc.* 14, 2900–2930. doi: 10.1038/s41596-019-0206-y

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Can. J. Stat.* 9, 139–158. doi: 10.2307/3314608

El Hentati, F. Z., Gruy, F., Iobagiu, C., and Lambert, C. (2010). Variability of CD3 membrane expression and T cell activation capacity. *Cytom. Part B Clin. Cytom.* 78, 105–114. doi: 10.1002/cyto.b.20496

Ferré, E. M., Break, T. J., Burbelo, P. D., Allgäuer, M., Kleiner, D. E., Jin, D., et al. (2019). Lymphocyte-driven regional immunopathology in pneumonitis caused by impaired central immune tolerance. *Sci. Transl. Med.* 11:eaav5597. doi: 10.1126/scitranslmed.aav5597

Fridman, W. H., Galon, J., Pagès, F., Tartour, E., Sautès-Fridman, C., and Kroemer, G. (2011). Prognostic and predictive impact of intra-and peritumoral immune infiltrates. *Cancer Res.* 71, 5601–5605. doi: 10.1158/0008-5472.CAN-11-1316

Gombin, J., Vaidyanathan, R., and Agafonik, V. (2017). *Concaveman: A Very Fast 2D Concave Hull Algorithm. R Package Version 1.1.0*. Available online at: https://cran.r-project.org/web/packages/concaveman/index.html (accessed September 30, 2020).

Gong, C., Anders, R. A., Zhu, Q., Taube, J. M., Green, B., Cheng, W., et al. (2018). Quantitative characterization of CD8+ T cell clustering and spatial heterogeneity in solid tumors. *Front. Oncol.* 8:649. doi: 10.3389/fonc.2018.00649

Gong, C., Milberg, O., Wang, B., Vicini, P., Narwal, R., Roskos, L., et al. (2017). A computational multiscale agent-based model for simulating spatio-temporal tumour immune response to PD1 and PDL1 inhibition. *J. R. Soc. Interface* 14, 20170320. doi: 10.1098/rsif.2017.0320

Guy, C. L., Weiss, E., Che, S., Jan, N., Zhao, S., and Rosu-Bubulac, M. (2019). Evaluation of image registration accuracy for tumor and organs at risk in the thorax for compliance with TG 132 recommendations. *Adv. Radiat. Oncol.* 4, 177–185. doi: 10.1016/j.adro.2018.08.023

Haanen, J., Baars, A., Gomez, R., Weder, P., Smits, M., De Gruijl, T., et al. (2006). Melanoma-specific tumor-infiltrating lymphocytes but not circulating melanoma-specific T cells may predict survival in resected advanced-stage melanoma patients. *Cancer Immunol. Immunother.* 55, 451–458. doi: 10.1007/s00262-005-0018-5

Habets, R. A., De Bock, C. E., Serneels, L., Lodewijckx, I., Verbeke, D., Nittner, D., et al. (2019). Safe targeting of T cell acute lymphoblastic leukemia by pathology-specific NOTCH inhibition. *Sci. Transl. Med.* 11:eaau6246. doi: 10.1126/scitranslmed.aau6246

Halama, N., Michel, S., Kloor, M., Zoernig, I., Benner, A., Spille, A., et al. (2011). Localization and density of immune cells in the invasive margin of human colorectal cancer liver metastases are prognostic for response to chemotherapy. *Cancer Res.* 71, 5670–5677. doi: 10.1158/0008-5472.CAN-11-0268

Hendry, S., Salgado, R., Gevaert, T., Russell, P. A., John, T., Thapa, B., et al. (2017). Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell

carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Adv. Anat. Pathol.* 24:311. doi: 10.1097/PAP.0000000000000161

Hess, R. (2007). *The Essential Blender: Guide to 3D Creation with the Open Source Suite Blender*. San Francisco, CA: No Starch Press.

Hoehme, S., Bertaux, F., Weens, W., Grasl-Kraupp, B., Hengstler, J. G., and Drasdo, D. (2018). Model prediction and validation of an order mechanism controlling the spatiotemporal phenotype of early hepatocellular carcinoma. *Bull. Math. Biol.* 80, 1134–1171. doi: 10.1007/s11538-017-0375-1

Jafarnejad, M., Gong, C., Gabrielson, E., Bartelink, I. H., Vicini, P., Wang, B., et al. (2019). A computational model of neoadjuvant PD-1 inhibition in non-small cell lung cancer. *AAPS J.* 21:79. doi: 10.1208/s12248-019-0350-x

Ji, Z., Zhao, W., Lin, H.-K., and Zhou, X. (2019). Systematically understanding the immunity leading to CRPC progression. *PLoS Comput. Biol.* 15:e1007344. doi: 10.1371/journal.pcbi.1007344

Kather, J. N., Hörner, C., Weis, C.-A., Aung, T., Vokuhl, C., Weiss, C., et al. (2019). CD163+ immune cell infiltrates and presence of CD54+ microvessels are prognostic markers for patients with embryonal rhabdomyosarcoma. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-019-45551-y

Kather, J. N., Poleszczuk, J., Suarez-Carmona, M., Krisam, J., Charoentong, P., Valous, N. A., et al. (2017). In silico modeling of immunotherapy and stroma-targeting therapies in human colorectal cancer. *Cancer Res.* 77, 6442–6452. doi: 10.1158/0008-5472.CAN-17-2006

Kawai, O., Ishii, G., Kubota, K., Murata, Y., Naito, Y., Mizuno, T., et al. (2008). Predominant infiltration of macrophages and CD8+ T cells in cancer nests is a significant predictor of survival in stage IV nonsmall cell lung cancer. *Cancer* 113, 1387–1395. doi: 10.1002/cncr.23712

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., and Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine* 27, 317–328. doi: 10.1016/j.ebiom.2017.12.026

Kim, P. S., Levy, D., and Lee, P. P. (2009). Modeling and simulation of the immune system as a self-regulating network. *Methods Enzymol.* 467, 79–109. doi: 10.1016/S0076-6879(09)67004-X

Klemm, F., and Joyce, J. A. (2015). Microenvironmental regulation of therapeutic response in cancer. *Trends Cell Biol.* 25, 198–213. doi: 10.1016/j.tcb.2014.11.006

Ladányi, A. (2015). Prognostic and predictive significance of immune cells infiltrating cutaneous melanoma. *Pigment Cell Melanoma Res.* 28, 490–500. doi: 10.1111/pcmr.12371

Li, X., Gruosso, T., Zuo, D., Omeroglu, A., Meterissian, S., Guiot, M.-C., et al. (2019). Infiltration of CD8+ T cells into tumor cell clusters in triple-negative breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 116, 3678–3687. doi: 10.1073/pnas.1817652116

Lianyuan, T., Dianrong, X., Chunhui, Y., Zhaolai, M., and Bin, J. (2018). The predictive value and role of stromal tumor-infiltrating lymphocytes in pancreatic ductal adenocarcinoma (PDAC). *Cancer Biol. Ther.* 19, 296–305. doi: 10.1080/15384047.2017.1416932

Long, G. V., Weber, J. S., Infante, J. R., Kim, K. B., Daud, A., Gonzalez, R., et al. (2016). Overall survival and durable responses in patients with BRAF V600–mutant metastatic melanoma receiving dabrafenib combined with trametinib. *J. Clin. Oncol.* 34, 871–878. doi: 10.1200/JCO.2015.62.9345

Ma, H., Wang, H., Sove, R. J., Jafarnejad, M., Tsai, C.-H., Wang, J., et al. (2020). A quantitative systems pharmacology model of T cell engager applied to solid tumor. *AAPS J.* 22:85. doi: 10.1208/s12248-020-00450-3

Mani, N. L., Schalper, K. A., Hatzis, C., Saglam, O., Tavassoli, F., Butler, M., et al. (2016). Quantitative assessment of the spatial heterogeneity of tumor-infiltrating lymphocytes in breast cancer. *Breast Cancer Res.* 18:78. doi: 10.1186/s13058-016-0737-x

McDonald, T. (2016). *SDraw: Spatially Balanced Sample Draws for Spatial Objects. R Package Version 2.1.13*. Available online at: https://cran.r-project.org/web/packages/SDraw/index.html (accessed September 30, 2020).

Melosky, B., Juergens, R., Hirsh, V., Mcleod, D., Leighl, N., Tsao, M. S., et al. (2020). Amplifying outcomes: checkpoint inhibitor combinations in first–line non–small cell lung cancer. *Oncologist* 25:64. doi: 10.1634/theoncologist.2019-0027

Milberg, O., Gong, C., Jafarnejad, M., Bartelink, I. H., Wang, B., Vicini, P., et al. (2019). A QSP model for predicting clinical responses to monotherapy, combination and sequential therapy following CTLA-4, PD-1, and PD-L1 checkpoint blockade. *Sci. Rep.* 9, 1–17. doi: 10.1038/s41598-019-47802-4

Motzer, R. J., Penkov, K., Haanen, J., Rini, B., Albiges, L., Campbell, M. T., et al. (2019). Avelumab plus axitinib versus sunitinib for advanced renal-cell carcinoma. *N. Engl. J. Med.* 380, 1103–1115. doi: 10.1056/NEJMoa1816047

Mrozek-Gorska, P., Buschle, A., Pich, D., Schwarzmayr, T., Fechtner, R., Scialdone, A., et al. (2019). Epstein–Barr virus reprograms human B lymphocytes immediately in the prelatent phase of infection. *Proc. Natl. Acad. Sci. U.S.A.* 116, 16046–16055. doi: 10.1073/pnas.1901314116

Norton, K.-A., Gong, C., Jamalian, S., and Popel, A. S. (2019). Multiscale agent-based and hybrid modeling of the tumor immune microenvironment. *Processes* 7:37. doi: 10.3390/pr7010037

Norton, K.-A., Jin, K., and Popel, A. S. (2018). Modeling triple-negative breast cancer heterogeneity: effects of stromal macrophages, fibroblasts and tumor vasculature. *J. Theor. Biol.* 452, 56–68. doi: 10.1016/j.jtbi.2018.05.003

Norton, K.-A., Wallace, T., Pandey, N. B., and Popel, A. S. (2017). An agent-based model of triple-negative breast cancer: the interplay between chemokine receptor CCR5 expression, cancer stem cells, and hypoxia. *BMC Syst. Biol.* 11:68. doi: 10.1186/s12918-017-0445-x

Pageon, S. V., Nicovich, P. R., Mollazade, M., Tabarin, T., and Gaus, K. (2016). Clus-DoC: a combined cluster detection and colocalization analysis for single-molecule localization microscopy data. *Mol. Biol. Cell* 27, 3627–3636. doi: 10.1091/mbc.E16-07-0478

Pages, F., Kirilovsky, A., Mlecnik, B., Asslaber, M., Tosolini, M., Bindea, G., et al. (2009). In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.* 27, 5944–5951. doi: 10.1200/JCO.2008.19.6147

Pebesma, E., and Bivand, R. S. (2005). *S classes and Methods for Spatial Data: The Sp Package. R Package Version 1.4-2*. Available online at: https://cran.r-project.org/web/packages/sp/index.html (accessed September 30, 2020).

Pourhasanzade, F., Sabzpoushan, S., Alizadeh, A. M., and Esmati, E. (2017). An agent-based model of avascular tumor growth: Immune response tendency to prevent cancer development. *Simulation* 93, 641–657. doi: 10.1177/0037549717699072

Renner, K., Bruss, C., and Färber, S. (2020). *Monitoring Human T Cell Activation in the Context of Immunotherapeutic Approaches*. Bremen, Germany: OLS OMNI Life Science GmbH & Co KG. Available online at: https://www.bulldog-bio.com/wp-content/uploads/2020/06/T-cell-activation-CASY-AppNote_OLS.pdf

Rieger, T. R., Allen, R. J., Bystricky, L., Chen, Y., Colopy, G. W., Cui, Y., et al. (2018). Improving the generation and selection of virtual populations in quantitative systems pharmacology models. *Prog. Biophys. Mol. Biol.* 139, 15–22. doi: 10.1016/j.pbiomolbio.2018.06.002

Rodríguez Casal, A., and Pateiro López, B. (2010). *Generalizing the Convex Hull of a Sample: the R Package Alphahull. R Package Version 2.2*. Available online at: https://cran.r-project.org/web/packages/alphahull/index.html (accessed September 30, 2020).

Royet, J.-P. (1991). Stereology: a method for analyzing images. *Prog. Neurobiol.* 37, 433–474. doi: 10.1016/0301-0082(91)90009-P

Santiago, I., Santinha, J., Ianus, A., Galzerano, A., Theias, R., Maia, J., et al. (2019). Susceptibility perturbation MRI maps tumor infiltration into mesorectal lymph nodes. *Cancer Res.* 79, 2435–2444. doi: 10.1158/0008-5472.CAN-18-3682

Schmid, P., Rugo, H. S., Adams, S., Schneeweiss, A., Barrios, C. H., Iwata, H., et al. (2020). Atezolizumab plus nab-paclitaxel as first-line treatment for unresectable, locally advanced or metastatic triple-negative breast cancer (IMpassion130): updated efficacy results from a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol.* 21, 44–59. doi: 10.1016/S1470-2045(19)30689-8

Schwen, L. O., Andersson, E., Korski, K., Weiss, N., Haase, S., Gaire, F., et al. (2018). Data-driven discovery of immune contexture biomarkers. *Front. Oncol.* 8:627. doi: 10.3389/fonc.2018.00627

Shi, Z. Z., Wu, C.-H., and Ben-Arieh, D. (2014). Agent-based model: a surging tool to simulate infectious diseases in the immune system. *Open J. Model. Simul.* 2:12. doi: 10.4236/ojmsi.2014.21004

Sové, R. J., Jafarnejad, M., Zhao, C., Wang, H., Ma, H., and Popel, A. S. (2020). QSP−IO: a quantitative systems pharmacology toolbox for mechanistic multi−scale modeling for immuno-oncology applications. *CPT Pharmacometrics Syst. Pharmacol.* doi: 10.1002/psp4.12546 [Epub ahead of print].

Stamatelos, S. K., Bhargava, A., Kim, E., Popel, A. S., and Pathak, A. P. (2019). Tumor ensemble-based modeling and visualization of emergent angiogenic heterogeneity in breast cancer. *Sci. Rep.* 9, 1–14. doi: 10.1038/s41598-019-40888-w

Strokotov, D. I., Yurkin, M. A., Gilev, K. V., Van Bockstaele, D. R., Hoekstra, A. G., Rubtsov, N., et al. (2009). Is there a difference between T-and B-lymphocyte morphology? *J. Biomed. Optics* 14:064036. doi: 10.1117/1.3275471

Suzuki, H., Chikazawa, N., Tasaka, T., Wada, J., Yamasaki, A., Kitaura, Y., et al. (2010). Intratumoral CD8+ T/FOXP3+ cell ratio is a predictive marker for survival in patients with colorectal cancer. *Cancer Immunol. Immunother.* 59, 653–661. doi: 10.1007/s00262-009-0781-9

Tanaka, U., Ogata, Y., and Stoyan, D. (2008). Parameter estimation and model selection for Neyman−Scott point processes. *Biom. J.* 50, 43–57. doi: 10.1002/bimj.200610339

Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016). "Visualizing large-scale and high-dimensional data," in *Proceedings of the 25th International Conference on World Wide Web*, ed. J. Bourdeau, (New York, NY: Association for Computing Machinery), 287–297. doi: 10.1145/2872427.2883041

Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika* 36, 18–25. doi: 10.2307/2332526

Tsakiroglou, A. M., Fergie, M., Oguejiofor, K., Linton, K., Thomson, D., Stern, P. L., et al. (2020). Spatial proximity between T and PD-L1 expressing cells as a prognostic biomarker for oropharyngeal squamous cell carcinoma. *Br. J. Cancer* 122, 539–544. doi: 10.1038/s41416-019-0634-z

Tsourkas, P. K., Baumgarth, N., Simon, S. I., and Raychaudhuri, S. (2007). Mechanisms of B-cell synapse formation predicted by Monte Carlo simulation. *Biophys. J.* 92, 4196–4208. doi: 10.1529/biophysj.106.094995

Turgeon, M. L. (2005). *Clinical Hematology: Theory and Procedures*. Philadelphia, PA: Lippincott Williams & Wilkins.

Tyekucheva, S., Bowden, M., Bango, C., Giunchi, F., Huang, Y., Zhou, C., et al. (2017). Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nat. Commun.* 8:420. doi: 10.1038/s41467-017-00460-4

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics* 63, 252–258. doi: 10.1111/j.1541-0420.2006.00667.x

Wang, H., Milberg, O., Bartelink, I. H., Vicini, P., Wang, B., Narwal, R., et al. (2019). In silico simulation of a clinical trial with anti-CTLA-4 and anti-PD-L1 immunotherapies in metastatic breast cancer using a systems pharmacology model. *R. Soc. Open Sci.* 6:190366. doi: 10.1098/rsos.190366

Wang, H., Sove, R. J., Jafarnejad, M., Rahmeh, S., Jaffee, E. M., Stearns, V., et al. (2020). Conducting a virtual clinical trial in HER2-negative breast cancer using a quantitative systems pharmacology model with an epigenetic modulator and immune checkpoint inhibitors. *Front. Bioeng. Biotechnol.* 8:141. doi: 10.3389/fbioe.2020.00141

Wang, J., Zhang, L., Jing, C., Ye, G., Wu, H., Miao, H., et al. (2013). Multi-scale agent-based modeling on melanoma and its related angiogenesis analysis. *Theor. Biol. Med. Model.* 10:41. doi: 10.1186/1742-4682-10-41

Weibel, E. R., Kistler, G. S., and Scherle, W. F. (1966). Practical stereological methods for morphometric cytology. *J. Cell Biol.* 30, 23–38. doi: 10.1083/jcb.30.1.23

Wells, D. K., Chuang, Y., Knapp, L. M., Brockmann, D., Kath, W. L., and Leonard, J. N. (2015). Spatial and functional heterogeneities shape collective behavior of tumor-immune networks. *PLoS Comput. Biol.* 11:e1004181. doi: 10.1371/journal.pcbi.1004181

Wong, P. F., Wei, W., Smithy, J. W., Acs, B., Toki, M. I., Blenman, K. R., et al. (2019). Multiplex quantitative analysis of tumor-infiltrating lymphocytes and immunotherapy outcome in metastatic melanoma. *Clin. Cancer Res.* 25, 2442–2449. doi: 10.1158/1078-0432.CCR-18-2652

Yuan, Y. (2016). Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harb. Perspect. Med.* 6:a026583. doi: 10.1101/cshperspect.a026583

Zhang, A. W., Mcpherson, A., Milne, K., Kroeger, D. R., Hamilton, P. T., Miranda, A., et al. (2018). Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. *Cell* 173, 1755–1769.e22. doi: 10.1016/j.cell.2018.03.073

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

BRIEF RESEARCH REPORT
published: 17 December 2020
doi: 10.3389/fphys.2020.533101

Metastasis Initiation Precedes Detection of Primary Cancer—Analysis of Metastasis Growth *in vivo* in a Colorectal Cancer Test Case

Gili Hochman[1], *Einat Shacham-Shmueli*[2], *Stephen P. Raskin*[2], *Sara Rosenbaum*[1] *and Svetlana Bunimovich-Mendrazitsky*[1]*

[1] Department of Mathematics, Ariel University, Ariel, Israel, [2] Sheba Medical Center, Tel Hashome, Israel

Most cases of deaths from colorectal cancer (CRC) result from metastases, which are often still undetectable at disease detection time. Even so, in many cases, shedding is assumed to have taken place before that time. The dynamics of metastasis formation and growth are not well-established. This work aims to explore CRC lung metastasis growth rate and dynamics. We analyzed a test case of a metastatic CRC patient with four lung metastases, with data of four serial computed tomography (CT) scans measuring metastasis sizes while untreated. We fitted three mathematical growth models—exponential, logistic, and Gompertzian—to the CT measurements. For each metastasis, a best-fitted model was determined, tumor doubling time (TDT) was assessed, and metastasis inception time was extrapolated. Three of the metastases showed exponential growth, while the fourth showed logistic restraint of the growth. TDT was around 93 days. Predicted metastasis inception time was at least 4–5 years before the primary tumor diagnosis date, though they did not reach detectable sizes until at least 1 year after primary tumor resection. Our results support the exponential growth approximation for most of the metastases, at least for the clinically observed time period. Our analysis shows that metastases can be initiated before the primary tumor is detectable and implies that surgeries accelerate metastasis growth.

Keywords: lung metastases, mathematical growth models, primary tumor resection, exponential growth, logistic growth, liver metastasectomy, colorectal cancer, clinical metastasis growth data

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, United States

Reviewed by:
Leonid Hanin,
Idaho State University, United States
Sebastien Benzekry,
Institut National de Recherche en Informatique et en Automatique (INRIA), France

***Correspondence:**
Svetlana Bunimovich-Mendrazitsky
svetlanabu@ariel.ac.il

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *06 February 2020*
Accepted: *20 November 2020*
Published: *17 December 2020*

Citation:
Hochman G, Shacham-Shmueli E, Raskin SP, Rosenbaum S and Bunimovich-Mendrazitsky S (2020) Metastasis Initiation Precedes Detection of Primary Cancer—Analysis of Metastasis Growth in vivo in a Colorectal Cancer Test Case. Front. Physiol. 11:533101. doi: 10.3389/fphys.2020.533101

INTRODUCTION

Colorectal cancer (CRC) is one of the most common causes of cancer-related deaths worldwide, and the primary cause for CRC patient death is the development of metastatic disease (Van Cutsem et al., 2014; Vatandoust et al., 2015). Statistical data are available on patterns of colorectal metastasis sites (Riihimäki et al., 2016; Stewart et al., 2018), but the dynamics of metastasis formation and growth are not well-established. It is assumed that a significant part of metastases is seeded at a very early stage, before primary tumor detection (Fisher et al., 1977; Fisher, 1980; Siegel et al., 2017). Surgery is now the main curative treatment in both local and metastatic diseases (Stein and Schlag, 2007).

The most common site of CRC metastases is the liver, and the next is the lungs. Liver resection is now the standard of care for patients with resectable hepatic metastases (Stewart et al., 2018). However, there is evidence that stress response aroused by surgery may accelerate metastasis growth (Behrenbruch et al., 2018; Zheng et al., 2018). The extrahepatic disease is considered a risk factor in terms of survival after hepatic metastasectomy (Stewart et al., 2018). Specifically, the presence of limited preoperative small pulmonary nodules in the lungs was claimed to be associated with shorter progression-free survival (PFS) after hepatic metastasectomy (Maithel et al., 2010). Data on the effects of such a surgery on the growth of remaining metastases is not available and cannot be deduced retrospectively. Mathematical models, providing reliable representation of the metastasis growth patterns, may shed light on the metastatic growth process, and help in optimizing treatments for the prevention of metastasis growth.

Mathematical growth models are used as simplified approximations to dynamics of the actual biological process. Such models were extensively studied for primary tumors (Brú et al., 2003; Kozusko and Bajzer, 2003), but much less for metastasis growth dynamics. When modeled, exponential growth is often assumed, at least for the first period of growth (Haeno et al., 2012; Benzekry et al., 2014; Hanin and Bunimovich-Mendrazitsky, 2014; Hanin et al., 2016), although logistic or Gompertzian models—which have the feature of upper limitation on growth—are biologically more plausible. Comparison of different growth laws had been done by modeling *in vivo* data of metastatic cancer in several works, starting with Iwata's model (Iwata et al., 2000). When this model was applied on hepatocellular carcinoma patient data, the Gompertzian growth showed the best fit for the dynamics and size distribution of multiple liver metastases. Other works that followed are mostly based on animal models, for which data of untreated metastases is easier to obtain than for humans. See Hartung et al. (2014), Baratchart et al. (2015), Benzekry et al. (2016), and lately, Vaghi et al. (2020), who suggested that the Gompertzian growth model is the most appropriate model to be used for predictions of the metastatic growth process.

However, such predictions are hard to prove in humans, since clinical data on untreated metastasis growth is rare. Added to the diversity between different patients and metastases, it increases the difficulty in finding reliable growth patterns to be used as predictors. Specifically, for pulmonary metastases, the available clinical data implies that in most cases, exponential growth is a good enough approximation for the time period of observation (Collins et al., 1956; Sabra et al., 2017). Yet, different types of pulmonary metastases may vary in their growth pattern, in the natural history of the disease, and also in the possible different effects of surgery on the growth of the remaining metastases. Hence, the analysis of longitudinal clinical data of specific metastases dynamics is essential in order to characterize metastasis growth and pave the way to individualized prognosis and therapy.

Lately, we have published an analysis of data from a rare test case of a metastatic CRC patient, with untreated growth of 10 lung metastases repeatedly measured over 3 years (Hochman et al., 2019). We have shown that exponential growth can be approximated to all metastases and that metastases were initiated at least 8–11 years before the primary disease detection. Here, we present another unique test case of a colon cancer patient with measured growth of untreated lung metastases. These metastases were first detected 2.6 years after primary tumor resection, and 1.7 years after a liver metastasis was resected in a second operation. This case is different from the former (Hochman et al., 2019) in the location of the primary tumor—sigmoid colon in this case, and rectal in the former case, and also in the fact that here there were two metastatic locations (liver and lungs), and two operations were conducted. These distinctions imply a different type of lung metastases, with a possibly different route of metastatic spread, which may induce a different course of the natural history of metastases. We analyze the current case in the same way as in the former case, examining the validity of exponential, logistic, and Gompertzian approximations, and estimate the natural history (i.e., time of onset) of metastases. We address the question of whether former conclusions are also valid for this case.

In addition, the primary tumor, in this case, was detected and removed relatively early, at the size of 0.68 cm^3, when the liver and lung metastases were still undetectable. This is compared to the former case, where at the time of disease detection there was a 6-cm^3 tumor in the rectum, a colonic polyp, and at least eight lung metastases were already detectable. Here, we re-examine the prediction of early metastasis onset time, not only for metastases observed at first detection of the primary (as in Hochman et al., 2019) but also for metastases that were occult at the time of disease detection and observed only 2 years afterward. In this case of late-detected metastases, we also faced a more difficult task as we wanted to study the effects of the two surgeries on metastasis growth.

METHODS

Data

A 59-year-old patient was diagnosed with sigmoid colon cancer and underwent resection, revealing stage T2N1M0. The measured volume of the primary tumor at surgery was 0.68 cm^3. Adjuvant chemotherapy was given (5FU for 6 months and oxaliplatin, which was stopped after one cycle because of allergic reaction). One year after primary tumor removal, a metastasis was discovered in the liver and was resected. CT scans at the time of the first diagnosis and at the time of liver metastasectomy did not detect any metastases in the lungs. After liver metastasectomy (1.7 years), chest CT showed four metastases. Three additional scans were conducted in the following 2 years. The measured volumes of metastases at these time points are reported in **Supplementary Table 1** and visualized in **Figure 1**. Metastases were peripheral, with no large vessels observed near them, which negates vascularization effects on metastasis growth rate. During this period, systemic treatment (chemotherapy, targeted treatment) was not administered due to patient preference.

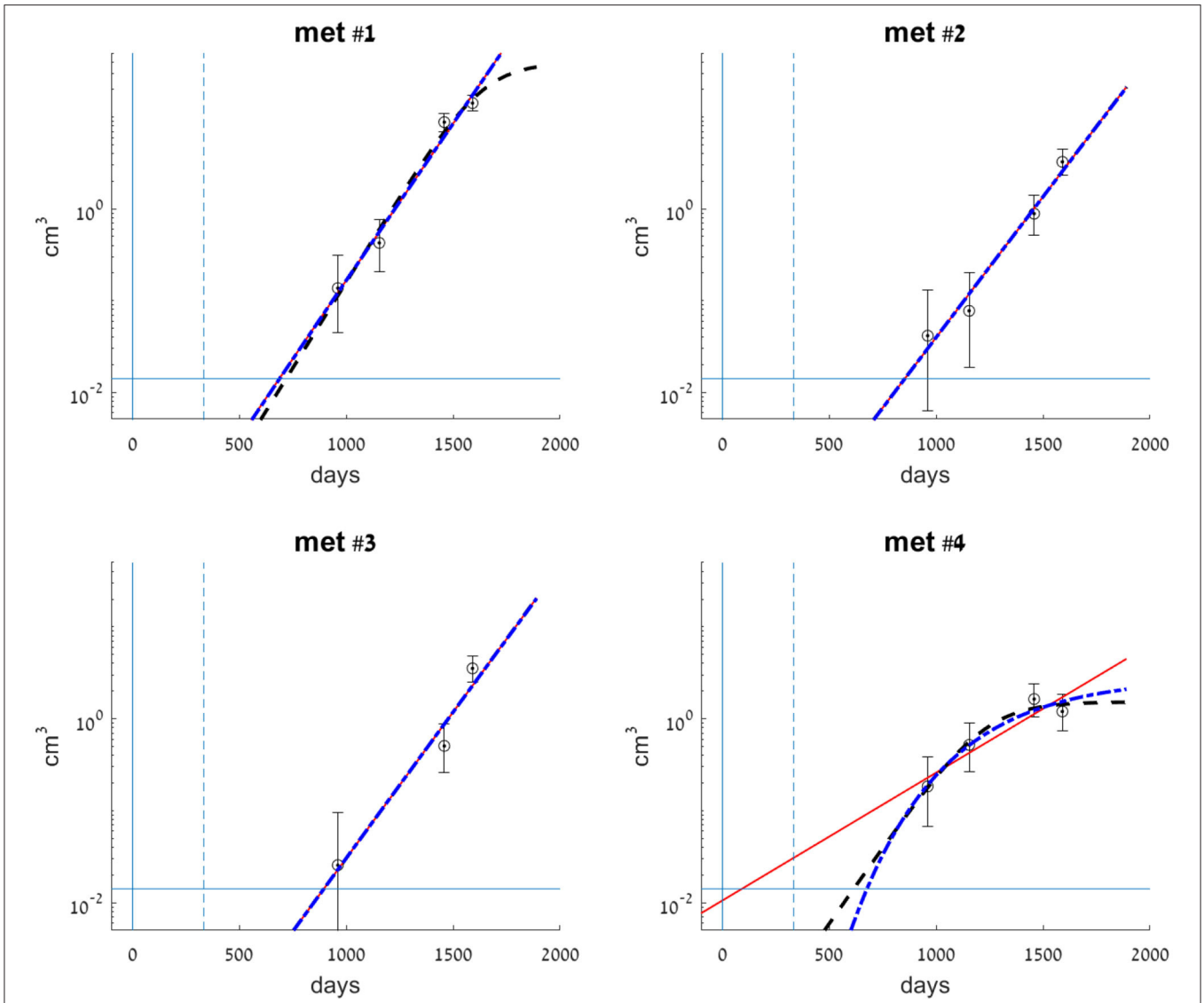

FIGURE 1 | Clinical data measurements (circles) for each of the observed metastases, compared with exponential (smooth red line), logistic (dashed black line), and Gompertzian (dashed-dotted blue line) growth laws fitted to the data. Metastases volumes are presented on a logarithmic scale. Model parameters for each of the metastases were fitted to its observed volumes (see **Supplementary Table 1**). Error bars for each of the clinical data measurements represent errors in volume, where the measurement error is ±2 mm in each dimension of the lesion. Vertical lines at $t = 0$ (smooth) and $t = 333$ (dashed) mark the days of primary colon tumor resection and liver metastasis resection, respectively. The horizontal line at 0.014 cm^3 marks the detection limit of the CT scan.

Modeling

Based on the measurement data obtained, we wanted to fit a growth model (exponential, logistic, or Gompertzian) for each of the metastases and assess growth rate parameter values using the same methods described in Hochman et al. (2019).

Exponential growth was modeled by:

$$\Psi\,(t) = N_0^{exp} e^{\lambda t}, \tag{1}$$

where Ψ (t) is the metastasis volume at time t, counted from the day of primary tumor resection, N_0^{exp} is the size of metastasis at $t = 0$, and λ is the growth rate parameter.

Logistic growth was modeled by:

$$\Theta\,(t) = \frac{K^{logistic}}{1 + \left(\frac{K^{logistic}}{N_0^{logistic}} - 1\right) e^{-\,rt}}, \tag{2}$$

where Θ (t) is metastasis volume at time t, $N_0^{logistic}$ is the size of metastasis at t=0, $K^{logistic}$ is the upper limit of tumor size (carrying capacity), and r is a rate parameter.

Gompertzian growth was modeled by:

$$\Phi\,(t) = K^{gomp} e^{\ln\left(\frac{N_0^{gomp}}{K^{gomp}}\right) e^{-\beta t}}, \tag{3}$$

where $\Phi(t)$ is metastasis volume at time t, N_0^{gomp} is the size of metastasis at t=0, K^{gomp} is the limiting tumor size, and β is a rate parameter.

Tumor doubling time (TDT) can be calculated, in the case of exponential growth, from the growth rate parameter (λ in Equation 1), using the equation:

$$TDT = ln(2)/\lambda. \quad (4)$$

In case of logistic growth, when $\Theta(t) \ll K^{logistic}$, an approximation of Equation 2 gives *TDT* the same as in Equation 4, with *r* instead of λ.

The direct fit of the data was carried out for each of the metastases separately, to optimize the parameter values by numerical minimization of the sum of squared errors (SSE) for model predictions compared to the log-volume of measured tumor sizes:

$$SSE = \sum_{i=1}^{n} \left(ln\left(f\left(t_i, p\right)\right) - ln\left(Y_i\right) \right)^2, \quad (5)$$

where n is the total number of available measurements, Y_i is the observed metastasis volume at time t_i, and $f\left(t_i, p\right)$ is the predicted metastasis volume at the same time, as calculated by each of the model equations (Equations 1–3), depending on the estimated parameters vector p, which includes the two or three parameters of the relevant model equation.

The search was limited to biologically feasible parameter values: $N_0^{exp} \geq 0$, $\lambda \geq 0$ In Equation 1, $N_0^{logistic} \geq 0$, $r \geq 0$, $K^{logistic} \geq 1$ *cell volume* in Equation 2, and $N_0^{gomp} \geq 0$, $\beta \geq 0$, $K^{gomp} \geq 10^{-9}\text{cm}^3$ in Equation 3. Note, that for all three models, values of $0 < N_0 < 10^{-9}\, cm^3 = 1\, cell\, volume$ mean that the time of inception of metastasis (defined as time of appearance of the first malignant cell) is after the time of tumor resection, defined at $t = 0$.

We also assume a minimal biologically plausible value for metastasis doubling time. According to the reported statistical data, the range of TDT values starts from 28.2 days (Tomimaru et al., 2018) or even 22 days (Chojniak and Younes, 2003), as measured in groups of 65 and 21 patients with CRC pulmonary metastases, respectively. Therefore, we limited the selection of parameter values to obtain a minimum TDT value of 25 days, from the time of the onset of metastasis until the time when the threshold volume for detection by CT scan was reached. This threshold is approximated as 0.014 cm^3, which is the volume of a spherical lesion with a 3-mm diameter (Bankier et al., 2017). The procedure was performed using the Matlab functions *lsqnonlin*, *nlinfit*, and *fmincon*.

Subsequently, the fitted models were used to estimate the time of onset of metastasis. For this purpose, the fitted curve with estimated parameters for each metastasis k was extrapolated backward to determine the time of onset of metastasis (T_k), defined as the time of appearance of the first malignant cell. In the same way, we assessed the earliest possible detection time (D_k), defined as the time of metastasis size reaching the threshold enabling detection by CT scan, defined above as 0.014 cm^3.

Error Estimation

The maximal experimental error in measuring the metastatic volumes was ±2 mm in each dimension of the lesion, which is assumed to be spherical. For each reported data point, we calculated the measured diameter of a sphere, and the measurement error in volume (reflected in the error bars in **Figure 1**) was estimated according to this measured diameter ±2 mm.

To assess the reliability of the fitted models within the measurement errors, a sensitivity analysis was conducted, by simulating 1,000 random samples of artificial data, uniformly distributed within these error bars. For each of these samples, we have performed the model fit and obtained parameter values. Then, we have analyzed the distribution of the resulting fitted parameter values and of the estimated times for metastasis formation (T_k) and metastasis earliest detection time (D_k), which are directly defined by the fitted parameters. The mean, median, relative standard error, and interdecile range (i.e., difference between the first and the ninth deciles, 10 and 90%) of the fitted parameter values were calculated.

RESULTS

Fitting and Comparing Growth Models

We have fitted each of the growth models examined to each of the four metastases. Values for the parameters of each of the three models were fitted to the dataset of all three or four available measurements in time. The parameters' optimal values are presented in **Table 1**. Fitted curves are presented, along with clinical measurements in **Figure 1**. The SSE score of the goodness of fit is also detailed in **Table 1**.

Metastases #1–3 were constantly growing over the entire time period examined. In general, the exponential growth model provided a good fit for these metastases (see **Figure 1**). Logistic and Gompertzian models were, in most cases, redundant; they converge with extremely high values of the parameter K (see **Table 1**), i.e., they essentially degenerate into an exponential model. For metastasis #1, the logistic model showed a slightly better fit than the exponential model (with lower SSE value, **Table 1**). However, since the exponential model is simpler, i.e., with one less parameter, and since the difference between the two models' predictions in the time period of interest (to the date of the last measure) was small, we considered also this metastasis as exponentially growing. Furthermore, sensitivity analysis has shown that logistic model parameter values are more sensitive to changes in the measured sizes within the measurement errors (see **Supplementary Table 3**, rows 5–9 compared to rows 1–4 and **Supplementary Figures 1–11**). Therefore, the exponential model is more reliable in this case.

For metastasis #4, the last measure showed growth had stopped; hence, the exponential model demonstrated poor accuracy. Gompertzian model is not reliable in this case, as its parameters optimal values are very sensitive to changes in data (see sensitivity analysis results, **Supplementary Table 3**, last five rows, and **Supplementary Figures 12–17**). The logistic model yielded the best fit to actual growth measurements (SSE value, **Table 1**). Note that the data shows a slight decrease in volume;

TABLE 1 | Values of estimated optimal parameters for each of the analyzed metastases, of the three fitted models, along with the value of SSE (Equation 5).

	Exponential			Logistic				Gompertz			
	N_0^{exp} [cm^3]	λ [years^{-1}]	SSE	$N_0^{logistic}$ [cm^3]	$K^{logistic}$ [cm^3]	r [years^{-1}]	SSE	N_0^{gomp} [cm^3]	K^{gomp} [cm^3]	β [years^{-1}]	SSE
met #1	6.35E−05	2.87	0.26	2.86E−05	39.23	3.15	0.22	5.91E−05	2.84E+296	4.21E−03	0.26
met #2	3.42E−05	2.58	0.36	3.42E−05	7.10E+06	2.58	0.36	3.24E−05	6.11E+307	3.63E−03	0.36
met #3	2.03E−05	2.67	0.49	2.03E−05	2.35E+07	2.67	0.49	1.92E−05	6.30E+307	3.76E−03	0.50
met #4	0.0106	1.17	0.36	1.29E−04	1.52	2.80	0.09	9.23E−12	2.81	0.87	0.14

In case of small or no difference between the Gompertzian or Logistic model and the exponential one, we chose the exponential model, which is simpler and more reliable. Note, that for metastasis #3, there were only three data points available; hence, models with three parameters (i.e., Gompertzian and logistic) should reproduce three data points exactly. However, since we have limited the numerical search to feasible values (see section Materials and Methods), the fit of Gompertzian and logistic models converged to exponential growth. Colored lines represent the chosen models, for which the sum of squared error (SSE) value was smallest.

however, this decrease is within the measurement error range. Therefore, the fitted logistic model shows that the metastasis' volume has reached its capacity, and the fitted value of $K^{logistic}$ (**Table 1**) is close to the last two measured values (and within the measurement error range, as shown in **Figure 1**).

In conclusion, for metastases #1–3, the exponential growth model is the preferable one, while for metastasis #4, the logistic model showed the best fit.

Metastasis Growth Rate

For the exponentially growing metastases (#1–3), the values of the exponent of the growth rate λ are all in the same order of magnitude, averaged 2.71 years^{-1}, with a standard deviation of 0.15 year^{-1}. This value corresponds to a tumor doubling time of 93 days (Equation 4). For metastasis #4, TDT can be approximated for the first period, when growth is close to exponential. In this case, the logistic growth rate is represented by the parameter r in Equation 2. Its fitted value was 2.80 years^{-1}, corresponding to TDT = 90 days, which is close to the exponential growth rate of metastases #1–3.

Assessing Metastasis Natural History

If we assume each metastasis has followed the same growth law since its inception, then the metastasis onset time (i.e., time of emergence of the first malignant clonogenic cell), T_k, can be estimated for each metastasis #k. Backward extrapolation of the fitted growth curves can be used to find the time when metastasis volume is one cell, according to the model. The earliest possible detection time (i.e., time of metastasis size reaching to the threshold enabling detection by CT scan), D_k, can be evaluated in the same way, extrapolating to the time when tumor size according to the model is 0.014 cm^3. This extrapolation, according to the best-fitted growth curve—logistic for metastasis #4 and exponential for the others—is presented in **Figure 2**. Calculated values for T_k and D_k for every metastasis, by each of the three fitted models, are presented in **Supplementary Table 2**. The results show that all metastases were formed around 4 years before the primary tumor was detected. Yet, the earliest possible time when a metastasis could be detected was only after the second (liver metastasis) resection, marked as D_1-D_4 in **Figure 2**.

Note that the calculated values for T_4 and D_4 based on the logistic model are more sensitive to the measurement error than those based on exponential fit (see **Supplementary Table 3**, rows 6–7 from the end). Hence, the latter conclusion should not be taken as certain for metastasis #4.

DISCUSSION

Understanding metastasis dynamics and growth is essential for improving cancer therapy, especially toward individualization of treatment. Retrospective statistical data can recognize patterns of metastasis growth in different subgroups of patients but cannot decipher the reasons for the difference between subgroups. Analysis of specific cases, particularly utilizing clinical dynamical data of metastasis growth, is necessary to gain a deeper understanding of the metastatic process, and eventually provide reliable individual prognosis and treatment plans. In this work, we used unique data of a metastatic CRC patient to explore the dynamics of untreated lung metastasis growth. We concluded the natural history of the disease and how it is affected by factors like surgical intervention.

In the test case examined here, three lung metastases (metastases #1–3) constantly grew, and for them, *exponential growth* was found to be a good approximation. The estimated exponential growth rates of all metastases were quite similar, implying that variability between metastasis growth rates can be neglected. This result agrees with the former analyzed case (Hochman et al., 2019). For the fourth metastasis, it seems that growth has stopped during the time period in which measures were taken (**Figure 1**), showing that the metastasis growth ability has reached a certain limit. This blockage of the increase was observed in parallel with cavitation formation in the lesion, as observed on the CT scans (**Figure 3**). In this case, the cavity volume is included in the reported measured volume. However, it forms a negligible portion of the lesion volume; hence, the halt in growth is not a direct effect of the cavitation. Yet, cavitary lesions may behave differently, as they are composed of heterogeneous tissue. Here, any unknown process that causes the observed deceleration of growth is implicitly modeled as a logistic decay of the metastasis growth rate.

In general, results imply that the metastatic growth is logistically bounded, although in most cases, exponential growth can be approximated for the time period of measures. This is in line with the exponential growth

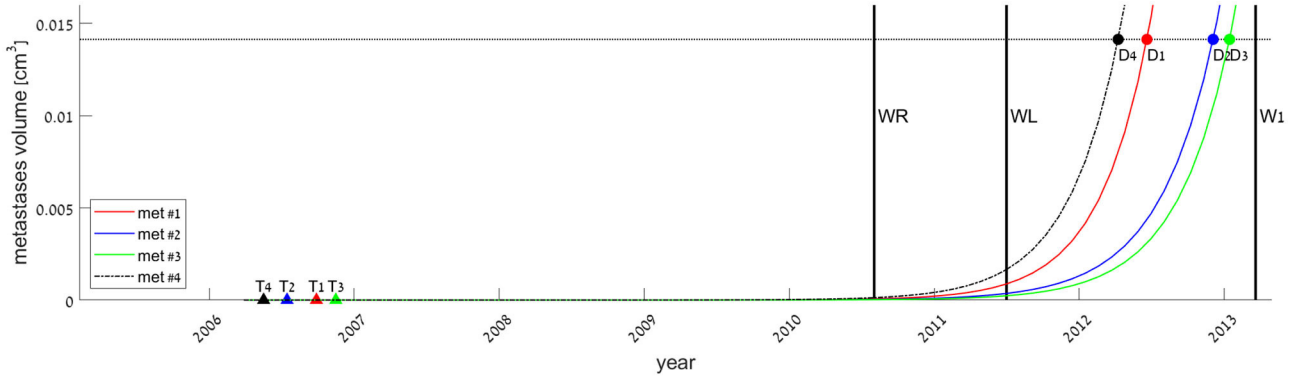

FIGURE 2 | Estimated times of onset of metastases, defined as the time of appearance of the first malignant cell (filled triangles, marked T_k for each metastasis #k), and estimated times of metastases' earliest possible detection time, defined as the time of metastasis size reaching the threshold enabling detection by CT scan (filled circles, D_k). T_k and D_k values were extrapolated from the fitted models, exponential (smooth lines, for metastases #1–3) or logistic (dashed-dotted, for metastasis #4). Vertical lines represent the dates of primary tumor resection (WR), liver metastasis resection (WL), and first measure of lung metastases (W1). The horizontal line at 0.014 cm^3 represents the detection limit of the CT scan.

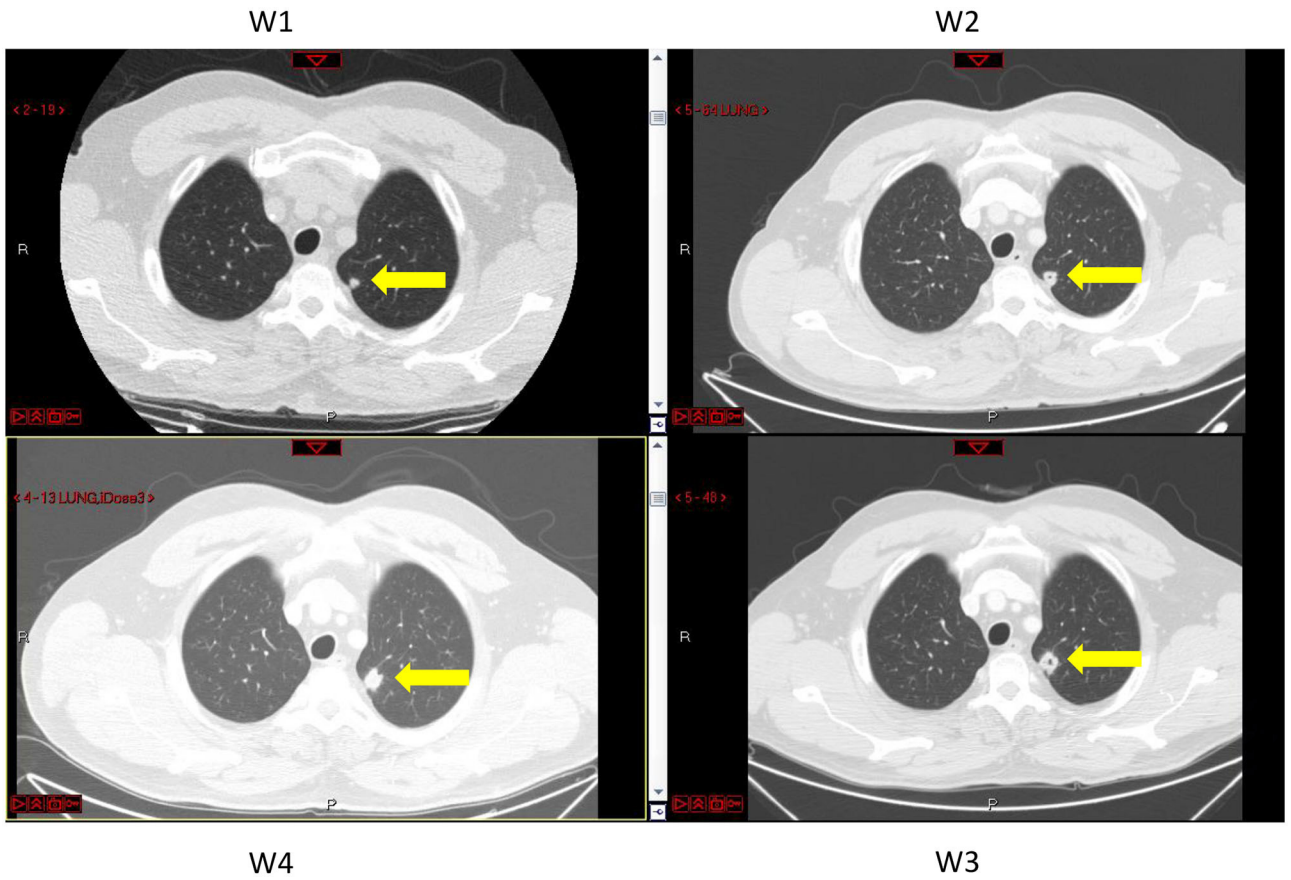

FIGURE 3 | CT images of the lungs taken at different times, marked W1 to W4, at which metastasis volumes were measured (see **Figure 1** and **Supplementary Table 1**). Metastasis #4 is marked by yellow arrows. A pore in the center of the lesion can be seen at times W2 and W3.

pattern that had been observed for pulmonary metastases from CRC (Collins et al., 1956) and thyroid cancer (Sabra et al., 2017).

Metastasis average growth rate, which is 2.71 $years^{-1}$ for the exponentially growing metastases here, corresponding to TDT of 93 days, seems higher than the previously reported

rate of 1.48 years^{-1} in Hochman et al. (2019) (TDT of 171 days). Heterogeneity of tumor aggressiveness is common among different patients. However, here we can surmise that the reason for this difference is related to the effect of the hepatic metastasectomy. In this case the patient underwent two surgeries, first to remove the primary tumor and later to remove the liver metastasis, in contrast to the formerly published case in which only one operation (for primary tumor resection) was conducted.

There is evidence that tumor resection has implications that accelerate the metastasis growth, both because of the stress response caused by surgery (Maida et al., 2009; Tohme et al., 2017; Behrenbruch et al., 2018; Zheng et al., 2018) and due to the removal of the inhibiting effect that the resected tumor had induced on metastases (Retsky et al., 2010; Benzekry et al., 2017; Hanin and Rose, 2018). From our data, we cannot determine what was the metastasis growth rate before the surgeries. However, we hypothesize that in the secondary (lung) metastases, during their growth, the patient undergoes two surgeries, growth rate increases even more than after a single surgery. Besides, the primary tumor was in the sigmoid colon; therefore, it is likely that metastatic spread was via the portal circulation to the liver, and not directly to the lungs (Riihimäki et al., 2016). We can hypothesize that metastases that have developed in a later stage, as a "metastasis of metastasis," might represent a more aggressive phenotype.

The aggressiveness of metastases can be caused by other unknown variables. Hence, a general conclusion from a comparison between cases is limited. With that being said, this case can be compared to statistical data available in the literature for the TDT of CRC pulmonary metastases. Reported mean values of TDT range between 109 (Spratt and Spratt, 1964) and 129 days (Tomimaru et al., 2018). In our case, the growth is faster than this reported range. It is close to the TDT range reported for liver metastases, which are known as more aggressive (Nomura et al., 1998). In summary, the notable aggressiveness of metastases in this case, after two surgeries, supports the assumption that each event of surgery leads to faster of metastases.

The natural history of the metastases is evaluated to estimate prognosis and develop an optimal individualized treatment plan. In this case, metastasis formation time (T) and the earliest possible detection time (D) were restored from models of growth in a later period. Note that T can be related to as a parameter of the model, and it can be negative or positive (i.e., before or after primary tumor detection time).

Backward extrapolation of the growth models fitted to data of a later period (after the surgeries) suggests that at least three of the four metastases were seeded about 4 years before, yet could not have been discovered until 1.7 years after primary tumor resection, at the earliest (**Figure 2**). This extrapolated result is true if the growth rate remains the same from the time of metastasis inception. However, this is quite unlikely, as the implications of the two resections, which were discussed above, may cause acceleration of the metastasis growth. If we consider our model to be correct only for the time *after* the second surgery and assume that growth was *slower* before this surgery, T would be even earlier (although D would not be affected). Therefore, the T values extrapolated here represent the *latest possible* estimated time of metastasis formation.

Formerly published works (Benzekry et al., 2014; Bilous et al., 2019; Vaghi et al., 2020) suggest that the Gompertzian model describes best the metastatic growth, and that considering Gompertzian growth instead of exponential may change extrapolation results, as the curves differ greatly at early times. However, in our case, the Gompertzian models for metastases #1–3 degenerate into exponent, which means that our data is given in an early period of time in the metastatic process when metastasis sizes had not reached their capacity. The Gompertzian model in this period coincides with the exponential curve; therefore, it would make no difference in the predicted value of T (see **Supplementary Table 2**). Hence, we can conclude that metastases were seeded about 4 years before disease detection and stayed occult until 2 years after it. This result agrees with the empirically supported notion that many metastases are seeded before the primary tumor is even detectable (Hanahan and Weinberg, 2011; Siegel et al., 2017). Lately, analysis of genomic exome-sequencing data has shown that liver and brain metastases in CRC are probably seeded at early stages of disease (Hu et al., 2019).

This work is based on the data of a single patient. Different growth patterns might apply to other cases with different primary tumors and metastatic sites because inter-patient variability is high. Since such clinical data of untreated metastasis growth is not common, robust conclusions on the metastatic growth pattern are difficult to achieve. An example of a way to deal with this challenge is a population model approach, which was used lately by Benzekry et al. to analyze clinical data from brain metastases in NSCLC (Bilous et al., 2019). Their model, comprising of metastasis dissemination and size distribution as a function of primary tumor size, suggests the Gompertzian growth law as most suitable to the data. Like in our case, it is shown that metastases have already been disseminated, but were still occult at the time of disease detection. Unlike in our case, the choice of Gompertzian model makes a significant difference in predicting metastasis onset time, because it differs from exponential curve at early times. This result may suggest that Gompertzian growth is more appropriate to use for prediction of metastasis natural history. Yet, it was achieved for a different type of cancer, and it is still necessary to analyze the clinical data of CRC pulmonary metastases, and more specifically, the data of different subtypes of CRC lung metastases, in order to understand the metastatic process in the relevant type of disease.

In order to reveal how personal characteristics affect metastasis growth pattern, different cases should be studied, thus the unique data of the test case published here are valuable. These data may be used for further analyses with different modeling methods. We intend to do this with a model that includes metastatic dissemination as a function of primary tumor size in a future publication.

In summary, from the unique clinical data of metastasis growth dynamics, we conclude that:

- Untreated lung metastasis growth is logistic, but in most cases, exponential law is a legitimate approximation, as metastases do not verge on the limitation on lesion growth.
- Metastases can be initiated before the primary tumor is detectable (in this case, at least 4–5 years before the primary tumor was detected).
- Surgical removal of the primary tumor or metastasectomy might lead to faster-growing metastases. This is implied by notably short TDT in this case after two surgeries.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SB-M designed the research. GH and SRo performed the research. SB-M and GH contributed analytic tools. SB-M, GH, and ES-S analyzed the data. SRa measured the metastases. GH and ES-S wrote the paper. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Helen Byrne for her advice about tools for examination of model reliability.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.533101/full#supplementary-material

REFERENCES

Bankier, A. A., MacMahon, H., Goo, J. M., Rubin, G. D., Schaefer-Prokop, C. M., and Naidich, D. P. (2017). Recommendations for measuring pulmonary nodules at CT: a statement from the fleischner society. *Radiology* 285, 584–600. doi: 10.1148/radiol.2017162894

Baratchart, E., Benzekry, S., Bikfalvi, A., Colin, T., Cooley, L. S., Pineau, R., et al. (2015). Computational modelling of metastasis development in renal cell carcinoma. *PLoS Comput. Biol.* 11:e1004626. doi: 10.1371/journal.pcbi.1004626

Behrenbruch, C., Shembrey, C., Paquet-Fifield, S., Mølck, C., Cho, H. J., Michael, M., et al. (2018). Surgical stress response and promotion of metastasis in colorectal cancer: a complex and heterogeneous process. *Clin. Exp. Metastasis* 35, 333–345. doi: 10.1007/s10585-018-9873-2

Benzekry, S., Lamont, C., Barbolosi, D., Hlatky, L., and Hahnfeldt, P. (2017). Mathematical modeling of tumor–tumor distant interactions supports a systemic control of tumor growth. *Cancer Res.* 77, 5183–5193. doi: 10.1158/0008-5472.CAN-17-0564

Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J. M., Hlatky, L., et al. (2014). Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput. Biol.* 10:e1003800. doi: 10.1371/journal.pcbi.1003800

Benzekry, S., Tracz, A., Mastri, M., Corbelli, R., Barbolosi, D., and Ebos, J. M. (2016). Modeling spontaneous metastasis following surgery: an *in vivo-in silico* approach. *Cancer Res.* 76, 535–547. doi: 10.1158/0008-5472.CAN-15-1389

Bilous, M., Serdjebi, C., Boyer, A., Tomasini, P., Pouypoudat, C., Barbolosi, D., et al. (2019). Quantitative mathematical modeling of clinical brain metastasis dynamics in non-small cell lung cancer. *Sci Rep.* 9:13018. doi: 10.1038/s41598-019-49407-3

Brú, A., Albertos, S., Subiza, J. L., García-Asenjo, J. L., and Brú, I. (2003). The universal dynamics of tumor growth. *Biophys. J.* 85, 2948–2961. doi: 10.1016/S0006-3495(03)74715-8

Chojniak, R., and Younes, R. N. (2003). Pulmonary metastases tumor doubling time: assessment by computed tomography. *Am. J. Clin. Oncol.* 26, 374–377. doi: 10.1097/01.COC.0000026481.38654.52

Collins, V. P., Loeffler, R. K., and Tivey, H. (1956). Observations on growth rates of human tumors. *Am. J. Roentgenol. Radium Ther. Nucl. Med.* 76, 988–1000.

Fisher, B. (1980). Laboratory and clinical research in breast cancer—a personal adventure: the David A. Karnofsky memorial lecture. *Cancer Res.* 40, 3863–3874.

Fisher, B., Montague, E., Redmond, C., Barton, B., Borland, D., Fisher, E. R., et al. (1977). Comparison of radical mastectomy with alternative treatments for primary breast cancer: a first report of results from a prospective randomized clinical trial. *Cancer* 39, 2827–2839.

Haeno, H., Gonen, M., Davis, M. B., Herman, J. M., Iacobuzio-Donahue, C. A., and Michor, F. (2012). Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies. *Cell* 148, 362–375. doi: 10.1016/j.cell.2011.11.060

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/J.CELL.2011.02.013

Hanin, L., and Bunimovich-Mendrazitsky, S. (2014). Reconstruction of the natural history of metastatic cancer and assessment of the effects of surgery: gompertzian growth of the primary tumor. *Math. Biosci.* 247, 47–58. doi: 10.1016/j.mbs.2013.10.010

Hanin, L., and Rose, J. (2018). Suppression of metastasis by primary tumor and acceleration of metastasis following primary tumor resection: a natural law? *Bull. Math. Biol.* 80, 519–539. doi: 10.1007/s11538-017-0388-9

Hanin, L., Seidel, K., and Stoevesandt, D. (2016). A "universal" model of metastatic cancer, its parametric forms and their identification: what can be learned from site-specific volumes of metastases. *J. Math. Biol.* 72, 1633–1662. doi: 10.1007/s00285-015-0928-6

Hartung, N., Mollard, S., Barbolosi, D., Benabdallah, A., Chapuisat, G., Henry, G., et al. (2014). Mathematical modeling of tumor growth and metastatic spreading: validation in tumor-bearing mice. *Cancer Res.* 74, 6397–6407. doi: 10.1158/0008-5472.CAN-14-0721

Hochman, G., Shacham-Shmueli, E., Heymann, T., Raskin, S., and Bunimovich-Mendrazitsky, S. (2019). Metastases growth patterns *in vivo*—a unique test case of a metastatic colorectal cancer patient. *Front. Appl. Math. Stat.* 5:56. doi: 10.3389/fams.2019.00056

Hu, Z., Ding, J., Ma, Z., Sun, R., Seoane, J. A., Shaffer, J. S., et al. (2019). Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* 51, 1113−1122. doi: 10.1038/s41588-019-0423-x

Iwata, K., Kawasaki, K., and Shigesada, N. A. (2000). Dynamical model for the growth and size distribution of multiple metastatic tumors. *J. theor. Biol.* 203, 177–186. doi: 10.1006/jtbi.2000.1075

Kozusko, F., and Bajzer, Ž. (2003). Combining gompertzian growth and cell population dynamics. *Math. Biosci.* 185, 153–167. doi: 10.1016/S0025-5564(03)00094-4

Maida, V., Ennis, M., Kuziemsky, C., and Corban, J. (2009). Wounds and survival in cancer patients. *Eur. J. Cancer* 45, 3237–3244. doi: 10.1016/J.EJCA.2009.05.014

Maithel, S. K., Ginsberg, M. S., D'Amico, F., DeMatteo, R. P., Allen, P. J., Fong, Y., et al. (2010). Natural history of patients with subcentimeter pulmonary nodules undergoing hepatic resection for metastatic colorectal cancer. *J. Am. Coll. Surg.* 210, 31–38. doi: 10.1016/j.jamcollsurg.2009.09.032

Nomura, K., Miyagawa, S., Harada, H., Kitamura, H., Seki, H., Shimada, R., et al. (1998). Relationship between doubling time of liver metastases from colorectal carcinoma and residual primary cancer. *Dig. Surg.* 15, 21–24. doi: 10.1159/000018581

Retsky, M., Demicheli, R., Hrushesky, W., Baum, M., and Gukas, I. (2010). Surgery triggers outgrowth of latent distant disease in breast cancer: an inconvenient truth? *Cancers* 2, 305–337. doi: 10.3390/cancers2020305

Riihimäki, M., Hemminki, A., Sundquist, J., and Hemminki, K. (2016). Patterns of metastasis in colon and rectal cancer. *Sci. Rep*. 6:29765. doi: 10.1038/srep29765

Sabra, M. M., Sherman, E. J., and Tuttle, R. M. (2017). Tumor volume doubling time of pulmonary metastases predicts overall survival and can guide the initiation of multikinase inhibitor therapy in patients with metastatic, follicular cell-derived thyroid carcinoma. *Cancer* 123, 2955–2964. doi: 10.1002/cncr.30690

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G., Barzi, A., et al. (2017). Colorectal cancer statistics, 2017. *CA. Cancer J. Clin*. 67, 177–193. doi: 10.3322/caac.21395

Spratt, J. S. Jr., and Spratt, T. L. (1964). Rates of growth of pulmonary metastases and host survival. *Ann. Surg.* 159, 161–171.

Stein, U., and Schlag, P. M. (2007). "Clinical, biological, and molecular aspects of metastasis in colorectal cancer," in *Targeted Therapies in Cancer*, ed M. Dietel, (Heidelberg: Springer), 61–80.

Stewart, C. L., Warner, S., Ito, K., Raoof, M., Wu, G. X., Lu, W. P., et al. (2018). Cytoreduction for colorectal metastases: liver, lung, peritoneum, lymph nodes, bone, brain. When does it palliate, prolong survival, and potentially cure? *Curr. Probl. Surg.* 55, 330–379. doi: 10.1067/j.cpsurg.2018.08.004

Tohme, S., Simmons, R. L., and Tsung, A. (2017). Surgery for cancer: a trigger for metastases. *Cancer Res.* 77, 1548–1552. doi: 10.1158/0008-5472.CAN-16-1536

Tomimaru, Y., Noura, S., Ohue, M., Okami, J., Oda, K., Higashiyama, M., et al. (2018). Metastatic tumor doubling time is an independent predictor of intrapulmonary recurrence after pulmonary resection of solitary pulmonary metastasis from colorectal cancer. *Dig. Surg.* 25, 220–225. doi: 10.1159/000140693

Vaghi, C., Rodallec, A., Fanciullino, R., Ciccolini, J., Mochel, J. P., Mastri, M., et al. (2020). Population modeling of tumor growth curves and the reduced Gompertz model improve prediction of the age of experimental tumors. *PLoS Comput. Biol.* 16:e1007178. doi: 10.1371/journal.pcbi.1007178

Van Cutsem, E., Cervantes, A., Nordlinger, B., Arnold, D., and ESMO Guidelines Working Group. (2014). Metastatic colorectal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 25(Suppl. 3), iii1–iii9. doi: 10.1093/annonc/mdu260

Vatandoust, S., Price, T. J., and Karapetis, C. S. (2015). Colorectal cancer: metastases to a single organ. *World J. Gastroenterol*. 21, 11767–11776. doi: 10.3748/wjg.v21.i41.11767

Zheng, J., Jia, L., Mori, S., and Kodama, T. (2018). Evaluation of metastatic niches in distant organs after surgical removal of tumor-bearing lymph nodes. *BMC Cancer* 18:608. doi: 10.1186/s12885-018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 December 2020
doi: 10.3389/fgene.2020.565206

Identification and Validation of Two Lung Adenocarcinoma-Development Characteristic Gene Sets for Diagnosing Lung Adenocarcinoma and Predicting Prognosis

Cheng Liu[1]†, Xiang Li[2†], Hua Shao[2] and Dan Li[2]*

[1]Department of Thoracic Surgery, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [2]Department of Neurology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Padhmanand Sudhakar, Katholieke Universiteit (KU) Leuven, Belgium
Priyanka Baloni, Institute for Systems Biology (ISB), United States

***Correspondence:**
Cheng Liu
liuchengdoctor@163.com

†These authors share first authorship

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: 01 July 2020
Accepted: 26 November 2020
Published: 21 December 2020

Citation:
Liu C, Li X, Shao H and Li D (2020) Identification and Validation of Two Lung Adenocarcinoma-Development Characteristic Gene Sets for Diagnosing Lung Adenocarcinoma and Predicting Prognosis. Front. Genet. 11:565206. doi: 10.3389/fgene.2020.565206

Background: Lung adenocarcinoma (LUAD) is one of the main types of lung cancer. Because of its low early diagnosis rate, poor late prognosis, and high mortality, it is of great significance to find biomarkers for diagnosis and prognosis.

Methods: Five hundred and twelve LUADs from The Cancer Genome Atlas were used for differential expression analysis and short time-series expression miner (STEM) analysis to identify the LUAD-development characteristic genes. Survival analysis was used to identify the LUAD-unfavorable genes and LUAD-favorable genes. Gene set variation analysis (GSVA) was used to score individual samples against the two gene sets. Receiver operating characteristic (ROC) curve analysis and univariate and multivariate Cox regression analysis were used to explore the diagnostic and prognostic ability of the two GSVA score systems. Two independent data sets from Gene Expression Omnibus (GEO) were used for verifying the results. Functional enrichment analysis was used to explore the potential biological functions of LUAD-unfavorable genes.

Results: With the development of LUAD, 185 differentially expressed genes (DEGs) were gradually upregulated, of which 84 genes were associated with LUAD survival and named as LUAD-unfavorable gene set. While 237 DEGs were gradually downregulated, of which 39 genes were associated with LUAD survival and named as LUAD-favorable gene set. ROC curve analysis and univariate/multivariate Cox proportional hazards analyses indicated both of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score were a biomarker of LUAD. Moreover, both of these two GSVA score systems were an independent factor for LUAD prognosis. The LUAD-unfavorable genes were significantly involved in p53 signaling pathway, Oocyte meiosis, and Cell cycle.

Conclusion: We identified and validated two LUAD-development characteristic gene sets that not only have diagnostic value but also prognostic value. It may provide new insight for further research on LUAD.

Keywords: lung adenocarcinoma, prognostic stratification system, The Cancer Genome Atlas, gene set variation analysis score, predicting prognosis

INTRODUCTION

Lung cancer is the most common cancer (11.6% of the total cases) among men and women in the world, which is also the main cause of cancer death (18.4% of the total cancer deaths; Bray et al., 2018). Non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancer cases (Govindan et al., 2006), and lung adenocarcinoma (LUAD) is one of the main subtypes of NSCLC. Smoking is currently considered to be the main cause of lung cancer. However, LUAD is more likely to occur in women who do not smoke, and the age of patients tends to become younger (Hecht, 1999; Donner et al., 2018). Early, LUAD can be treated by surgery; however, most patients with LUAD are often diagnosed with advanced cancer (Ding et al., 2008). Although target therapy is effective for selected advanced LUAD, the overall survival of patients is poor due to the emergence of drug resistance. Therefore, it has become one of the hot spots in clinical research to find the diagnosis and prognosis indexes of LUAD.

In recent years, high-throughput sequencing technology and gene database have been widely used in the study of cancer diagnosis and prognosis (Feng et al., 2016; Dama et al., 2017; Zhao et al., 2018a; He et al., 2019). For example, EGFR, KRAS, BRAF, and ERBB 2 have been shown to be associated with treatment efficacy and prognosis (Naoki et al., 2002; Mendelsohn and Baselga, 2003; Guan et al., 2013). Moreover, DGCR 5 has been found to be a prognostic indicator and therapeutic target for the diagnosis and treatment of LUAD (Dong et al., 2018). Overexpression of Rcc 2 induces epithelial-mesenchymal metastasis in LUAD, enhances cell mobility, and promotes tumor metastasis (Pang et al., 2017). Overexpression of KIF20A confers malignant phenotype of LUAD by promoting cell proliferation and inhibiting apoptosis (Zhao et al., 2018b). However, most studies do not take the simultaneous changes of multiple genes into account. Moreover, there are few studies on the LUAD-development characteristic gene sets.

In present study, we identified two LUAD-development characteristic gene sets named as LUAD-unfavorable gene set and LUAD-favorable gene set. Gene set variation analysis (GSVA) was used to score individual samples against the two gene sets. Survival analysis and receiver operating characteristic (ROC) curve analysis were used to identify the diagnostic and prognostic capabilities of two gene sets GSVA score, respectively. Both of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score were reliable biomarkers for diagnosing LUAD and independent biomarkers for predicting prognosis.

MATERIALS AND METHODS

The Cancer Genome Atlas (TCGA; Tomczak et al., 2015)[1] and Gene Expression Omnibus (GEO; Barrett et al., 2013)[2] are the international genetic databases, which are publicly accessible and freely available to researchers. In our study, a total of 512 LUAD samples and 57 healthy lung tissue samples were downloaded from TCGA, including 281 stage I LUADs, 121 stage II LUADs, 84 stage III LUADs, and 26 stage IV LUADs. In addition, GSE10072 based on GPL96 platform was downloaded from GEO, including 58 LUAD samples and 49 healthy lung tissue samples. GSE31210 based on GPL570 platform was downloaded from GEO, including 226 LUAD samples and 20 healthy lung tissue samples. The two data sets were used to verify the prognostic value. The "normalizeBetweenArrays" function in the limma package (Ritchie et al., 2015) was used to normalize the gene expression profiles. If a gene responds to multiple probes, the average value of these probes is considered to be the expression value of the corresponding gene. The flow of this study is shown in **Figure 1**.

Differential Expression Analysis and Short Time-Series Expression Miner

In TCGA, the RNA sequencing expression profile of LUAD was displayed as read counts, which was subsequently normalized by voom function (Law et al., 2014) in limma package. Differentially expressed genes (DEGs) in four stages of LUAD were identified using limma package, respectively. $p < 0.01$ adjusted by the false discovery rate (FDR) and $|\log$ fold change(FC)$| > 1.5$ were considered as significance. In the developing of LUAD, if a DEG was gradually upregulated (logFCstage I vs. control < logFCstage II vs. control < logFCstage III vs. control < logFCstage IV vs. control) or gradually downregulated (logFCstage I vs. control > logFCstage II vs. control > logFCstage III vs. control > logFCstage IV vs. control), and then it was considered to be LUAD-development characteristic gene. These genes were organized into different clusters based on expression patterns using short time-series expression miner (STEM; Ernst and Bar-Joseph, 2006).

Survival Analysis and LUAD-Development Characteristic Gene Set

We used the median expression value of each LUAD-development characteristic gene as the cutoff point to dichotomize patients into high-expression group and low-expression group. Moreover, Kaplan Meier survival analysis and log rank method were performed to explore whether the expression level of the LUAD-development characteristic gene is related to the overall survival (OS) time. Survival analysis was performed using survival package[3] in R, and $p < 0.01$ was considered as significance. In our study, a LUAD-development characteristic gene which gradually upregulated with the development of LUAD and associated with poor prognosis of LUAD was considered to be LUAD-unfavorable gene. On the contrary, a LUAD-development characteristic gene which gradually downregulated with the development of LUAD and associated with good prognosis of LUAD was considered to be LUAD-favorable gene.

[1]https://www.cancer.gov/
[2]https://www.ncbi.nlm.nih.gov/geo/
[3]https://CRAN.R-project.org/package=survival

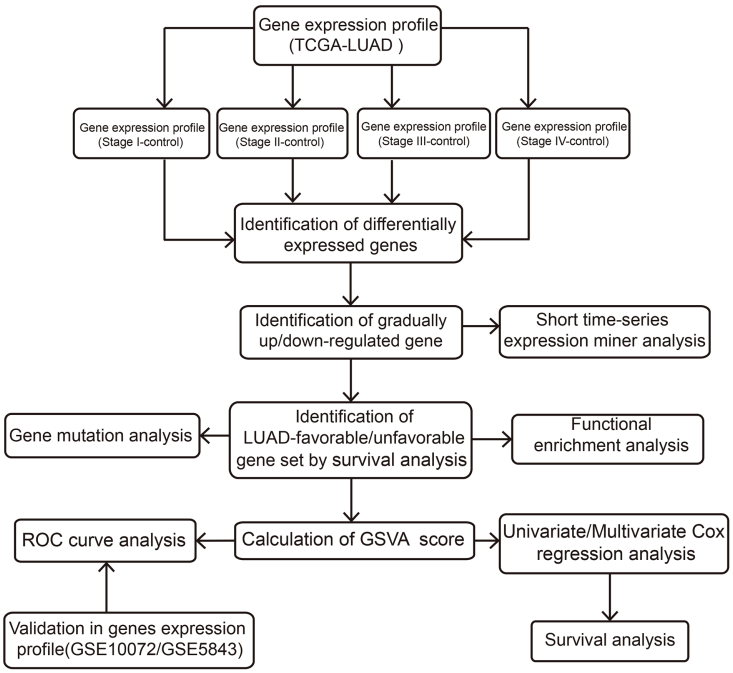

FIGURE 1 | Flowchart of this study.

LUAD-unfavorable genes and LUAD-favorable genes constituted LUAD-unfavorable genes set and LUAD-favorable genes set, respectively.

Calculation of LUAD-Development Characteristic GSVA Score

Gene set variation analysis is a popular method of scoring individual samples for molecular characteristics or gene sets. GSVA package (Hanzelmann et al., 2013) in R was used to calculate LUAD-unfavorable GSVA score and LUAD-favorable GSVA score for individual samples.

ROC Curve Analysis and Univariate/Multivariate Cox Proportional Hazards Analyses

The pROC package (Robin et al., 2011) was used to conduct ROC curve analysis of LUAD-unfavorable GSVA score and LUAD-unfavorable GSVA score to evaluate their ability to diagnose LUAD. Univariate/multivariate Cox proportional hazards analyses were used to compare the relative prognostic value of the two GSVA score systems with that of routine clinicopathological features.

Functional Enrichment Analysis

To further explore the biological function of LUAD-unfavorable genes, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed using the clusterProfiler package (Yu et al., 2012) in R. $p < 0.05$ was considered as significance.

Gene Mutation Analysis and Validation of Differential Expression of LUAD-Unfavorable Genes at Protein Level

In order to explore the potential mechanism about differential expression of LUAD-unfavorable genes, the TCGAbiolinks package (Mounir et al., 2019) was used to download and scan the alteration statuses of LUAD-unfavorable genes. In addition, we randomly selected 10 genes from LUAD-unfavorable gene set and scanned the Human Protein Atlas[4] (Colwill et al., 2011) web tool to validate whether the LUAD-unfavorable genes are upregulated at protein level, compared with normal lung tissue.

RESULTS

Multiple Genes Were Defined as LUAD-Development Characteristic Genes

Compared to normal lung tissue samples, there were 3,082 DEGs in stage I LUADs, 3,437 DEGs in stage II LUADs, 3,518 DEGs in stage III LUADs, and 3,510 DEGs stage IV LUADs (**Figure 2A**). It indicated that the gene expression patterns were various with the development of LUAD. A total of 2,658 common DEGs was in stage I-IV LUADs (**Figure 2B**). Among of them, 185 DEGs were gradually upregulated and 237 DEGs were gradually downregulated with the development of LUAD, which maybe play a crucial role in the LUAD development. The result of STEM demonstrated that two gene clusters were

[4]https://v15.proteinatlas.org/

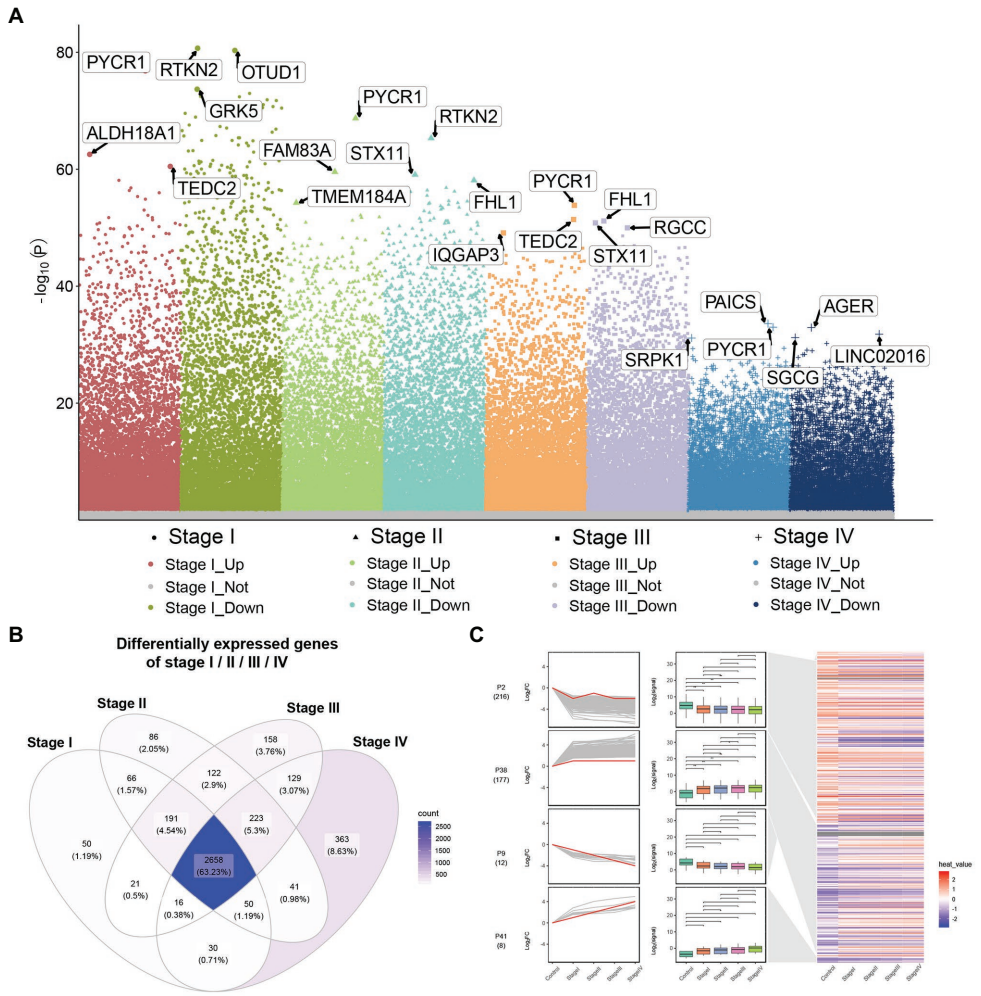

FIGURE 2 | Differential expression gene analysis and short time-series expression miner (STEM) analysis. **(A)** Manhattan plot showed differentially expressed genes (DEGs) in different stage of lung adenocarcinoma (LUAD). Genes with significant differences are highlighted. **(B)** Common DEGs in LUAD stage I–IV. **(C)** STEM results. Line plots (left panels) and box plots (right panels) are used to show fold changes (log2 scale) and absolute expression levels (log2 scale), respectively. In each line plot, one representative gene is highlighted in red.

significantly upregulated, while two gene clusters were significantly downregulated with the development of LUAD (**Figure 2C**).

LUAD-Development Characteristic Genes Were Associated With LUAD Prognosis

The result of survival analysis showed a total of 84 LUAD-development characteristic genes that are gradually upregulated with the development of LUAD and associated with poor prognosis, while a total of 39 LUAD-development characteristic genes that are gradually downregulated with the development of LUAD and associated with good prognosis (**Table 1**). This means that not all LUAD-development characteristic genes are associated with the prognosis of LUAD. In the LUAD-unfavorable gene set, NEK2, CENPK, CDC25C, PLK4, LYPD3, FAM72D, NEIL3, GTSE1, CDK1, and KIF14 were the ten genes with most significant association with poor prognosis (**Figure 3A**). While in the LUAD-favorable gene set, OR7E47P, MS4A2, RAB44, BMP5, ARHGEF6, JAML, TRPC2, HPGDS, HPSE2, and KLK11 were the ten genes with most significant association with good prognosis (**Figure 3B**).

LUAD-Unfavorable GSVA Score and LUAD-Favorable GSVA Score Are Biomarker of LUAD and LUAD Prognosis

As shown in **Figure 4A**, LUAD-favorable GSVA score was gradually downregulated with the development of LUAD, while LUAD-unfavorable GSVA score was gradually upregulated with the development of LUAD. Moreover, the result of ROC curve analysis indicated that both LUAD-unfavorable GSVA score and LUAD-favorable GSVA score are a biomarker of LUAD with AUC = 0.982 and AUC = 0.994, respectively

(**Figure 4B**). Furthermore, the two GSVA score systems were also validated in GSE10072 (**Figure 4C**) and GSE31210 (**Figure 4D**), respectively. According to median GSVA score, all LUAD patients in TCGA were separated into low-score group and high-score group. And both the two GSVA score systems were significantly associated with LUAD prognosis (**Figure 4E**). Patients with high LUAD-unfavorable GSVA score had worse prognosis, while patients with high LUAD-favorable GSVA score had better prognosis. Univariate and multivariate Cox analysis showed that the two GSVA score systems were the independent factors for LUAD prognosis compared with clinicopathological features (**Tables 2** and **3**). Moreover, the two GSVA score systems were also significantly associated with LUAD prognosis in GSE31210 (**Figure 4F**).

TABLE 1 | LUAD-unfavorable gene set and LUAD-favorable gene set.

Gene set	Gene symbol
LUAD-unfavorable gene set	ARHGAP11A, ASPM, BLM, C5orf34, CA9, CCNA2, CDC25C, CDC6, CDCA2, CDK1, CENPF, CENPK, CHAF1B, CLSPN, DDX11-AS1, DEPDC1, DNMT3B, DTL, E2F7, ECT2, EGLN3, ESCO2, EXO1, FAM111B, FAM57B, FAM72D, FAM83D, FANCI, FBXO43, GAL, GTSE1, HASPIN, HELLS, HMMR, KIF11, KIF14, KIFC1, KNL1, KNTC1, KREMEN2, KRT6A, KRT81, LINC01269, LOC101929128, LYPD3, MAD2L1, MELK, MIR924HG, MKI67, MYO19, NCAPG, NDC80, NEIL3, NEK2, NUF2, NUSAP1, OIP5, ORC1, ORC6, PAICS, PARPBP, PCLAF, PIMREG, PLK1, PLK4, POLQ, PRC1, PTPRN, RAD51, RRM2, SGO1, SLC2A1-AS1, SPAG5, SPOCK1, TEDC2, TESMIN, TGFBR3L
LUAD-favorable gene set	TICRR, TROAP, TTK, TYMS, UBE2T, UCA1, ZWINT, ACKR1, ADAMTS8, ADGRF5, ARHGEF6, ATP13A4, BMP5, CASS4, CCDC69, CLEC3B, COL6A6, CTSG, FAM189A2, FBP1, FCER1A, FLI1, GCSAML, GIMAP4, GIMAP7, HPGDS, HPSE2, INMT, JAML, KLK11, LSAMP, LY86, MAL, MS4A2, OR7E47P, P2RY12, RAB44, RTN1, SCN2B, SIGLEC17P, SLCO4C1, SPN, TM6SF1, TRPC2, UNC45B, ZEB2

The Differential Expression of LUAD-Unfavorable Gene May Not Result From Mutation

Only 52 (9.17%) of 567 samples had an alteration in one or several LUAD-unfavorable genes and most samples did not have genetical alteration (**Figure 5A**). Moreover, compared with normal lung tissue, ten genes (ASPM, BLM, CDC25C, CDK1, DEPDC1, KIF11, KIF14, LYPD3, NEK2, and PLK4) of LUAD-unfavorable gene set were included in The Human Protein Atlas and were highly expressed in LUAD (**Figure 5B**).

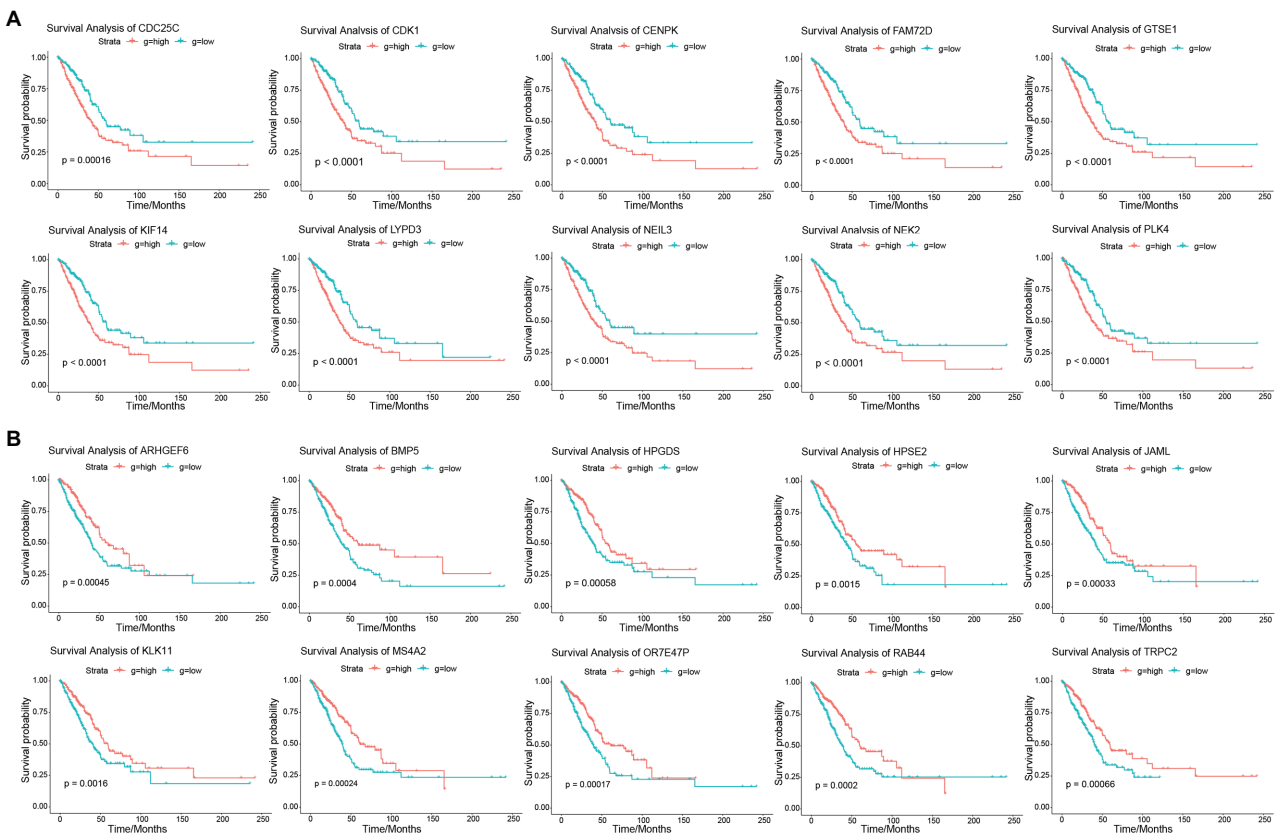

FIGURE 3 | Survival analysis. **(A)** Survival curves of 10 genes most significantly correlated with LUAD in LUAD-unfavorable gene set. **(B)** Survival curves of 10 genes most significantly correlated with LUAD in LUAD-favorable gene set.

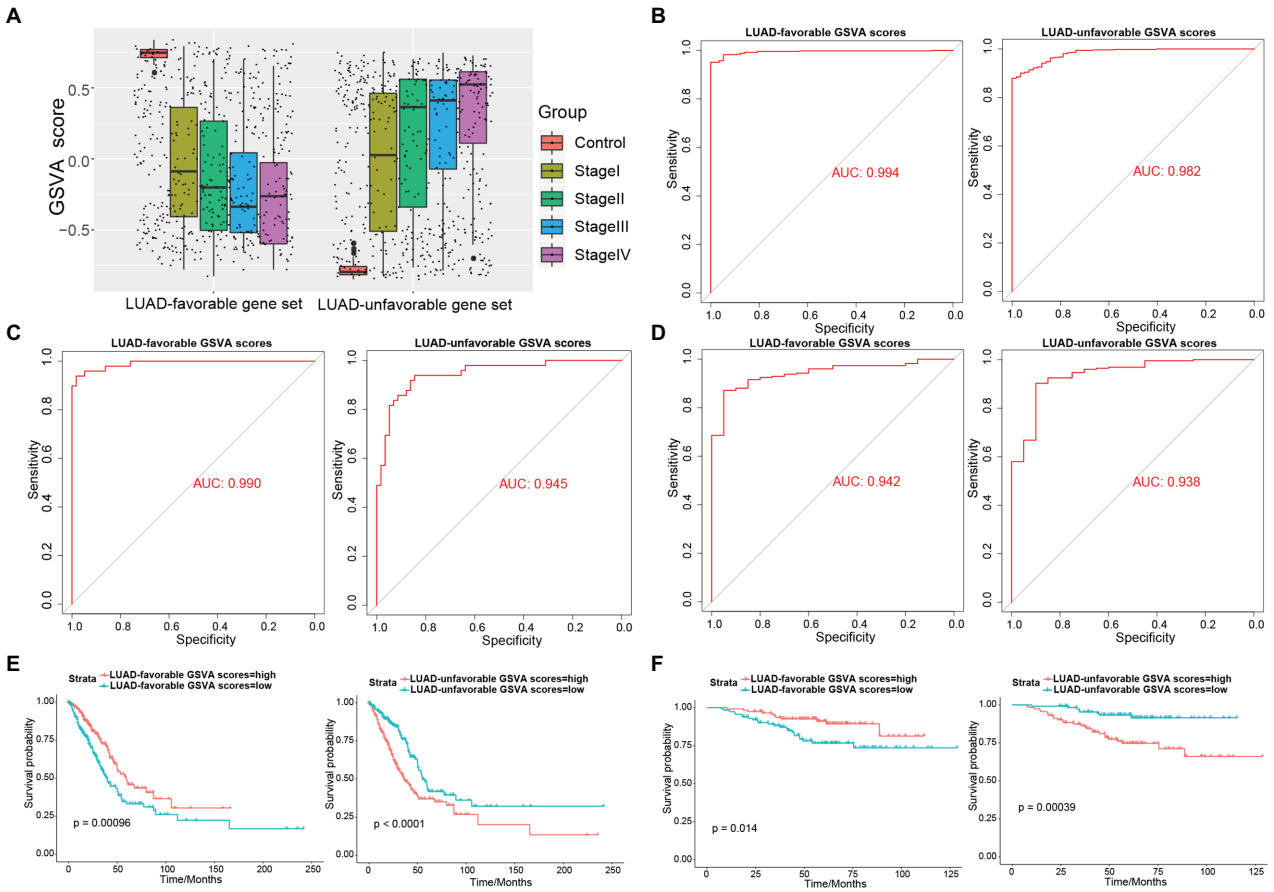

FIGURE 4 | Exploring the diagnostic and prognostic abilities of LUAD-unfavorable gene set and LUAD-favorable gene set. **(A)** The box plots of LUAD-unfavorable gene set gene set variation analysis (GSVA) score and LUAD-favorable gene set GSVA score. **(B)** Receiver operating characteristic (ROC) curves analysis of LUAD-unfavorable gene set GSVA score and LUAD-favorable gene set GSVA score. **(C)** ROC curves analysis of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score in GSE10072. **(D)** ROC curves analysis of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score in GSE31210. **(E)** Survival analysis of LUAD-unfavorable gene set GSVA score and LUAD-favorable gene set GSVA score in LUAD from The Cancer Genome Atlas (TCGA). **(F)** Survival analysis of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score in GSE31210.

TABLE 2 | Univariate and multivariate analyses of LUAD-unfavorable GSVA score.

Factor	Univariate Cox analysis			Multivariate Cox analysis		
	β	*p*	HR (95% CI)	*β*	*p*	HR (95% CI)
Gender (female/male)	0.025	0.867	0.763–1.378			
Age (>65 years/≦65 years)	0.178	0.243	0.886–1.610			
T stage (T3–4/T1–2)	0.821	0.000	1.543–3.346	0.589	0.016	1.114–2.914
Lymph node stage (N2–3/N0–1)	0.818	0.000	1.582–3.243	0.121	0.757	0.523–2.437
Metastasis (M1/M0)	0.749	0.006	1.234–3.626	0.109	0.799	0.482–2.580
Pathological stage (III–IV/I–II)	0.967	0.000	1.924–3.592	0.509	0.211	0.749–3.695
LUAD-unfavorable GSVA score (high/low)	0.614	0.000	1.365–2.500	0.575	0.002	1.230–2.565

LUAD-Unfavorable Genes Involved in Multiple Cancer-Related Pathways

In order to explore the biological functions of LUAD-unfavorable genes, LUAD-unfavorable genes were performed functional enrichment analysis. The results showed that these genes are mainly related to nuclear division, organelle fission, mitotic nuclear division, nuclear chromosome segregation, and chromosome segregation (**Figure 6A**). Moreover, LUAD-unfavorable genes were significantly involved in many pathways, such as Fanconi anemia pathway, p53 signaling pathway, Oocyte meiosis, Cell cycle, and Progesterone-mediated oocyte maturation (**Figure 6B**).

TABLE 3 | Univariate and multivariate analyses of LUAD-favorable GSVA score.

Factor	Univariate Cox analysis			Multivariate Cox analysis		
	β	*p*	HR (95% CI)	*β*	*p*	HR (95% CI)
Gender (female/male)	0.025	0.867	0.763–1.378			
Age (>65 years/≦65 years)	0.178	0.243	0.886–1.610			
T stage (T3–4/T1–2)	0.821	0.000	1.543–3.346	0.557	0.028	1.062–2.869
Lymph node stage (N2–3/N0–1)	0.818	0.000	1.582–3.243	0.420	0.254	0.739–3.134
Metastasis (M1/M0)	0.749	0.006	1.234–3.626	0.263	0.518	0.586–2.886
Pathological stage (III–IV/I–II)	0.967	0.000	1.924–3.592	0.365	0.361	0.659–3.152
LUAD-favorable GSVA score (high/low)	−0.489	0.001	0.454–0.828	−0.434	0.017	0.453–0.926

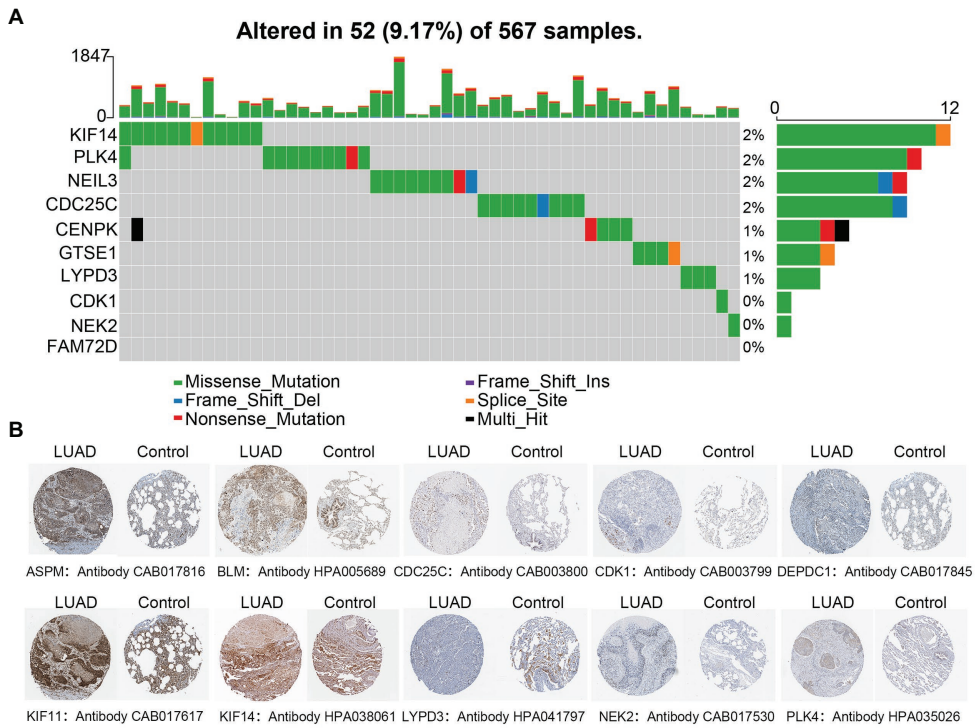

FIGURE 5 | Genetical alteration analysis and immunohistochemistry analysis. **(A)** Genetical alteration analysis of LUAD-unfavorable genes. **(B)** Identification of LUAD-unfavorable gene at protein level. Lung cancer samples are on the left and normal lung tissue samples are on the right.

DISCUSSION

In the world, lung cancer is the main cause of cancer-related death. Even with surgical treatment, the recurrence rate of lung cancer still is very high (Scott et al., 2007). Therefore, it is of great significance to explore biomarkers which can accurately diagnose lung cancer and predict prognosis for the treatment and management of lung cancer. A large number of studies have shown that abnormal expression of genes in lung cancer (including LUAD) is closely related to prognosis, and can be used as a potential biomarker of prognosis (Xu et al., 2013; Cui et al., 2015; Giatromanolaki et al., 2015).

In the present study, we found a number of genes were differentially expressed in LUAD different stages. This indicated gene expression patterns were various with the LUAD development. Compared to normal lung tissue, a gene may be differentially expressed in early LUAD but not in advanced stage. We identified 422 LUAD-development characteristic genes, including 185 genes gradually upregulated and 237 genes gradually downregulated with LUAD-development. The development of LUAD results from synergistic effects of multiple genes. Notably, not all LUAD-development characteristic genes are associated with the prognosis of LUAD. LUAD-unfavorable gene set contained 84 gradually upregulated DEGs and LUAD-favorable gene set contained 39 gradually downregulated DEGs. Unsurprisingly, previous studies have suggested that some of them are associated with LUAD development. NEK2 is overexpressed in a variety of malignant tumors and is closely related to tumor drug resistance, rapid recurrence, and poor prognosis (Zhou et al., 2013; Fang and Zhang, 2016; Li et al., 2017). KIF14 has also been found to

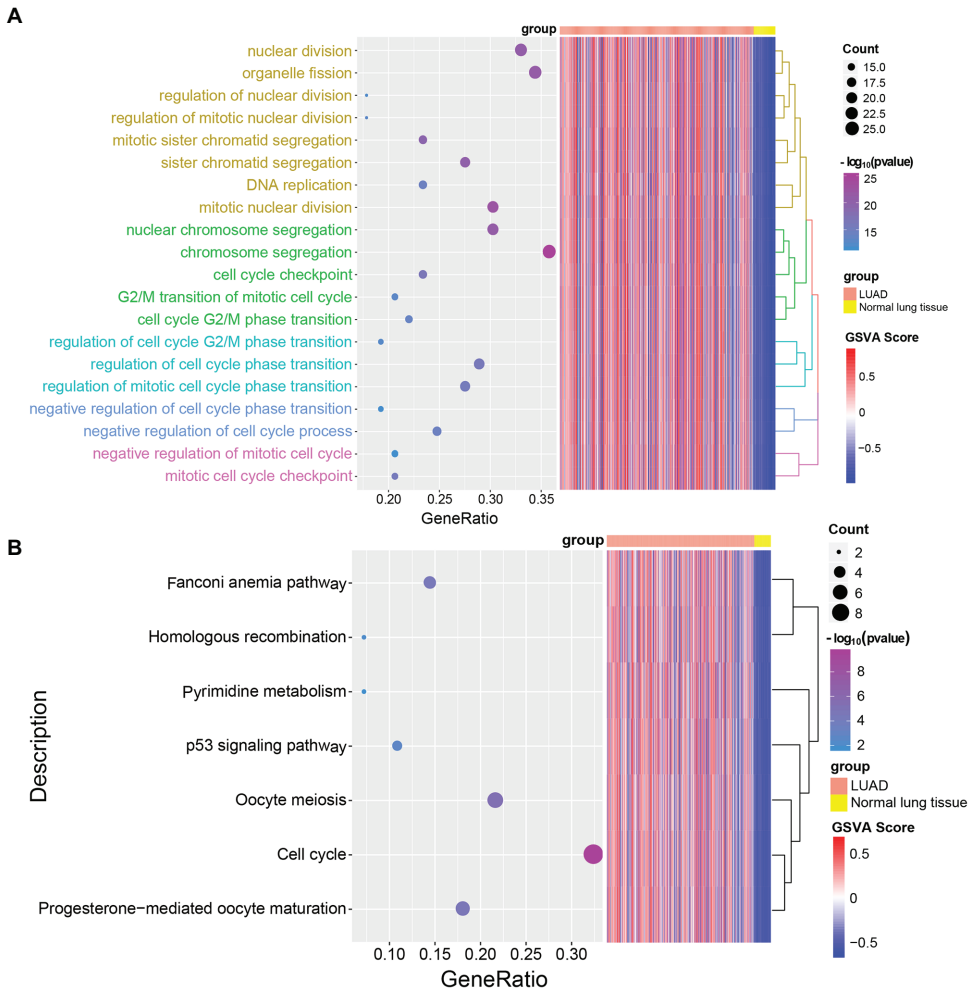

FIGURE 6 | Gene Ontology (GO) biological process (BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis of LUAD-unfavorable genes. **(A)** Biological process of LUAD-unfavorable genes. **(B)** KEGG pathway analysis of LUAD-unfavorable genes.

be associated with poor prognosis in a variety of cancers (O'Hare et al., 2016; Zhang et al., 2017). While in the LUAD-favorable gene set, genes which were significantly associated with LUAD survival included OR7E47P, MS4A2, RAB44, BMP5, ARHGEF6, and KLK11. Among them, KLK11 was found to be a diagnostic and prognostic indicator of NSCLC (Xu et al., 2014). These result confirmed the possibility that the LUAD-unfavorable gene set and LUAD-unfavorable gene set can be used as a prognostic model for LUAD.

All samples were calculated LUAD-unfavorable GSVA scores and LUAD-favorable GSVA scores. This is obviously different from the gene signatures in other previous studies (Li et al., 2014; Shi et al., 2018; Liu et al., 2019). In the previous studies, a gene often got a coefficient from a Cox regression analysis or other method in the training set. However, due to the limitations of the sample size and the heterogeneity of the tumor, we may never know the true coefficient of a gene. Therefore, GSVA was used to score individual samples against gene sets (LUAD-unfavorable gene set and LUAD-favorable gene set) in our study. ROC curve analysis suggested that both LUAD-unfavorable GSVA score and LUAD-favorable GSVA score exhibited strong diagnostic capacity of LUAD and which was verified in other two independent data sets. Univariate and multivariate Cox regression analysis suggested that LUAD-unfavorable GSVA score and LUAD-unfavorable gene set were independent prognostic factors for LUAD's overall survival. This result was also verified in an independent data set.

Moreover, we found that the mutation rate of most genes is very low, indicating that the differential expression of genes may not be caused by mutation. Additionally, functional enrichment analysis indicates that LUAD-unfavorable genes are significantly involved in p53 signaling pathway, Cell cycle, and other pathways. It is suggested that LUAD-unfavorable genes may be involved in the occurrence and development of LUAD through these pathways. However, further studies are needed to investigate and validate the functions of these genes.

In the present study, although we provided new insights into the LUAD prognostic stratification system, several limitations were notable. Firstly, the two gene sets may be too large. Their application to the clinic still needs to wait for further decline in sequencing costs. Secondly, the synergy between the genes of these two gene sets to promote LUAD development still requires molecular experimental validation.

CONCLUSION

In conclusion, we identified and validated two LUAD-development characteristic gene sets that not only have diagnostic value but also prognostic value. It may provide new insight for further research on LUAD.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

CL and XL conducted the experiments. HS and DL designed the experiments and wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

The study was supported by Subject of Education Department of Heilongjiang Provincial (no. 12531258 and 12511264).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.researchsquare.com/article/rs-12465/v1 (Liu et al., 2020). We would like to thank the Bioinforcraftsman Technology Research and Development Co., Ltd for generously assisting with science research experience and bioinformatics analysis.

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Colwill, K., Renewable Protein Binder Working Group, and Graslund, S. (2011). A roadmap to generate renewable protein binders to the human proteome. *Nat. Methods* 8, 551–558. doi: 10.1038/nmeth.1607

Cui, Y., Liu, J., Yin, H. B., Liu, Y. F., and Liu, J. H. (2015). Fibulin-1 functions as a prognostic factor in lung adenocarcinoma. *Jpn. J. Clin. Oncol.* 45, 854–859. doi: 10.1093/jjco/hyv094

Dama, E., Melocchi, V., Dezi, F., Pirroni, S., Carletti, R. M., Brambilla, D., et al. (2017). An aggressive subtype of stage I lung adenocarcinoma with molecular and prognostic characteristics typical of advanced lung cancers. *Clin. Cancer Res.* 23, 62–72. doi: 10.1158/1078-0432.CCR-15-3005

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., Mclellan, M. D., Cibulskis, K., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069–1075. doi: 10.1038/nature07423

Dong, H. X., Wang, R., Jin, X. Y., Zeng, J., and Pan, J. (2018). LncRNA DGCR5 promotes lung adenocarcinoma (LUAD) progression via inhibiting hsa-mir-22-3p. *J. Cell. Physiol.* 233, 4126–4136. doi: 10.1002/jcp.26215

Donner, I., Katainen, R., Sipila, L. J., Aavikko, M., Pukkala, E., and Aaltonen, L. A. (2018). Germline mutations in young non-smoking women with lung adenocarcinoma. *Lung Cancer* 122, 76–82. doi: 10.1016/j.lungcan.2018.05.027

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191. doi: 10.1186/1471-2105-7-191

Fang, Y., and Zhang, X. (2016). Targeting NEK2 as a promising therapeutic approach for cancer treatment. *Cell Cycle* 15, 895–907. doi: 10.1080/15384101.2016.1152430

Feng, A., Tu, Z., and Yin, B. (2016). The effect of HMGB1 on the clinicopathological and prognostic features of non-small cell lung cancer. *Oncotarget* 7, 20507–20519. doi: 10.18632/oncotarget.7050

Giatromanolaki, A., Kalamida, D., Sivridis, E., Karagounis, I. V., Gatter, K. C., Harris, A. L., et al. (2015). Increased expression of transcription factor EB (TFEB) is associated with autophagy, migratory phenotype and poor prognosis in non-small cell lung cancer. *Lung Cancer* 90, 98–105. doi: 10.1016/j.lungcan.2015.07.008

Govindan, R., Page, N., Morgensztern, D., Read, W., Tierney, R., Vlahiotis, A., et al. (2006). Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: analysis of the surveillance, epidemiologic, and end results database. *J. Clin. Oncol.* 24, 4539–4544. doi: 10.1200/JCO.2005.04.4859

Guan, J. L., Zhong, W. Z., An, S. J., Yang, J. J., Su, J., Chen, Z. H., et al. (2013). KRAS mutation in patients with lung cancer: a predictor for poor prognosis but not for EGFR-TKIs or chemotherapy. *Ann. Surg. Oncol.* 20, 1381–1388. doi: 10.1245/s10434-012-2754-z

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

He, S. Y., Xi, W. J., Wang, X., Xu, C. H., Cheng, L., Liu, S. Y., et al. (2019). Identification of a combined RNA prognostic signature in adenocarcinoma of the lung. *Med. Sci. Monit.* 25, 3941–3956. doi: 10.12659/MSM.913727

Hecht, S. S. (1999). Tobacco smoke carcinogens and lung cancer. *J. Natl. Cancer Inst.* 91, 1194–1210. doi: 10.1093/jnci/91.14.1194

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29

Li, X., Shi, Y., Yin, Z., Xue, X., and Zhou, B. (2014). An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma. *J. Transl. Med.* 12:159. doi: 10.1186/1479-5876-12-159

Li, G., Zhong, Y., Shen, Q., Zhou, Y., Deng, X., Li, C., et al. (2017). NEK2 serves as a prognostic biomarker for hepatocellular carcinoma. *Int. J. Oncol.* 50, 405–413. doi: 10.3892/ijo.2017.3837

Liu, C., Li, X., Shao, H., and Li, D. (2020). Identification and validation of two LUAD-development characteristic gene sets for diagnosing lung adenocarcinoma and predicting prognosis [Preprint]. doi: 10.21203/rs.2.21884/v1

Liu, C., Li, Y., Wei, M., Zhao, L., Yu, Y., and Li, G. (2019). Identification of a novel glycolysis-related gene signature that can predict the survival of patients with lung adenocarcinoma. *Cell Cycle* 18, 568–579. doi: 10.1080/15384101.2019.1578146

Mendelsohn, J., and Baselga, J. (2003). Status of epidermal growth factor receptor antagonists in the biology and treatment of cancer. *J. Clin. Oncol.* 21, 2787–2799. doi: 10.1200/JCO.2003.01.504

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., et al. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.* 15:e1006701. doi: 10.1371/journal.pcbi.1006701

Naoki, K., Chen, T. H., Richards, W. G., Sugarbaker, D. J., and Meyerson, M. (2002). Missense mutations of the BRAF gene in human lung adenocarcinoma. *Cancer Res.* 62, 7001–7003.

O'Hare, M., Shadmand, M., Sulaiman, R. S., Sishtla, K., Sakisaka, T., and Corson, T. W. (2016). Kif14 overexpression accelerates murine retinoblastoma development. *Int. J. Cancer* 139, 1752–1758. doi: 10.1002/ijc.30221

Pang, B., Wu, N., Guan, R., Pang, L., Li, X., Li, S., et al. (2017). Overexpression of RCC2 enhances cell motility and promotes tumor metastasis in lung adenocarcinoma by inducing epithelial-mesenchymal transition. *Clin. Cancer Res.* 23, 5598–5610. doi: 10.1158/1078-0432.CCR-16-2909

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Scott, W. J., Howington, J., Feigenberg, S., Movsas, B., and Pisters, K., and American College of Chest Physicians (2007). Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest* 132, 234S–242S. doi: 10.1378/chest.07-1378

Shi, X., Tan, H., Le, X., Xian, H., Li, X., Huang, K., et al. (2018). An expression signature model to predict lung adenocarcinoma-specific survival. *Cancer Manag. Res.* 10, 3717–3732. doi: 10.2147/CMAR.S159563

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Xu, P., Liu, L., Wang, J., Zhang, K., Hong, X., Deng, Q., et al. (2013). Genetic variation in BCL2 3'-UTR was associated with lung cancer risk and prognosis in male Chinese population. *PLoS One* 8:e72197. doi: 10.1371/journal.pone.0072197

Xu, C. H., Zhang, Y., and Yu, L. K. (2014). The diagnostic and prognostic value of serum human kallikrein-related peptidases 11 in non-small cell lung cancer. *Tumour Biol.* 35, 5199–5203. doi: 10.1007/s13277-014-1674-x

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, Y., Yuan, Y., Liang, P., Zhang, Z., Guo, X., Xia, L., et al. (2017). Overexpression of a novel candidate oncogene KIF14 correlates with tumor progression and poor prognosis in prostate cancer. *Oncotarget* 8, 45459–45469. doi: 10.18632/oncotarget.17564

Zhao, K., Li, Z., and Tian, H. (2018a). Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. *OncoTargets Ther.* 11, 3415–3424. doi: 10.2147/OTT.S158638

Zhao, X., Zhou, L. L., Li, X., Ni, J., Chen, P., Ma, R., et al. (2018b). Overexpression of KIF20A confers malignant phenotype of lung adenocarcinoma by promoting cell proliferation and inhibiting apoptosis. *Cancer Med.* 7, 4678–4689. doi: 10.1002/cam4.1710

Zhou, W., Yang, Y., Xia, J., Wang, H., Salama, M. E., Xiong, W., et al. (2013). NEK2 induces drug resistance mainly through activation of efflux drug pumps and is associated with poor prognosis in myeloma and other cancers. *Cancer Cell* 23, 48–62. doi: 10.1016/j.ccr.2012.12.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Physiology

ORIGINAL RESEARCH
published: 23 August 2021
doi: 10.3389/fphys.2021.596194

Computational Reconstruction of Clonal Hierarchies From Bulk Sequencing Data of Acute Myeloid Leukemia Samples

Thomas Stiehl[1,2] *and Anna Marciniak-Czochra*[2]*

[1] *Institute for Computational Biomedicine – Disease Modeling, RWTH Aachen University, Aachen, Germany,* [2] *Institute of Applied Mathematics, Interdisciplinary Center for Scientific Computing and Bioquant Center, Heidelberg University, Heidelberg, Germany*

OPEN ACCESS

Edited by:
Doron Levy, University of Maryland, College Park, United States

Reviewed by:
Alexandra Jilkine, University of Notre Dame, United States
Torbjörn Lundh, Chalmers University of Technology, Sweden

***Correspondence:**
Anna Marciniak-Czochra
anna.marciniak@iwr.uni-heidelberg.de

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *18 August 2020*
Accepted: *26 July 2021*
Published: *23 August 2021*

Citation:
Stiehl T and Marciniak-Czochra A (2021) Computational Reconstruction of Clonal Hierarchies From Bulk Sequencing Data of Acute Myeloid Leukemia Samples. Front. Physiol. 12:596194. doi: 10.3389/fphys.2021.596194

Acute myeloid leukemia is an aggressive cancer of the blood forming system. The malignant cell population is composed of multiple clones that evolve over time. Clonal data reflect the mechanisms governing treatment response and relapse. Single cell sequencing provides most direct insights into the clonal composition of the leukemic cells, however it is still not routinely available in clinical practice. In this work we develop a computational algorithm that allows identifying all clonal hierarchies that are compatible with bulk variant allele frequencies measured in a patient sample. The clonal hierarchies represent descendance relations between the different clones and reveal the order in which mutations have been acquired. The proposed computational approach is tested using single cell sequencing data that allow comparing the outcome of the algorithm with the true structure of the clonal hierarchy. We investigate which problems occur during reconstruction of clonal hierarchies from bulk sequencing data. Our results suggest that in many cases only a small number of possible hierarchies fits the bulk data. This implies that bulk sequencing data can be used to obtain insights in clonal evolution.

Keywords: computational algorithm, acute myeloid leukemia, clonal evolution, clonal hierarchy, clonal pedigree, phylogenetic tree, bulk sequencing, stem cell

INTRODUCTION

Acute myeloid leukemia (AML) is an aggressive cancer of the blood forming system. It is characterized by expansion of malignant cells and impairment of healthy blood cell formation (Röllig et al., 2011; Döhner et al., 2017; Roloff and Griffiths, 2018). AML originates from a small population of malignant stem-like cells, referred to as leukemic stem cells (LSC) or leukemia initiating cells (LIC). A hallmark of AML is its poor prognosis and the high rate of relapse (Röllig et al., 2011; Döhner et al., 2017; Roloff and Griffiths, 2018).

The main reason for the high risk of relapse is the clonal heterogeneity of the disease. Sequencing studies reveal that the AML cell population is composed of multiple clones. Contributions of the individual clones to the total malignant cell burden vary over time

(Ding et al., 2012; Cancer Genome Atlas Research Network, 2013; Greif et al., 2018; Cocciardi et al., 2019; Ediriwickrema et al., 2020). Due to the high number of different clones, the probability is high that a subset of clones has a low sensitivity to chemotherapy, survives treatment and initiates relapse (Ding et al., 2012; Cancer Genome Atlas Research Network, 2013; Stiehl et al., 2014; Greif et al., 2018; Cocciardi et al., 2019; Ediriwickrema et al., 2020).

The clinical course of the disease shows a significant among-patient variability which can only be partially predicted based on currently existing risk-stratifications (Stiehl et al., 2014, 2015, 2020; Döhner et al., 2017; Wang et al., 2017; Roloff and Griffiths, 2018). To better understand the mechanism of relapse and to identify patients at risk, a quantitative understanding of clonal dynamics is required (Ding et al., 2012; Cancer Genome Atlas Research Network, 2013; Stiehl et al., 2014; Greif et al., 2018; Banck and Görlich, 2019; Cocciardi et al., 2019; Lorenzi et al., 2019; Ediriwickrema et al., 2020).

Next-generation sequencing studies have revealed a high number of genetic hits involved in AML pathogenesis. Genetic variability among different patients is considerable and new mutations are acquired during disease evolution (Ding et al., 2012; Cancer Genome Atlas Research Network, 2013; Greif et al., 2018; Cocciardi et al., 2019; Ediriwickrema et al., 2020). Correlation of mutations with clinical outcome has resulted in a genetics-based risk-stratification (Grimwade et al., 1998; Röllig et al., 2011; Döhner et al., 2017). However, the effect of many mutations on cell dynamics remains unclear (Bacher et al., 2008; Ding et al., 2012).

Relating genetic data to patient prognosis and malignant cell properties is challenging, since different genetic hits may enhance or inhibit each other (Grimwade et al., 1998; Bacher et al., 2008; Cancer Genome Atlas Research Network, 2013; Stiehl et al., 2014; Greif et al., 2018; Roloff and Griffiths, 2018). Furthermore, potentially unknown or undetected hits may impact the aberrations that are observed in clinical routine. Mathematical and computational models are important to link genetic data to functional cell properties such as proliferation and self-renewal of leukemic stem cells, both of which are of prognostic relevance (Stiehl et al., 2014, 2015, 2016, 2018, 2020; Banck and Görlich, 2019; Lorenzi et al., 2019).

Such models allow to estimate which leukemic cell properties correspond to the clinical course of an individual patient and to link the estimates to mutation data (Stiehl et al., 2014, 2015, 2020). This provides insights into the impact of different mutations and leads to new hypotheses about the underlying biological mechanisms and genotype-phenotype correlation.

Leukemic stem cell dynamics are governed by two key properties: proliferation rate and fraction of self-renewal. The proliferation rate describes how often LSC divide per unit of time. Upon division a LSC gives rise to two progeny which can either be LSC or of a more differentiated progenitor type. The fraction of self-renewal corresponds to the fraction of LSC among the progeny (Lutz et al., 2013; Stiehl and Marciniak-Czochra, 2017). Mathematical and computational models suggest that stem cell properties at diagnosis differ from those at relapse. Particularly, LSC at diagnosis are characterized by an increased self-renewal fraction and a higher proliferation rate compared to healthy cells. LSC at relapse are characterized by a slow proliferation rate and a further increase of the self-renewal fraction (Stiehl et al., 2014, 2016). Computer simulations and model analysis indicate that increased self-renewal leads to a competitive advantage of the respective clones and that clones appearing later in the course of the disease have a higher self-renewal compared to clones emerging earlier (Stiehl et al., 2014, 2016; Busse et al., 2016; Banck and Görlich, 2019; Lorenzi et al., 2019).

Single cell sequencing technology allows to detect mutations that are present in a single cell. Sequencing of a sufficiently large number of single cells allows to reconstruct the order of mutation acquisition and to visualize it as a so-called clonal hierarchy, clonal pedigree or phylogenetic tree (Kuipers et al., 2017; Ediriwickrema et al., 2020). Computational models have led to the hypothesis that the position of a clone in the phylogenetic tree correlates with its fraction of self-renewal (Stiehl et al., 2016). Therefore, phylogenetic trees may contain important information about cell properties that could be used to decipher the impact of mutations on the malignant cell kinetics.

In contrast to the single cell sequencing approach, bulk sequencing analyses a mixture of DNA of multiple cells, to which each cell contributes its specific (either mutated or non-mutated) alleles. Since in most cases each cell carries two versions of each allele, the bulk sample from n cells is a mixture of $2n$ allele versions. The so-called variant allele frequency (VAF) is the percentage of allele versions that is mutated. Bulk sequencing quantifies the frequency of a mutated allele in a cell population however does not determine how the detected mutations are distributed among the different clones (Roth et al., 2014; Kuipers et al., 2017; Brierley and Mead, 2020).

Single cell sequencing is a relatively new and costly technology that so far is not used in clinical routine (Brierley and Mead, 2020). To deduce clinically relevant knowledge from genetic data large patient groups have to be studied due to the high inter-individual heterogeneity of the detected mutations and their unknown interaction. For this reason, it is a relevant question whether clonal hierarchies can be deduced from bulk sequencing data which are routinely obtained after initial diagnosis of AML (Roth et al., 2014; Brierley and Mead, 2020), although most of the diagnostic sequencing is targeted on limited panels of "typical" driver mutations.

In this work we propose an algorithm that systematically constructs all phylogenetic trees that are in agreement with bulk sequencing data of an individual patient. This algorithm provides a tool to better understand the ambiguity of such reconstructions and their sensitivity to measurement errors.

To test our approach, we choose a recently published set of single cell sequencing data as a gold standard (ground truth) (Ediriwickrema et al., 2020). Based on the single cell sequencing data we calculate the variant allele frequency of the different mutations in a bulk sequencing sample and test whether the "real" clonal pedigree, i.e., the pedigree deduced from single cell sequencing data, can be reconstructed from it. We investigate how the correctness and uniqueness of the reconstruction depend on sampling and measurement errors.

Different approaches have been developed to track the order of mutations in AML. They include population based cross sectional studies (Delhommeau et al., 2009; Abdel-Wahab et al., 2010; Papaemmanuil et al., 2016), targeted and deep sequencing of paired samples taken at different time points (Bachas et al., 2012; Ding et al., 2012), single cell sequencing (Ediriwickrema et al., 2020) and others such as fluorescence-*in situ*-hybridization, xenografting, cell cultures or IPS technology (Anderson et al., 2011; Ran et al., 2012; Jonas, 2017; Nobile et al., 2019; Herudkova et al., 2020; Sandén et al., 2020). From the computational side, a range of tools have been developed to fit models and extract quantitative information from data in the context of AML (Attolini et al., 2010; Nobile et al., 2019) and other cancers, see e.g., (Roth et al., 2014; Caravagna et al., 2020) for statistical approaches using variant allele frequencies, (Attolini et al., 2010) for a population-based model and (Nobile et al., 2019) for xenotransplant data. These approaches are complemented by process-based models (Stiehl et al., 2014, 2016; Rahman et al., 2018; Banck and Görlich, 2019; Dinh et al., 2019, 2020; Salichos et al., 2020).

MATERIALS AND METHODS

Aim

We use variant allele frequencies from bulk sequencing as input data. The output we want to obtain are all clonal hierarchies that are compatible with the input data.

Assumptions

We assume that each mutation is only acquired once. Variant alleles cannot mutate back to wild type alleles. We only consider heterozygous mutations. We rescale the measurements such that the variant allele frequency of the most abundant mutation is equal to 100%.

Computational Methods

The method is summarized in **Figure 1**. Assuming that each mutation is irreversible and only acquired once, clonal pedigrees have the structure of labeled rooted trees. An (unrooted) tree is an undirected acyclic connected graph (Diestel, 2017). If one node of the tree is designated as root, a rooted tree is obtained. In a rooted tree we naturally assign directions to the edges pointing from the root towards the leaves. If a unique label is assigned to each node, the tree is referred to as a labeled tree (Diestel, 2017). The root of the tree corresponds to a genetic trait that is present in all clones. If the disease originates from a single founding mutation that is present in all malignant cells, the root can be identified with the founding mutation. This configuration applies to most leukemic patients. If there exist multiple founding mutations the root of the tree corresponds to the healthy phenotype. Each node in the tree corresponds to one clone. The label assigned to a node indicates the mutational events that gave rise to the clone. The edge pointing towards the node indicates which ancestor clone acquired the mutational event indicated by the label.

The tree structures can be mapped to matrices. We consider a tree with n nodes, corresponding to n clones denoted by *clone* 1 to *clone n*. Since each clone differs from its ancestor by exactly one new mutation, there exist n different mutations, which we number from 1 to n. Denote by $A_{1=i,j=n}$ a matrix. We set $a_{ij} = 1$ if clone j carries mutation i, otherwise we set $a_{ij} = 0$. We number the clones starting from the root (= *clone* 1) and proceed with increasing depth, i.e., if the depth of *clone i* is higher than the depth of *clone j*, then $j < i$. We denote the founding mutation as *mutation* 1 and the mutation that is present in *clone j* but not in its direct ancestor as *mutation j*. Then $A_{1=i,j=n}$ is an upper triangular matrix, with $a_{ii} = 1$, and a_{ij} from the set {0,1}.

We aim to solve the linear system of equations $Ax = b$, where b_i is the measured frequency of *mutation i* in the bulk sample and x_i is the abundance of *clone i* in the sample. We note that A has determinant 1 and therefore this system of equations has a unique solution. The solution is biologically feasible if all x_i are non-negative. The existence of a non-negative solution can be easily checked since the solutions of $Ax = b$ are given by $x_n = b_n$, $x_j = b_j - a_{jj+1}\, x_{j+1} - \ldots - a_{jn} x_n$. We say that the dataset b is compatible with the clonal hierarchy represented by matrix A if $Ax = b$ has a non-negative solution.

The founder mutation is denoted as *mutation* 1. It is present in all clones and, therefore, is the most abundant mutation in the bulk sample. This implies that $a_{1j} = 1$ for $1 \leq j \leq n$. Since we normalized the frequency of the most abundant mutation to

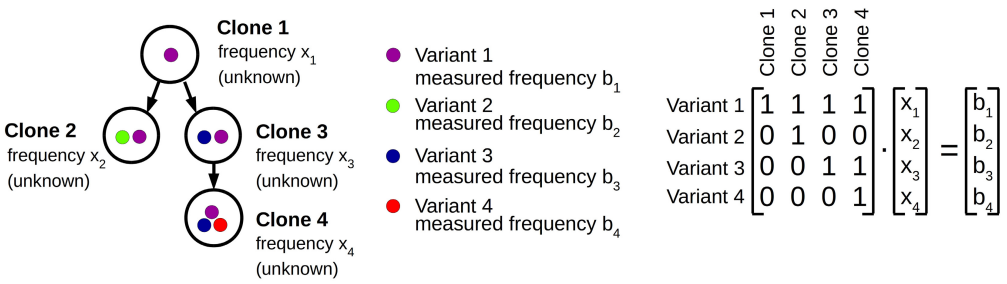

FIGURE 1 | Computational approach. Clonal hierarchies are rooted trees. The root of the tree either corresponds to wild type cells or to the AML founder clone. The colored dots represent different alleles or mutations. From bulk sequencing the allele frequencies b_i in the sample are known. The frequencies x_i of the different clones are unknown. The tree structure is represented by a triangular matrix. The measured data is compatible with the tree structure if the system $Ax = b$ has a non-negative solution.

100% it holds $b_1 = 100$. This implies that the sum over the x_i is equal to 100.

To systematically generate all possible trees, we use Pruefer sequences, a classical concept to bijectively map unrooted trees with n nodes to sequences of length n-2 (Prüfer, 1918). Each unrooted labeled tree with n nodes then corresponds to a sequence of length n-2 with elements from {1, ..., n}. This implies that there exist n^{n-2} unrooted labeled trees. Since each of the n nodes can be designated as root, there exist n^{n-1} labeled rooted trees.

Interpretation

If a biologically feasible solution of the system $Ax = b$ exists, the measured bulk allele frequencies b can be explained by the tree structure that corresponds to the matrix A. This means that the bulk allele frequencies b are obtained by mixing the different clones from the tree in appropriate proportions (the abundance of *clone i* has to equal x_i). For each pair A, b a biologically feasible solution can exist or cannot exist. For example, a tree with founder mutation X (i.e., each clone carries mutation X) cannot match to samples where the abundance of X in non-maximal.

Measurement Errors

If the measured data b are exact, non-existence of a biologically feasible solution indicates a mismatch of the tree structure and the allele frequencies. In case of experimental data, the non-existence of a biologically feasible solution can alternatively arise from measurement errors. For this reason it may be necessary to also consider solutions fulfilling $|| Ax\text{-}b|| < \varepsilon$ for an appropriate ε, where $||.||$ denotes e.g., the Euclidean norm.

To find such solutions, especially in the case where no biologically feasible (i.e., exact non-negative) solutions exist, we use an optimization approach to obtain a non-negative solution that reproduces the data as good as possible. For each matrix A that corresponds to a tree structure we minimize $||$ Ax-b$||$ under the constraints $x_i \geq 0$ (i = 1,...n), $x_1 + ... + x_n = 100$. If the measured VAF have different confidence intervals, we minimize the weighted error function $|| W(Ax\text{-}b)||$, where W is a diagonal matrix with entries related to the confidence intervals.

Solving the minimization problem for each possible tree structure allows to rank the tree structures based on the mismatch $|| Ax\text{-}b||$ and to identify which tree optimally fits to the data. A solution is referred to as exact if $|| Ax\text{-}b|| <$ 10E-16. We say that the tree structure corresponding to matrix $\tilde{A}$ is optimal if it holds $|| \tilde{A}x\text{-}b|| \leq || Ax\text{-}b||$ for all matrices A that represent a suitable tree and vectors x fulfilling $x_i \geq 0$ (i = 1,...n), $x_1 + ... + x_n = 100$. The optimization was carried out using the python cvxopt package (Andersen et al., 2018).

The impact of measurement errors in b on the reconstructed clonal frequencies x can be calculated based on Cramer's rule. For two vectors b, b' and the corresponding solutions x, x' we obtain $A(x\text{-}x') = b\text{-}b'$. Since the determinant of A is equal to one, Cramer's rule implies $x_i\text{-}x_i' = det(A_i)$, where A_i denotes the matrix A with the ith column replaced by $b\text{-}b'$. Consequently, $x_n\text{-}x_n' = b_n\text{-}b_n'$ and $|x_{n-1}\text{-}x_{n-1}'| \leq |b_n\text{-}b_n'| + |b_{n-1}\text{-}b_{n-1}'|$. Analogous formulas can be derived for $i < n\text{-}1$. However, depending on the structure of A, they can be lengthy. Therefore, the use of the condition number of A seems to be more convenient to estimate the errors. It quantifies how perturbations in b impact on the changes of x. For all considered tree structures, the condition numbers of the related matrices computed in the l^2 norm are provided in Section 2 of the Supplement.

Data

We plan to investigate if it is possible to reconstruct clonal hierarchies from bulk sequencing samples. This requires that the "true" clonal hierarchy is known, so that we can compare the result of our algorithm with reality. To know the "true" hierarchy we use single cell sequencing data from ref. (Ediriwickrema et al., 2020). We understand the clonal hierarchy and the clonal frequencies obtained from the single cell sequencing as ground truth. Since for the samples analyzed in Ediriwickrema et al. (2020) no bulk data are available, we calculate the bulk allele frequencies based on the single cell data. For simplicity we assume that the considered sample only contains leukemic cells and we exclude all sequenced wild type cells from the data. We calculate the bulk VAF of variant allele i as $a_{i1}f_1 + ... + a_{in}f_n$, where f_i is the frequency of clone i in the single cell data set and $a_{ij} = 1$ if clone j carries variant allele i and 0 otherwise. Since we consider a purely leukemic sample, the calculated VAF are normalized such that the frequency of the most abundant variant allele is 100%. We consider all patients from Ediriwickrema et al. (2020) that carry only heterozygous mutations and for whom data at diagnosis and relapse is available.

RESULTS

Exact Input Data Often Result in Unique Clonal Hierarchies

As gold standard we use the single cell sequencing data from Ediriwickrema et al. (2020), which provide the true clonal hierarchy and hence can be used to test the proposed algorithm. Based on the single cell data we calculate the variant allele frequencies in the bulk sample. The first question we ask is how many clonal hierarchies are compatible with the bulk variant allele frequencies of a given patient. **Figure 2** shows for each patient which hierarchies exactly fit to the data at diagnosis. We observe that, for 5 out of 6 patients, only one hierarchy exactly fits the bulk data. For one patient 6 hierarchies are consistent with the bulk data.

Similar observations hold for the relapse samples of the considered patients, **Figure 3**. Here all samples, except one (Patient 5) lead to unique tree configurations. In case of patient five all sequenced cells belong to the same clone, which makes it impossible to infer the order of mutations. In the next step we combine the diagnosis and relapse sample of each patient. For each patient **Figure 4** shows the tree configurations that are compatible with the data at both time points. We have uniqueness in all except one case.

To provide insights into the question whether the structure of the true hierarchy (e.g., linear vs. branched) determines how many tree configurations fit to the bulk dataset, we perform a computational experiment. We consider all tree structures for n = 4. For each of them, we generate 10000 bulk data sets by a random distribution among the different clones. Then, for

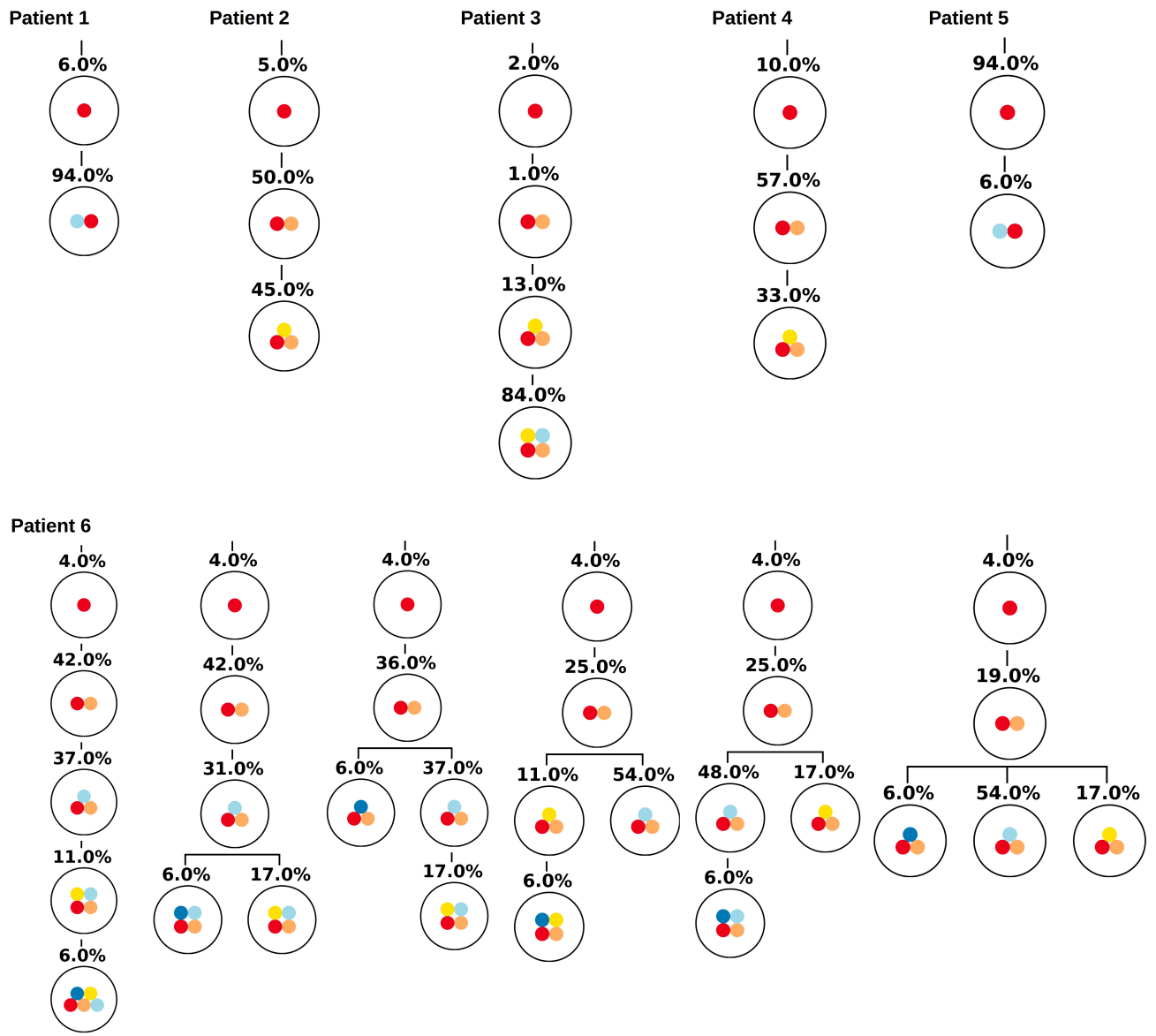

FIGURE 2 | Clonal hierarchies compatible with the bulk allele frequencies measured at diagnosis. For each considered patient all clonal hierarchies are depicted that are compatible with the bulk variant allele frequencies measured at diagnosis. The root of the tree corresponds to the founder mutation that is present in all leukemic cells. The percentages indicate the frequencies of the respective clones that have to be mixed to obtain the measured bulk VAF. We observe that in most cases the hierarchies are unique.

each randomly generated dataset, we check how many other tree structures reproduce the data without an error. The results are shown in the Supplement (**Supplementary Figure 1**). They suggest that linear hierarchies or hierarchies where branches appear only at nodes with a high depth exhibit uniqueness in many cases. The structures with branches near the root often admit multiple reconstructions. However, if data at two time-points are available, e.g., at diagnosis and relapse, in 50–70% of the cases only one or two configurations exactly fit to the bulk data.

Sampling Error Has Little Impact on the Uniqueness of Clonal Hierarchies

If the frequency of different clones in a large population is estimated based on a small sample, sampling errors can occur. To study the impact of sampling errors on the reconstructed clonal hierarchies we again use the single cell sequencing data from Ediriwickrema et al. (2020). We assume that the single cell data reflect the true frequencies of the clones in the malignant cell bulk of the respective patient. For an arbitrary patient k we

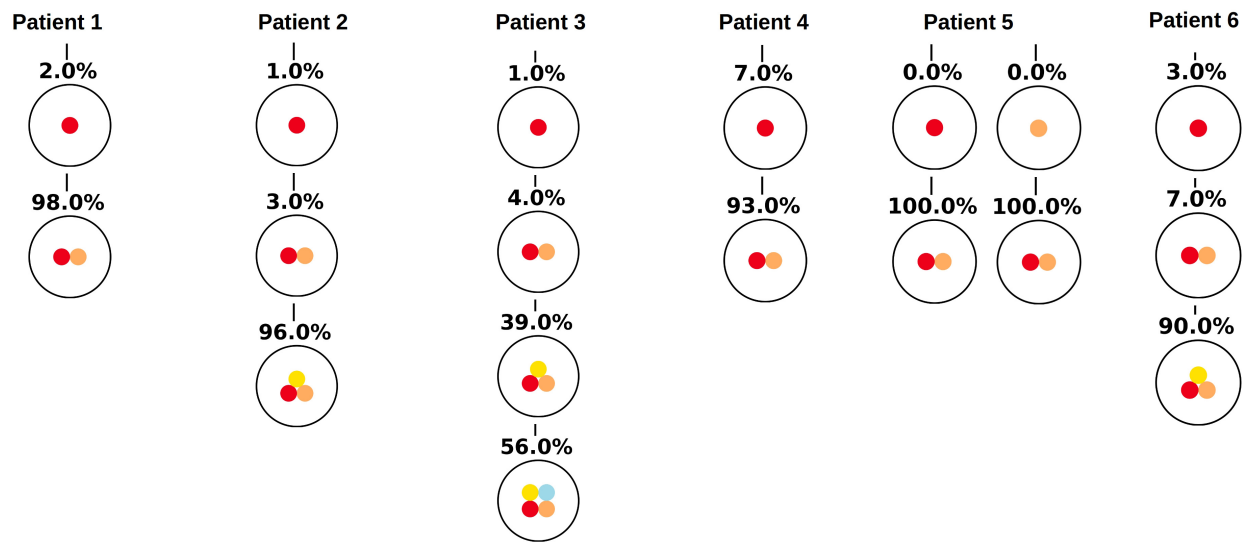

FIGURE 3 | Clonal hierarchies compatible with the bulk allele frequencies measured at relapse. For each considered patient, the figure shows which clonal hierarchies are compatible with the bulk variant allele frequencies measured at relapse. The root of the tree corresponds to the founder mutation. We observe that in most cases the hierarchies are unique.

know the total number n_k of sequenced leukemic single cells. Furthermore we know the frequencies $f_{i,k}$ of each clone that has been detected (here $f_{i,k}$ denotes the frequency of clone i in the sample of patient k). To study the impact of sampling on the bulk variant allele frequencies and on the reconstructed hierarchies for patient k, we draw 1,000 random samples of size n_k from a multinomial distribution with probabilities $p_i = f_{i,k}$. This approach is referred to as resampling (Gigli, 1996). For each of these 1,000 random samples we calculate the bulk variant allele frequencies and apply our algorithm to reconstruct the clonal hierarchies. The results are shown in **Figure 5**. In all cases except one the hierarchies fitting exactly to the data remain unique and are identical to the hierarchies obtained based on the exact data. For one patient in some of the resampled datasets the number of hierarchies matching the data increases by one. These results imply that the sampling error has a negligible impact on the clonal pedigrees that fit to the data. The sampling error also affects the clonal frequencies x_i obtained from the reconstruction. The reconstruction is based on linear equations. Therefore, if many samples are drawn from the same patient, the mean over the reconstructed frequencies approximates the true frequencies of the respective clones. We have assessed the standard deviations of the reconstructed clonal frequencies numerically based on 1,000 re-samplings. In all cases they were less than 4%, in patients 1 to 5 they were less than 1.5%.

Impact of Measurement Errors on Reconstruction of Clonal Hierarchies

Inaccuracies in sequencing are another possible source of error. To study their impact on the reconstructed clonal hierarchies, we add a normally distributed error to the bulk frequency of each allele. Such errors can have different impacts on the reconstructed clonal hierarchies. For each patient we considered 1,000 randomly perturbed versions of the original data. If the standard deviation of the error distribution is 0.5% (i.e., in 68% of cases the error is less or equal 0.5%, in 95% of cases the error is less or equal to 1%) the reconstruction algorithm works reliably in the sense that the true configuration is an optimal configuration, see **Figure 6**. In 5 out of 6 considered patients the optimum is unique. We repeated the simulation for a normally distributed error with a standard deviation of 5%, i.e., in 95% of cases the error is less than 10%, see **Figure 6**. For an error of this magnitude the true configuration not always remains an optimal configuration. Examples illustrating this observation are provided in the Supplement (**Supplementary Figure 2**). This especially applies to patients in whom the frequency of the founding clone is small (i.e., patients 2, 3 and 6). If the error is larger than the frequency of the founder clone it becomes impossible to reliably detect which hit occurs first. However, also in a single cell sequencing approach, rare clones can remain undetected due to sampling or sequencing errors, implying that the first hit remains unknown. In terms of variant allele frequencies this implies that trees cannot be reliably reconstructed if the difference between the two most abundant allele frequencies is of the order of magnitude of the sequencing error. In patients with many clones, our algorithm can often rule out most of the possible hierarchies and identify a small number of configurations fitting the data. In case of Patient 5 the true configuration is always among the upper 12% of the best fitting configurations (i.e., the best or second best), and in patient 6 among the upper 3.3% of the best fitting configurations (i.e., 4 out of 125). In case of small clone numbers such as for patients 2 or 4, the true configuration is always among the two best fitting hierarchies.

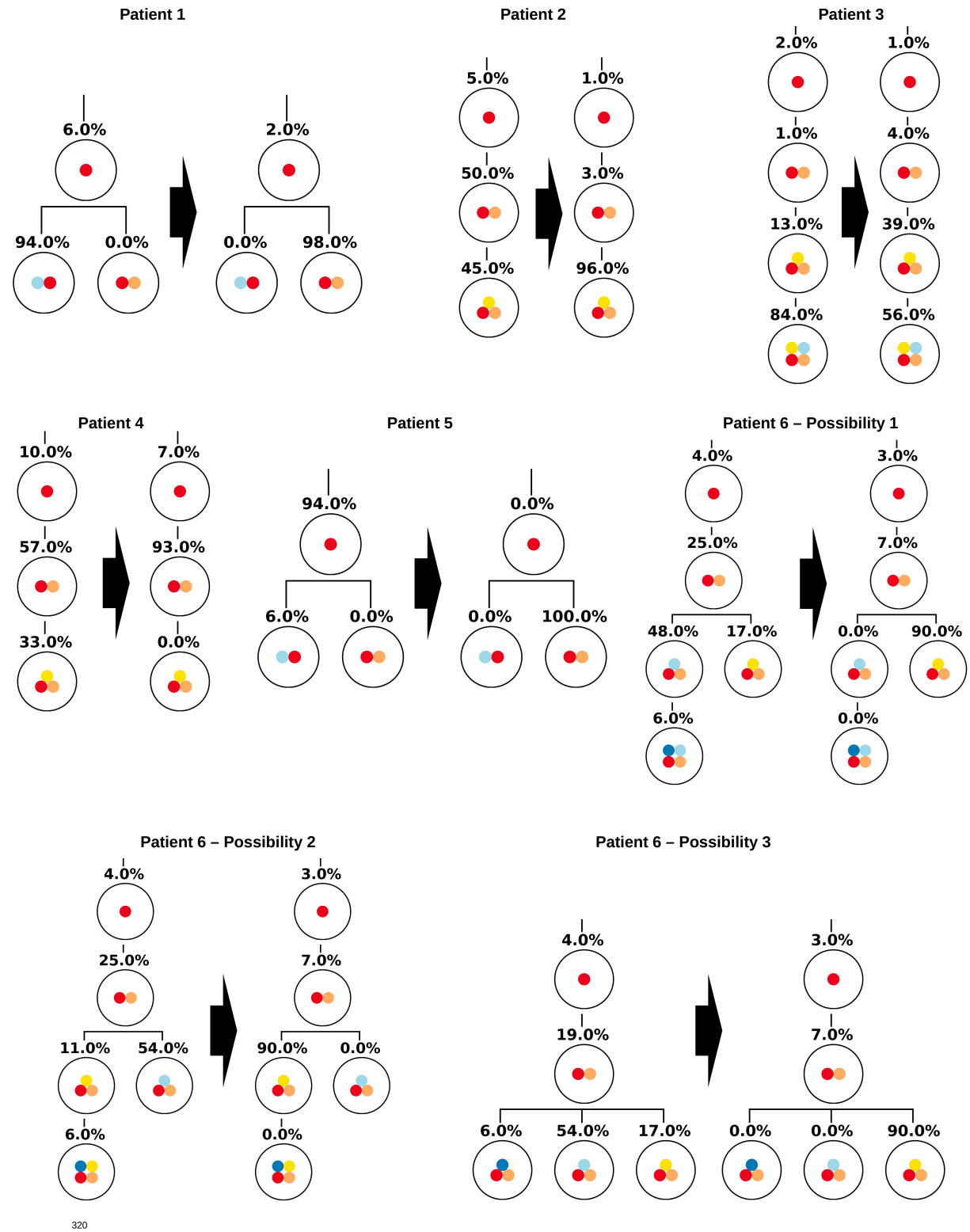

FIGURE 4 | Clonal hierarchies compatible with the bulk allele frequencies measured at diagnosis and relapse. For each patient, the figure shows all clonal hierarchies that are compatible with the bulk VAFs measured at diagnosis and relapse. The root of the tree corresponds to the founder mutation. We observe that in case of patient 6, the number of hierarchies compatible with the data is reduced compared to **Figure 2**. For patients 1–5, the reconstructed hierarchies coincide with the result from single cell sequencing. For patient 6, Possibility 3 corresponds to the true configuration.

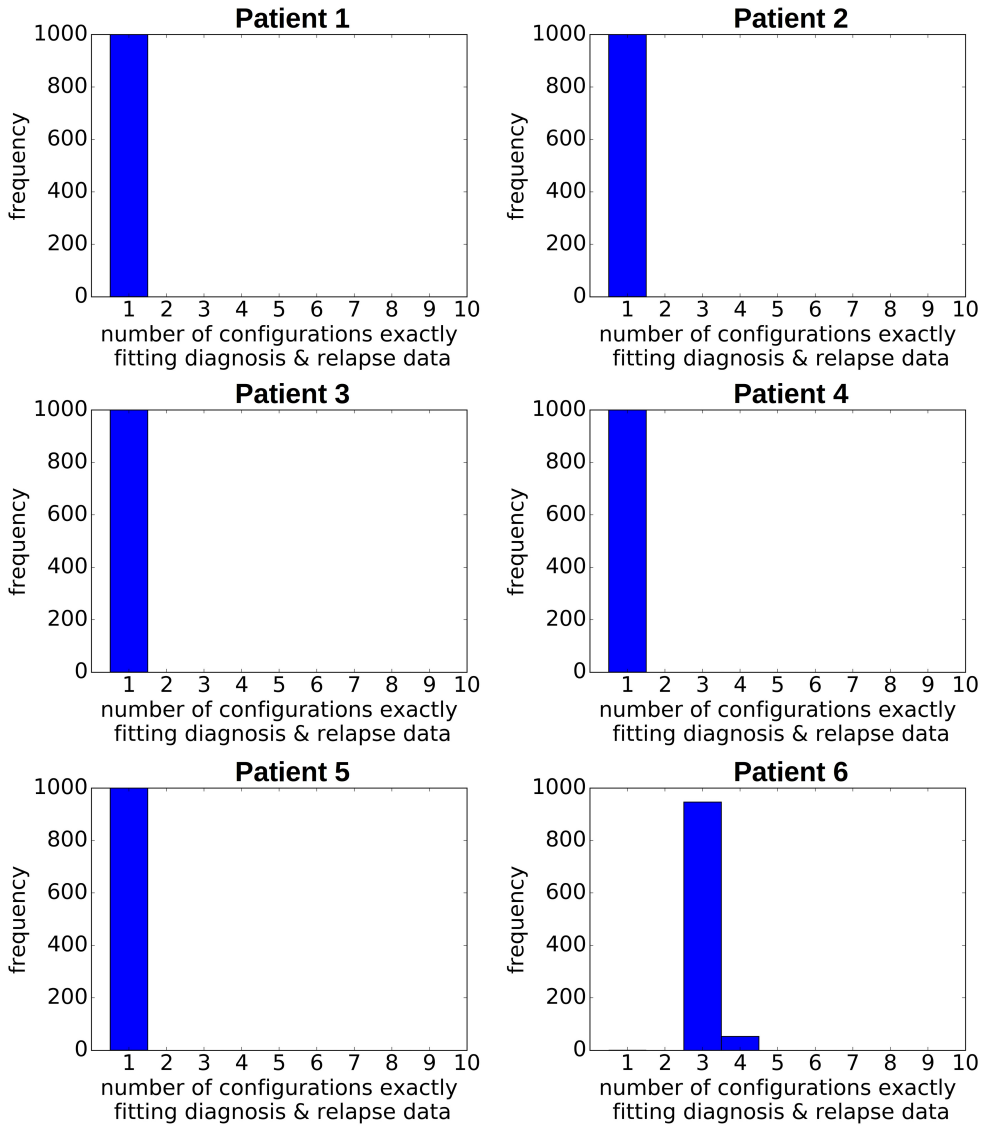

FIGURE 5 | Impact of sampling error on the reconstructed hierarchies. For each patient, we generate 1,000 random pairs of diagnosis and relapse samples from a multivariate distribution. The probabilities of the multivariate distribution equal the clonal frequencies in the single cell data. The size of the samples equals the number of sequenced leukemic cells. We recorded for each pair of randomly generated diagnosis/relapse samples the number of clonal hierarchies compatible with the resampled diagnosis and relapse data. The vertical axis shows how many of the 1,000 samples were compatible with 1, 2, 3, . . . hierarchies, respectively.

An Example of a Patient With Two Founder Clones

We now consider an example of a patient with two different founder clones. This scenario either corresponds to the rare case where the AML cell population originates from clones with different initial mutations or it corresponds to the case where the common founding mutation has not been detected. The latter may especially occur in the setting of targeted sequencing, where only a predefined subset of mutations is considered. Such a scenario occurs if in a purified AML sample (i.e., in a sample without healthy cells) all bulk VAF are significantly different from the expected maximum of 50% (for heterozygous mutations) or 100% (for homozygous mutations).

The proposed algorithm can take this scenario into account by considering healthy cells as the root of the tree. This means a healthy reference allele that is present in all cells is added to the list of variant allele frequencies, to obtain a single tree with a unique root. **Figure 7** shows all tree structures that are compatible with the measured data. The tree structures can be divided into two classes. In the first class of solutions, the frequency of healthy cells is zero at diagnosis and relapse (possibilities 1–2), in the second class the frequency of the healthy cells is positive (here 15%) at least one time point (possibilities 3–7). Solutions of the first class imply that there exist two founding clones (or an undetected unique founder mutation), solutions of the second class may imply that the sample contains a mixture of healthy and

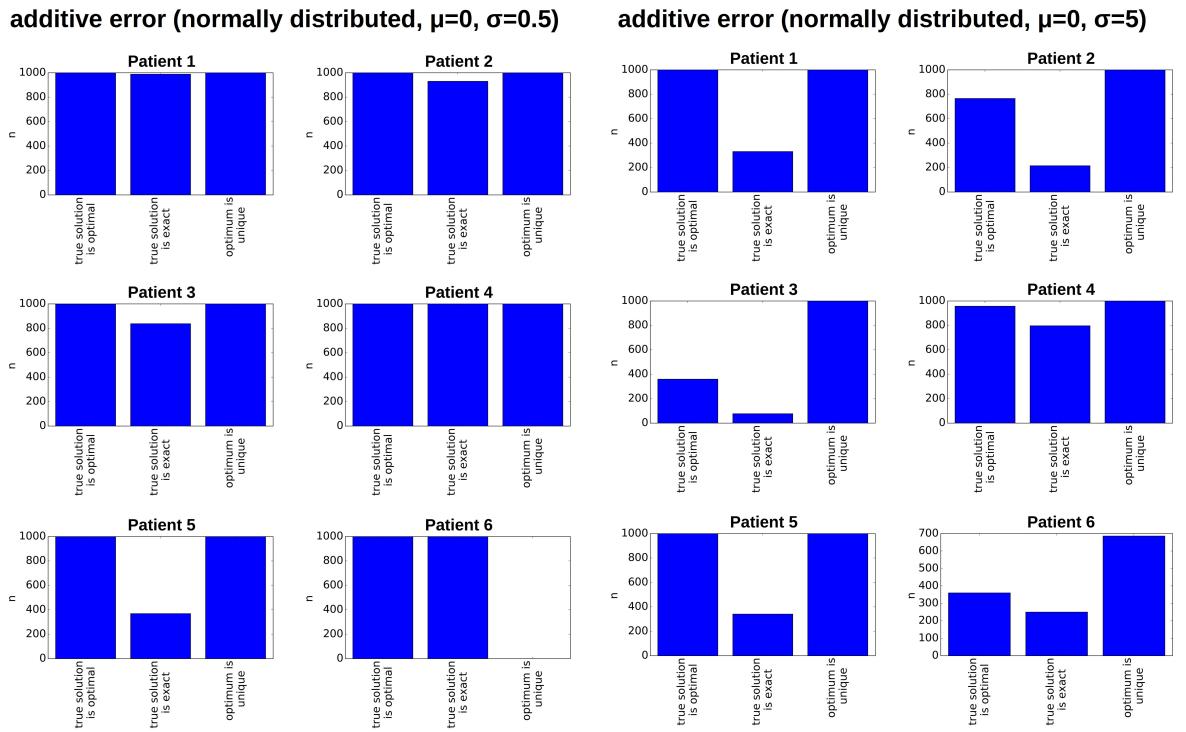

FIGURE 6 | Impact of measurement errors on the reconstructed hierarchies. For each patient, we generate 1,000 randomly perturbed pairs of diagnosis and relapse samples. The additive random perturbations were drawn from a normal distribution with mean zero and standard deviation 0.5% (left) or 5% (right), respectively. Perturbations leading to a VAF of less than zero or more than 100% were excluded. For each of the perturbed diagnosis/relapse pairs, we reconstructed all compatible clonal hierarchies. The figure indicates for how many of the 1,000 perturbed samples the true hierarchy optimally fits the perturbed data (compared to all other existing hierarchies), whether the true hierarchy can exactly reproduce the perturbed data and whether the optimal configuration is unique.

leukemic cells. If we can be sure that the experimental procedures prevent healthy cells from being sequenced (e.g., by FACS sorting for a leukemia specific surface marker before sequencing), only two possible tree structures remain.

As for the other patients the sampling error only leads to small changes in the numbers of clonal hierarchies that fit the data, **Figure 8A**. However, already small errors added to the bulk VAFs (normally distributed with a standard deviation of 0.5%) imply that in a majority of cases the true solution is no longer optimal, **Figures 8B,C**. The reason for this observation is as follows (see **Figure 9**). In the exact scenario there exist two founder mutations. The frequencies of both founder mutations add up to 100%. In presence of errors, it can happen that the frequencies of both founder mutations do not add up to exactly 100%. If their sum is slightly less than 100% the true hierarchy still leads to an exact solution (to compensate for the error the exact solution contains a small number of healthy cells). If due to the random error the sum over both sub-trees is slightly more than 100%, an exact solution is no longer possible. To circumvent this, we can relax the dataset by artificially adding a small number of healthy cells, e.g., x% to the dataset. In this case, for measurements where the frequencies of both founding clones add up to less than 100% + x% the true configuration still is an exact solution. We see in **Figure 10** that this relaxation increases the number of cases where the true solution is an optimal solution.

DISCUSSION

The aim of this study is to investigate the ambiguity of clonal hierarchies that are reconstructed from bulk sequencing data. For this purpose, we develop an algorithm that systematically tests which subset of all clonal hierarchies optimally fits a given dataset. We test this algorithm using bulk VAFs that have been calculated based on cell sequencing data sets. Since single cell sequencing reveals the true clonal hierarchy, this approach enables us to compare the output of our algorithm to the real configuration (Kuipers et al., 2017; Brierley and Mead, 2020).

First, we assume that the input data is exact, i.e., neither sampling nor measurement errors occur. Then for most of the considered patient samples exactly one clonal hierarchy optimally fits the bulk VAF. This clonal hierarchy is identical to the hierarchy obtained from single cell sequencing. In two of the considered patients, even for exact input data more than one clonal hierarchy is compatible with the bulk allele frequencies. The true hierarchy obtained from single cell sequencing is among them. This finding implies that even in absence of measurement error, the clonal hierarchy may not be uniquely defined by the bulk VAF.

When drawing multiple samples from the same malignant cell population the variant allele frequencies may differ from one sample to another. This may be caused by sampling error, or it

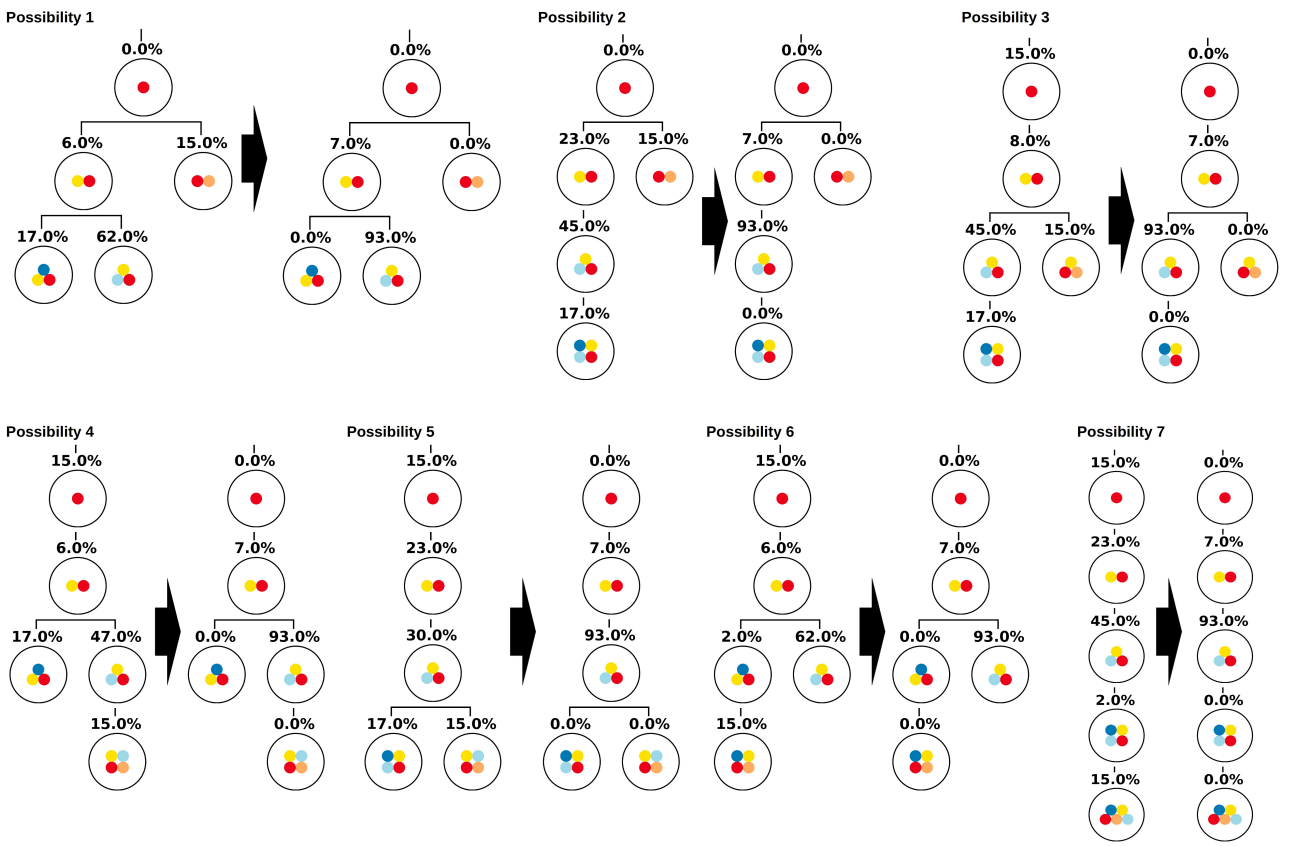

FIGURE 7 | Example of a patient with multiple founder clones. In this figure the root corresponds to wild type cells. The two founding events are indicated by yellow and orange circles. The Figure shows all hierarchies that are compatible with VAFs at diagnosis and relapse. Possibility 1 coincides with the hierarchy obtained from single cell sequencing.

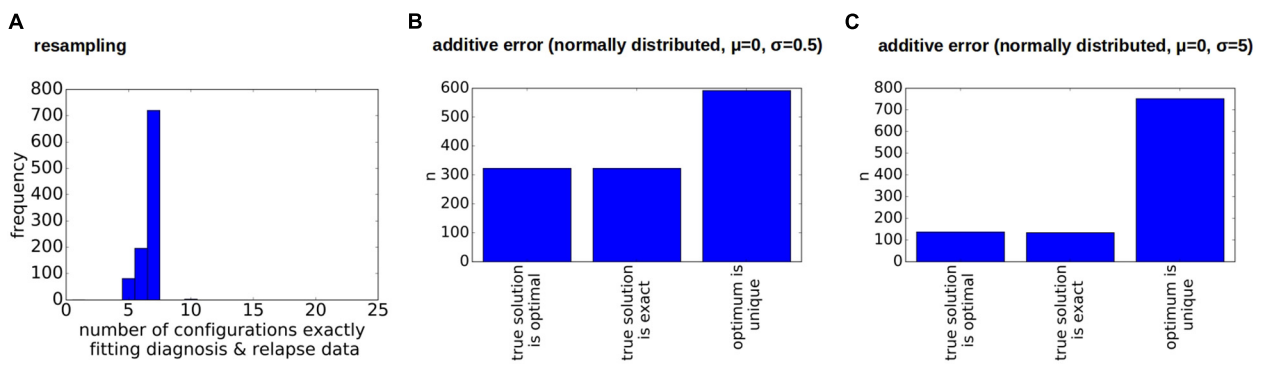

FIGURE 8 | Impact of errors on the reconstructions for a sample with two founders. **(A)** Impact of sampling error on the number of clonal hierarchies compatible with diagnosis and relapse data. The vertical axis shows how many of the 1,000 multinomial respamplings were compatible with 1, 2, 3, . . . hierarchies respectively. **(B,C)** Impact of measurement errors on the reconstructed hierarchies. We considered 1,000 perturbed versions of the original data. Additive perturbations of the VAF were drawn from a normal distribution with mean zero and standard deviation 0.5% **(B)** or 5% **(C)**. We observe that in the majority of cases the true configuration is not optimal.

may reflect inhomogeneity of the tumor. Assuming the tumor to be homogeneous, we aim to quantify the impact of sampling error on the reconstructed hierarchies. For each patient, the number of sequenced leukemic single cells n and the frequencies f_i of the different clones are known. To simulate the sampling error, for each patient we generate 1,000 random samples of size n

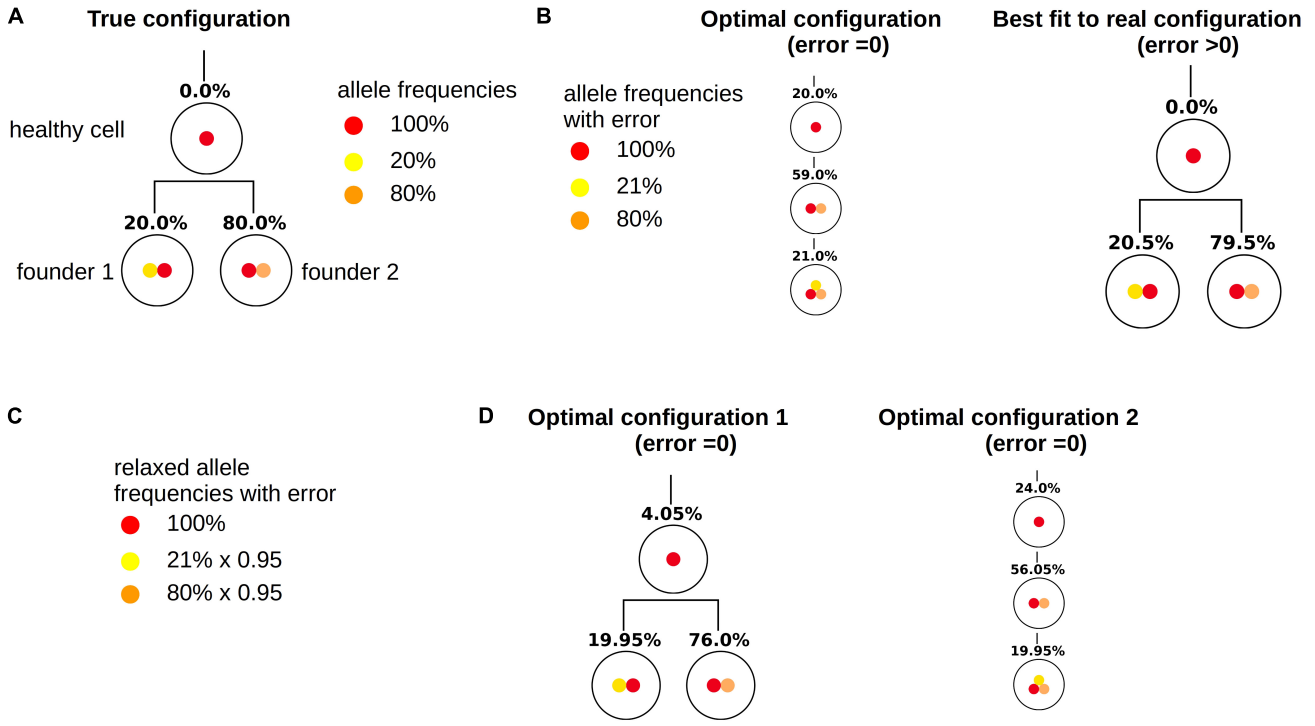

FIGURE 9 | Relaxation of measurements. This figure demonstrates how relaxation of the VAF can help to make the true configuration an optimal configuration. **(A)** True configuration with two founder clones. Wild type cells are added to the tree to have a unique root. Since the sample only contains leukemic cells the frequency of wild type cells equals zero. The frequencies of the two founder clones add up to 100%. **(B)** Variant allele frequencies are perturbed by a measurement error. The frequencies of the two founding mutations no longer add up to 100%. Therefore the true hierarchy is no longer an optimal hierarchy. **(C)** By multiplying the frequencies of mutated alleles with 0.95 we artificially add 5% of healthy cells to the sample. **(D)** If we reconstruct the clonal hierarchies for the modified data the true hierarchy is among the optimal hierarchies.

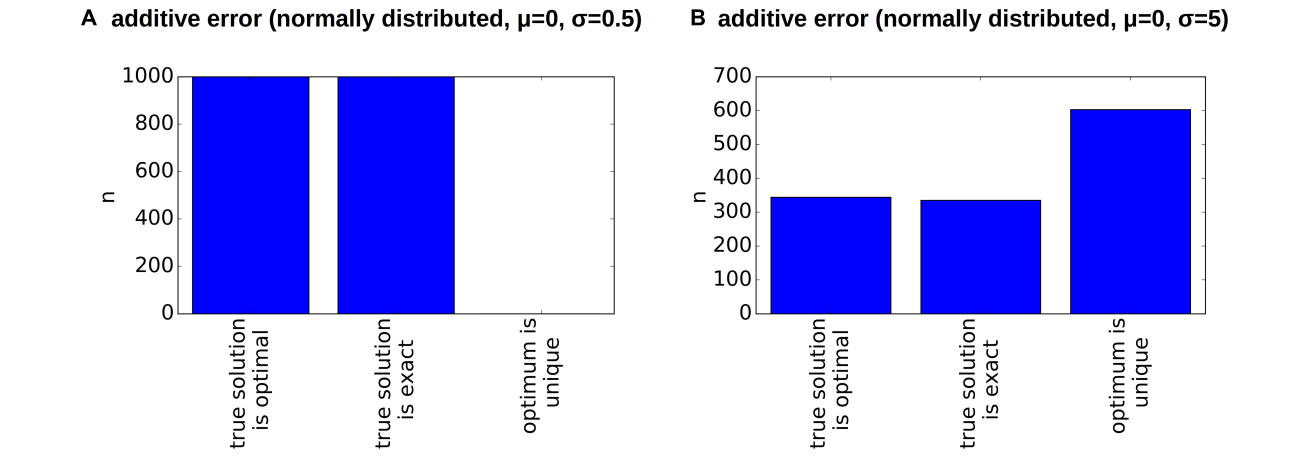

FIGURE 10 | Clonal hierarchies fitting the relaxed dataset. Here, we consider the relaxed version of the dataset from Patient 7. The relaxation makes the reconstruction more robust to errors. Additive perturbations of the VAF were drawn from a normal distribution with mean zero and standard deviation 0.5% **(A)** or 5% **(B)**. In comparison to **Figures 8B,C** the true hierarchy is in many cases an optimal hierarchy.

drawn from a multinomial distribution with probabilities $p_i = f_i$. For each of these random samples we calculate the bulk allele frequencies and construct all clonal hierarchies compatible with them. Based on results of this exercise we conclude that the sampling error has a negligible impact on the obtained clonal hierarchies, at least for the data at our disposal.

We test the robustness of the reconstruction by adding normally distributed errors of different amplitude to the bulk VAFs calculated from the single cell sequencing data. This takes into account potential misreads during the sequencing, amplification errors or impurities of the sample. We observe that for errors of about 5–10% the true hierarchy not necessarily

remains optimal. This especially applies to data sets where the frequency of the founding clone is of the order of magnitude of the error. However, even in this case, the true clonal structure is among the upper 3–15% of the best fitting hierarchies. This implies that also in the presence of relevant errors, our algorithm allows to rule out most tree configurations and results in a small subset of possible clonal hierarchies fitting to a data sample.

Mathematical models indicate that tree characteristics, e.g., the depth of the tree, correlate with clonal properties such as self-renewal and proliferation rate (Stiehl et al., 2016). In this context it can be sufficient to have an estimate of the depth of the true clonal hierarchy to draw conclusions about the effect of a mutation on cell kinetics or patient prognosis. This implies that in the case of non-unique clonal hierarchies, biological conclusions can be drawn if the potential hierarchies are sufficiently similar to each other.

Having measurements of bulk VAFs provided, our computational approach can be used to rank all possible clonal hierarchies based on their compatibility with the data (i.e., the smaller the error when fitting the dataset to a given hierarchy, the better the rank of the respective hierarchy). For all datasets considered in this study the real hierarchy is among the upper 3–15% of this ranking. Taking into account that in case of n clones n^{n-1} possible hierarchies exist our algorithm allows to rule out a significant number of them. Our algorithm can also be applied to scenarios in which the disease is derived from multiple founding clones. However, due to its combinatorial nature the algorithm can only be applied to relatively small clone numbers.

Our computational approach can be used to study how sensitive the reconstructed hierarchies are to perturbations of the input data. By adding random errors to the input data obtained from an experiment and by repeating the reconstruction with the perturbed input data it turns out that some datasets are robust with respect to the perturbations. This means that the obtained optimal clonal hierarchies do not change if the input data is perturbed. For other datasets perturbations of the input data leads to a change of the reconstructed hierarchies, indicating that the reconstruction may be affected by measurement errors. The robustness of a given dataset can be checked using our proposed framework. It is straightforward to take into account that the measured frequencies may have different confidence intervals. In principle our approach can also be applied to clustered single nucleotide variants (SNVs). Since the number of detected SNVs is usually high, the variants are grouped into clusters according to their allele frequencies. Each cluster comprises all SNVs with a similar allele frequency. The cluster center is defined as the average allele frequency of all SNVs that belong to the respective cluster. In this setting our algorithm can be applied using cluster centers as input data.

Mechanistic mathematical models allow to extract relevant information from clonal hierarchies, such as estimates of proliferation rates and self-renewal of the different clones (Whichard et al., 2010; Stiehl et al., 2016). Correlating these estimates with detected mutations and clinical observations may provide new insights into AML pathophysiology (Stiehl et al., 2014, 2016). The proposed framework is a first attempt to quantify the ambiguity emerging during reconstruction of clonal hierarchies from bulk sequencing data. It allows to identify when such reconstructions are reliable and can be used as input data for mechanistic models. This knowledge helps to make available routine clinical data to studies that require clonally resolved input (Leung et al., 2017).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Ediriwickrema et al. (2020).

AUTHOR CONTRIBUTIONS

TS and AM-C designed the research, discussed the results, and wrote the manuscript. TS implemented the algorithm and run simulations. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by research funding from the German Research Foundation DFG (SFB 873; subproject B08).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2021.596194/full#supplementary-material

REFERENCES

Abdel-Wahab, O., Manshouri, T., Patel, J., Harris, K., Yao, J., Hedvat, C., et al. (2010). Genetic analysis of transforming events that convert chronic myeloproliferative neoplasms to leukemias. *Cancer Res.* 70, 447–452. doi: 10.1158/0008-5472.can-09-3783

Anderson, K., Lutz, C., van Delft, F. W., Bateman, C. M., Guo, Y., Colman, S. M., et al. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature.* 469, 356–361. doi: 10.1038/nature09650

Andersen, M., Dahl, J., and Vandenberghe, L. (2018). *CVXOPT: A Python Package for Convex Optimization, Version 1.2.0*. Available online at: http://cvxopt.org/

Attolini, C. S., Cheng, Y. K., Beroukhim, R., Getz, G., Abdel-Wahab, O., Levine, R. L., et al. (2010). A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. U S A.* 107, 17604–17609. doi: 10.1073/pnas.1009117107

Bachas, C., Schuurhuis, G. J., Assaraf, Y. G., Kwidama, Z. J., Kelder, A., Wouters, F., et al. (2012). The role of minor subpopulations within the leukemic blast compartment of AML patients at initial diagnosis in the development of relapse. *Leukemia* 26, 1313–1320. doi: 10.1038/leu.2011.383

Bacher, U., Haferlach, C., Kern, W., Haferlach, T., and Schnittger, S. (2008). Prognostic relevance of FLT3-TKD mutations in AML: the combination matters–an analysis of 3082 patients. *Blood* 111, 2527–2537. doi: 10.1182/blood-2007-05-091215

Banck, J. C., and Görlich, D. (2019). In-silico comparison of two induction regimens (7 + 3 vs 7 + 3 plus additional bone marrow evaluation) in acute myeloid leukemia treatment. *BMC Syst. Biol.* 13:18. doi: 10.1186/s12918-019-0684-0

Brierley, C. K., and Mead, A. J. (2020). Single-cell sequencing in hematology. *Curr. Opin. Oncol.* 32, 139–145. doi: 10.1097/cco.0000000000000613

Busse, J. E., Gwiazda, P., and Marciniak-Czochra, A. (2016). Mass concentration in a nonlocal model of clonal selection. *J. Math. Biol.* 73, 1001–1033. doi: 10.1007/s00285-016-0979-3

Cancer Genome Atlas Research Network. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* 368, 2059–2074. doi: 10.1056/nejmoa1301689

Caravagna, G., Sanguinetti, G., Graham, T. A., and Sottoriva, A. (2020). The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data. *BMC Bioinform.* 21:531. doi: 10.1186/s12859-020-03863-1

Cocciardi, S., Dolnik, A., Kapp-Schwoerer, S., Rücker, F. G., Lux, S., Blätte, T. J., et al. (2019). Clonal evolution patterns in acute myeloid leukemia with NPM1 mutation. *Nat. Commun.* 10:2031.

Delhommeau, F., Dupont, S., Della Valle, V., James, C., Trannoy, S., Massé, A., et al. (2009). Mutation in TET2 in myeloid cancers. *N. Engl. J. Med.* 360, 2289–2301.

Diestel, R. (2017). *Graph Theory*. Cham: Springer.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481, 506–510.

Dinh, K., Corey, S. J., and Kimmel, M. (2019). Predicting minimal residual disease in acute myeloid leukemia through stochastic modeling of clonality. *Blood* 134:1448. doi: 10.1182/blood-2019-127457

Dinh, K. N., Corey, S. J., and Kimmel, M. (2020). Application of the moran model in estimating selection coefficient of mutated CSF3R clones in the evolution of severe congenital neutropenia to myeloid neoplasia. *Front. Physiol.* 11:806. doi: 10.3389/fphys.2020.00806

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., et al. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* 129, 424–447. doi: 10.1182/blood-2016-08-733196

Ediriwickrema, A., Aleshin, A., Reiter, J. G., Corces, M. R., Köhnke, T., Stafford, M., et al. (2020). Single-cell mutational profiling enhances the clinical evaluation of AML MRD. *Blood Adv.* 4, 943–952. doi: 10.1182/bloodadvances.2019001181

Gigli, A. (1996). Efficient bootstrap methods: A review. *J. Ital. Statist. Soc.* 1, 99–127. doi: 10.1007/bf02589584

Greif, P. A., Hartmann, L., Vosberg, S., Stief, S. M., Mattes, R., Hellmann, I., et al. (2018). Evolution of cytogenetically normal acute myeloid leukemia during therapy and relapse: an exome sequencing study of 50 patients. *Clin. Cancer Res.* 24, 1716–1726. doi: 10.1158/1078-0432.ccr-17-2344

Grimwade, D., Walker, H., Oliver, F., Wheatley, K., Harrison, C., Harrison, G., et al. (1998). The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The medical research council adult and children's leukaemia working parties. *Blood* 92, 2322–2333. doi: 10.1182/blood.v92.7.2322

Herudkova, Z., Culen, M., Folta, A., Jeziskova, I., Cerna, J., Loja, T., et al. (2020). Clonal hierarchy of main molecular lesions in acute myeloid leukaemia. *Br. J. Haematol.* 190, 562–572. doi: 10.1111/bjh.16341

Jonas, B. A. (2017). From MDS/AML to iPSC and back again. *Sci. Transl. Med.* 9:eaam9861. doi: 10.1126/scitranslmed.aam9861

Kuipers, J., Jahn, K., and Beerenwinkel, N. (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochim. Biophys. Acta Rev. Cancer* 1867, 127–138. doi: 10.1016/j.bbcan.2017.02.001

Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., et al. (2017). Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.* 27, 1287–1299. doi: 10.1101/gr.209973.116

Lorenzi, T., Marciniak-Czochra, A., and Stiehl, T. (2019). A structured population model of clonal selection in acute leukemias with multiple maturation stages. *J. Math Biol.* 79, 1587–1621. doi: 10.1007/s00285-019-01404-w

Lutz, C., Hoang, V. T., Buss, E., and Ho, A. D. (2013). Identifying leukemia stem cells–is it feasible and does it matter? *Cancer Lett.* 338, 10–14. doi: 10.1016/j.canlet.2012.07.014

Nobile, M. S., Vlachou, T., Spolaor, S., Bossi, D., Cazzaniga, P., Lanfrancone, L., et al. (2019). Modeling cell proliferation in human acute myeloid leukemia xenografts. *Bioinformatics* 35, 3378–3386. doi: 10.1093/bioinformatics/btz063

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* 374, 2209–2221.

Prüfer, H. (1918). Neuer Beweis eines Satzes über Permutationen. *Arch. Math. Phys.* 27, 742–744.

Rahman, M. S., Nicholson, A. E., and Haffari, G. (2018). HetFHMM: A novel approach to infer tumor heterogeneity using factorial hidden markov models. *J. Comput. Biol.* 25, 182–193. doi: 10.1089/cmb.2017.0101

Ran, D., Schubert, M., Taubert, I., Eckstein, V., Bellos, F., Jauch, A., et al. (2012). Heterogeneity of leukemia stem cell candidates at diagnosis of acute myeloid leukemia and their clinical significance. *Exp. Hematol.* 40, 155–165. doi: 10.1016/j.exphem.2011.10.005

Röllig, C., Bornhäuser, M., Thiede, C., Taube, F., Kramer, M., Mohr, B., et al. (2011). Long-term prognosis of acute myeloid leukemia according to the new genetic risk classification of the European LeukemiaNet recommendations: evaluation of the proposed reporting system. *J. Clin. Oncol.* 29, 2758–2765. doi: 10.1200/jco.2010.32.8500

Roloff, G. W., and Griffiths, E. A. (2018). When to obtain genomic data in Acute Myeloid Leukemia (AML) and which mutations matter. *Blood Adv.* 2, 3070–3080. doi: 10.1182/bloodadvances.2018020206

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., et al. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* 11, 396–398. doi: 10.1038/nmeth.2883

Salichos, L., Meyerson, W., Warrell, J., and Gerstein, M. (2020). Estimating growth patterns and driver effects in tumor evolution from individual samples. *Nat. Commun.* 11:732.

Sandén, C., Lilljebjörn, H., Orsmark Pietras, C., Henningsson, R., Saba, K. H., Landberg, N., et al. (2020). Clonal competition within complex evolutionary hierarchies shapes AML over time. *Nat. Commun.* 11:579.

Stiehl, T., Baran, N., Ho, A. D., and Marciniak-Czochra, A. (2014). Clonal selection and therapy resistance in acute leukaemias: mathematical modelling explains different proliferation patterns at diagnosis and relapse. *J. R. Soc. Interface* 11:20140079. doi: 10.1098/rsif.2014.0079

Stiehl, T., Baran, N., Ho, A. D., and Marciniak-Czochra, A. (2015). Cell division patterns in acute myeloid leukemia stem-like cells determine clinical course: a model to predict patient survival. *Cancer Res.* 75, 940–949. doi: 10.1158/0008-5472.can-14-2508

Stiehl, T., Ho, A. D., and Marciniak-Czochra, A. (2018). Mathematical modeling of the impact of cytokine response of acute myeloid leukemia cells on patient prognosis. *Sci. Rep.* 8:2809.

Stiehl, T., Lutz, C., and Marciniak-Czochra, A. (2016). Emergence of heterogeneity in acute leukemias. *Biol. Direct.* 11:51.

Stiehl, T., and Marciniak-Czochra, A. (2017). Stem cell self-renewal in regeneration and cancer: Insights from mathematical modeling. *Curr. Opin. Syst. Biol.* 5, 112–120. doi: 10.1016/j.coisb.2017.09.006

Stiehl, T., Wang, W., Lutz, C., and Marciniak-Czochra, A. (2020). Mathematical modeling provides evidence for niche competition in human AML and serves as a tool to improve risk stratification. *Cancer Res.* 80, 3983–3992. doi: 10.1158/0008-5472.can-20-0283

Wang, W., Stiehl, T., Raffel, S., Hoang, V. T., Hoffmann, I., Poisa-Beiro, L., et al. (2017). Reduced hematopoietic stem cell frequency predicts outcome in acute myeloid leukemia. *Haematologica* 102, 1567–1577. doi: 10.3324/haematol.2016.163584

Whichard, Z. L., Sarkar, C. A., Kimmel, M., and Corey, S. J. (2010). Hematopoiesis and its disorders: a systems biology approach. *Blood* 115, 2339–2347. doi: 10.1182/blood-2009-08-215798

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership