

# ADVANCES IN MATHEMATICAL AND COMPUTATIONAL ONCOLOGY, VOLUME II

EDITED BY: George Bebis, Max A. Alekseyev, Heyrim Cho, Jana Gevertz,  
David A. Hormuth, II and Maria Rodriguez Martinez  
PUBLISHED IN: Frontiers in Oncology, Frontiers in Genetics,  
Frontiers in Immunology, Frontiers in Medicine and  
Frontiers in Artificial Intelligence





# frontiers

## Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88976-413-6

DOI 10.3389/978-2-88976-413-6

## 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](http://frontiersin.org/about/contact)



# ADVANCES IN MATHEMATICAL AND COMPUTATIONAL ONCOLOGY, VOLUME II

Topic Editors:

**George Bebis**, University of Nevada, Reno, United States

**Max A. Alekseyev**, George Washington University, United States

**Heyrim Cho**, University of California, Riverside, United States

**Jana Gevertz**, The College of New Jersey, United States

**David A. Hormuth, II**, The University of Texas at Austin, United States

**Maria Rodriguez Martinez**, IBM Research - Zurich, Switzerland

**Citation:** Bebis, G., Alekseyev, M. A., Cho, H., Gevertz, J., Hormuth, D. A.,  
Martinez, M. R., eds. (2022). Advances in Mathematical and Computational  
Oncology, Volume II. Lausanne: Frontiers Media SA.  
doi: 10.3389/978-2-88976-413-6

# Table of Contents

- 04 Integrating Tumor Stroma Biomarkers With Clinical Indicators for Colon Cancer Survival Stratification**  
Yong Chen, Wenlong Wang, Bo Jiang, Lei Yao, Fada Xia and Xinying Li
- 24 Identification of a Ubiquitination-Related Gene Risk Model for Predicting Survival in Patients With Pancreatic Cancer**  
Hao Zuo, LuoJun Chen, Na Li and Qibin Song
- 36 Development and Validation of a Combined Model for Preoperative Prediction of Lymph Node Metastasis in Peripheral Lung Adenocarcinoma**  
Qi Li, Xiao-qun He, Xiao Fan, Chao-nan Zhu, Jun-wei Lv and Tian-you Luo
- 46 Artificial Intelligence Can Cut Costs While Maintaining Accuracy in Colorectal Cancer Genotyping**  
Alec J. Kacew, Garth W. Strohbehn, Loren Saulsberry, Neda Laiteerapong, Nicole A. Cipriani, Jakob N. Kather and Alexander T. Pearson
- 53 The Immune Subtypes and Landscape of Gastric Cancer and to Predict Based on the Whole-Slide Images Using Deep Learning**  
Yan Chen, Zepang Sun, Wanlan Chen, Changyan Liu, Ruoyang Chai, Jingjing Ding, Wen Liu, Xianzhen Feng, Jun Zhou, Xiaoyi Shen, Shan Huang and Zhongqing Xu
- 67 Immunogenomic Analyses of the Prognostic Predictive Model for Patients With Renal Cancer**  
Tao Feng, Jiahui Zhao, Dechao Wei, Pengju Guo, Xiaobing Yang, Qiankun Li, Zhou Fang, Ziheng Wei, Mingchuan Li, Yongguang Jiang and Yong Luo
- 89 Mathematical Modeling of Locoregional Recurrence Caused by Premalignant Lesions Formed Before Initial Treatment**  
Mitsuaki Takaki and Hiroshi Haeno
- 98 Re-Identification of Patient Subgroups in Uveal Melanoma**  
Thi Hai Yen Nguyen, Tin Nguyen, Quang-Huy Nguyen and Duc-Hau Le
- 107 SMRT: Randomized Data Transformation for Cancer Subtyping and Big Data Analysis**  
Hung Nguyen, Duc Tran, Bang Tran, Monikrishna Roy, Adam Cassell, Sergiu Dascalu, Sorin Draghici and Tin Nguyen
- 118 Gene Co-Expression in Breast Cancer: A Matter of Distance**  
Alfredo González-Espinoza, Jose Zamora-Fuentes, Enrique Hernández-Lemus and Jesús Espinal-Enríquez
- 131 Data-Driven Discovery of Mathematical and Physical Relations in Oncology Data Using Human-Understandable Machine Learning**  
Daria Kurz, Carlos Salort Sánchez and Cristian Axenie
- 150 Discriminative Localized Sparse Approximations for Mass Characterization in Mammograms**  
Sokratis Makrogiannis, Keni Zheng and Chelsea Harris
- 164 From Fitting the Average to Fitting the Individual: A Cautionary Tale for Mathematical Modelers**  
Michael C. Luo, Elpiniki Nikolopoulou and Jana L. Gevertz



# Integrating Tumor Stroma Biomarkers With Clinical Indicators for Colon Cancer Survival Stratification

Yong Chen, Wenlong Wang, Bo Jiang, Lei Yao, Fada Xia and Xinying Li\*

Department of General Surgery, Xiangya Hospital, Central South University, Changsha, China

## OPEN ACCESS

### Edited by:

George Bebis,  
University of Nevada, Reno,  
United States

### Reviewed by:

Robert Sebra,  
Icahn School of Medicine at Mount  
Sinai, United States  
Simon Barry,  
AstraZeneca, United Kingdom

### \*Correspondence:

Xinying Li  
lixinyingcn@126.com

### Specialty section:

This article was submitted to  
Precision Medicine,  
a section of the journal  
Frontiers in Medicine

**Received:** 18 July 2020

**Accepted:** 12 November 2020

**Published:** 07 December 2020

### Citation:

Chen Y, Wang W, Jiang B, Yao L, Xia F  
and Li X (2020) Integrating Tumor  
Stroma Biomarkers With Clinical  
Indicators for Colon Cancer Survival  
Stratification. *Front. Med.* 7:584747.  
doi: 10.3389/fmed.2020.584747

The tumor stroma plays an important role in tumor progression and chemotherapeutic resistance; however, its role in colon cancer (CC) survival prognosis remains to be investigated. Here, we identified tumor stroma biomarkers and evaluated their role in CC prognosis stratification. Four independent datasets containing a total of 1,313 patients were included in this study and were divided into training and testing sets. Stromal scores calculated using the estimation of stromal and immune cells in malignant tumors using expression data (ESTIMATE) algorithm were used to assess the tumor stroma level. Kaplan-Meier curves and the log-rank test were used to identify relationships between stromal score and prognosis. Tumor stroma biomarkers were identified by cross-validation of multiple datasets and bioinformatics methods. Cox proportional hazards regression models were constructed using four prognosis factors (age, tumor stage, the ESTIMATE stromal score, and the biomarker stromal score) in different combinations for prognosis prediction and compared. Patients with high stromal scores had a lower overall survival rate ( $p = 0.00016$ ), higher risk of recurrence ( $p < 0.0001$ ), and higher probability of chemotherapeutic resistance ( $p < 0.0001$ ) than those with low scores. We identified 16 tumor stroma biomarkers and generated a new prognosis indicator termed the biomarker stromal score (ranging from 0 to 16) based on their expression levels. Its addition to an age/tumor stage-based model significantly improved prognosis prediction accuracy. In conclusion, the tumor stromal score is significantly negatively associated with CC survival prognosis, and the new tumor stroma indicator can improve CC prognosis stratification.

**Keywords:** colon cancer, microenvironment, tumor stroma, immune cells, prognosis stratification

## INTRODUCTION

Colorectal cancer is the world's fourth most deadly cancer, accounting for ~10% of global cancer-related deaths each year (1, 2). Risk stratification and prognosis prediction of patients with colorectal cancer mainly rely on the tumor, lymph node, metastasis (TNM) classification system of the American Joint Committee on Cancer (3). However, this system provides useful but incomplete prognostic information, and additional clinicopathological and molecular characteristics should be considered to improve its prediction accuracy, such as mutation status, immune score, stromal components, and the presence of microsatellite instability (4–8).

Malignant solid tumors like colon cancer (CC) consist of not only tumor cells but also the tumor microenvironment (TME), which includes infiltrating immune cells, tumor stroma components,

and other normal epithelial cells (9). The tumor stroma and immune cells are increasingly thought to play important roles in CC progression and drug resistance (10, 11); however, the specific molecules involved and their mechanisms remain unclear, particularly for the tumor stroma. Pagès et al. (12) developed a new indicator, termed an “immunoscore,” which could effectively predict CC prognosis. It measures the density of CD3+ and CD8+ T-cell effectors within the tumor and its invasive margins to assess the levels of infiltrating immune cells. We hypothesized that adding an additional indicator based on the tumor stroma into the current classification system would further improve CC prognosis stratification.

Estimation of stromal and immune cells in malignant tumors using expression data (ESTIMATE) is a newly developed algorithm that assesses the levels of the tumor stroma and infiltrating immune cells using the transcriptional profiles of cancer tissues, by detecting the specific gene expression signatures of stromal and immune cells (13). This method has been applied to several cancers and has proved helpful for prognosis stratification (14, 15); however, it has not been applied to CC. Based on this method, the purpose of this study was to develop a new specific tumor stroma indicator to improve the risk stratification and prognosis prediction of patients with CC.

## MATERIALS AND METHODS

### Data Preparation

Normalized gene expression matrices and matched clinical information for GSE39582 and GSE17538, which contain 556 and 232 patients with CC, respectively, were downloaded from the Gene Expression Omnibus database. These microarray datasets, both acquired on Affymetrix Human Genome U133 Plus 2.0 Arrays, were combined for further analysis by correcting batch effects using the ComBat method implemented in the “SVA” package. Normalized mRNA expression and protein/phosphorylation expression matrices and matched clinical information from a dataset containing 106 patients with CC were obtained from the cBioPortal database (<http://www.cbioportal.org/>). TCGA project-COAD level 3 gene expression and micro (mi)RNA expression matrices, normalized by fragments per kilobase of exon per million reads mapped fragments (FPKM) and reads per million mapped reads (RPM), respectively, and a corresponding DNA methylation beta matrix were downloaded using the R package “TCGAbiolinks.” Inclusion criteria for patients were: (1) complete information regarding survival status and time; and (2) a follow-up time  $\geq 1$  month. Human reference genome annotation data (version: GRCh38.p13) and human binding motif data (version: GRCh38.p13) were downloaded from the Ensembl BioMart database (<https://useast.ensembl.org/index.html>) to predict transcription factors (TFs) regulating target genes.

### Correlations Between the ESTIMATE Stromal Score and Clinical Prognosis

The ESTIMATE algorithm was applied to calculate the stromal score of each CC patient using gene expression profiles. To identify the most significant stromal score threshold for

patient grouping, we used the method “maximally selected rank statistics” in the R package “maxstat” (16). Patients were divided into high and low stromal score groups according to the threshold value. Then, Kaplan-Meier (KM) analysis and a log-rank test were used to identify survival differences between the high and low stromal score groups in the training set, and validation was performed using the testing sets. Moreover, we performed Wilcoxon rank-sum and/or Kruskal-Wallis tests to identify relationships between the ESTIMATE stromal score and clinical features, including T, N, and M pathological results and the tumor stage.

### Correlations Between the ESTIMATE Stromal Score and Chemotherapy Resistance

A subset of 540 patients from GSE39582 with information regarding adjuvant chemotherapy was divided into three groups based on their treatment regimens and stromal scores: patients who were not treated with chemotherapy, patients with low stromal scores who were treated with chemotherapy, and patients with high stromal scores who were treated with chemotherapy. Then, we performed Wilcoxon rank-sum and Kruskal-Wallis tests to identify differences in the stromal score distribution between the three groups. KM analyses and log-rank tests were used to identify survival differences.

### Identification of Specific Differentially Expressed Genes (SDEGs)

To identify SDEGs in the high stromal score group vs. the low stromal score group, we analyzed differences between the groups in three independent datasets (the training set and the testing sets). The R package “limma” was used to identify differentially expressed genes (DEGs), based on thresholds of log fold change  $> 1$  and adjusted  $p$  ( $\text{adj}P$ )  $< 0.05$ . Then, we performed overlap analysis of the top 30 DEGs from each dataset to identify SDEGs that were significantly increased in the high stromal score group compared to the low stromal score group.

### Identification of Clinically Significant Modules

We conducted weighted co-expression network analysis (WGCNA) to identify modules most relevant to the tumor stroma and characterize the correlation patterns among module genes using the R package “WGCNA.” The mRNA weighted co-expression network was constructed using the mRNA expression profile in the training set and the top 10,000 variable genes measured by median absolute deviation. The “WGCNA” package function `pickSoftThreshold` was used to select an appropriate soft-thresholding power value, which was applied to construct a scale-free topology matrix. Parameters used to construct the co-expression gene modules were as follows: a `deepSplit` of 2, a `minModuleSize` of 30, a `maxBlockSize` of 20,000, and merging of highly similar modules when the module eigengene height in the clustering was  $< 0.25$ . Finally, we related the modules to clinical features to identify the module whose genes were most relevant to the stromal score.

## Module Preservation Analysis and Functional Annotation

To examine the stability of the identified stroma-related module, we performed module preservation analysis using the function `modulePreservation` (17) in the “WGCNA” package and the two mRNA expression profiles in the testing sets, with the parameter `nPermutation` set to 200. The preservation  $Z$ -summary ( $Z$ ) was used to estimate module preservation between different datasets, with  $Z > 10$ ,  $5 < Z \leq 10$ , and  $Z \leq 5$  indicating high, median, and low preservation, respectively. Then, to explore the biological functions of the genes in the stroma-related module, we performed gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses using the R package “clusterProfiler.”  $\text{Adj}P < 0.01$  was considered statistically significant.

## Hub Gene Identification

Hub genes within modules are genes that have a high degree of connectivity in the associated interaction network and play important roles in related clinical features. To identify hub genes in the stroma-related module, we first constructed a protein-protein interaction (PPI) network containing all genes in the module using the online database STRING (<https://string-db.org/>). Then, we imported the PPI network into Cytoscape (version 3.71) to calculate the degree of each node. Candidate hub genes had degrees  $> 90$ . We also performed overlap analysis between candidate hub genes and the three DEG sets to further filter the hub genes.

## Biomarker Identification

In this study, tumor stroma biomarkers were defined as closely related to the stromal score and significantly negatively correlated with survival prognosis. All identified SDEGs and hub genes were initially selected as candidate biomarkers. We first conducted  $t$ -tests to further validate the expression differences of these genes between the high and low stromal score groups at the protein level using the protein/phosphorylation expression matrix. Protein features containing  $> 30\%$  missing values were excluded prior to the  $t$ -test. The criterion for filtering was  $p < 0.05$ . Next, we conducted Pearson correlation analyses using the mRNA expression profile from the training set to determine the relationships between candidate biomarkers and the stromal score. The criteria for screening were  $p < 0.01$  and  $r > 0.5$ . The results were verified by the same method using the testing sets.

## Correlations Between Biomarkers and Prognosis

We divided patients in the training set into high and low expression groups according to the optimal cutoff of each biomarker’s mRNA expression, as determined by the R package “maxstat.” Then, we performed KM analysis and log-rank tests to determine survival differences between the two groups based on each biomarker. Statistical significance was defined as  $p < 0.05$ . We validated the results in the same manner using the testing sets. Biomarkers that produced statistically significant differences in both the training set and the testing sets were retained for further analysis.

## Construction of the Prognosis Model

In addition to the stromal score calculated by ESTIMATE, we created another new indicator for risk stratification, termed the biomarker stromal score, a cumulative measure of the number of biomarkers that were significantly higher in each patient. We divided the patients into low-, median-, and high-risk groups based on their biomarker stromal scores using the R package “maxstat,” then performed KM analysis and log-rank tests to determine survival differences between the three groups using survival information from all patients in the training set and the testing sets. Moreover, to estimate and compare the stratification ability of each prognosis feature, we performed time-dependent receiver operating characteristic (ROC; 3-year and 5-year) analysis with  $1,000 \times$  bootstrap resampling for each feature (age, pathology T, pathology N, pathology M, tumor stage, ESTIMATE stromal score, and biomarker stromal score) separately. Finally, we performed multivariate regression analyses to construct three multivariable Cox proportional hazards models using the prognosis features age, tumor stage, ESTIMATE stromal score, and biomarker stromal score in different combinations. Two evaluation methods [time-dependent ROC curves (area under the curve (AUC) and the concordance index (C-index)] were used to measure the prediction accuracy of each prognosis model with  $1,000 \times$  bootstrap resampling, and their performance was compared using the  $p$ -value of the likelihood ratio. In addition, to use the prediction model clinically, a nomogram was developed to predict the 1–5-year survival rates of patients with CC, and calibration curves were used to test its performance.

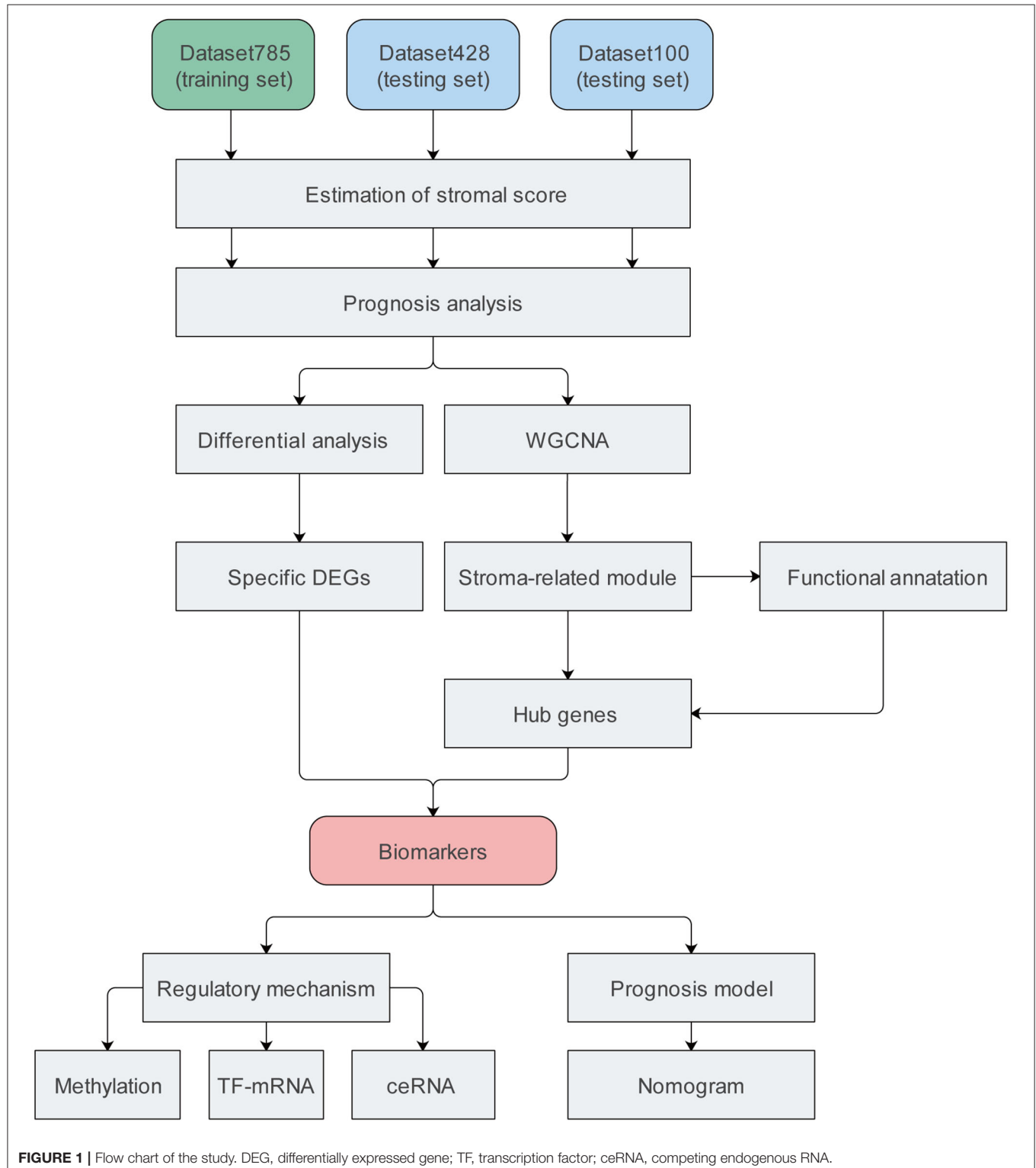
## Construction of the Direct Regulatory Network

To explore potential regulatory mechanisms of the biomarkers, we examined their methylation status, TFs, and competing endogenous RNA (ceRNA) networks. We analyzed methylation differences between the high and low stromal score groups using the R packages “ChAMP” and a methylation beta matrix containing 281 patients to detect CpG sites with significant changes in methylation. The thresholds for statistical significance were  $\text{adj}P < 0.05$  and  $\text{deltaBeta} < -0.05$ .

We used the human reference genome annotation dataset and human binding motif dataset, which uses the position weight matrix method to predict potential TF binding sites, to predict TFs that interact with target gene promoters. Binding sites with scores  $< 0$  were filtered out of the binding motif dataset, and the promoter region of a gene was defined as the region between 1,000 bp upstream and 200 bp downstream of the transcriptional start site in the genome annotation dataset. We further filtered the TFs according to their differential expression in the high and low stromal score groups using the protein/phosphorylation expression matrix. Moreover, to improve the confidence of the TF assignments, we performed Pearson correlation analysis to identify associations between TFs and target genes using a subset from dataset100 containing 96 patients with both mRNA expression and protein/phosphorylation expression profiles, with thresholds of  $p < 0.05$  and  $r > 0.3$ .

Finally, given the positive regulatory associations between long non-coding (lnc)RNAs and mRNAs in ceRNA networks, we first performed Pearson correlation analysis to examine associations between the expression of lncRNAs and the

biomarkers using a dataset containing 453 patients with both lncRNA and mRNA expression profiles. The criteria for filtering lncRNAs were  $r > 0.65$  and  $p < 0.01$ . We predicted direct miRNA-mRNA interactions using the online





database StarBase (<http://starbase.sysu.edu.cn/>). For inclusion, interactions needed to be validated at least once by cross-linking immunoprecipitation (CLIP), and predicted by at least three of the PITA, RNA22, miRmap, miR0T, miRanda, PicTar, and TargetScan databases. Direct lncRNA-miRNA interactions were predicted using the starBase miRanda tool. For inclusion, interactions needed to be validated at least once by CLIP. Then, we merged the lncRNA-miRNA and miRNA-mRNA networks to generate the direct lncRNA-miRNA-mRNA regulatory network. Finally, we performed KM analysis and log-rank tests to identify survival differences based on the expression levels of lncRNAs and miRNAs in the ceRNA network, using lncRNA and miRNA expression matrices containing 428 and 413 patients, respectively. Statistically significant ( $p < 0.05$ ) lncRNAs and miRNAs were retained. The network was constructed and visualized using Cytoscape.

## Statistical Analyses

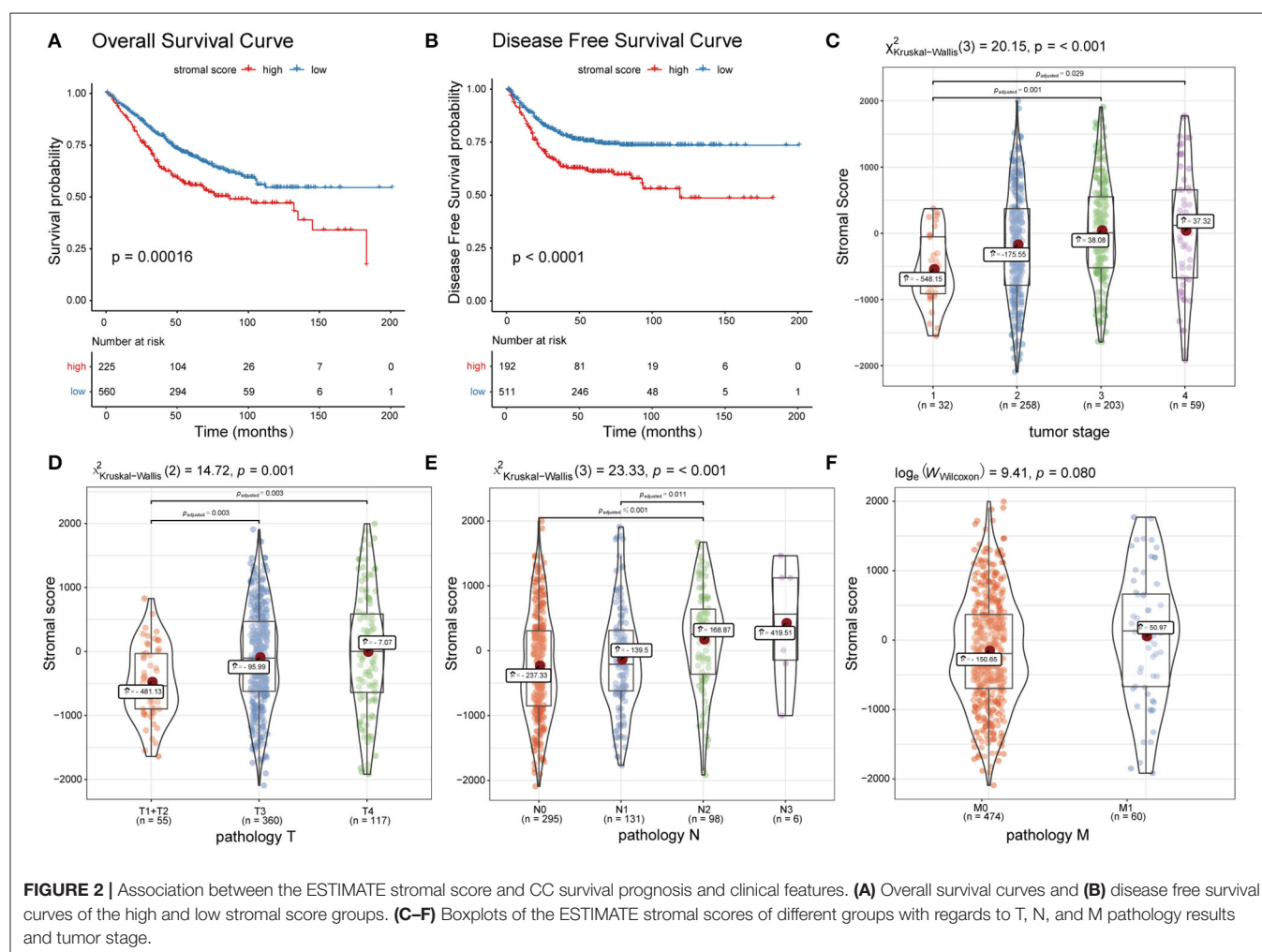
All statistical analyses in this study were completed in R version 3.6.3 (<https://www.r-project.org/>). Appropriate R packages were used for different analyses. For these, specific parameters used

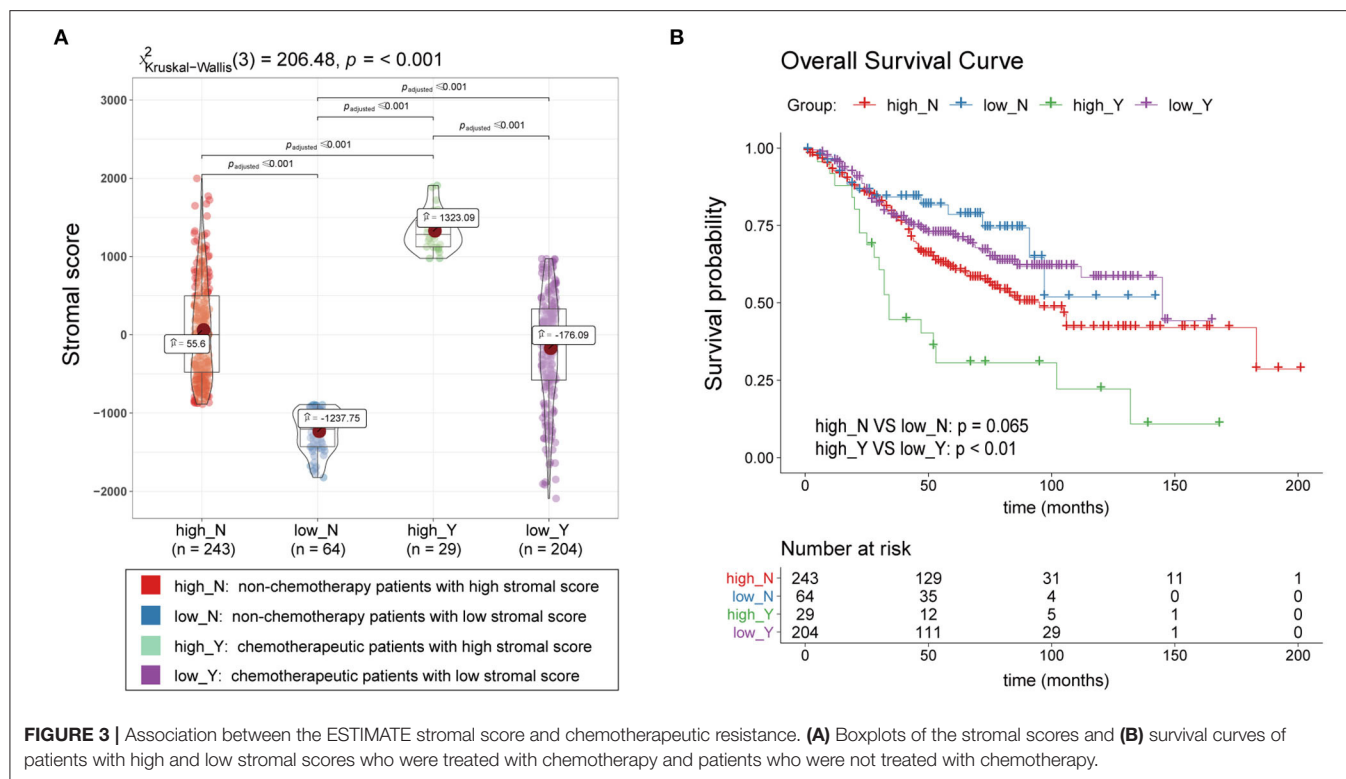
are listed in their respective sections, while default parameters are not listed. The threshold of statistical significance varied among different statistical analyses but was at least  $p < 0.05$ .

## RESULTS

### Data Collection

We included four datasets containing nine expression matrices and a total of 1,313 patients with primary CC. Different expression matrices in the same dataset shared the same patients; however, the number of patients in the matrices were not necessarily the same. The GSE39582 and GSE17538 datasets were combined into a training set with an mRNA expression matrix containing 785 patients, termed dataset785. This set was mainly used to mine data in our study. A dataset containing 100 patients was obtained from the cBioPortal database was termed dataset100. It consisted of mRNA and protein/phosphorylation expression matrices. Another dataset, obtained from The Cancer Genome Atlas (TCGA), containing 428 patients and mRNA, lncRNA, miRNA, and methylation expression matrices, was termed dataset428. Dataset100 and dataset428 were defined





**FIGURE 3 |** Association between the ESTIMATE stromal score and chemotherapeutic resistance. **(A)** Boxplots of the stromal scores and **(B)** survival curves of patients with high and low stromal scores who were treated with chemotherapy and patients who were not treated with chemotherapy.

as testing datasets mainly used for verification and molecular mechanism analysis. Details regarding the datasets are provided in **Supplementary Table 1**, and the complete workflow of the study is displayed in **Figure 1**.

## Correlations Between the ESTIMATE Stromal Score and Clinical Prognosis

Patients in dataset785 were divided into high and low stromal score groups based on the determined optimal cutoff. KM analysis and a log-rank test revealed that patients with low scores had significantly better overall survival (OS;  $p = 0.00016$ ; **Figure 2A**) and disease-free survival (DFS;  $p < 0.0001$ ; **Figure 2B**) than patients with high scores. These results were validated using dataset100 and/or dataset428 (**Supplementary Figures 1A–C**). Wilcoxon rank-sum and Kruskal-Wallis tests identified statistically significant relationships between the stromal score and clinical features, including the T, N, M pathology results and tumor stage (**Figures 2C–F**). The results were verified using dataset100 (**Supplementary Figures 2A–D**). Therefore, these results indicate that tumor stroma is closely associated with tumor progression and survival prognosis.

## Correlations Between the ESTIMATE Stromal Score and Chemotherapy Resistance

A total of 540 patients with information regarding adjuvant chemotherapy were included and divided into four groups based on the ESTIMATE stromal score and adjuvant chemotherapy information. Wilcoxon rank-sum and Kruskal-Wallis tests

showed the distribution of the ESTIMATE stromal score between groups were significantly different and the details were shown in **Figure 3A**. Patients treated with chemotherapy who had high stromal scores had a lower OS rate than those with low stromal scores and patients who were not treated with chemotherapy ( $p < 0.01$ ); however, there was no significant difference in survival between chemotherapy patients who had low stromal scores and patients not treated with chemotherapy (**Figure 3B**). Therefore, our findings indicate that patients treated with chemotherapy who have high stromal scores are more vulnerable to the development of chemotherapeutic tolerance and have a poor survival prognosis.

## SDEG Identification

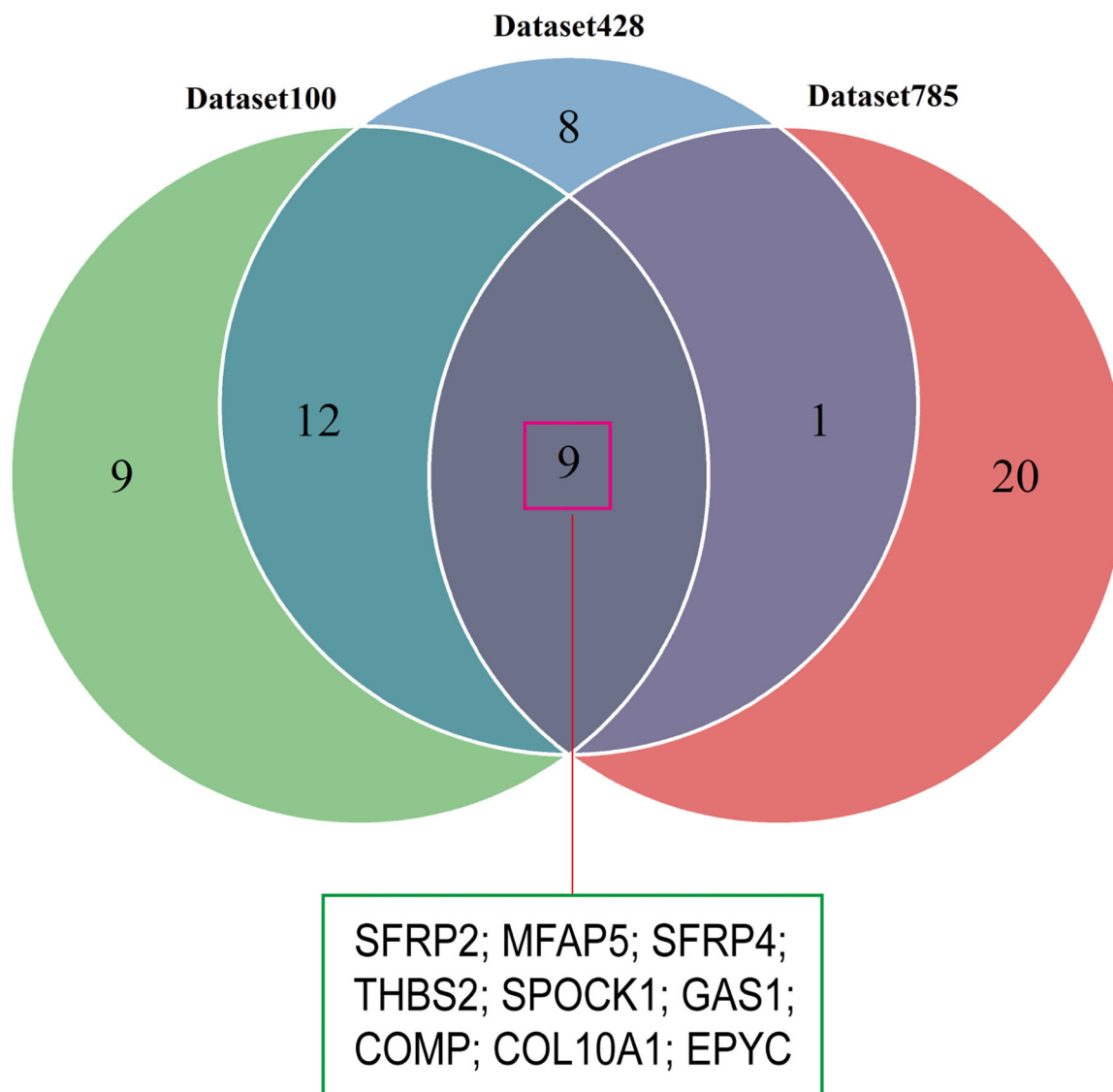
We conducted differential analyses between the high and low stromal score groups using the mRNA expression profiles in dataset785, dataset100, and dataset428, and identified 246, 501, and 2,313 DEGs, respectively (**Supplementary Figures 3A–C**). Overlap analysis of the top 30 DEGs from each DEG set (based on the log fold change) produced nine SDEGs (**Figure 4** and **Table 1**). Notably, among these nine SDEGs, gene SFRP2 had the biggest logFC. These SDEGs may play an important role in tumor stroma-induced promotion of tumor progression.

## Identification of Stroma-Related Modules Through WGCNA

To construct the mRNA co-expression network, we selected 6 as the appropriate sort-thresholding power value, which generated 21 mRNA modules (**Supplementary Figures 4A–C**). Association analysis between the modules and clinical features revealed that



## Overlap analysis for SDEGs



**FIGURE 4 |** Identification of SDEGs. The Venn diagram shows the overlap between the top 30 DEGs from the three datasets.

the yellow module (containing 1,173 genes) was most related to the stromal score ( $r = 0.929$ ;  $p = 0$ ; **Figures 5A,B**). This indicates that genes in the yellow module, particularly its hub genes, may play important roles in the tumor stroma-induced promotion of tumor progression and drug resistance.

### Module Preservation Analysis and Functional Annotation

To examine the stability of the stroma-related module (yellow) in the training set identified above, we performed module preservation analyses using the two testing sets (dataset428 and

dataset100). As shown in **Figures 5C,D**, the horizontal dashed lines indicate the Zsummary (Z) thresholds for strong evidence of conservation ( $>10$ ) and for low to moderate evidence of conservation ( $>2$ ), so we can see the yellow module had good performance with  $Z > 10$  in both dataset428 and dataset100, which means that genes in the yellow module have high consistency in the training set and testing sets. Moreover, to determine the functional involvement of the tumor stroma, the 1,173 genes in the yellow module were subjected to GO and KEGG pathway enrichment analyses. As shown in **Figures 6A,B**, enriched biological processes (BPs), molecular functions (MFs),

**TABLE 1** | Differential analysis statistics of the nine SDEGs.

Gene	Dataset785		Dataset428		Dataset100	
	logFC	adj. P.Val	logFC	adj. P.Val	logFC	adj. P.Val
SFRP2	3.33	5.2E-114	4.14	7.2E-34	3.25	2.2E-06
COL10A1	2.50	2.3E-85	3.47	6.0E-28	2.40	4.6E-06
SFRP4	2.37	1.3E-93	3.32	1.2E-31	2.06	9.9E-05
THBS2	2.11	1.6E-82	2.88	4.0E-38	1.91	1.1E-06
SPOCK1	2.11	1.6E-104	2.98	2.3E-36	1.96	7.2E-07
MFAP5	1.98	5.4E-96	2.96	4.4E-37	1.89	5.5E-07
COMP	1.96	1.7E-79	3.66	1.4E-31	2.76	3.6E-05
EPYC	1.74	1.3E-49	3.65	5.0E-27	2.92	2.6E-05
GAS1	1.70	9.6E-115	3.13	7.1E-40	2.42	4.9E-07

SDEGs, specific differentially expressed genes; logFC, log fold change; adj.P.Val, adjusted p-value.

and cellular components were all significantly focused on the extracellular matrix (ECM). Most of genes in the six most statistically significant signaling pathways were overexpressed in the high stromal score group, and the three pathways “ECM-receptor interactions,” “focal adhesions,” and “PI3K-Akt signaling pathway” shared a significant number of genes (Figures 6C,D). This indicates that these biological processes and signaling pathways are closely related to tumor stroma function.

## Identification of Hub Genes

Hub genes were defined as genes with high degrees of connectivity in a PPI network of the yellow module. The interaction network contained 1,105 nodes and 8,927 edges, and node degrees ranged from 1 to 220 (Supplementary Figure 5). We selected 20 candidate hub genes based on a degree threshold of  $\geq 90$ . Overlap analysis of candidate hub genes from the three DEG sets identified 11 hub genes (Figure 7 and Table 2). This means that these 11 hub genes may have important impacts on the function of tumor stroma.

## Identification of Tumor Stroma Biomarkers

The nine SDEGs and 11 hub genes were considered candidate tumor stroma biomarkers and were all significantly related to the stromal score based on thresholds of  $r > 0.5$  and  $p < 0.01$  (Figure 8). The results were verified using the two testing sets (Supplementary Figures 6A,B). *t*-tests using the protein expression matrix revealed that 16/20 genes were significantly overexpressed in the high stromal score group compared to the low stromal score group (Supplementary Table 2). Four genes [epiphycan (EPYC), growth arrest specific 1 (GAS1), SPARC (osteonectin), cwcw- and kazal-like domains proteoglycan 1 (SPOCK1), and secreted phosphoprotein 1 (SPP1)] were not included in the protein expression matrix; therefore, we are unable to determine whether they display differential protein expression. Given their significant differential mRNA expression, these four genes were retained, resulting in 20 candidate biomarkers for further analysis. Among these candidate biomarkers, four collagen family members

(COL1A1, COL1A2, COL3A1, COL10A1) which are well-known to be closely related to the function of the stroma were contained, which also strongly supports the reliability of our results. However, due to the heterogeneity of the tumor microenvironment, even the same markers may play different roles in different cancers. Therefore, although some markers found in our study were closely related to the prognosis in CC, they may play different roles in other cancers.

## Correlations Between Tumor Stroma Biomarkers and Survival Prognosis

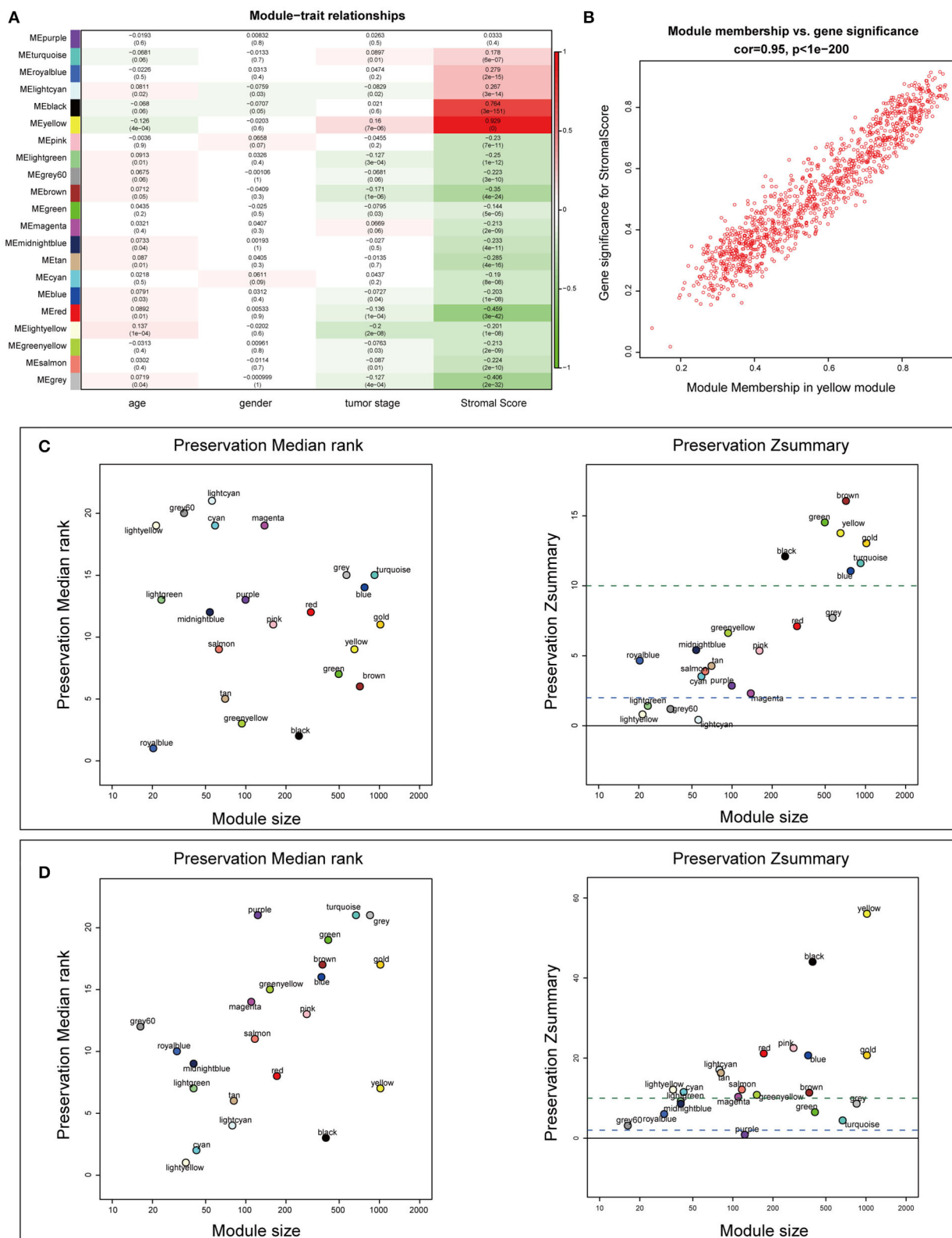
Survival analyses using the training set showed that when applied separately, each candidate biomarker generated a significant survival difference between the high and low score groups (Figure 9). However, only 14/20 were validated in the testing set based on a threshold of  $p < 0.05$  (Supplementary Figure 7). Two more were marginally significant [microfibril associated protein 5 ( $p = 0.056$ ) and thrombospondin 2 ( $p = 0.051$ )] were retained. Therefore, we finally identified 16 tumor stroma biomarkers in this study, this suggests that these genes are closely related to the tumor stromal function and survival prognosis of CC patients.

## Identification of a New Prognosis Indicator for Risk Stratification

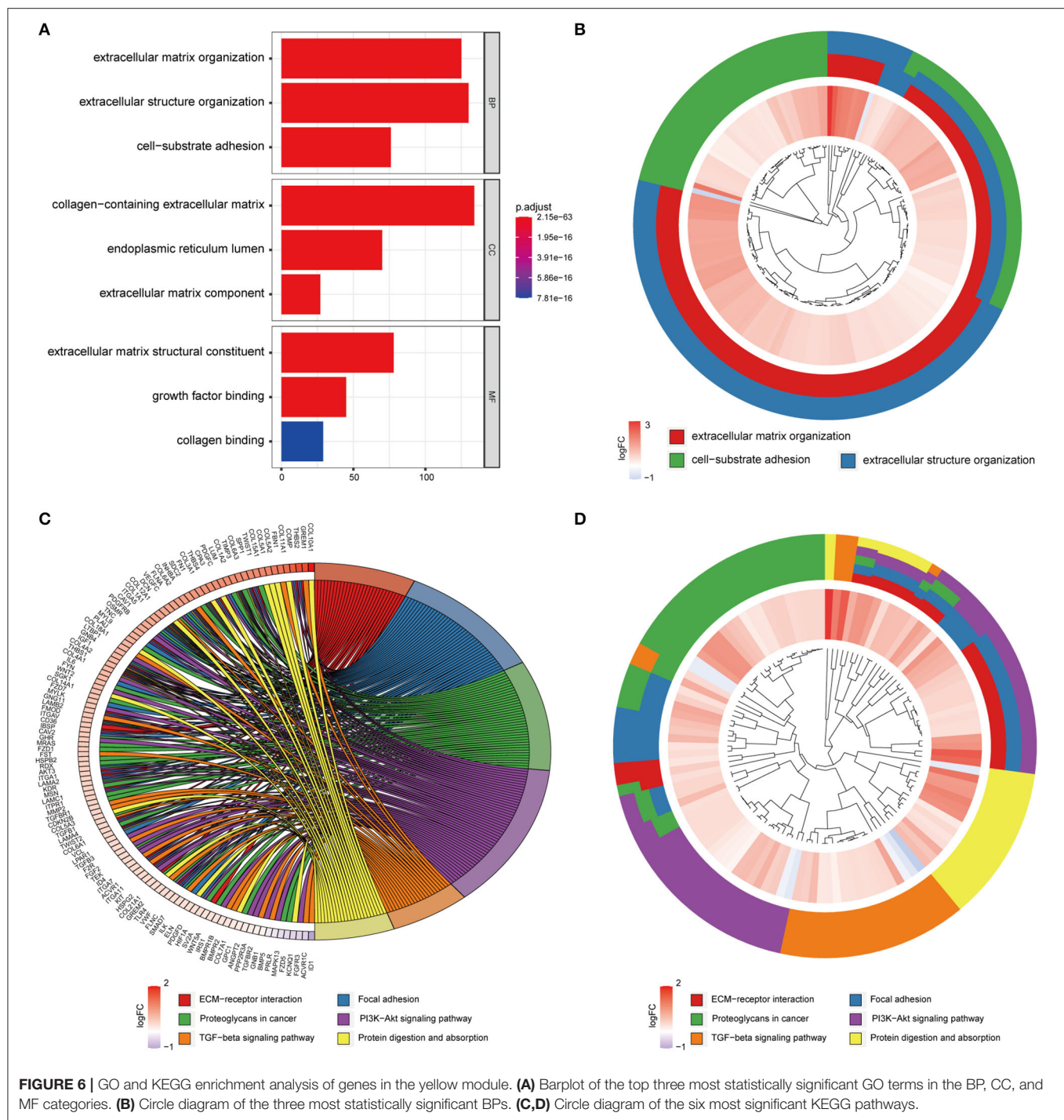
We next generated a new prognosis indicator based on the 16 tumor stroma biomarkers, termed the biomarker stromal score, which ranged from 0 to 16. We divided 1,313 patients (all with complete OS information; 787 also had complete DFS information) into three risk groups based on thresholds of 0, 1–9, and 10–16. OS and DFS analyses both revealed significant survival differences between the three risk groups (Figures 10A,B). Time-dependent ROC analyses showed that the ability of the biomarker stromal score to predict 3- and 5-year OS was superior to the features of patient age and ESTIMATE stromal score and had similar AUC values to the T, N, and M pathology results. The tumor stage had the best prediction accuracy (Figure 11A). Therefore, the biomarker stromal score is a comparably effective prognosis indicator to known clinical features.

## Construction of the Prognosis Model

We used four prognosis factors (age, tumor stage, the ESTIMATE stromal score, and the biomarker stromal score) in different combinations to construct prognosis models, and data on 1,295 patients with complete age, tumor stage, ESTIMATE stromal score, and biomarker stromal score information were used in multivariable regression analyses. Three prognosis models were constructed: model 1 included age and tumor stage; model 2 included age, tumor stage, and the ESTIMATE stromal score; and model 3 included age, tumor stage, and the biomarker stromal score. Time-dependent ROC (3- and 5-year) and C-index results revealed that model 3 had the best prediction accuracy (Figure 11B). The hazard ratios of each feature in model 3 are shown in Figure 12A. Model 3 risk scores ranged from 0.104 to 12.539, and patients were divided into five risk groups based on thresholds of  $\leq 0.556$ ,



**FIGURE 5 |** Identification of stroma-relevant mRNA modules and module preservation analysis. **(A)** Heatmap of module-trait relationships. **(B)** Scatter plot of correlations between gene module membership and gene significance in the yellow module. **(C,D)** Preservation medianRank and Zsummary graphs of the testing sets dataset100 and dataset428. Dashed blue and green lines show the thresholds  $Z = 2$  and  $Z = 10$ , respectively.



0.557–0.896, 0.897–1.27, 1.28–3.99, and >3.99. Significant survival differences were observed between the five groups (Figure 12B). In the nomogram plot, weighted scores calculated based on the age, tumor stage, and biomarker stromal score were used to predict the 1–5-year OS rate of patients with CC (Figure 12C). The calibration curve demonstrated good performance for the nomogram plot compared to an ideal model (Supplementary Figure 8). Therefore, our findings suggest that

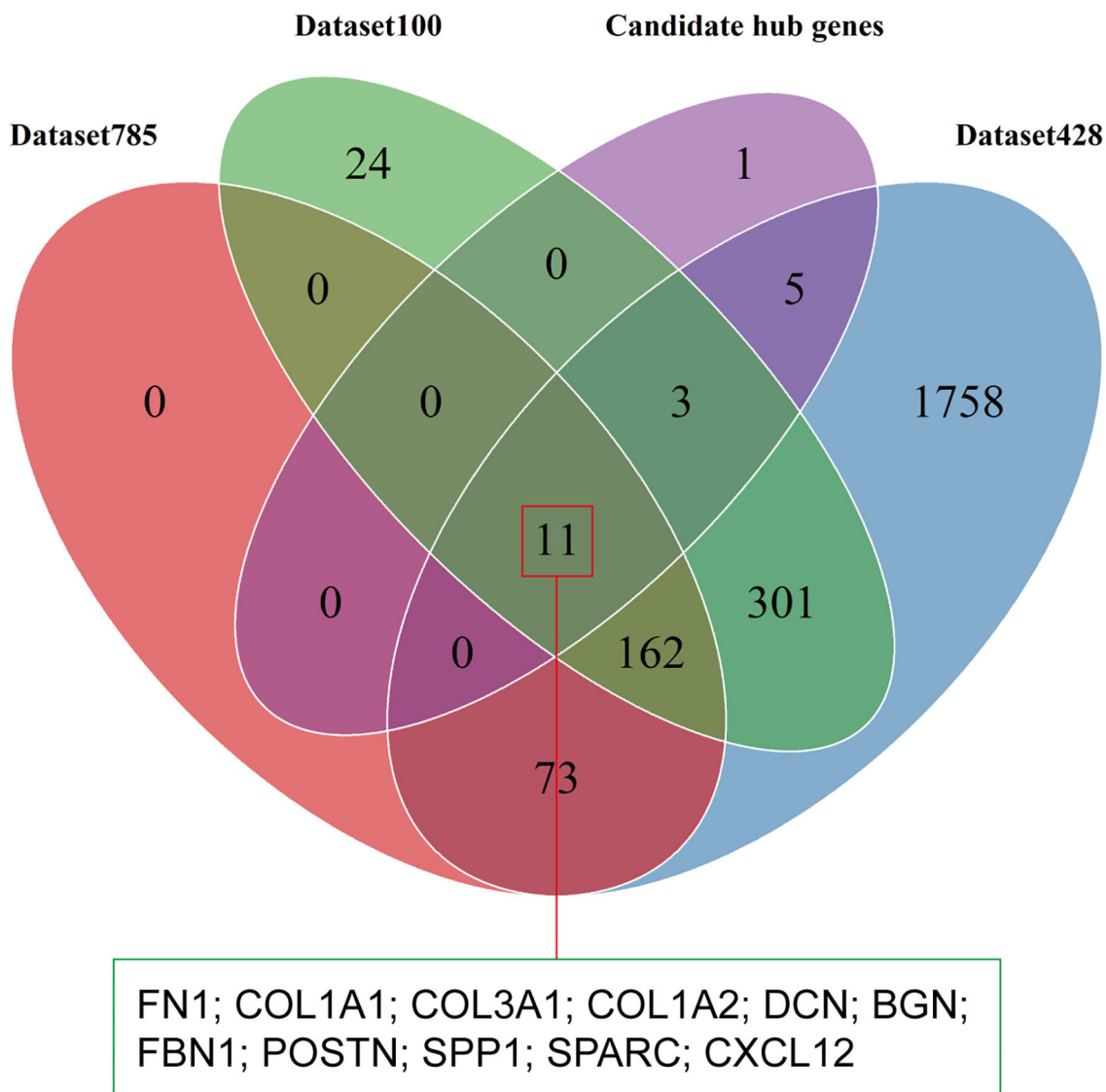
the biomarker stromal score can improve CC survival prognosis prediction accuracy.

## Construction of Biomarker Regulatory Networks

Differential analysis of the methylation beta matrix between the high and low stromal score groups revealed that 9/16 biomarkers contained at least one significantly demethylated CpG site (a total



## Overlap analysis for hub genes



**FIGURE 7** | Identification of hub genes. The Venn diagram shows the overlap between the 20 candidate hub genes with degrees of connectivity >90 and the three DEG sets.

of 66 CpG probes based on thresholds of  $\Delta\text{Beta} < -0.05$  and  $\text{adj}P < 0.05$ ; **Figure 13A** and **Supplementary Table 3**). Among them, secreted frizzled related protein 2 (*SFRP2*) had the most demethylated sites (29; with a mean  $\Delta\text{Beta}$  of  $-0.098$ ). This suggests that increased demethylation contributes to the high expression of the biomarkers in the high stromal score group. Besides, survival analyses showed that 15/66 probes could make significant survival differences based on the optimal cutoff of each probes ( $p < 0.05$ ; **Supplementary Table 3**). We next constructed a TF-mRNA regulatory network consisting of 4 TFs, 12 mRNAs,

and a total of 19 edges; the interaction details are shown in **Figure 13B** and **Supplementary Table 4**. Interestingly, RUNX family transcription factor 2 (*RUNX2*) could regulate 11/12 mRNAs in the network. We also constructed a ceRNA network consisting of 7 lncRNAs, 26 miRNAs, and 10 mRNAs, with a total of 53 edges (**Figure 13B** and **Supplementary Table 5**). Survival analyses based on the lncRNAs and miRNAs included in the networks are shown in **Supplementary Figures 9A,B**. In summary, these regulatory networks provide new insights into the mechanism of tumor stroma biomarkers of CC.

**TABLE 2 |** Statistics of the 11 hub genes in the yellow module.

Genes	Degrees of connectivity	Dataset785		Dataset428		Dataset100	
		logFC	adj. P.Val	logFC	adj. P.Val	logFC	adj. P.Val
FN1	220	1.24	4.1E-61	2.53	2.9E-31	1.74	1.2E-04
COL1A1	137	1.04	3.9E-64	2.29	1.2E-36	1.60	9.1E-04
COL3A1	119	1.29	2.4E-69	2.22	9.3E-41	1.33	1.6E-06
COL1A2	116	1.37	3.6E-62	2.19	1.4E-38	1.36	4.0E-05
DCN	105	1.07	2.1E-99	2.17	6.1E-36	1.22	5.8E-07
BGN	100	1.42	1.1E-87	2.16	2.5E-41	1.78	4.6E-06
FBN1	98	1.69	3.5E-113	2.21	1.4E-42	1.30	1.8E-07
POSTN	98	1.29	1.2E-76	2.56	1.2E-33	1.70	1.5E-07
SPP1	98	1.39	2.1E-42	3.03	2.8E-25	2.06	7.1E-05
SPARC	94	1.40	4.5E-91	1.83	2.2E-42	1.26	4.1E-08
CXCL12	91	1.43	2.7E-87	1.85	8.9E-40	1.54	4.7E-08

Degrees of connectivity, edge counts of genes in the PPI network calculated by Cytoscape; logFC, log fold change; adj.P.Val, adjusted p-value.

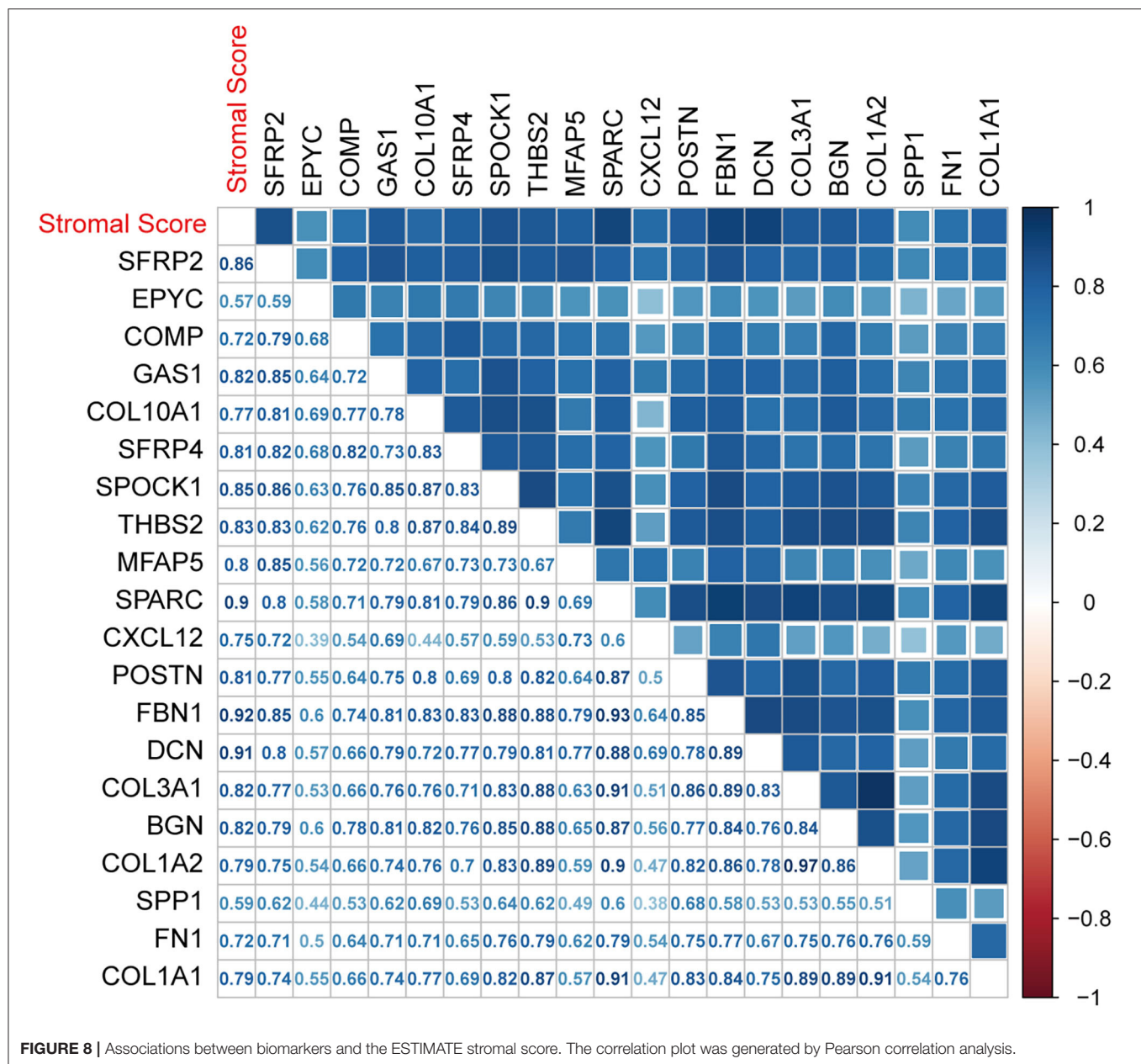
## DISCUSSION

Presently, risk stratification and prognosis prediction for patients with CC is mainly based on clinical and pathological characteristics (3, 4). In a recent study, Pagès et al. (12) demonstrated that a new indicator, the immunoscore, can effectively improve the accuracy of prognosis prediction for patients with CC. In this study, we have identified 16 tumor stroma biomarkers for primary CC and created a new indicator for risk stratification and prognosis prediction based on them. Our findings indicate that the tumor stroma is significantly negatively associated with survival prognosis, and that our new tumor stroma indicator could significantly improve the OS prediction accuracy of the currently used classification system.

It is well-known that interactions between cancer cells and the TME play important roles in tumor progression and therapeutic resistance (18, 19). While tumor cells have historically been the main therapeutic target of cancer treatment, different components of the TME, such as immune cells and angiogenic factors, have been recently targeted as well (20–23). However, these studies took limited notice of stromal components, and acquiring further insight into the interactions between cancer cells and the tumor stroma may provide novel biomarkers for stroma-targeted therapies as well as an increased understanding of drug resistance. Furthermore, there remains a lack of uniform criteria to assess tumor stroma condition. In this study, we assessed the CC tumor stroma by assigning scores based on stromal signatures generated using the ESTIMATE algorithm (13), and found that patients with high stromal scores had worse survival prognosis than patients with low stromal scores. Our findings in CC are consistent with results for several other cancers, such as gastric cancer, prostate cancer, and early-stage non-small cell lung cancer (14, 24–26). This indicates that the scores generated by this method may be a good tool to assess the CC tumor stroma condition and could be used as a prognosis factor for CC. In addition, we also found that the stroma score was significantly negatively correlated with the survival

prognosis of chemotherapy patients, which may be caused by the resistance of tumor stroma to chemotherapy. Previous studies (27, 28) showed tumor-stromal architecture has been associated with modulation of the response to anti-angiogenic therapy, and combined therapy of chemotherapy and anti-angiogenesis was more effective than monotherapy. Therefore, the role of tumor stroma on anti-angiogenic therapy deserves further study.

While the notion that therapies targeting cancer cells and the TME are equally important is widely accepted (29), specific biomarkers of the tumor stroma are still lacking, and the molecular mechanisms by which the stroma affects the tumor remain unclear, because of its heterogeneity and complexity (30, 31). In this study, to clarify the biological processes and signaling pathways affected by the tumor stroma in the promotion of CC progression and chemotherapy resistance, we conducted enrichment analysis on the tumor stroma-related genes. Interestingly, in GO and KEGG analyses, the most statistically significant terms and pathways were related to the ECM: the BP “ECM organization” (adjP = 5.22E-59) and the KEGG pathway “ECM-receptor interaction” (adjP = 6.47E-11), respectively. Tumor progression results in ECM component changes and remodeling. This makes the ECM more conducive to promoting the growth, survival, and migration of cancer cells (32), and can increase drug resistance in various ways. For instance, the buildup of a rigid ECM surrounding tumor cells creates a physical barrier that reduces the diffusion of therapeutic agents (33, 34). Cancer cells can also evade chemotherapy by strongly adhering to ECM proteins through a process known as cell adhesion-mediated drug resistance (35–37). Our findings suggest that the ECM plays an important role in the progression and therapeutic resistance of CC. Two proven key signaling pathways related to tumor progression and chemotherapy resistance, the phosphatidylinositol 3-kinase (PI3K)-AKT serine/threonine kinase 1 (AKT1) and transforming growth factor  $\beta$ 1 pathways (38–40), were also significantly enriched in our study. Most of the genes enriched in these two pathways were highly expressed in the high stromal score group.

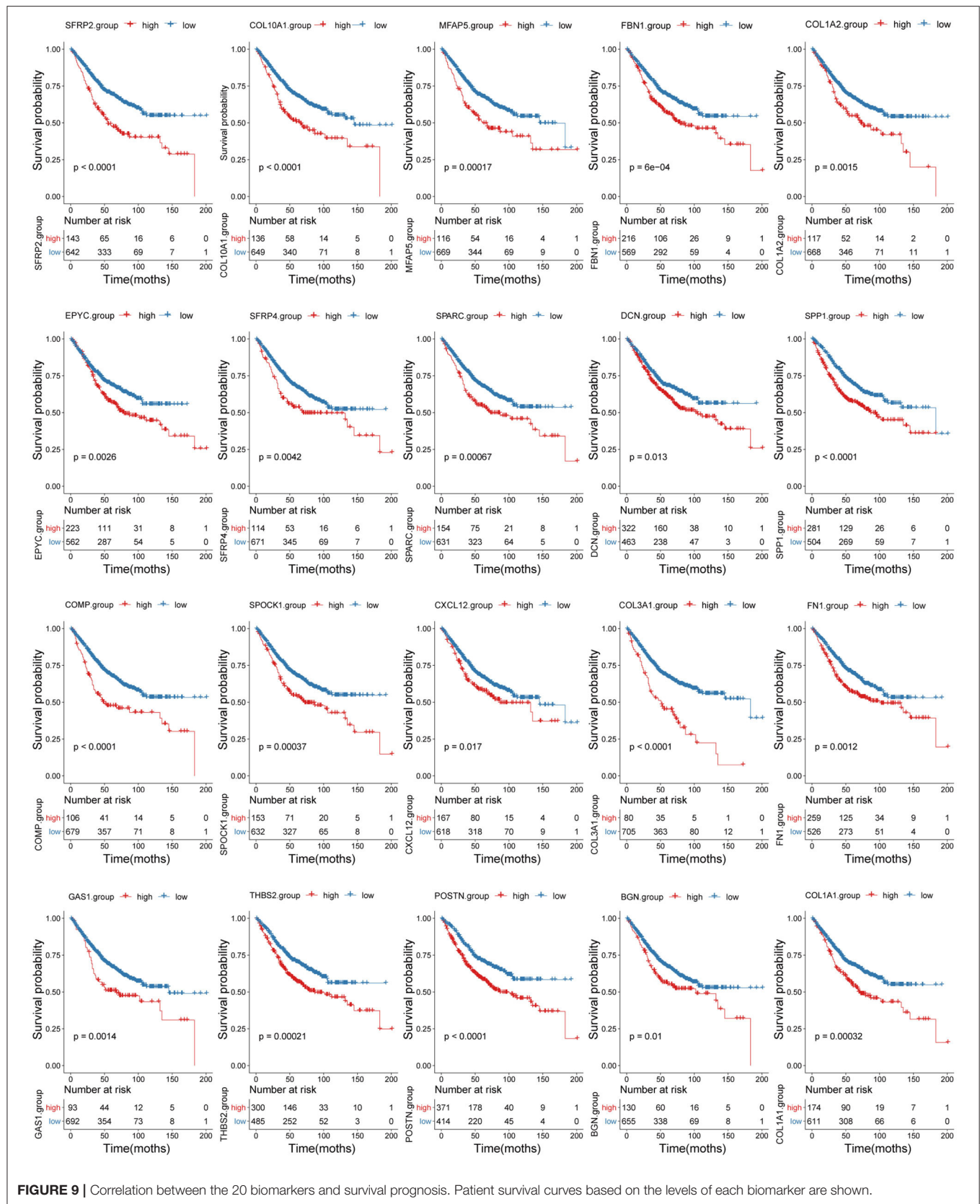


**FIGURE 8 |** Associations between biomarkers and the ESTIMATE stromal score. The correlation plot was generated by Pearson correlation analysis.

Our results therefore identify major biological processes and key signaling pathways related to the effects of the tumor stroma on CC, providing valuable clues for its treatment.

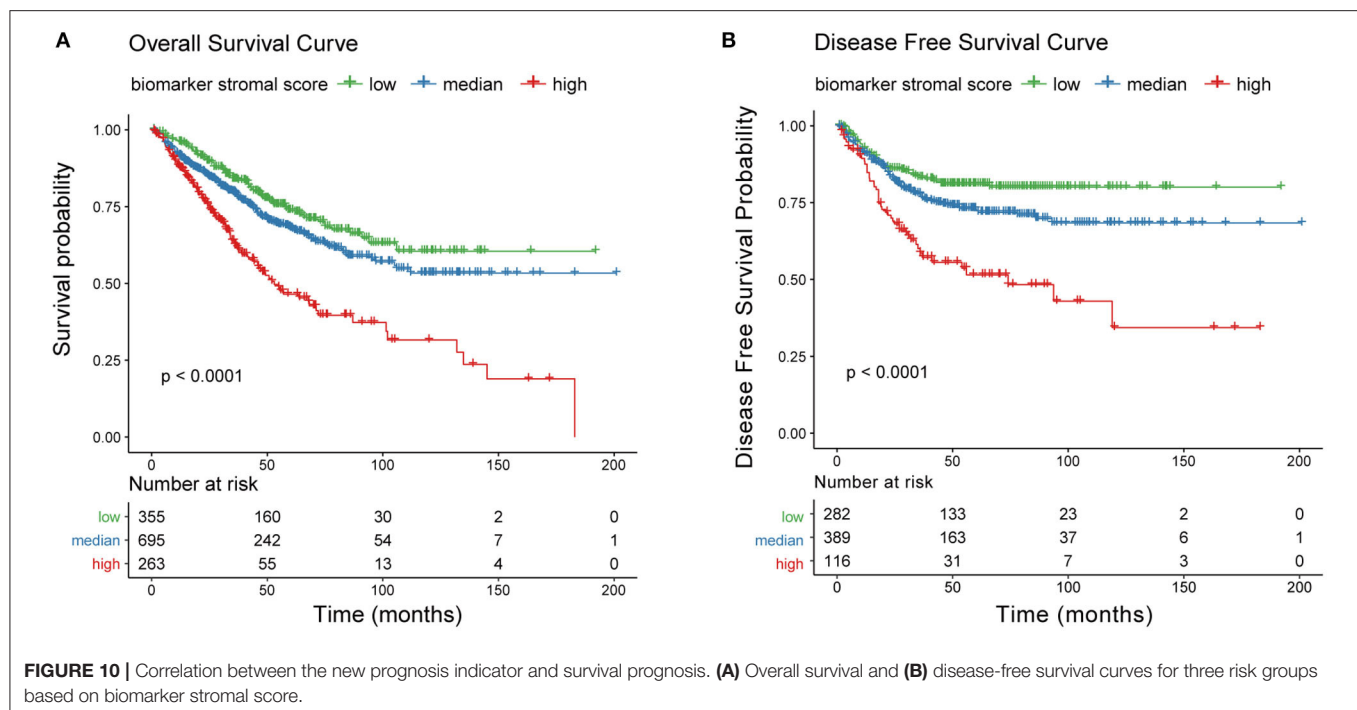
We identified 16 tumor stroma biomarkers that were closely related to the survival prognosis of patients with CC, and some have previously reported associations with CC tumor progression. For instance, fibronectin 1 (FN1) had the highest degree of connectivity in the PPI network. Xie et al. (41) showed that inhibiting FN1-SRC proto-oncogene, non-receptor tyrosine kinase/protein tyrosine kinase 2-guanosine triphosphatase (GTPase) signaling could inhibit CC metastasis, and Cai et al. (42) reported that FN1 depletion could inhibit colorectal carcinogenesis by suppressing proliferation, migration,

and invasion. The significant DEGs *SFRP2* and *SFRP4*, and especially *SFRP2*, had the most demethylated sites and the biggest logFC values in our study, and are involved in the biological processes of “extracellular matrix organization” and “extracellular structure organization.” Vincent et al. (43) reported that *SFRP2* and *SFRP4* are typically associated with poor prognosis concomitant with epithelial-to-mesenchymal transition (EMT). Nfonsam et al. (44) found that patients with CC that overexpress *SFRP4* have poor OS. In these patients, *SFRP4* levels were negatively correlated with the levels of the EMT suppressors claudin 4 (*CLDN4*), claudin 7 (*CLDN7*), tight junction protein 3 (*TJP3*), mucin 1, cell surface associated (*MUC1*), and cadherin 1 (*CDH1*). Klement et al. (45)



**FIGURE 9 |** Correlation between the 20 biomarkers and survival prognosis. Patient survival curves based on the levels of each biomarker are shown.





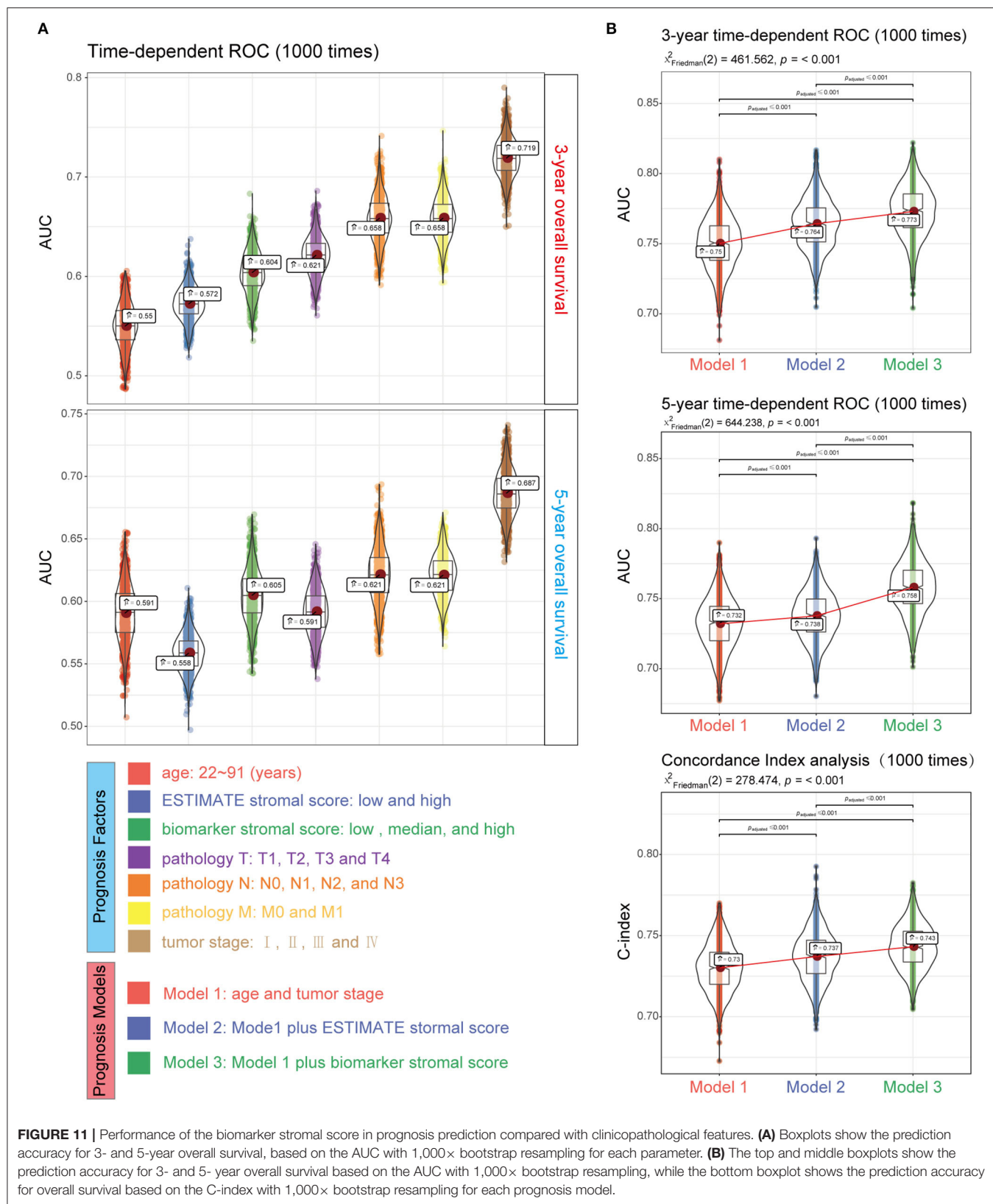
demonstrated that high SPP1 expression was associated with decreased OS by acting as an immune checkpoint to suppress T cell activation. C-X-C motif chemokine ligand 12 (CXCL12), secreted by fibroblasts, can promote the proliferation and invasion of CC via the PTEN/PI3K/Akt and MAPK/PI3K/AP-1 signaling pathways (46–48). In addition, its receptor C-X-C motif chemokine receptor 4 (CXCR4) has been used as an effective therapeutic target in prostate cancer (49–51). Thus, the findings of these studies further support our results.

Regarding the regulatory mechanisms of the biomarkers, we were surprised to find that RUNX2 could regulate 11/12 mRNAs in the TF-mRNA network. Increasing evidence has highlighted the importance of RUNX2 in a variety of cancers. For instance, it is highly expressed in metastatic prostate cancer cells and may play an important role in prostate cancer-derived metastatic bone disease (52, 53). RUNX2 plays an oncogenic role in esophageal carcinoma by activating the PI3K/AKT1 and extracellular-regulated kinase signaling pathways (54). Targeting RUNX2 represses cell growth and metastasis in lung cancer cells (55) and inhibits the progression of breast cancer to metastatic bone disease (56). Besides, regarding the regulatory function of RUNX2 in the network, Francisco et al. (57) reported elevated RUNX2 may transcriptionally activate genes mediating osteosarcoma progression and metastasis by targeting SPP1. Toshihisa et al. (58) reported that Runx2 could induce the expression of major bone matrix protein genes, including COL1A1, SPP1, and FN1, *in vitro*. Besides, Toshihisa et al. (59) also reported Runx2 plays an important role in the bone metastasis of breast and prostate cancers by up-regulating SPP1. Although some regulatory relationships in the network have been verified by previous studies, there are still many waiting for further verification. However, despite increasing evidence of the

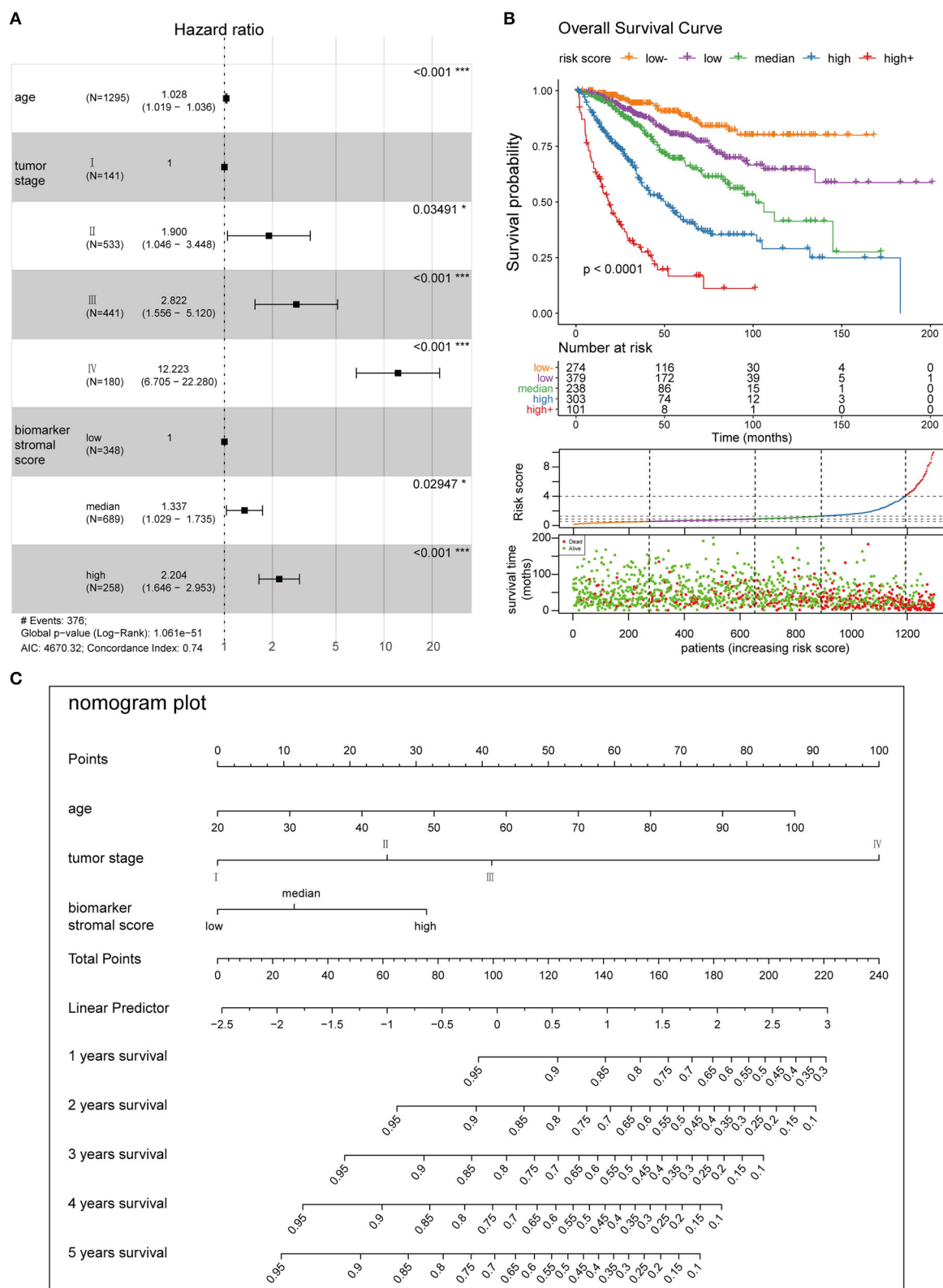
importance of RUNX2 in various cancers, there are no reports about its relevance in CC. Our results suggest that the role of RUNX2 in CC is worthy of further study.

The current risk classification for cancers is mainly based on the TNM staging system (3); however, for a deeper understanding of tumor progression, more prognosis factors should be considered. For instance, Weiser et al. showed that an extended prognosis model including TNM staging, the tumor grade, the number of collected metastatic lymph nodes, age, and sex had higher sensitivity and specificity for CC (the C-index rose from 0.60 to 0.68) than a model using the TNM system alone (4). Pagès et al. (12) showed that adding an immunoscore to a model combining clinical variables can significantly improve OS prediction accuracy of AUC from 0.6 to 0.62. In this study, we created a new prognosis indicator based on tumor stroma biomarkers. Adding this indicator to a prognosis model based on age and tumor stage also significantly improved the prediction accuracy, with a similar degree of improvement to Pagès's immunoscore (3-year AUC raised from 0.75 to 0.773; 5-year AUC raised from 0.732 to 0.758). In addition, as the new indicator is based on only 16 biomarkers, testing will be easier, more effective, and more economically feasible for patients with CC vs. the ESTIMATE stromal score, which is based on 141 signatures.

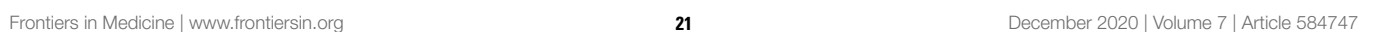
Our study demonstrates the important role of the tumor stroma in CC tumor progression and chemotherapy resistance and provides novel candidates for targeted CC therapies. However, the data available for this study is limited, and our findings are mainly obtained through bioinformatics analysis of high-throughput data which have inevitable batch differences between different datasets due to sequencing technologies, so these findings will require further validation with more clinical data and molecular experiments. In our future work, we will



**FIGURE 11 |** Performance of the biomarker stromal score in prognosis prediction compared with clinicopathological features. **(A)** Boxplots show the prediction accuracy for 3- and 5-year overall survival, based on the AUC with 1,000× bootstrap resampling for each parameter. **(B)** The top and middle boxplots show the prediction accuracy for 3- and 5- year overall survival based on the AUC with 1,000× bootstrap resampling, while the bottom boxplot shows the prediction accuracy for overall survival based on the C-index with 1,000× bootstrap resampling for each prognosis model.



**FIGURE 12 |** Clinical application of the best multivariable hazards model. **(A)** Forest plot of hazard ratios for the three prognosis features in model 3. **(B)** Survival curves and scatter plots of patients in five different risk groups, based on the risk score. **(C)** A nomogram plot was constructed with the three prognosis features to predict the 1–5-year overall survival rates of patients with CC.



test additional clinical datasets and perform additional molecular experimental verification on the identified biomarkers. Notable, this is the first study to consider the tumor stroma in CC risk stratification, and the new prognosis indicator and prognosis model created in this study will increase the accuracy of risk stratification and survival prediction, improving the outcomes of patients with CC.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

This study was performed in collaboration among all authors. YC and XL contributed to the study design. YC, WW, and BJ downloaded the datasets and performed the statistical analyses. LY and FX analyzed the results. YC drafted the manuscript, and all authors contributed to revision of the final manuscript. LY supervised the study and manuscript preparation.

## REFERENCES

- Li G, Li M, Liang X, Xiao Z, Zhang P, Shao M, et al. Identifying DCN and HSPD1 as potential biomarkers in colon cancer using 2D-LC-MS/MS combined with iTRAQ technology. *J. Cancer*. (2017) 8:479–89. doi: 10.7150/jca.17192
- Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. *Lancet*. (2019) 394:1467–80. doi: 10.1016/S0140-6736(19)32319-0
- Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA Cancer J. Clin.* (2017) 67:93–9. doi: 10.3322/caac.21388
- Weiser MR, Gonen M, Chou JF, Kattan MW, Schrag D. Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. *J. Clin. Oncol.* (2011) 29:4796–802. doi: 10.1200/JCO.2011.36.5080
- 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
- Devarakonda S, Rotolo F, Tsao MS, Lanc I, Brambilla E, Masood A, et al. Tumor mutation burden as a biomarker in resected non-small-cell lung cancer. *J. Clin. Oncol.* (2018) 36:2995–3006. doi: 10.1200/JCO.2018.78.1963
- Zhou R, Zhang J, Zeng D, Sun H, Rong X, Shi M, et al. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I–III colon cancer. *Cancer Immunol. Immunother.* (2019) 68:433–42. doi: 10.1007/s00262-018-2289-7
- Yang Y, Shi Z, Bai R, Hu W. Heterogeneity of MSI-H gastric cancer identifies a subtype with worse survival. *J. Med. Genet.* (2020). doi: 10.1136/jmedgenet-2019-106609. [Epub ahead of print].
- Mu Y, Chen Y, Zhang G, Zhan X, Li Y, Liu T, et al. Identification of stromal differentially expressed proteins in the colon carcinoma by quantitative proteomics. *Electrophoresis*. (2013) 34:1679–92. doi: 10.1002/elps.201200596
- Koliariaki V, Pallangyo CK, Gretten FR, Kollis G. Mesenchymal cells in colon cancer. *Gastroenterology*. (2017) 152:964–79. doi: 10.1053/j.gastro.2016.11.049
- Tauriello DVE, Palomo-Ponce S, Stork D, Berenguer-Llargo A, Badia-Ramentol J, Iglesias M, et al. TGF $\beta$  drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature*. (2018) 554:538–43. doi: 10.1038/nature25492

## FUNDING

This work was funded by the National Natural Science Foundation of China (grant number 81672885) and the Hunan Province Natural Science Foundation (grant number 2019JJ40475). The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## ACKNOWLEDGMENTS

We thank Central South University for technical support. We acknowledge the cBioPortal, Gene Expression Omnibus, and The Cancer Genome Atlas databases for contributing the data used in this study. We also thank Editage for English language editing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2020.584747/full#supplementary-material>

- Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet*. (2018) 391:2128–39. doi: 10.1016/S0140-6736(18)30789-X
- Yoshihara K, Shahmoradgolli M, Martínez E, Vegesna R, Kim H, Torres-García W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* (2013) 4:1–11. doi: 10.1038/ncomms3612
- Wang H, Wu X, Chen Y. Stromal-immune score-based gene signature: a prognosis stratification tool in gastric cancer. *Front. Oncol.* (2019) 9:1212. doi: 10.3389/fonc.2019.01212
- Xu WH, Xu Y, Wang J, Wan FN, Wang HK, Cao DL, et al. Prognostic value and immune infiltration of novel signatures in clear cell renal cell carcinoma microenvironment. *Aging*. (2019) 11:6999–7020. doi: 10.18632/aging.102233
- Hothorn T, Zeileis A. Generalized maximally selected statistics. *Biometrics*. (2008) 64:1263–9. doi: 10.1111/j.1541-0420.2008.00995.x
- Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* (2011) 7:e1001057. doi: 10.1371/journal.pcbi.1001057
- Wang M, Zhao J, Zhang L, Wei F, Lian Y, Wu Y, et al. Role of tumor microenvironment in tumorigenesis. *J. Cancer*. (2017) 8:761–73. doi: 10.7150/jca.17648
- Jiang X, Wang J, Deng X, Xiong F, Ge J, Xiang B, et al. Role of the tumor microenvironment in PD-L1/PD-1-mediated tumor immune escape. *Mol. Cancer*. (2019) 18:10. doi: 10.1186/s12943-018-0928-4
- Emens LA, Silverstein SC, Khleif S, Marincola FM, Galon J. Toward integrative cancer immunotherapy: targeting the tumor microenvironment. *J. Transl. Med.* (2012) 10:70. doi: 10.1186/1479-5876-10-70
- Ngambenjawong C, Gustafson HH, Pun SH. Progress in tumor-associated macrophage (TAM)-targeted therapeutics. *Adv. Drug Deliv. Rev.* (2017) 114:206–21. doi: 10.1016/j.addr.2017.04.010
- Salmaninejad A, Valilou SE, Soltani A, Ahmadi S, Abarghan YJ, Rosengren RJ, et al. Tumor-associated macrophages: role in cancer development and therapeutic implications. *Cell. Oncol.* (2019) 42:591–608. doi: 10.1007/s13402-019-00453-z
- Xie C, Ji N, Tang Z, Li J, Chen Q. The role of extracellular vesicles from different origin in the microenvironment of head and neck cancers. *Mol. Cancer*. (2019) 18:83. doi: 10.1186/s12943-019-0985-3



24. Gentles AJ, Bratman SV, Lee LJ, Harris JP, Feng W, Nair RV, et al. Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage non-small cell lung cancer. *J. Natl. Cancer Inst.* (2015) 107:djv211. doi: 10.1093/jnci/djv211
25. Mo F, Lin D, Takhar M, Ramnarine VR, Dong X, Bell RH, et al. Stromal gene expression is predictive for metastatic primary prostate cancer. *Eur. Urol.* (2018) 73:524–32. doi: 10.1016/j.eururo.2017.02.038
26. Mao M, Yu Q, Huang R, Lu Y, Wang Z, Liao L. Stromal score as a prognostic factor in primary gastric cancer and close association with tumor immune microenvironment. *Cancer Med.* (2020) 9:4980–90. doi: 10.1002/cam4.2801
27. Frentzas S, Simoneau E, Bridgeman VL, Vermeulen PB, Foo S, Kostaras E, et al. Vessel co-option mediates resistance to anti-angiogenic therapy in liver metastases. *Nat. Med.* (2016) 22:1294–302. doi: 10.1038/nm.4197
28. Kuczyński EA, Yin M, Bar-Zion A, Lee CR, Butz H, Man S, et al. Co-option of liver vessels and not sprouting angiogenesis drives acquired sorafenib resistance in hepatocellular carcinoma. *J. Natl. Cancer Inst.* (2016) 108:djw030. doi: 10.1093/jnci/djw030
29. Valkenburg KC, de groot AE, Pienta KJ. Targeting the tumour stroma to improve cancer therapy. *Nat. Rev. Clin. Oncol.* (2018) 15:366–381. doi: 10.1038/s41571-018-0007-1
30. Laplagne C, Domagala M, Le Naour A, Quemerais C, Hamel D, Fournié JJ, et al. Latest advances in targeting the tumor microenvironment for tumor suppression. *Int. J. Mol. Sci.* (2019) 20:4719. doi: 10.3390/ijms20194719
31. Fridman WH, Miller I, Sautès-Fridman C, Byrne AT. Therapeutic targeting of the colorectal tumor stroma. *Gastroenterology.* (2020) 158:303–21. doi: 10.1053/j.gastro.2019.09.045
32. Vennin C, Murphy KJ, Morton JP, Cox TR, Pajic M, Timpson P. Reshaping the tumor stroma for treatment of pancreatic cancer. *Gastroenterology.* (2018) 154:820–38. doi: 10.1053/j.gastro.2017.11.280
33. Leight JL, Wozniak MA, Chen S, Lynch ML, Chen CS. Matrix rigidity regulates a switch between TGF- $\beta$ 1-induced apoptosis and epithelial-mesenchymal transition. *Mol. Biol. Cell.* (2012) 23:781–91. doi: 10.1091/mbc.e11-06-0537
34. Provenzano PP, Cuevas C, Chang AE, Goel VK, Von Hoff DD, Hingorani SR. Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma. *Cancer Cell.* (2012) 21:418–29. doi: 10.1016/j.ccr.2012.01.007
35. Yu M, Tannock IF. Targeting tumor architecture to favor drug penetration: a new weapon to combat chemoresistance in pancreatic cancer? *Cancer Cell.* (2012) 21:327–9. doi: 10.1016/j.ccr.2012.03.002
36. Fei M, Hang Q, Hou S, He S, Ruan C. Adhesion to fibronectin induces p27(Kip1) nuclear accumulation through down-regulation of Jab1 and contributes to cell adhesion-mediated drug resistance (CAM-DR) in RPMI 8,226 cells. *Mol. Cell. Biochem.* (2014) 386:177–87. doi: 10.1007/s11010-013-1856-7
37. Ham IH, Oh HJ, Jin H, Bae CA, Jeon SM, Choi KS, et al. Targeting interleukin-6 as a strategy to overcome stroma-induced resistance to chemotherapy in gastric cancer. *Mol. Cancer.* (2019) 18:68. doi: 10.1186/s12943-019-0972-8
38. Okkenhaug K, Graupera M, Vanhaesebroeck B. Targeting PI3K in cancer: impact on tumor cells, their protective stroma, angiogenesis, and immunotherapy. *Cancer Discov.* (2016) 6:1090–105. doi: 10.1158/2159-8290.CD-16-0716
39. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. TGF $\beta$  attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature.* (2018) 554:544–8. doi: 10.1038/nature25501
40. American Association for Cancer Research. TGF $\beta$  promotes immune evasion to limit the efficacy of anti-PD-1/PD-L1. *Cancer Discov.* (2018) 8:OF10. doi: 10.1158/2159-8290.CD-RW2018-034
41. Xie Y, Liu C, Qin Y, Chen J, Fang J. Knockdown of IRE1 $\gamma$  suppresses metastatic potential of colon cancer cells through inhibiting FN1-Src/FAK-GTPases signaling. *Int. J. Biochem. Cell Biol.* (2019) 114:105572. doi: 10.1016/j.biocel.2019.105572
42. Cai X, Liu C, Zhang TN, Zhu YW, Dong X, Xue P. Down-regulation of FN1 inhibits colorectal carcinogenesis by suppressing proliferation, migration, and invasion. *J. Cell. Biochem.* (2018) 119:4717–28. doi: 10.1002/jcb.26651
43. Vincent KM, Postovit LM. A pan-cancer analysis of secreted frizzled-related proteins: re-examining their proposed tumour suppressive function. *Sci. Rep.* (2017) 7:42719. doi: 10.1038/srep42719
44. Nfonsam LE, Jandova J, Jecius HC, Omesiete PN, Nfonsam VN. SFRP4 expression correlates with epithelial mesenchymal transition-linked genes and poor overall survival in colon cancer patients. *World J. Gastrointest. Oncol.* (2019) 11:589–98. doi: 10.4251/wjgo.v11.i8.589
45. Klement JD, Paschall AV, Redd PS, Ibrahim ML, Lu C, Yang D, et al. An osteopontin/CD44 immune checkpoint controls CD8+ T cell activation and tumor immune evasion. *J. Clin. Invest.* (2018) 128:5549–60. doi: 10.1172/JCI123360
46. Ma J-C, Sun X-W, Su H, Chen Q, Guo T-K, Li Y, et al. Fibroblast-derived CXCL12/SDF-1 $\alpha$  promotes CXCL6 secretion and co-operatively enhances metastatic potential through the PI3K/Akt/mTOR pathway in colon cancer. *World J. Gastroenterol.* (2017) 23:5167–78. doi: 10.3748/wjg.v23.i28.5167
47. Ma J, Su H, Yu B, Guo T, Gong Z, Qi J, et al. CXCL12 gene silencing down-regulates metastatic potential via blockage of MAPK/PI3K/AP-1 signaling pathway in colon cancer. *Clin. Transl. Oncol.* (2018) 20:1035–45. doi: 10.1007/s12094-017-1821-0
48. Ma J, Sun X, Wang Y, Chen B, Qian L, Wang Y. Fibroblast-derived CXCL12 regulates PTEN expression and is associated with the proliferation and invasion of colon cancer cells via PI3K/Akt signaling. *Cell Commun. Signal.* (2019) 17:119. doi: 10.1186/s12964-019-0432-5
49. Domanska UM, Timmer-Bosscha H, Nagengast WB, Oude Munnink TH, Kruizinga RC, Ananias HJK, et al. CXCR4 inhibition with AMD3100 sensitizes prostate cancer to docetaxel chemotherapy. *Neoplasia.* (2012) 14:709–18. doi: 10.1593/neo.12324
50. Domanska UM, Boer JC, Timmer-Bosscha H, van Vugt MATM, Hoving HD, Kliphuis NM, et al. CXCR4 inhibition enhances radiosensitivity, while inducing cancer cell mobilization in a prostate cancer mouse model. *Clin. Exp. Metastasis.* (2014) 31:829–39. doi: 10.1007/s10585-014-9673-2
51. D'Alterio C, Buoncervello M, Ieranò C, Napolitano M, Portella L, Rea G, et al. Targeting CXCR4 potentiates anti-PD-1 efficacy modifying the tumor microenvironment and inhibiting neoplastic PD-1. *J. Exp. Clin. Cancer Res.* (2019) 38:432. doi: 10.1186/s13046-019-1420-8
52. Akech J, Wixted JJ, Bedard K, Van Der Deen M, Hussain S, Guise TA, et al. Runx2 association with progression of prostate cancer in patients: mechanisms mediating bone osteolysis and osteoblastic metastatic lesions. *Oncogene.* (2010) 29:811–21. doi: 10.1038/onc.2009.389
53. Baniwal SK, Khalid O, Gabet Y, Shah RR, Purcell DJ, Mav D, et al. Runx2 transcriptome of prostate cancer cells: insights into invasiveness and bone metastasis. *Mol. Cancer.* (2010) 9:258. doi: 10.1186/1476-4598-9-258
54. Lu H, Jiang T, Ren K, Li ZL, Ren J, Wu G, et al. RUNX2 plays an oncogenic role in esophageal carcinoma by activating the PI3K/AKT and ERK signaling pathways. *Cell. Physiol. Biochem.* (2018) 49:217–25. doi: 10.1159/000492872
55. Bai X, Meng L, Sun H, Li Z, Zhang X, Hua S. MicroRNA-196b inhibits cell growth and metastasis of lung cancer cells by targeting Runx2. *Cell. Physiol. Biochem.* (2017) 43:757–67. doi: 10.1159/000481559
56. Taipaleenmäki H, Browne G, Akech J, Zustin J, Van Wijnen AJ, Stein JL, et al. Targeting of Runx2 by miR-135 and miR-203 impairs progression of breast cancer and metastatic bone disease. *Cancer Res.* (2015) 75:1433–44. doi: 10.1158/0008-5472.CAN-14-1026
57. Villanueva F, Araya H, Briceño P, Varela N, Stevenson A, Jerez S, et al. The cancer-related transcription factor RUNX2 modulates expression and secretion of the matricellular protein osteopontin in osteosarcoma cells to promote adhesion to endothelial pulmonary cells and lung metastasis. *J. Cell. Physiol.* (2019) 234:13659–79. doi: 10.1002/jcp.28046
58. Komori T. Regulation of proliferation, differentiation and functions of osteoblasts by runx2. *Int. J. Mol. Sci.* (2019) 20:1694. doi: 10.3390/ijms20071694
59. Komori T. Runx2, an inducer of osteoblast and chondrocyte differentiation. *Histochem. Cell Biol.* (2018) 149:313–23. doi: 10.1007/s00418-018-1640-6

**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.

Copyright © 2020 Chen, Wang, Jiang, Yao, Xia and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Identification of a Ubiquitination-Related Gene Risk Model for Predicting Survival in Patients With Pancreatic Cancer

Hao Zuo<sup>1,2†</sup>, LuoJun Chen<sup>1,2†</sup>, Na Li<sup>1,2\*</sup> and Qibin Song<sup>1,2\*</sup>

<sup>1</sup> Cancer Center, Renmin Hospital of Wuhan University, Wuhan, China, <sup>2</sup> Hubei Provincial Research Center for Precision Medicine of Cancer, Wuhan, China

## OPEN ACCESS

### Edited by:

Maria Rodriguez Martinez,  
IBM Research - Zürich, Switzerland

### Reviewed by:

Arsheed A. Ganaie,  
University of Minnesota Twin Cities,  
United States  
Edmund Ui-Hang Sim,  
Universiti Malaysia, Sarawak, Malaysia

### \*Correspondence:

Na Li  
nalirenmin@163.com  
Qibin Song  
qibinsong@163.com

<sup>†</sup> These authors have contributed  
equally to this work and share first  
authorship

### Specialty section:

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

**Received:** 30 September 2020

**Accepted:** 30 November 2020

**Published:** 22 December 2020

### Citation:

Zuo H, Chen L, Li N and Song Q  
(2020) Identification of a  
Ubiquitination-Related Gene Risk  
Model for Predicting Survival  
in Patients With Pancreatic Cancer.  
Front. Genet. 11:612196.  
doi: 10.3389/fgene.2020.612196

Pancreatic cancer is known as “the king of cancer,” and ubiquitination/deubiquitination-related genes are key contributors to its development. Our study aimed to identify ubiquitination/deubiquitination-related genes associated with the prognosis of pancreatic cancer patients by the bioinformatics method and then construct a risk model. In this study, the gene expression profiles and clinical data of pancreatic cancer patients were downloaded from The Cancer Genome Atlas (TCGA) database and the Genotype-tissue Expression (GTEx) database. Ubiquitination/deubiquitination-related genes were obtained from the gene set enrichment analysis (GSEA). Univariate Cox regression analysis was used to identify differentially expressed ubiquitination-related genes selected from GSEA which were associated with the prognosis of pancreatic cancer patients. Using multivariate Cox regression analysis, we detected eight optimal ubiquitination-related genes (RNF7, NPEPPS, NCCRP1, BRCA1, TRIM37, RNF25, CDC27, and UBE2H) and then used them to construct a risk model to predict the prognosis of pancreatic cancer patients. Finally, the eight risk genes were validated by the Human Protein Atlas (HPA) database, the results showed that the protein expression level of the eight genes was generally consistent with those at the transcriptional level. Our findings suggest the risk model constructed from these eight ubiquitination-related genes can accurately and reliably predict the prognosis of pancreatic cancer patients. These eight genes have the potential to be further studied as new biomarkers or therapeutic targets for pancreatic cancer.

**Keywords:** pancreatic cancer, bioinformatics, prognosis, ubiquitination-related genes, risk model

## INTRODUCTION

Pancreatic cancer is a highly fatal disease, with 43,090 deaths every 5 years (Siegel et al., 2017), the 5-year overall survival rate is only 6% (Miller et al., 2019). Many factors contribute to low survival rates for pancreatic cancer. The most important factor may be that more than half of patients are diagnosed with advanced pancreatic cancer, and the 5-year survival rate of advanced pancreatic cancer is only 3% (Ilic and Ilic, 2016). Pancreatic cancer is characterized not only by early recurrence and invasion but also by chemical and radiation resistance (Adamska et al., 2018). In recent years, targeted therapy and emerging immunotherapy have opened up new prospects for the treatment of pancreatic cancer. However, the exploration of new therapeutic targets and prognostic biomarkers for pancreatic cancer still needs to be further carried out. Over the past decade, numerous studies have identified some sensitive and effective biomarkers for pancreatic cancer.

Ubiquitination/deubiquitination is an ATP-dependent reversible reaction that binds small ubiquitin molecules to the target protein through multi-step reactions involving ubiquitin-activating enzyme E1, ubiquitin-binding enzyme E2, and ubiquitin-ligase E3 (Hershko et al., 1979). ATP provides energy, E1 is activated, and the Glycine residue at the Carboxy terminal of ubiquitin and the active Cysteine of E1 forms a thioester bond. Next, E1 transfers the ubiquitin to the cysteine residue of the ubiquitin carrier protein E2. E3 is specific in that it coordinates ubiquitin covalently to specific target proteins. The way ubiquitin molecules bind plays an important role in the function of the modified protein (Dikic et al., 2009). Ubiquitination produces a protein that is either monoubiquitinated or polyubiquitinated when one of the seven Lysine residues of ubiquitin binds to the C-terminal Glycine of another ubiquitin. The reverse process of ubiquitination is called deubiquitination. Ubiquitination is best known for its role in mediating protein degradation. Besides, ubiquitination is also involved in the processes of meiosis, autophagy, DNA repair, immune response, and apoptosis. Ubiquitinated proteasome pathway is involved in almost all intracellular molecular biological processes, affecting gene expression and signal transduction in the regulation of DNA damage repair, participating in the differentiation of senescent cells, regulating tumor progression of malignant transformation, and mediating therapeutic resistance (Welchman et al., 2005).

Previous studies have shown that ubiquitination/deubiquitination play important roles in pancreatic cancer. Lian et al. (2020) found that ubiquitin specific peptidase 5 (USP5) enhances STAT3 signaling and promotes migration and invasion in pancreatic cancer. Chen et al. (2020) found that E3 ubiquitin ligase UBR5 promotes pancreatic cancer growth and aerobic glycolysis by downregulating FBP1 via destabilization of C/EBP $\alpha$ . Yang et al. (2019) found that USP44 suppresses pancreatic cancer progression and overcomes gemcitabine resistance by deubiquitinating FBP1. There is no doubt that ubiquitination/deubiquitination is closely related to the progression of pancreatic cancer. Exploration of ubiquitination/deubiquitination related genes in pancreatic cancer is also necessary.

In this study, by analyzing the dataset from TCGA and GTEx database, we aim to study and verify the expression characteristics of ubiquitination-related genes. We then selected several ubiquitination-related genes that were significantly associated with the prognosis of pancreatic cancer patients through a series of statistical methods. Finally, we established a new and reliable risk model to predict the prognosis of pancreatic cancer patients based on the screened risk genes.

## MATERIALS AND METHODS

### Databases

To download the transcriptome data of 178 patients (The Cancer Genome Atlas database, TCGA database) with pancreatic cancer and the transcriptome data of 36 cases of normal pancreatic tissue (Genotype-Tissue Expression database, GTEx database) from the

UCSC XENA website<sup>1</sup>. Clinical information of pancreatic cancer patients was obtained from the TCGA database. All data are processed using R software<sup>2</sup>. The clinical features of pancreatic cancer patients, include age, gender, pathological grade, T-stage, N-stage, M-stage, and TNM-stage.

### Gene Set Enrichment Analysis

GSEA<sup>3</sup> was used to explore whether the transcriptome data showed statistically significant difference between the two groups (normal and tumor). The expression data of mRNAs, including 36 normal pancreatic tissue and 178 pancreatic cancer samples were analyzed. Normalized P value ( $P < 0.05$ ) and normalized enrichment score (NES) were used to determine what functions had to be selected for further analysis.

### Screening for Differentially Expressed Genes (DEGs)

We screened DEGs from these ubiquitination/deubiquitination related genes obtained from GSEA analysis. The “limma” package was used to screen out the DEGs ( $\text{Log}_2$  fold change  $\neq 0$ ,  $P < 0.05$ ).

### GO Analysis and KEGG Analysis

Gene Ontology (GO) database is a kind of free and open database, the database includes three aspects of information: biological process, cellular component, and molecular function. The biological functions of genes can be classified and these genes included in the functions that we selected can be further understood through the GO analysis. DAVID online tool<sup>4</sup> was used for GO analysis (Xia et al., 2015). Kyoto Encyclopedia of Genes and Genomes (KEGG) database is a database that systematically analyzes the metabolic pathways of gene products in cells and the functions of these gene products. The database is useful for studying gene and expression information as a whole network. KEGG integrates the data of genomic chemical molecules and biochemical systems, including the sequence and genome of metabolic pathways, drugs, and diseases. We used the “clusterProfiler” package (Yu et al., 2012) from Bioconductor to do KEGG analysis of these DEGs.  $P$ -value  $< 0.05$  was used as the inclusion standard in the analysis.

### Identification and Inclusion of Prognostic Ubiquitination-Related Genes for the Construction of a Risk Model

As in previous studies (Li et al., 2020), univariate Cox regression ( $p < 0.05$ ) was used to screen out the ubiquitination-related genes that were significantly associated with the prognosis of pancreatic cancer patients from the DRGs. Multivariate Cox proportional hazards regression analysis (with forwarding selection and backward selection) was then used to analyze these ubiquitination-related genes selected by univariate Cox regression. Finally, optimal ubiquitination-related genes (risk

<sup>1</sup><https://xenabrowser.net/>

<sup>2</sup><https://www.r-project.org/>

<sup>3</sup><http://www.broadinstitute.org/gsea/index.jsp>

<sup>4</sup><http://david.ncifcrf.gov/>



genes) were obtained to be incorporated into the risk model. The alteration of these risk genes is shown online<sup>5</sup>.

## Construction of the Prognostic Risk Model in Pancreatic Cancer Cohort

Multivariate Cox proportional hazards regression analysis was used to select the optimal risk genes and construct the Cox regression model. In this process, we can obtain the estimated regression coefficients of each gene. The expression levels of mRNA and estimated regression coefficients of the risk genes were used to calculate a risk score for each pancreatic cancer patients. The risk score model was established with the following formula: Risk score = expression level of Gene1 \*  $\beta_1$  + expression level of Gene2 \*  $\beta_2$  + ... + expression level of Gene<sub>n</sub> \*  $\beta_n$ ; where  $\beta$  is the estimated regression coefficient calculated by the multivariate Cox regression model.

The risk model was used to measure the prognostic risk for each pancreatic cancer patient. The median risk score was used as the cut-off value to divide all the pancreatic cancer patients into two groups: the high-risk group and the low-risk group. The low-risk group has a better prognosis.

## Independent Prognostic Value of the Risk Model in the Pancreatic Cancer Cohort

Next, univariate and multivariate Cox regression analysis were performed to assess whether the risk model was independent of other clinical features (age, gender, pathological grade, T-stage, and N-stage) as a prognostic factor for pancreatic cancer patients ( $p < 0.05$ ). The X-tile software was used to identify the optimal cut-off value of the age significantly correlated to the prognosis of pancreatic cancer patients. Because there are too many patients in M0-stage and too few patients in stage III/IV, we excluded these two clinical features (M-stage and TNM-stage) from this analysis. Besides, cases with incomplete clinical information were also excluded. Then, we constructed receiver operating curves (ROC) and calculated the area under the curve (AUC) to assess whether our model accurately predicted the overall survival (OS) of pancreatic cancer patients. C-index value of 0.75 or greater were considered to have excellent predictive value, and value of 0.6 or greater were considered acceptable for survival predictions (Cho et al., 2019).

## Validation of the Eight-mRNA Model in Predicting Survival Using Kaplan–Meier Curves

Kaplan–Meier curves and the log-rank test were used to validate the prognostic significance of the risk model ( $p < 0.05$ ).

## Validation of the Risk Genes in Protein Level

Furthermore, the Human Protein Atlas database<sup>6</sup> was used to validate the protein expression level of these risk genes compared to the level of gene transcription.

<sup>5</sup><http://www.cbioportal.org/>

<sup>6</sup><https://www.proteinatlas.org/>

## RESULTS

### Gene Set Enrichment Analysis

Expression data set for 55242 mRNAs from the TCGA database and GTEx database were analyzed. Five ubiquitination/deubiquitination-related gene sets we validated by GSEA analysis and there were two gene sets, including REACTOME\_ANTIGEN\_PROCESSING\_UBIQUITINATION\_PROTEASOME\_DEGRADATION, and REACTOME\_PROTEIN\_UBIQUITINATION were significantly enriched (Table 1 and Figure 1). These 441 ubiquitination-related genes in the two functions were selected for the subsequent analysis.

### GO Analysis and KEGG Analysis

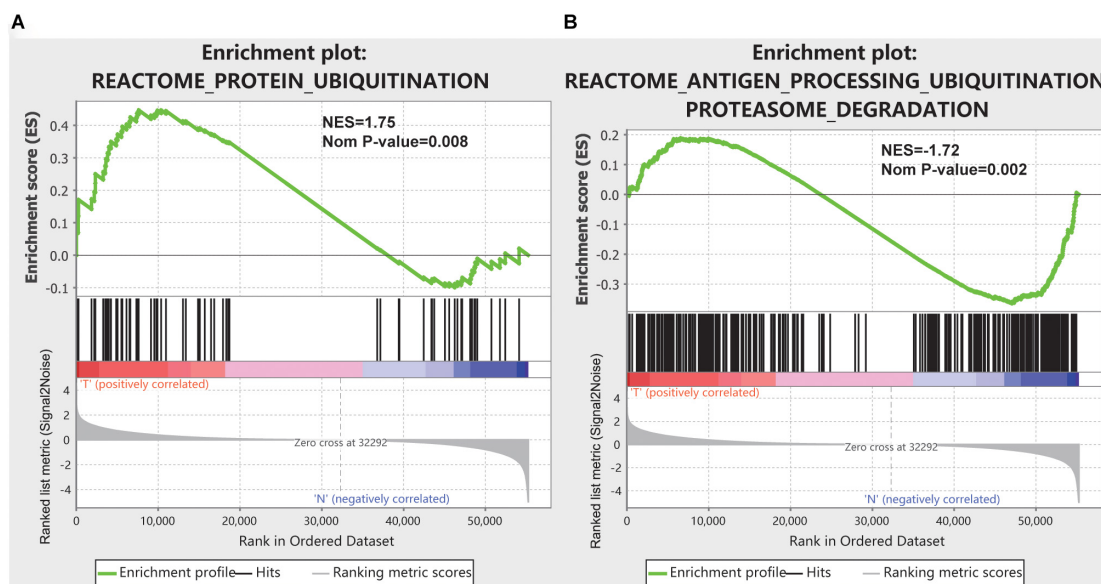
Of these 441 ubiquitination-related genes in the two functions, 134 DEGs were screened. These 134 ubiquitination-related DEGs were used to do the GO analysis and KEGG analysis. The results of the GO analysis showed that the functions of the ubiquitination-related genes were concentrated in the functions of the protein polyubiquitination, post-translational protein modification, and proteasome-mediated ubiquitin-dependent protein catabolic process, as shown in Table 2. The results of KEGG analysis showed that the functions of the ubiquitination-related genes were concentrated in ubiquitin-mediated proteolysis, proteasome, and cell cycle, as shown in Table 2.

## Identification and Inclusion of Prognostic Ubiquitination-Related Genes for the Construction of a Risk Model

Sixty-three ubiquitination-related genes significantly correlated with the prognosis of pancreatic cancer patients were screened through the univariate Cox regression analysis from the 134 DEGs. Next, eight optimal ubiquitination-related genes (risk gene) obtained by multivariate Cox analysis were used to construct a risk model (Table 3): RNF7, NPEPPS, NCCRP1, BRCA1, TRIM37, RNF25, CDC27, and UBE2H. The effect of the expression value of these genes on the prognosis of pancreatic cancer is shown in Figures 2A–H. Then, the alteration of the seven genes in 175 clinical pancreatic cancer samples was analyzed in the cBioPortal database. Results showed that there were 33(19%) sequenced cases among the 175 pancreatic cancer samples with the eight genes altering. The alterations of the

TABLE 1 | Gene sets enriched in pancreatic cancer.

Gene sets enriched in pancreatic cancer		
GS follow link to MSigDB	NES	NOM $p$ -value
GO_UBIQUITIN_DEPENDENT_ERAD_PATHWAY	−1.46	0.075
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	−1.45	0.067
REACTOME_ANTIGEN_PROCESSING_UBIQUITINATION_PROTEASOME_DEGRADATION	−1.72	0.002
REACTOME_DEUBIQUITINATION	1.38	0.085
REACTOME_PROTEIN_UBIQUITINATION	1.75	0.008



**FIGURE 1** | GSEA results for the enrichment plots of two gene sets (REACTOME\_ANTIGEN\_PROCESSING\_UBIQUITINATION\_PROTEASOME\_DEGRADATION, and REACTOME\_PROTEIN\_UBIQUITINATION) that were significantly differentiated in normal and pancreatic cancer tissues based on TCGA database. **(A)** Enrichment plot of the REACTOME\_ANTIGEN\_PROCESSING\_UBIQUITINATION\_PROTEASOME\_DEGRADATION gene set. **(B)** Enrichment plot of the REACTOME\_PROTEIN\_UBIQUITINATION gene set.

**TABLE 2** | Result of GO and KEGG analysis for these ubiquitination-related DEGs.

ID	Description	P-adjust	Q-value
<b>GO analysis</b>			
GO:0000209	Protein polyubiquitination	<0.001	<0.001
GO:0043687	Post-translational protein modification	<0.001	<0.001
GO:0043161	Proteasome-mediated ubiquitin-dependent protein catabolic process	<0.001	<0.001
GO:0010498	Proteasomal protein catabolic process	<0.001	<0.001
GO:0031145	Anaphase-promoting complex-dependent catabolic process	<0.001	<0.001
GO:0006513	Protein monoubiquitination	<0.001	<0.001
GO:0031146	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	<0.001	<0.001
<b>KEGG analysis</b>			
hsa04120	Ubiquitin mediated proteolysis	<0.001	<0.001
hsa03050	Proteasome	<0.001	<0.002
hsa04110	Cell cycle	<0.001	<0.003
hsa04141	Protein processing in endoplasmic reticulum	<0.001	<0.004
hsa04114	Oocyte meiosis	0.017	0.016
hsa05017	Spinocerebellar ataxia	0.018	0.016
hsa04144	Endocytosis	0.020	0.018
hsa05169	Epstein-Barr virus infection	0.023	0.020
hsa04115	p53 signaling pathway	0.043	0.038

eight genes are shown in **Figure 3A**. We also investigated the different expressions of the eight genes between pancreatic cancer tissues and normal pancreatic tissues. Among the eight genes,

five genes (BRCA1, TRIM37, RNF25, CDC27, and UBE2H) were significantly upregulated and three genes (RNF7, NPEPPS, and NCCRP1) were significantly down regulated in the tumor tissues (**Figure 3B**).

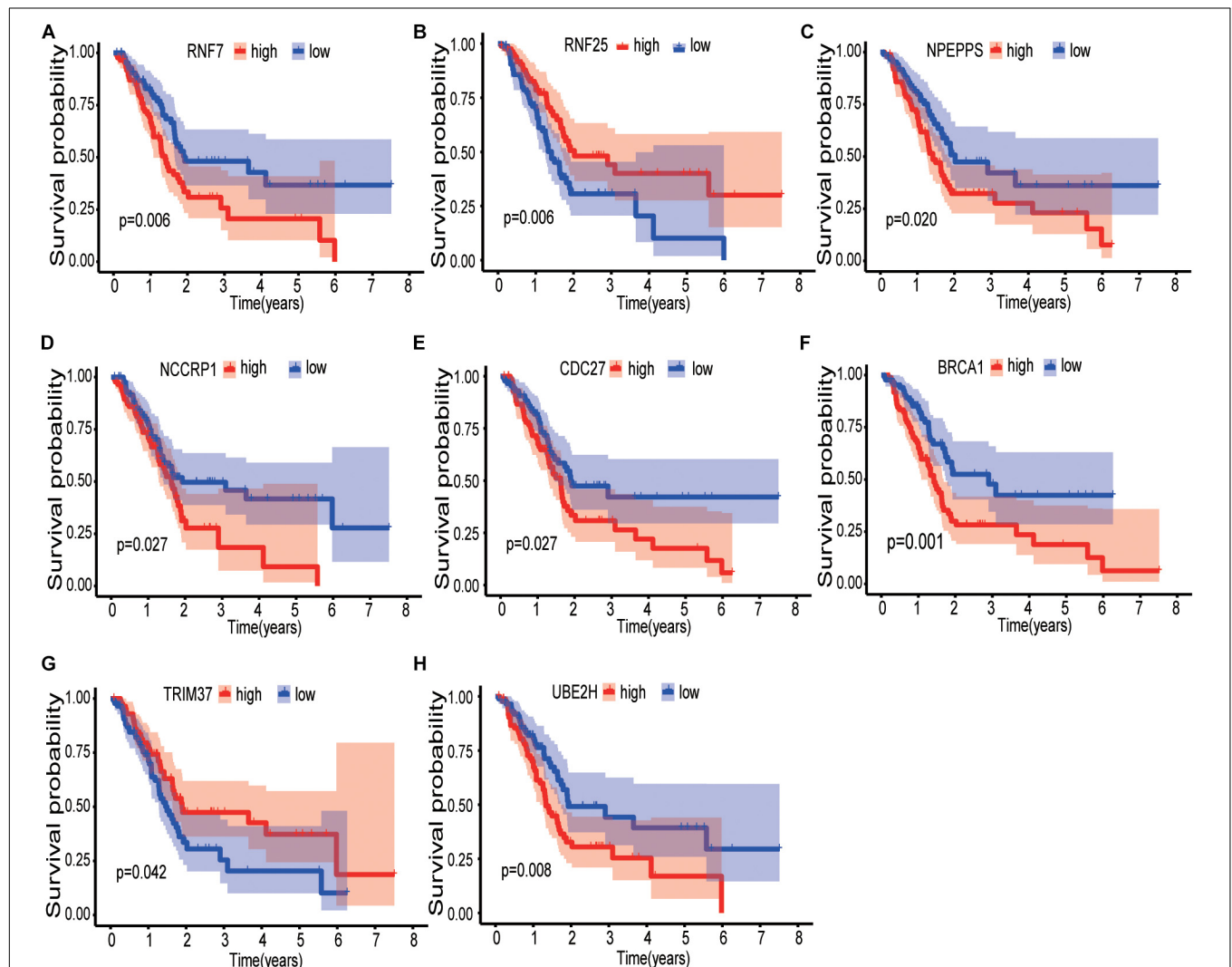
## Construction of the Prognostic Risk Model in Pancreatic Cancer Cohort

Finally, 171 pancreatic cancer patients were included in the risk model. The computational formula was as follows: Risk score =  $(2.3538 \times \text{expression of RNF7}) + (-1.0029 \times \text{expression of NPEPPS}) + (0.2271 \times \text{expression of NCCRP1}) + (1.1898 \times \text{expression of BRCA1}) + (-1.6370 \times \text{expression of TRIM37}) + (-1.5668 \times \text{expression of RNF25}) + (1.9902 \times \text{expression of CDC27}) + (1.0606 \times \text{expression of UBE2H})$ .

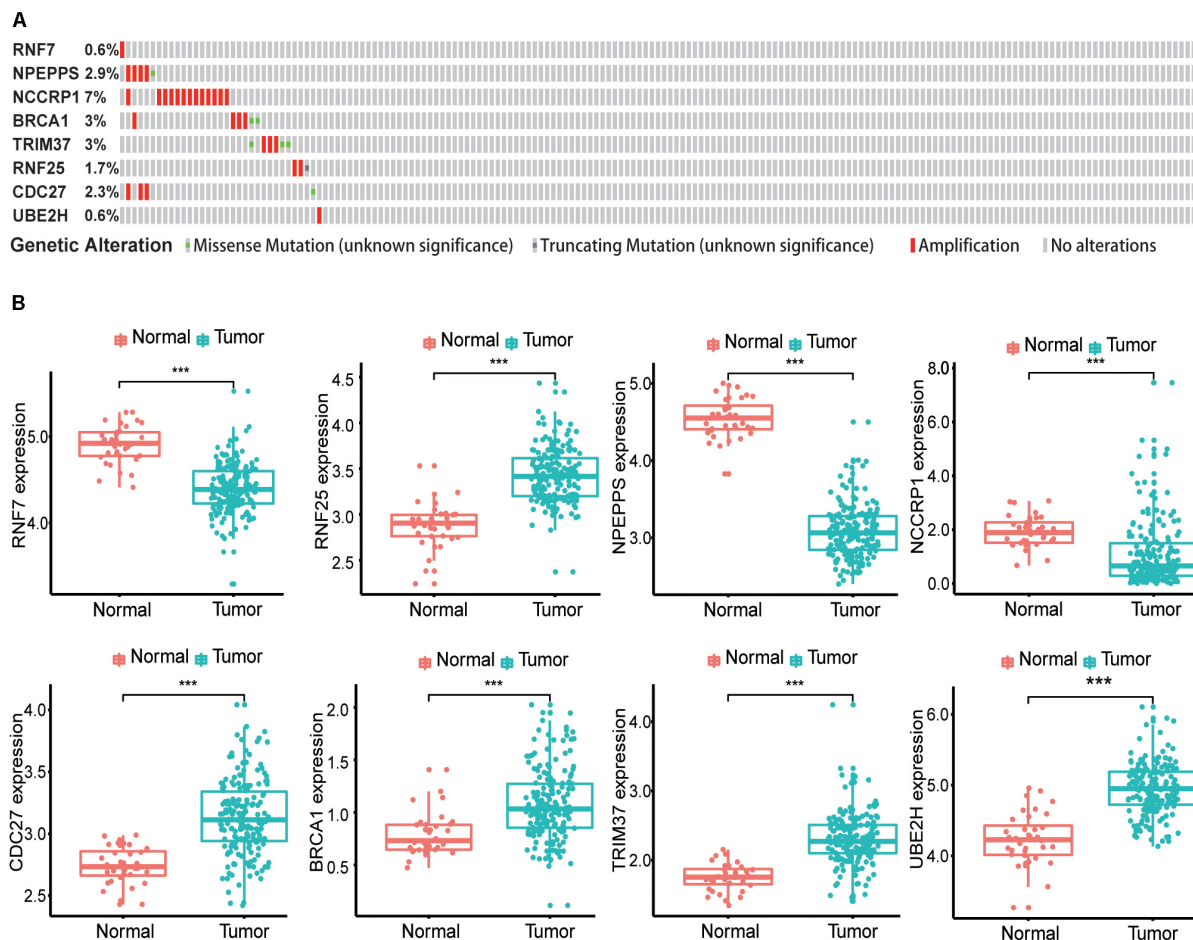
Patients were divided into two groups, the high-risk group ( $n = 85$ ) and the low-risk group ( $n = 86$ ). The high-risk group had a worse outcome than the low-risk group ( $p < 0.001$ ). The 1- and 3-year OS of pancreatic cancer patients in the high-risk group were 87.7 and 64.7%, respectively, while the corresponding OS in the low-risk group was 57.5 and 17.9%, respectively. The AUC (ROC) value of the risk model in 1-year, and 3-year were 0.756, and 0.810, respectively (**Figures 4A,B**). Then, risk scores of these pancreatic cancer patients were ranked and their distribution was analyzed. We divided pancreatic cancer patients into low-risk and high-risk groups by the median risk score for all patients enrolled in the study (**Figure 4C**). The survival status of each patient in the pancreatic cancer patients was shown in **Figure 4D**. As can be intuitively seen from **Figure 4D**, the higher the risk score, the shorter the OS of pancreatic cancer patients.

**TABLE 3 |** The detailed information of eight prognostic mRNAs significantly associated with the prognosis of pancreatic cancer patients.

mRNA	Official name	Ensemble ID	Location	$\beta$ (Cox)	HR (95% CI)	p-value
RNF7	Ring finger protein 7	ENSG00000114125	Chr3: 141, 738, 209-141, 747, 560	2.3538	10.526 (3.759, 29.475)	<0.001
NPEPPS	Aminopeptidase puromycin sensitive	ENSG00000141279	Chr17: 47, 522, 933-47, 623, 276	-1.0029	0.367 (0.154, 0.875)	0.024
NCCRP1	NCCRP1, F-box associated domain containing	ENSG00000188505	Chr19: 39, 196, 964-39, 201, 884	0.2271	1.255 (1.061, 1.484)	0.008
BRCA1	BRCA1 DNA repair associated	ENSG0000012048	Chr17: 43, 044, 295-43, 125, 364	1.1898	3.286 (1.543, 7.001)	0.002
TRIM37	Tripartite motif containing 37	ENSG00000108395	Chr17: 58, 968, 010-59, 106, 880	-1.6370	0.195 (0.081, 0.469)	<0.001
RNF25	Ring finger protein 25	ENSG00000163481	Chr2: 218, 663, 874-218, 672, 002	-1.5668	0.209 (0.087, 0.502)	<0.001
CDC27	Cell division cycle 27	ENSG00000004897	Chr17: 47, 117, 703-47, 189, 295	1.9902	7.317 (2.672, 20.037)	<0.001
UBE2H	Ubiquitin conjugating enzyme E2 H	ENSG00000186591	Chr7: 129, 830, 732-129, 952, 960	1.0606	2.888 (1.392, 5.991)	0.004



**FIGURE 2 |** Kaplan-Meier curves of the effect of the gene expression level of the risk genes (RNF7, NPEPPS, NCCRP1, BRCA1, TRIM37, RNF25, CDC27, and UBE2H) on the prognosis of pancreatic cancer patients. **(A)** Kaplan-Meier curve of the effect of RNF7 gene expression level. **(B)** Kaplan-Meier curve of the effect of RNF25 gene expression level. **(C)** Kaplan-Meier curve of the effect of NPEPPS gene expression level. **(D)** Kaplan-Meier curve of the effect of NCCRP1 gene expression level. **(E)** Kaplan-Meier curve of the effect of CDC27 gene expression level. **(F)** Kaplan-Meier curve of the effect of BRCA1 gene expression level. **(G)** Kaplan-Meier curve of the effect of TRIM37 gene expression level. **(H)** Kaplan-Meier curve of the effect of UBE2H gene expression level.



**FIGURE 3 |** Identification of mRNAs associated with patient survival. **(A)** The alteration proportion for the eight selected genes in 175 clinical samples of pancreatic cancer in the cBioPortal database. **(B)** Different expression of eight genes in the normal pancreatic tissues and tumor tissues based on TCGA database. (\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ).

## Independent Prognostic Value of the Risk Model in the Entire Pancreatic Cancer Cohort

A total of 163 pancreatic cancer patients were included in this analysis. Results of the univariate analysis showed that age, pathological grade, T-stage, N-stage, and risk score were significantly correlated with the prognosis of pancreatic cancer patients. The result of multivariate analysis showed that the risk score was independently correlated with the OS for patients with pancreatic cancer (Table 4).

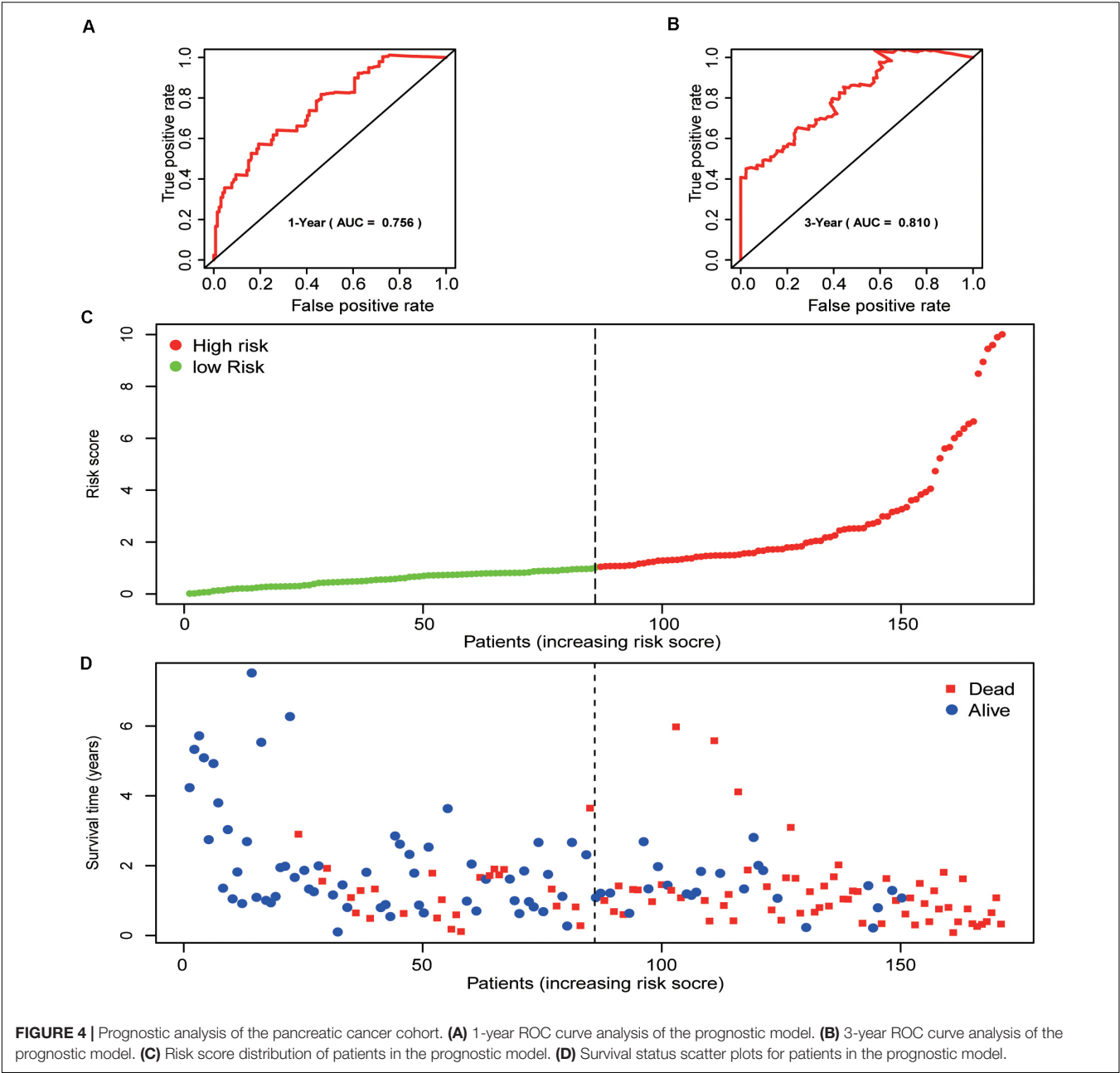
## Validation of the Eight-mRNA Signature in Predicting Survival Using Kaplan–Meier Curves

The results of the univariate analysis showed that age was an independent prognostic factor for pancreatic cancer, and the X-tile software found that 62 and 76 were the optimal cut-off values for the prognosis of pancreatic cancer patients (Supplementary Figure 1). The result of Kaplan–Meier curves

showed the effects of age, gender, histological grade, T-stage, N-stage, and risk score on the prognosis of pancreatic cancer patients (Figures 5A–F). The result of Kaplan–Meier curves showed that our risk model was a stable predictive tool for the prognosis of pancreatic cancer patients stratified by age (<62, 62–76, and >76), gender (male and female), pathological grade (G1/2, or G3/4), T-stage (T1/2, or T3/4), and N-stage (N0 or N1) (Figures 6A–K). Patients with pancreatic cancer in the high-risk group had significantly shorter OS than those in the low-risk group when the patients were stratified into different subgroups based on age, gender, pathological grade, T-stage, and N-stage.

## Validation of the Risk Genes

The protein levels of immunohistochemistry (IHC) staining obtained from the HPA database showed that the expression of the protein in four risk genes (BRCA1, TRIM37, RNF25, and UBE2H) was significantly higher in pancreatic cancer tissues than in normal pancreatic tissues, three genes (RNF7, NPEPPS, and NCCRP1) do the opposite, which was consistent with that at the transcriptional level. Only CDC27 protein expression levels was

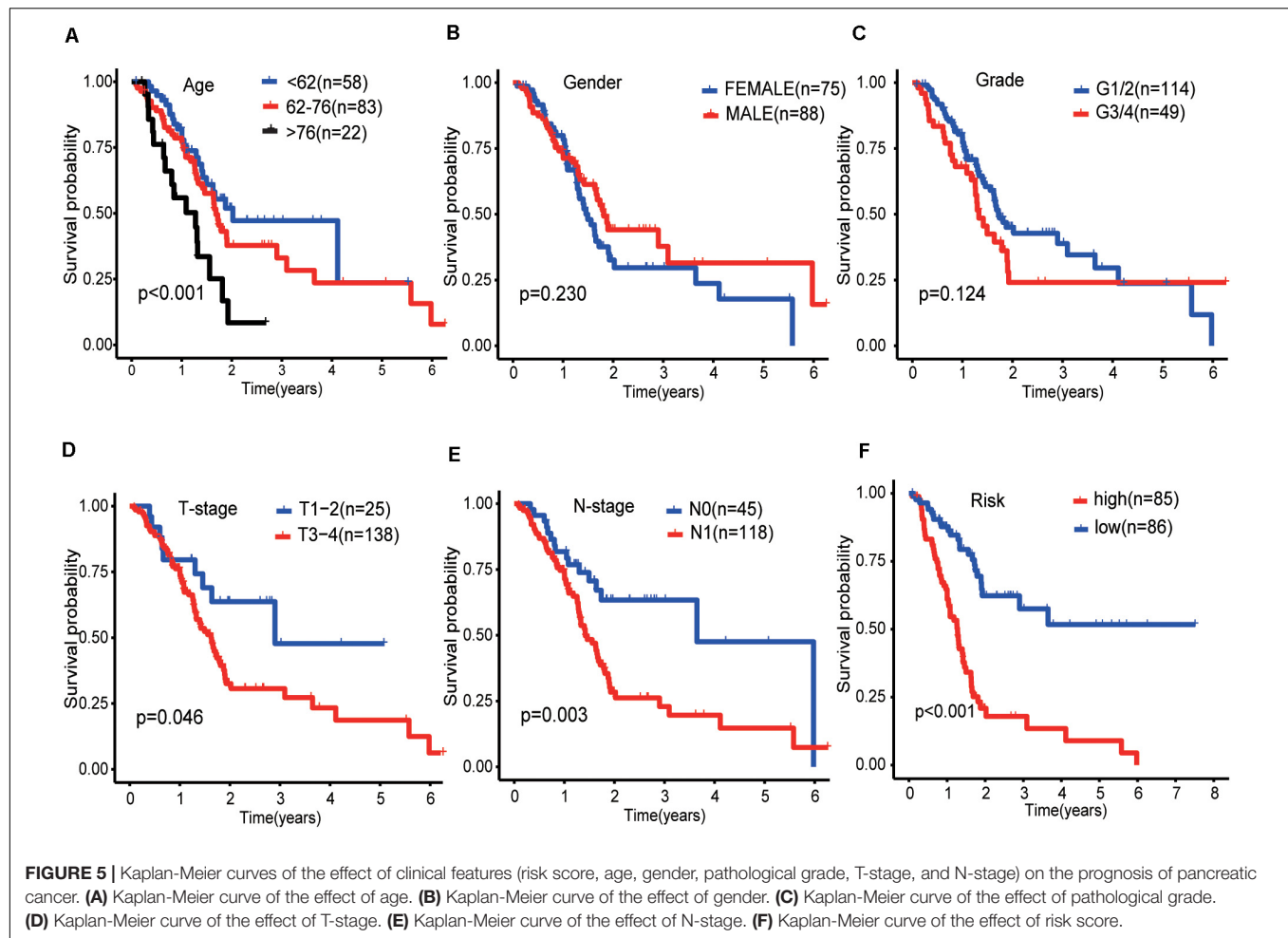


**FIGURE 4 |** Prognostic analysis of the pancreatic cancer cohort. **(A)** 1-year ROC curve analysis of the prognostic model. **(B)** 3-year ROC curve analysis of the prognostic model. **(C)** Risk score distribution of patients in the prognostic model. **(D)** Survival status scatter plots for patients in the prognostic model.

**TABLE 4 |** Effects of various clinical features on pancreatic cancer patients.

Clinical feature	Number	Univariate analysis			Multivariate analysis		
		HR	95% CI	p-value	HR	95% CI	p-value
Age (<62/62–76/>76)	58/83/22	1.028	1.005–1.052	0.016	1.550	0.919–2.614	0.100
Gender (female/male)	75/88	0.768	0.499–1.183	0.232	0.825	0.530–1.285	0.394
Grade (G1/2/G3/4)	114/49	1.387	1.001–1.924	0.049	1.243	0.777–1.990	0.364
N-stage (N0/N1)	45/118	2.004	0.999–4.021	0.050	1.229	0.589–2.563	0.583
T-stage (T1/2/T3/4)	25/138	2.222	1.286–3.838	0.004	1.598	0.890–2.869	0.116
Risk score (low/high)	86/85	1.284	1.207–1.367	<0.001	1.250	1.172–1.333	<0.001





high in both the normal and tumor group in the HPA database (Figures 7A–H).

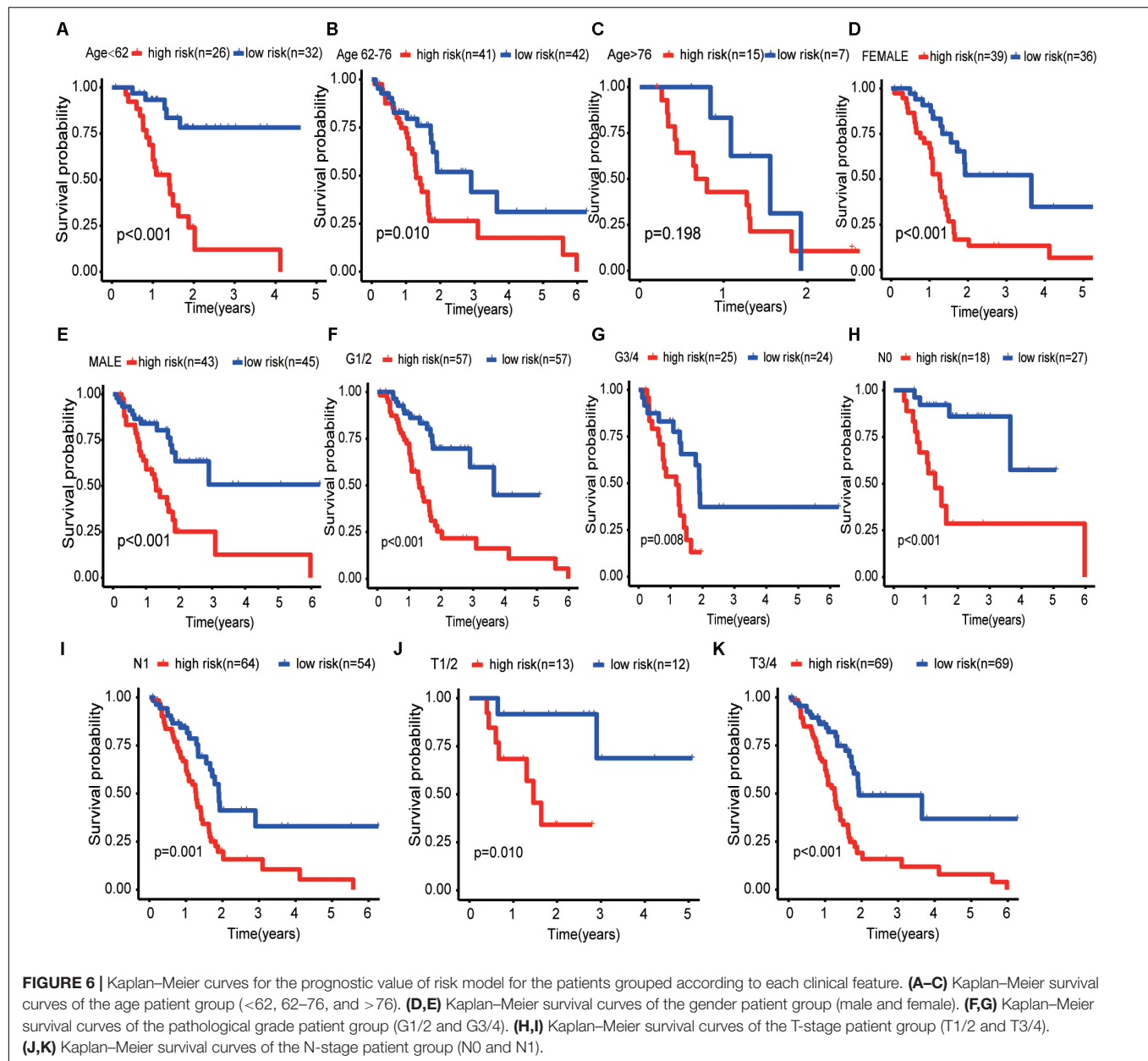
## DISCUSSION

One or more pathway data sets are used to assess the ranking list of statistically significant genes/proteins using GSEA. GSEA can not only detect statistically significant genes and proteins group-wise but also enrich the previous research characteristics of gene sets in functional genomes in a large database of pathway gene sets (Subramanian et al., 2005; Wu et al., 2014). In our study, mRNA expression data from 178 patients with pancreatic cancer and 36 normal pancreatic tissues were used for GSEA analysis, and significant differences were found in two functions. These two functions are all related to ubiquitination, indicating that ubiquitination changes significantly in the development of pancreatic cancer. And then, these ubiquitination-related genes in the two functions were selected for subsequent analysis.

Combined with GO enrichment analysis and KEGG enrichment analysis, the results suggest that these genes are closely related to the ubiquitination process of pancreatic

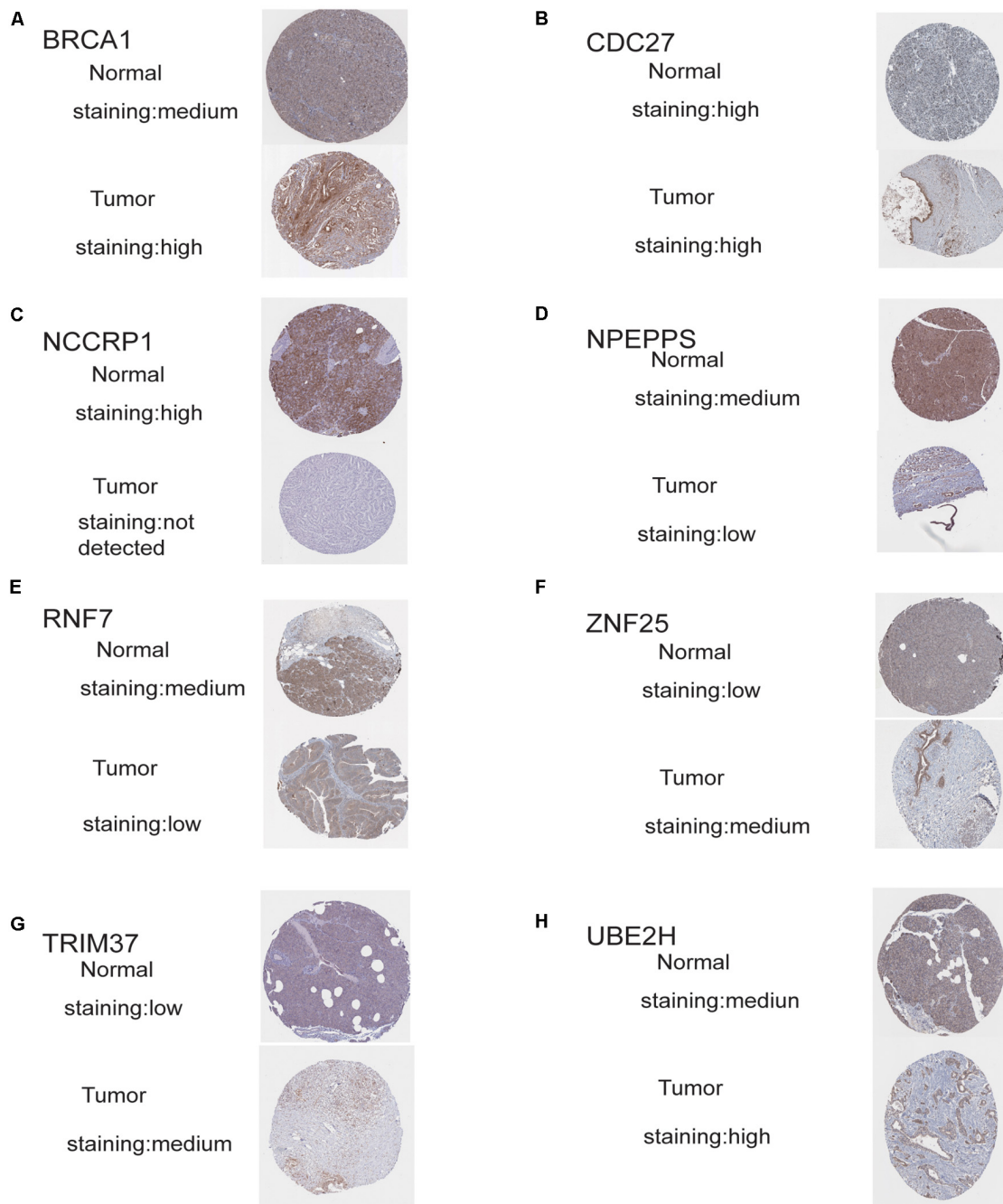
cancer. Next, eight optimal ubiquitination-related genes were identified via multivariate Cox proportional hazards regression analysis, and they were used to construct a risk model. The reliability and stability of the model were further validated. The results showed that the model could accurately distinguish pancreatic cancer patients with different survival outcomes. The results of univariate and multivariate analysis showed that our model could independently predict the outcome of pancreatic cancer patients. The result of Kaplan–Meier curves shows that our risk model has excellent stability and reliability in predicting the prognosis of pancreatic cancer at all age, gender, pathological grade, T-stage, and N-stage. Therefore, our risk model can screen high-risk patients for personalized treatment. Finally, the eight risk genes were validated by the HPA database, and the results showed that the protein expression level of the eight genes was generally consistent with those at the transcriptional level. These results suggest that the genes we identified deserve further study.

Of the eight genes identified, seven genes (RNF7, NCCRP1, BRCA1, TRIM37, RNF25, CDC27, and UBE2H) have been reported to play roles in ubiquitination (Asamitsu et al., 2003; Kallio et al., 2011; Link et al., 2016; Cho et al., 2018; Lim and Joo, 2020; Meitinger et al., 2020; Zhang et al., 2020). It



has not been reported that NPEPPS directly participates in the process of ubiquitination, but NPEPPS is also known to degrade the tau protein, which accumulates and polymerizes in some neurodegenerative diseases (Kudo et al., 2011). In our study, the expression of these ubiquitination-related genes was significantly associated with the prognosis of patients with pancreatic cancer, providing us with a new key to the study of pancreatic cancer. Among these genes, some have been studied as biomarkers for cancer. For example, BRCA has been proved to be a biomarker in many cancers, and its mutation or not has a guiding role in the application of targeted drugs, such as pancreatic cancer (Wu and Shi, 2020). RNF7, an apoptosis-sensitive gene, has been shown in several previous studies to play an important role in the development and progression

of tumors such as prostate cancer and lung cancer (Li et al., 2014; Tan et al., 2016). There are also relevant studies showing that RNF7 regulates ionizing radiation-induced apoptosis in pancreatic cancer (Kim et al., 2011). TRIM37 has also been shown to promote the proliferation, invasion and migration in breast cancer, lung cancer, gastric cancer, glioma, and pancreatic cancer (Jiang et al., 2016; Li et al., 2018; Tang et al., 2018; Hu et al., 2019; Fu et al., 2020). CDC27 promotes the progression and affects PD-L1 expression in T-cell lymphoblastic lymphoma, and also promotes epithelial-to-mesenchymal transition in colorectal cancer (Qiu et al., 2017; Song et al., 2020). There are few studies on the role of NCCRP1, RNF25, and UBE2H in cancer, but the existing research results suggest that these three genes also have the potential to become new tumor biomarkers



**FIGURE 7 |** Validation of risk genes at the translational level. **(A)** Validation of BRCA1 by The Human Protein Atlas database (IHC). **(B)** Validation of CDC27 by The Human Protein Atlas database (IHC). **(C)** Validation of NCCRP1 by The Human Protein Atlas database (IHC). **(D)** Validation of NPEPPS by The Human Protein Atlas database (IHC). **(E)** Validation of RNF7 by The Human Protein Atlas database (IHC). **(F)** Validation of ZNF25 by The Human Protein Atlas database (IHC). **(G)** Validation of TRIM37 by The Human Protein Atlas database (IHC). **(H)** Validation of UBE2H by The Human Protein Atlas database (IHC).

or targets for cancer (Miwa et al., 2017; Cho et al., 2018; Zhu et al., 2018).

Of the eight genes we identified, three genes (RNF7, NPEPPS, and NCCRP1) were down-regulated and the remaining five (BRCA1, TRIM37, RNF25, CDC27, and UBE2H) were up-regulated in tumor tissue compared to normal

pancreatic tissue. But we found that even though some genes (RNF7, NPEPPS, and NCCRP1) were down-regulated in tumor group, patients with pancreatic cancer with high expression of these genes had a worse prognosis. Some genes are up-regulated (TRIM37 and RNF25), but high expression of these genes has a better prognosis. So we suspect that these genes play an opposite



role in the development and progression of pancreatic cancer. For example, NPEPPS may inhibit tumor formation in normal tissue but may promote tumor progression once the tumor has formed. This phenomenon has been reported in previous literature. In retrospect, the study has shown that TGF- $\beta$  is a key negative regulator of cell proliferation, but the abnormal function of retinoblastoma protein can lead to the inhibition of the function of TGF- $\beta$  and promote the progression of pancreatic cancer (Gore et al., 2014). Another study showed that Daple is also a tumor-suppressor gene, although it appears only in the early stages of cancer to function as a tumor-suppressor gene. In the later stages of cancer, when cancer cells escape from their primary sites and circulate in the blood, the expression of Daple makes cancer cells more aggressive and more likely to spread (Aznar et al., 2015).

Many previous studies have explored new potential biomarkers and therapeutic targets for pancreatic cancer through bioinformatic methods. Wu et al. (2019) screened nine DEGs (MET, KLK10, COL17A1, CEP55, ANKRD22, ITGB6, ARNTL2, MCOLN3, and SLC25A45) through the joint analysis of GEO and TCGA databases and construct a risk score model. They also analyzed the relationship between the nine gene models and tumor immune infiltration. Wei et al. (2019) constructed a risk model to predict the prognosis of pancreatic cancer patients by screening nine immune-related lncRNAs from the TCGA database. Compared with the previous studies, we use GSEA enrichment analysis to explore the function of ubiquitination in pancreatic cancer, and on this basis, identify eight ubiquitination-related genes to construct a risk model. There has been no previous study on the bioinformatics related to the ubiquitination of pancreatic cancer, and our study provides a new idea for relevant studies on the progression of pancreatic cancer.

Of course, our study also has some shortcomings. First, our study was a retrospective study based on a public database. The data we used has not been validated by prospective clinical trials. Besides, the identified mechanism of ubiquitination-related genes affecting the development of pancreatic cancer needs further support from basic experimental studies. Next, we need to collect clinical specimens and data for subsequent studies.

## CONCLUSION

Using GSEA enrichment analysis, we found that the ubiquitination-related functions of pancreatic cancer were

significantly different from those of normal pancreatic tissues. Subsequently, we extracted and screened the genes in these functions, and finally selected eight genes significantly related to the prognosis of pancreatic cancer patients as risk genes to construct a risk model. This model has a good predictive effect on the prognosis of pancreatic cancer patients. Moreover, these eight genes have the potential to be further studied as new biomarkers or therapeutic targets for pancreatic cancer.

## DATA AVAILABILITY STATEMENT

The datasets supporting the conclusions of this article are obtained from The Cancer Genome Atlas (TCGA) portal website (<https://portal.gdc.cancer.gov/>) and the Genotype-Tissue Expression (GTEx) database (<https://xenabrowser.net/>). The alteration of these genes is from an online database (<http://www.cbioportal.org/>). The authors did not have special access privileges.

## AUTHOR CONTRIBUTIONS

NL and QS: conception and design. HZ and LC: data acquisition, data analysis and interpretation, and article writing and revision. All authors have read and agreed to the published version of the manuscript.

## FUNDING

This research was supported by National Natural Science Foundation of China (81802980).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.612196/full#supplementary-material>

**Supplementary Figure 1** | The result of the best cut-off values for the age by using the X-tile software.

## REFERENCES

- Adamska, A., Elaskalani, O., Emmanouilidi, A., Kim, M., Abdol Razak, N. B., Metharom, P., et al. (2018). Molecular and cellular mechanisms of chemoresistance in pancreatic cancer. *Adv. Biol. Regul.* 68, 77–87. doi: 10.1016/j.jbior.2017.11.007
- Asamitsu, K., Tetsuka, T., Kanazawa, S., and Okamoto, T. (2003). RING finger protein AO7 supports NF-kappaB-mediated transcription by interacting with the transactivation domain of the p65 subunit. *J. Biol. Chem.* 278, 26879–26887. doi: 10.1074/jbc.M211831200
- Aznar, N., Midde, K. K., Dunkel, Y., Lopez-Sanchez, I., Pavlova, Y., Marivin, A., et al. (2015). Daple is a novel non-receptor GEF required for trimeric G protein activation in Wnt signaling. *eLife* 4:e07091. doi: 10.7554/eLife.07091
- Chen, L., Yuan, R., Wen, C., Liu, T., Feng, Q., Deng, X., et al. (2020). E3 ubiquitin ligase UBR5 promotes pancreatic cancer growth and aerobic glycolysis by downregulating FBP1 via destabilization of C/EBPalpha. *Oncogene* doi: 10.1038/s41388-020-01527-1
- Cho, J. H., You, Y. M., Yeom, Y. I., Lee, D. C., Kim, B. K., Won, M., et al. (2018). RNF25 promotes gefitinib resistance in EGFR-mutant NSCLC cells by inducing NF-kappaB-mediated ERK reactivation. *Cell Death Dis.* 9:587. doi: 10.1038/s41419-018-0651-5

- Cho, S. H., Pak, K., Jeong, D. C., Han, M. E., Oh, S. O., and Kim, Y. H. (2019). The AP2M1 gene expression is a promising biomarker for predicting survival of patients with hepatocellular carcinoma. *J. Cell Biochem.* 120, 4140–4146. doi: 10.1002/jcb.27699
- Dikic, I., Wakatsuki, S., and Walters, K. J. (2009). Ubiquitin-binding domains - from structures to functions. *Nat. Rev. Mol. Cell Biol.* 10, 659–671. doi: 10.1038/nrm2767
- Fu, T., Ji, K., Jin, L., Zhang, J., Wu, X., Ji, X., et al. (2020). ASB16-AS1 up-regulated and phosphorylated TRIM37 to activate NF-kappaB pathway and promote proliferation, stemness, and cisplatin resistance of gastric cancer. *Gas. Cancer* doi: 10.1007/s10120-020-01096-y [Epub ahead of print].
- Gore, A. J., Deitz, S. L., Palam, L. R., Craven, K. E., and Korc, M. (2014). Pancreatic cancer-associated retinoblastoma 1 dysfunction enables TGF-beta to promote proliferation. *J. Clin. Invest.* 124, 338–352. doi: 10.1172/JCI71526
- Hershko, A., Ciechanover, A., and Rose, I. A. (1979). Resolution of the ATP-dependent proteolytic system from reticulocytes: a component that interacts with ATP. *Proc. Natl. Acad. Sci. U.S.A.* 76, 3107–3110. doi: 10.1073/pnas.76.7.3107
- Hu, X., Xiang, D., Xie, Y., Tao, L., Zhang, Y., Jin, Y., et al. (2019). LSD1 suppresses invasion, migration and metastasis of luminal breast cancer cells via activation of GATA3 and repression of TRIM37 expression. *Oncogene* 38, 7017–7034. doi: 10.1038/s41388-019-0923-2
- Ilic, M., and Ilic, I. (2016). Epidemiology of pancreatic cancer. *World J. Gastroenterol.* 22, 9694–9705. doi: 10.3748/wjg.v22.i44.9694
- Jiang, J., Tian, S., Yu, C., Chen, M., and Sun, C. (2016). TRIM37 promoted the growth and migration of the pancreatic cancer cells. *Tumour Biol.* 37, 2629–2634. doi: 10.1007/s13277-015-4078-7
- Kallio, H., Tolvanen, M., Janis, J., Pan, P. W., Laurila, E., Kallioniemi, A., et al. (2011). Characterization of non-specific cytotoxic cell receptor protein 1: a new member of the lectin-type subfamily of F-box proteins. *PLoS One* 6:e27152. doi: 10.1371/journal.pone.0027152
- Kim, S. Y., Yang, E. S., Lee, Y. S., Lee, J., and Park, J. W. (2011). Sensitive to apoptosis gene protein regulates ionizing radiation-induced apoptosis. *Biochimie* 93, 269–276. doi: 10.1016/j.biochi.2010.09.020
- Kudo, L. C., Parfenova, L., Ren, G., Vi, N., Hui, M., Ma, Z., et al. (2011). Puromycin-sensitive aminopeptidase (PSA/NPEPPS) impedes development of neuropathology in hPSA/TAU(P301L) double-transgenic mice. *Hum. Mol. Genet.* 20, 1820–1833. doi: 10.1093/hmg/ddr065
- Li, H., Tan, M., Jia, L., Wei, D., Zhao, Y., Chen, G., et al. (2014). Inactivation of SAG/RBX2 E3 ubiquitin ligase suppresses KrasG12D-driven lung tumorigenesis. *J. Clin. Invest.* 124, 835–846. doi: 10.1172/JCI70297
- Li, J. P., Li, R., Liu, X., Huo, C., Liu, T. T., Yao, J., et al. (2020). A seven immune-related lncRNAs model to increase the predicted value of lung Adenocarcinoma. *Front. Oncol.* 10:560779. doi: 10.3389/fonc.2020.560779
- Li, Y., Deng, L., Zhao, X., Li, B., Ren, D., Yu, L., et al. (2018). Tripartite motif-containing 37 (TRIM37) promotes the aggressiveness of non-small-cell lung cancer cells by activating the NF-kappaB pathway. *J. Pathol.* 246, 366–378. doi: 10.1002/path.5144
- Lian, J., Liu, C., Guan, X., Wang, B., Yao, Y., Su, D., et al. (2020). Ubiquitin specific peptidase 5 enhances STAT3 signaling and promotes migration and invasion in pancreatic cancer. *J. Cancer* 11, 6802–6811. doi: 10.7150/jca.48536
- Lim, K. H., and Joo, J. Y. (2020). Predictive potential of circulating Ube2h mRNA as an E2 ubiquitin-conjugating enzyme for diagnosis or treatment of Alzheimer's disease. *Int. J. Mol. Sci.* 21:3398. doi: 10.3390/ijms21093398
- Link, L. A., Howley, B. V., Hussey, G. S., and Howe, P. H. (2016). PCBP1/HNRNP E1 protects chromosomal integrity by translational regulation of CDC27. *Mol. Cancer Res.* 14, 634–646. doi: 10.1158/1541-7786.MCR-16-0018
- Meitinger, F., Ohta, M., Lee, K. Y., Watanabe, S., Davis, R. L., Anzola, J. V., et al. (2020). TRIM37 controls cancer-specific vulnerability to PLK4 inhibition. *Nature* 585, 440–446. doi: 10.1038/s41586-020-2710-1
- Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., et al. (2019). Cancer treatment and survivorship statistics, 2019. *CA Cancer J. Clin.* 69, 363–385. doi: 10.3322/caac.21565
- Miwa, T., Kanda, M., Koike, M., Iwata, N., Tanaka, H., Umeda, S., et al. (2017). Identification of NCCRP1 as an epigenetically regulated tumor suppressor and biomarker for malignant phenotypes of squamous cell carcinoma of the esophagus. *Oncol. Lett.* 14, 4822–4828. doi: 10.3892/ol.2017.6753
- Qiu, L., Tan, X., Lin, J., Liu, R. Y., Chen, S., Geng, R., et al. (2017). CDC27 induces metastasis and invasion in colorectal cancer via the promotion of epithelial-to-mesenchymal transition. *J. Cancer* 8, 2626–2635. doi: 10.7150/jca.19381
- Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387
- Song, Y., Song, W., Li, Z., Song, W., Wen, Y., Li, J., et al. (2020). CDC27 promotes tumor progression and affects PD-L1 expression in T-Cell Lymphoblastic lymphoma. *Front. Oncol.* 10:488. doi: 10.3389/fonc.2020.00488
- 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
- Tan, M., Xu, J., Siddiqui, J., Feng, F., and Sun, Y. (2016). Depletion of SAG/RBX2 E3 ubiquitin ligase suppresses prostate tumorigenesis via inactivation of the PI3K/AKT/mTOR axis. *Mol. Cancer* 15:81. doi: 10.1186/s12943-016-0567-6
- Tang, S. L., Gao, Y. L., and Wen-Zhong, H. (2018). Knockdown of TRIM37 suppresses the proliferation, migration and invasion of glioma cells through the inactivation of PI3K/Akt signaling pathway. *Biomed. Pharmacother.* 99, 59–64. doi: 10.1016/j.biopha.2018.01.054
- Wei, C., Liang, Q., Li, X., Li, H., Liu, Y., Huang, X., et al. (2019). Bioinformatics profiling utilized a nine immune-related long noncoding RNA signature as a prognostic target for pancreatic cancer. *J. Cell Biochem.* 120, 14916–14927. doi: 10.1002/jcb.28754
- Welchman, R. L., Gordon, C., and Mayer, R. J. (2005). Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell Biol.* 6, 599–609. doi: 10.1038/nrm1700
- Wu, B., and Shi, L. (2020). Cost-effectiveness of maintenance olaparib for germline BRCA-mutated metastatic pancreatic cancer. *J. Natl. Compr. Cancer Netw.* 18, 1528–1536. doi: 10.6004/jnccn.2020.7587
- Wu, M., Li, X., Zhang, T., Liu, Z., and Zhao, Y. (2019). Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. *Front. Oncol.* 9:996. doi: 10.3389/fonc.2019.00996
- Wu, X., Hasan, M. A., and Chen, J. Y. (2014). Pathway and network analysis in proteomics. *J. Theor. Biol.* 362, 44–52. doi: 10.1016/j.jtbi.2014.05.031
- Xia, J., Gill, E. E., and Hancock, R. E. (2015). Network analyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* 10, 823–844. doi: 10.1038/nprot.2015.052
- Yang, C., Zhu, S., Yang, H., Deng, S., Fan, P., Li, M., et al. (2019). USP44 suppresses pancreatic cancer progression and overcomes gemcitabine resistance by deubiquitinating FBP1. *Am. J. Cancer Res.* 9, 1722–1733.
- 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, F., Yan, P., Yu, H., Le, H., Li, Z., Chen, J., et al. (2020). L ARP7 Is a BRCA1 ubiquitinase substrate and regulates genome stability and Tumorigenesis. *Cell Rep.* 32:107974. doi: 10.1016/j.celrep.2020.107974
- Zhu, Y. C., Wang, W. X., Song, Z. B., Zhang, Q. X., Xu, C. W., Chen, G., et al. (2018). MET-UBE2H fusion as a novel mechanism of acquired EGFR resistance in lung Adenocarcinoma. *J. Thorac. Oncol.* 13, e202–e204. doi: 10.1016/j.jtho.2018.05.009

**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.

Copyright © 2020 Zuo, Chen, Li and Song. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Development and Validation of a Combined Model for Preoperative Prediction of Lymph Node Metastasis in Peripheral Lung Adenocarcinoma

Qi Li<sup>1†</sup>, Xiao-qun He<sup>1†</sup>, Xiao Fan<sup>2</sup>, Chao-nan Zhu<sup>3</sup>, Jun-wei Lv<sup>3‡</sup> and Tian-you Luo<sup>1\*‡</sup>

## OPEN ACCESS

### Edited by:

David A. Hormuth, II,  
The University of Texas at Austin,  
United States

### Reviewed by:

Shaofeng Duan,  
GE Healthcare, China  
Maria F. Chan,  
Memorial Sloan Kettering Cancer  
Center, United States

### \*Correspondence:

Tian-you Luo  
lty89011721@sina.com

<sup>†</sup>These authors have contributed  
equally to this work and share  
first authorship

<sup>‡</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Thoracic Oncology,  
a section of the journal  
Frontiers in Oncology

Received: 04 March 2021

Accepted: 23 April 2021

Published: 24 May 2021

### Citation:

Li Q, He X-q, Fan X, Zhu C-n, Lv J-w  
and Luo T-y (2021) Development and  
Validation of a Combined Model for  
Preoperative Prediction of Lymph  
Node Metastasis in Peripheral  
Lung Adenocarcinoma.  
Front. Oncol. 11:675877.  
doi: 10.3389/fonc.2021.675877

<sup>1</sup> Department of Radiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, <sup>2</sup> Department of Radiology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing, China, <sup>3</sup> Hangzhou YITU Healthcare Technology, Hangzhou, China

**Background:** Based on the “seed and soil” theory proposed by previous studies, we aimed to develop and validate a combined model of machine learning for predicting lymph node metastasis (LNM) in patients with peripheral lung adenocarcinoma (PLADC).

**Methods:** Radiomics models were developed in a primary cohort of 390 patients (training cohort) with pathologically confirmed PLADC from January 2016 to August 2018. The patients were divided into the LNM (–) and LNM (+) groups. Thereafter, the patients were subdivided according to TNM stages N0, N1, N2, and N3. Radiomic features from unenhanced computed tomography (CT) were extracted. Radiomic signatures of the primary tumor (R1) and adjacent pleura (R2) were built as predictors of LNM. CT morphological features and clinical characteristics were compared between both groups. A combined model incorporating R1, R2, and CT morphological features, and clinical risk factors was developed by multivariate analysis. The combined model's performance was assessed by receiver operating characteristic (ROC) curve. An internal validation cohort containing 166 consecutive patients from September 2018 to November 2019 was also assessed.

**Results:** Thirty-one radiomic features of R1 and R2 were significant predictors of LNM (all  $P < 0.05$ ). Sex, smoking history, tumor size, density, air bronchogram, spiculation, lobulation, necrosis, pleural effusion, and pleural involvement also differed significantly between the groups (all  $P < 0.05$ ). R1, R2, tumor size, and spiculation in the combined model were independent risk factors for predicting LNM in patients with PLADC, with area under the ROC curves (AUCs) of 0.897 and 0.883 in the training and validation cohorts, respectively. The combined model identified N0, N1, N2, and N3, with AUCs ranging from 0.691–0.927 in the training cohort and 0.700–0.951 in the validation cohort, respectively, thereby indicating good performance.

**Conclusion:** CT phenotypes of the primary tumor and adjacent pleura were significantly associated with LNM. A combined model incorporating radiomic signatures, CT morphological features, and clinical risk factors can assess LNM of patients with PLADC accurately and non-invasively.

**Keywords:** radiomics, lymph node metastasis, computed tomography, lung adenocarcinoma, machine learning

## INTRODUCTION

Despite advances in early detection, diagnosis, staging, and treatment, lung cancer still remains the leading cause of death worldwide (1). Additionally, peripheral lung adenocarcinoma (PLADC), defined as adenocarcinoma occurring below segmental bronchus, is the most common histological subtype of lung cancer (2). Evaluating the status of lymph node metastasis (LNM) accurately is of great benefit to the treatment strategy decision and prognosis of patients with PLADC.

Previous studies (3, 4) have reported a significant association between LNM and computed tomography (CT) features and clinicopathological variables, including tumor centrality, consolidation-to-tumor ratio, age, papillary/micropapillary predominant subtype, and more advanced T stage in non-small cell lung cancer. Some researchers have reported that pleural involvement on preoperative CT images had a moderate correlation with visceral pleural invasion (5, 6). Chang et al. (7) concluded that lymphatic and visceral pleural surface invasion could be used to predict LNM. In other words, previous studies have concluded that pleural involvement was closely related to LNM (5–7). Therefore, we hypothesized that the primary tumor is a “seed,” adjacent pleura is the “soil,” and tumor cells could inseminate systematically through subpleural lymphatics owing to abundant lymphatic and vascular networks within the sub-pleura. Although previous studies have shown that several histological parameters can be predictors of LNM, these evaluation parameters are only available postoperatively. Preoperative knowledge of LNM can provide valuable information for determining the scope of surgical resection and the need of adjuvant therapy (8–10).

Radiomics, the high-throughput extraction of advanced quantitative imaging features from radiographic images, has attracted increased attention of physicians in recent years and has shown promise in characterizing tumor phenotypes,

including imaging diagnosis, treatment, and prediction of prognosis and treatment efficacy of tumors (11–13). Recent studies have recognized the contribution of radiomics in the preoperative assessment of lymph node status in lung cancer (14–17). However, these studies predicted LNM of lung cancer mainly by extracting the quantitative information of the tumor itself. To the best of our knowledge, whether the combination of the radiomic signatures of the primary tumor (R1) and those of adjacent pleura (R2) can produce a superior prediction of LNM for patients with PLADC have not yet been established.

Therefore, the study aim was to develop and validate a combined model that incorporates R1, R2, and CT morphological features and identify clinical risk factors for predicting LNM in patients with PLADC.

## METHODS

### Patient Selection

This study obtained ethical approval from the institutional review board in our hospital, and the need for informed consent was waived due to the retrospective nature of the study. A total of 390 patients with pathologically confirmed PLADC during January 2014 to August 2018 were included as a training cohort. **Data Supplement A1** presents the patient recruitment flowchart as well as the inclusion and exclusion criteria of this study.

Patients in the training cohort were divided into the LNM (+) group ( $n = 228$ ) and LNM (–) group ( $n = 162$ ), with an average age of  $60.36 \pm 9.86$  years (range: 24–83 years). Additionally, all patients were subdivided into N0 ( $n = 162$ , no regional node metastasis), N1 ( $n = 56$ , metastasis in ipsilateral pulmonary or hilar nodes), N2 ( $n = 156$ , metastasis in ipsilateral mediastinal/subcarinal nodes), and N3 ( $n = 13$ , metastasis in contralateral mediastinal/hilar, or supraclavicular nodes) according to the 8th edition of the Tumor–Node–Metastasis (TNM) classification. Clinical characteristics, including age, sex, and smoking history, were collected. In addition, data from 166 consecutive patients with PLADC (N0 = 75, N1 = 19, N2 = 61, N3 = 11) with a mean age of  $60.51 \pm 9.19$  years (range: 42–81 years) in our institution during September 2018 to November 2019 were collected and included as an internal validation cohort.

### CT Image Acquisition and Morphological Features Analysis

Chest CT scan was performed with Discovery 750 HD CT (GE Health care, Milwaukee, WI, USA), and the original images were reconstructed using a medium sharp reconstruction algorithm

**Abbreviations:** LNM, Lymph node metastasis; PLADC, Peripheral lung adenocarcinoma; CT, Computed tomography; R1, Radiomic signature of the primary tumor; R2, Radiomic signature of the adjacent pleura; TNM, Tumor–Node–Metastasis; N0, No regional node metastasis; N1, Metastasis in ipsilateral pulmonary or hilar nodes; N2, Metastasis in ipsilateral mediastinal/subcarinal nodes; N3, Metastasis in contralateral mediastinal/hilar, or supraclavicular nodes; PACS, Picture Archiving and Communication System; HU, Hounsfield units; ROIs, Regions of interest; ICCs, Intraclass correlation coefficients; AIC, Akaike information criterion; ROC, Receiver operating characteristic; AUCs, Area under the ROC curves; RS1, Radiomics score of the primary tumor; RS2, Radiomics score of the adjacent pleura; MRI, Magnetic resonance imaging; FDG-PET/CT, Fluorodeoxyribose positron emission tomography combined with CT; DWI, Diffusion-weighted magnetic resonance imaging.



with a thickness of 0.625–1.25 mm and transmitted to the Picture Archiving and Communication System (PACS). CT features were reviewed in both lung window images (window width: 1600 Hounsfield units [HU]; window level: –600 HU) and mediastinal window images (window width: 400 HU; window level: 40 HU).

A senior radiologist (with 18 years of work experience in thoracic imaging diagnosis) and a junior radiologist (with 13 years of work experience in thoracic imaging diagnosis) reviewed the CT images to reach a consensus. Tumor size (the longest diameter of the tumor on cross-sectional images), tumor density (solid or sub-solid), air space, air bronchogram, lobulation, spiculation, pleural effusion, necrosis, and pleural involvement were measured and evaluated. Referring to the standards established in previous research (3), pleural involvement was classified into three types (**Figures 1–4**): Type I, which manifested as one or more linear shadows between tumor and pleura on lung window images but was not observed on mediastinal window images; Type II, which manifested as linear or cord-like shadows between the tumor and pleura observed in both lung windows and mediastinal window images; and Type III, which were tumors attached to the pleura with a broad base. For tumors with concurrent Type I,

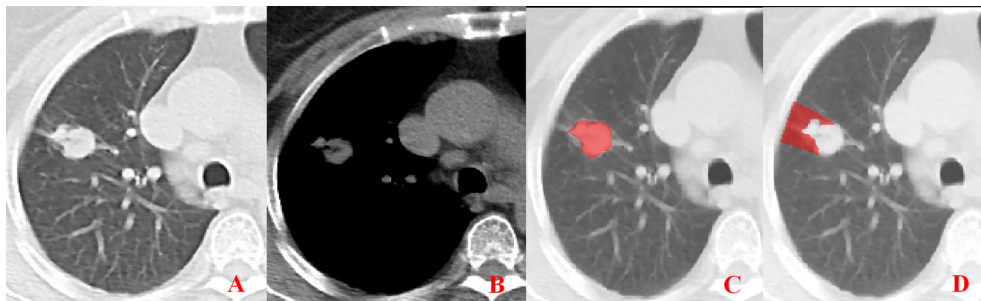
Type II, or Type III presentation, the pleural involvement was recorded as the latter type.

## Radiomic Feature Selection and Signature Building

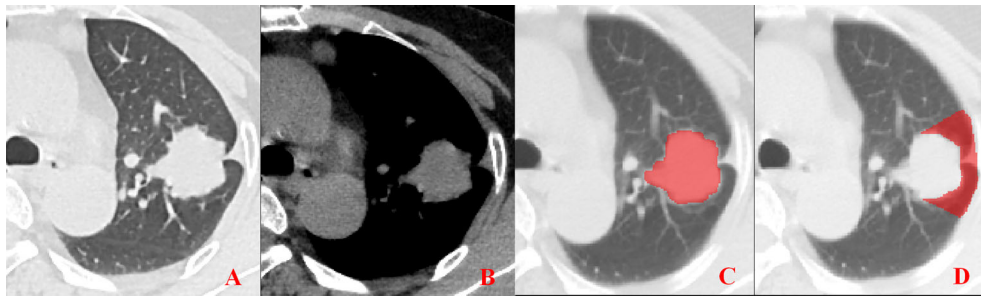
Unenhanced CT images of PLADC were extracted from PACS and then exported to the ITK-SNAP software (version 2.2.0, [www.itk-snap.org](http://www.itk-snap.org)) for manual segmentation. Considering that LNM depends on the synergies of the primary tumor and nearby pleura, both of them are investigated. For the primary tumor, the largest slice of tumor was selected from axial CT images, and regions of interest (ROIs) were carefully drawn on it and adjacent two slices, covering the whole contour of tumor. For all nearby pleura delineation, we tried to avoid the soft tissue and ribs of the chest wall; additionally, all pleural ROI delineation was defined as two lines tangent to the edges of the tumor, intersecting the visceral pleura at 90°. If there was no pleural involvement, ROI was drawn on the region between the primary tumor and pleura on the largest slice of tumor and adjacent two slices; if there was pleural involvement of Type I, Type II, and Type III, three adjacent slices showing the sign of pleural involvement most clearly were selected and delineated (**Figures 1–4**). To ensure consistency, these delineations were performed three times, and



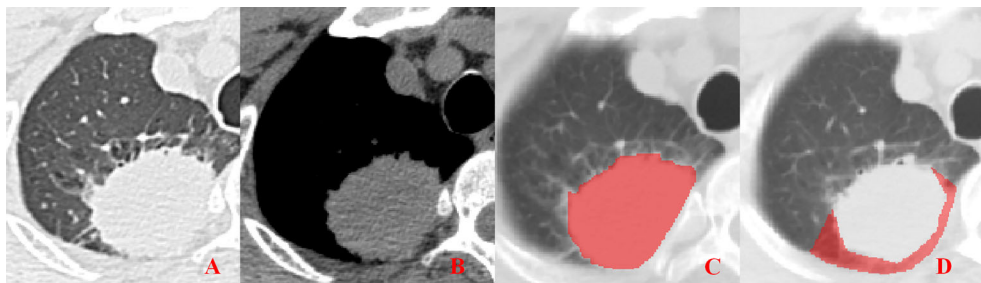
**FIGURE 1** | Representative image showing no pleural involvement. (A, B) No pleural involvement is seen in either the lung window or mediastinal window images. (C, D) ROI delineation of the primary tumor and nearby pleura.



**FIGURE 2** | Pleural involvement of Type I. (A, B) One or more linear shadows are observed between the tumor and pleura in the lung window images but are not observed in the mediastinal window images. (C, D) ROI delineation of the primary tumor and nearby pleura.



**FIGURE 3 |** Pleural involvement of Type II. **(A, B)** Linear or cord-like shadows are observed between the tumor and pleura in both the lung window and mediastinal window images. **(C, D)** ROI delineation of the primary tumor and nearby pleura.



**FIGURE 4 |** Pleural involvement of Type III. **(A, B)** Tumor attached to the pleura with a broad base observed in both the lung window and mediastinal window images. **(C, D)** ROI delineation of the primary tumor and nearby pleura.

reproducibility assessment on intra-reader agreement were assessed by intraclass correlation coefficients (ICCs) for radiomics feature extraction after ROI delineation, ICC > 0.75 were retained as they showed good agreement between different segmentations.

Radiomic feature extraction was performed on PyRadiomic platform implemented in Python (<https://pyradiomics.readthedocs.io/en/latest/>), which can extract radiomic features from CT images *via* an algorithm with a large panel of engineered hard-coded features, such as morphological features (ROI size, volume, surface area, etc.), first-order features (geometric morphology and histogram features), second-order texture features (gray level co-occurrence matrix, gray level long matrix, gray level generation matrix, and neighborhood gray difference matrix), and other features based on filtering and transformation (wavelet transform).

As shown in **Supplementary Figure A2**, radiomic feature selection and signature building of R1 and R2, including these steps, were performed. First, we normalized the resolution feature matrix. For each vector, we calculated the L2 norm and divided it. The feature vector was then mapped to a unit vector. Second, we compared the similarity of each feature pair due to the high dimensionality of the radiomic features space. If the Pearson correlation coefficient of a feature pair was greater than 0.90, we randomly removed one feature pair. Third, we

combined the optimal subset method with a minimum Akaike's Information Criterion (AIC) to select the best combination of features. The optimal subset method can provide the corresponding  $\chi^2$  value in the case where all feature number combinations are different, but it cannot identify the best combination. Therefore, the corresponding AIC values under various combinations could be calculated to find the smallest corresponding AIC value. We built a final logistic regression model using a combination of features under the minimum AIC correspondence. Using this method, we selected features to build the R1 and R2 models. Finally, after traversing five machine-learning algorithms, we chose multinomial logistic regression as the final classifier.

## Radiomics Model Construction and Evaluation

R1 and R2 models that reflected the radiomics signature of the primary tumor and adjacent pleura were established; an R1+R2 model was also constructed as a whole ROI to explore the ability to predict LNM in patients with PLADC. A combined model, including R1 and R2, CT morphological features, and clinical risk factors, was developed by multivariate logistic regression analysis. Moreover, a combined nomogram based on the logistic regression model was then plotted. Hosmer–Lemeshow goodness of fit test was applied to evaluate the calibration of



the combined model, and the results were represented by a calibration curve.

## Lymph Node Status Ascertainment

All patients underwent lobectomy or a more extensive resection. Systematic lymph node dissection was performed in all patients according to the European Society of Thoracic Surgeons guidelines (18, 19). The minimal number of dissected lymph nodes was six and at least three mediastinal lymph nodal stations and subcarinal stations had to be included. The hilar and intrapulmonary lymph nodes were excised as well. All surgical specimens and lymph nodes were fixed in 10% formalin and then sliced at the maximum dimension, and all sections were embedded in paraffin. Two experienced pathologists blindly evaluated all slices and lymph nodes together, and any disagreement was resolved by consensus. Pathological TNM stage, histological type, and lymph node station were evaluated according to the 8th edition of the TNM classification of lung cancer (2017) provided by the International Union against Cancer and the American Joint Commission on Cancer (20, 21).

## Statistical Analysis

Statistical analysis was performed using SPSS statistics (version 24; IBM, Armonk, NY, USA) and R software (version 3.6.1; <http://www.Rproject.org>). For continuous variables of clinical characteristics and CT morphological features, independent t-test or Mann–Whitney U test was performed; for categorical variables, Chi-square test was used for comparisons between the two groups. The combined model was constructed with multivariate logistic regression analysis and the performance of the combined model was evaluated using receiver operating characteristic (ROC) curve. A combined nomogram and a calibration curve of the combined model were then plotted. A calibration curve showing discrete experimental points close to or nearly coinciding with the diagonal

would indicate that the calibration of the combined model was high. A two-sided  $P$  value  $< 0.05$  was considered to be indicative of statistical significance.

## RESULTS

### Clinical Characteristics and CT Morphological Features

Males ( $P = 0.025$ ) and smokers ( $P = 0.005$ ) were more common in the LNM (+) group than in the LNM (–) group. However, no significant difference in age was observed between the two groups ( $P = 0.794$ ). Tumor size, density, air bronchogram, spiculation, lobulation, necrosis, pleural effusion, and pleural involvement were found to be associated with LNM (all  $P < 0.05$ ). Tumor size was larger in the LNM (+) group than that in the LNM (–) group ( $P < 0.001$ ). Tumors with solid density, air bronchogram, spiculation, lobulation, necrosis, and pleural effusion were more common in the LNM (+) group than in the LNM (–) (all  $P < 0.05$ ). However, there were no significant differences in air space and vascular convergence between the two groups (all  $P > 0.05$ , **Table 1**).

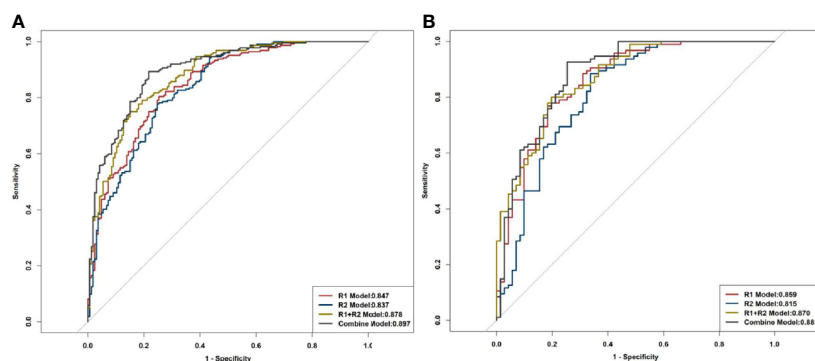
### Radiomics Model Construction

The R1 model was built with 13 features, including original first-order variance, wavelet transform, gray histogram features, gradient, and lbp.3D.k glszm small-area emphasis; the areas under the ROC curves (AUCs) for predicting LNM were 0.847 and 0.859 in training cohort and validation cohort, respectively (**Figure 5**). The R2 model was built with 19 features, including wavelet, square root, logarithm, and gradient, with AUCs of 0.837 and 0.815 for the prediction of LNM in the training cohort and validation cohort, respectively. In total, 1300 features were extracted from both the primary tumor and pleura. After ranking these features, 31 features from R1 and R2 were found to be

**TABLE 1** | Comparison of the clinical characteristics and CT morphological features between the LNM (–) and LNM (+) groups (n, %).

Characteristics	LNM(–) group (237)	LNM(+) group (319)	Sig.	P value
Age (years)	60.28 ± 10.21	60.50 ± 9.33	0.261	0.794 <sup>a</sup>
Sex (male)	111 (46.84%)	180 (56.43%)	5.014	0.025 <sup>b</sup>
Smoker	82 (34.60%)	147 (46.08%)	7.931	0.005 <sup>b</sup>
Tumor size (mm)	23.00 (16.00, 30.00)	32.00 (24.00, 42.00)	9.023	< 0.001 <sup>c</sup>
Density			82.686	< 0.001 <sup>b</sup>
Solid	162 (68.35%)	308 (96.55%)		
Sub-solid	75 (31.65%)	11 (3.45%)		
Air space	74 (31.22%)	81 (25.39%)	2.300	0.129 <sup>b</sup>
Air bronchogram	53 (22.36%)	32 (10.03%)	15.966	< 0.001 <sup>b</sup>
Spiculation	56 (23.63%)	134 (42.01%)	20.415	< 0.001 <sup>b</sup>
Lobulation	209 (88.19%)	304 (95.30%)	9.639	0.002 <sup>b</sup>
Necrosis	20 (8.44%)	64 (20.06%)	14.325	< 0.001 <sup>b</sup>
Vascular convergence	54 (22.78%)	69 (21.63%)	0.105	0.746 <sup>b</sup>
Pleural effusion	2 (0.84%)	13 (4.08%)	5.409	0.020 <sup>b</sup>
Pleural involvement			44.470	< 0.001 <sup>b</sup>
Absent	23 (9.70%)	20 (6.27%)		$P^{\#}$
Type I	144 (60.76%)	114 (35.74%)		$P^*$
Type II	30 (12.66%)	82 (25.71%)		$P^*$
Type III	40 (16.88%)	103 (32.29%)		$P^*$

<sup>a</sup>independent t-test; <sup>b</sup>Chi-squared test; <sup>c</sup>Mann–Whitney U test;  $P^{\#}$  means  $P > 0.05$  and  $P^*$  means  $P < 0.05$  for further pairwise comparison between two groups. LNM, lymph node metastasis; CT, computed tomography.



**FIGURE 5** | ROC curves of R1, R2, R1+R2, and the combined model for distinguishing LNM. **(A)** Training cohort. **(B)** Validation cohort.

significantly associated with LNM (all  $P < 0.05$ ), and AUCs of R1+R2 model were 0.878 and 0.870 in the training and validation cohorts, respectively (**Figure 5**). Furthermore, the combined model was also developed with AUCs of 0.897 and 0.883 for the training and validation cohorts, respectively (**Figure 5**).

## Evaluation of the Radiomics Models

Multivariable analysis revealed that long diameter, presence of spiculation, radiomics score of the primary tumor (RS1), and radiomics score of the pleura around the tumor (RS2) were significant predictors (**Table 2**). Therefore, they were fused as a radiomics nomogram (**Figure 6A**). The calibration curve showed that the discrete experimental points were similar to or the same

as the diagonal, which indicated that the calibration of the combined model was high (**Figure 6B**).

## Radiomics Model for Identifying N0, N1, N2, and N3

Radiomic signatures also showed good performance in identifying the lymph node stage of N0, N1, N2, and N3 (**Supplementary Figure A3**) as shown by the following AUCs: for the R1 model, 0.839, 0.691, 0.768, and 0.864 in the training cohort and 0.870, 0.700, 0.769, and 0.845 in the validation cohort, respectively; for the R2 model, 0.808, 0.783, 0.763, and 0.885 in the training cohort, and 0.810, 0.777, 0.752, and 0.943 in the validation cohort, respectively; for the R1+R2 model, 0.866, 0.812, 0.824, and 0.927 in the training cohort and 0.841, 0.794, 0.815, and 0.951 in the validation cohort, respectively; and for the combined model, 0.916, 0.797, 0.823, and 0.927 in the training cohort and 0.860, 0.773, 0.832, and 0.859 in the validation cohort, respectively (**Table 3**, **Figure 7**).

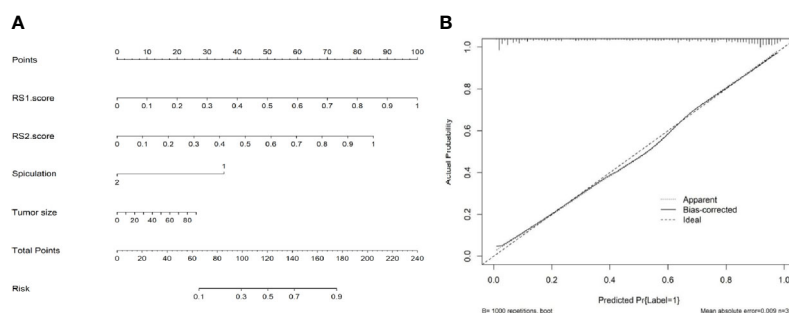
**TABLE 2** | Variables and coefficients of the radiomics nomogram.

Variables	$\beta$	Adjusted OR (95% CI)	P value
RS1 (per 0.1 increase)	3.9867	53.88 (14.89–215.1)	< 0.0001
RS2 (per 0.1 increase)	3.4074	30.19 (8.73–112.25)	< 0.0001
Tumor size	0.0117	1.01 (0.99–1.04)	0.3856
Spiculation	–1.4176	0.24 (0.13–0.45)	< 0.0001
Intercept	–1.9709	0.14 (0.04–0.44)	0.0009

RS1, radiomics score of the primary tumor; RS2, radiomics score of the adjacent pleura.

## DISCUSSION

Radiologic examinations, including CT, magnetic resonance imaging (MRI), and positron emission tomography combined



**FIGURE 6** | Nomogram and calibration curve of radiomic models. **(A)** Nomogram of the combined model. **(B)** Calibration curve showing that the discrete experimental points are coincident with the diagonal, which indicates that the calibration of the combined model is high.

**TABLE 3 |** AUCs of radiomics models for evaluating lymph node staging.

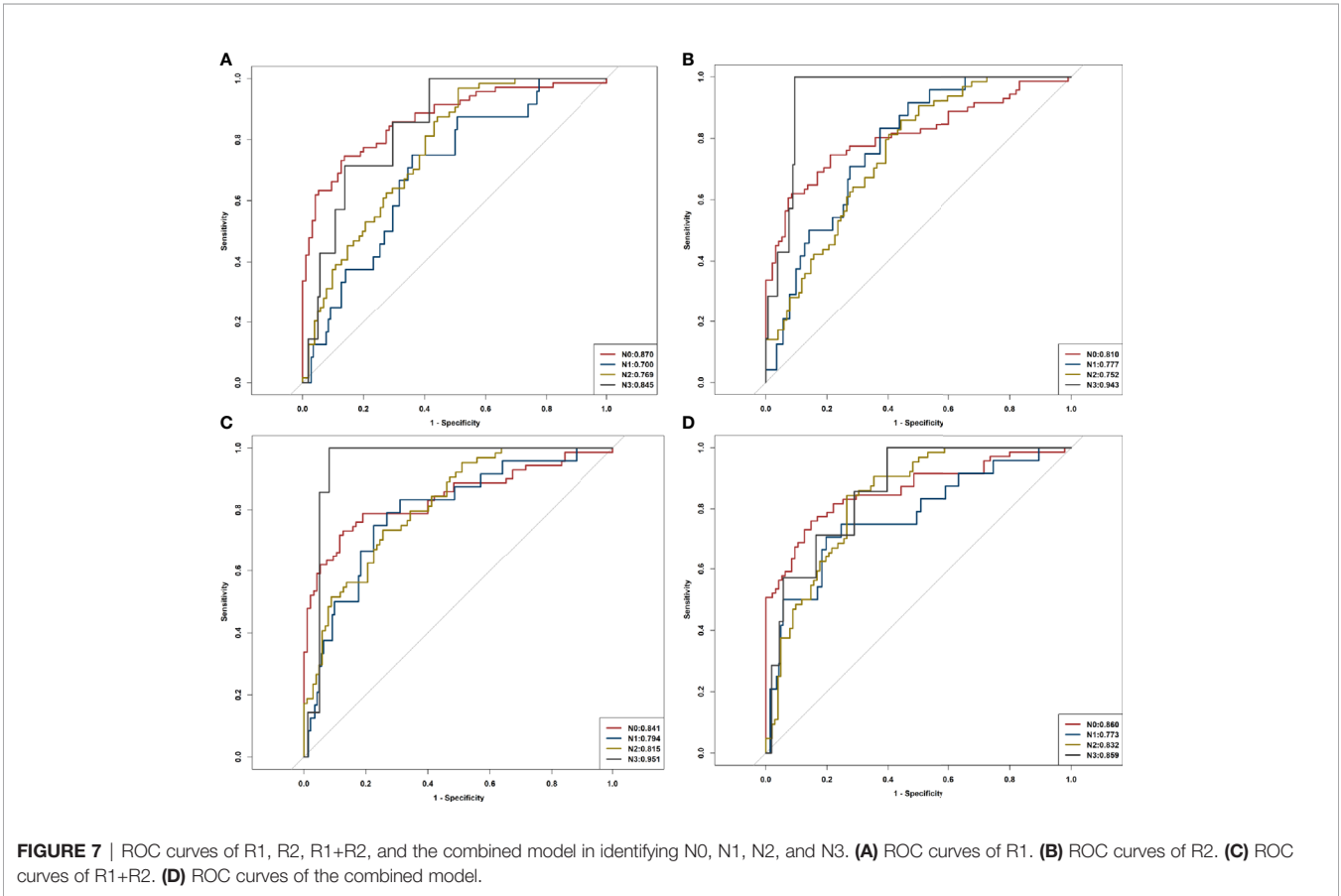
Models	Training cohort				Validation cohort			
	N0	N1	N2	N3	N0	N1	N2	N3
R1	0.839	0.691	0.768	0.864	0.870	0.700	0.769	0.845
R2	0.808	0.783	0.763	0.885	0.810	0.777	0.752	0.943
R1+R2	0.866	0.812	0.824	0.927	0.841	0.794	0.815	0.951
Combined model	0.916	0.797	0.823	0.927	0.860	0.773	0.832	0.859

N0, No regional node metastasis; N1, Metastasis in ipsilateral pulmonary or hilar nodes; N2, Metastasis in ipsilateral mediastinal/subcarinal nodes; N3, Metastasis in contralateral mediastinal/hilar or supraclavicular nodes.

with CT (FDG-PET/CT), can be used for pretherapeutic lymph node assessments (22–24). As an alternative, CT is an important part of the PLADC staging process in clinical practice. However, some previous studies have observed low sensitivity and specificity of CT, and others have shown that CT was severely limited when relying solely on a short-axis diameter of  $\leq 10$  mm of the thoracic lymph nodes in accurately evaluating malignant nodes (25, 26). Diffusion-weighted magnetic resonance imaging (DWI) of MRI has been applied in lung cancer staging for the last two decades; however, further development of protocols and more clinical trials for lymph node evaluation are still needed (23). FDG-PET/CT has been reported to be superior to CT for evaluating LNM of lung cancer, but high false-positive rate and radiation dosage have restricted its clinical application (27). Therefore, preoperative imaging for noninvasive evaluation of

the status of lymph nodes is highly desirable. In the present study, we developed and validated a radiomics signature-based model that incorporates radiomic signatures of both the primary tumor and adjacent pleura, CT morphological features, and clinical factors for prediction of LNM in patients with PLADC.

In this study, R1, which reflects radiomic signatures of the primary tumor had AUCs of 0.847 and 0.859 for predicting LNM in the training and validation cohorts, respectively, suggesting a huge potential for radiomics in predicting LNM. Consistent with our results, previous researchers have also reported that radiomic signatures were of great value in predicting LNM in lung cancer (15, 28); Wang et al. (17) confirmed that radiomic signatures from peritumoral lung parenchyma would increase the prediction efficiency of LNM in clinical stage T1 lung adenocarcinoma. Additionally, R2, which showed radiomic



signatures of pleura around the tumor, was associated with LNM in patients with PLADC, and yielded AUCs of 0.837 and 0.815 for predicting LNM in the training and validation cohorts, respectively. To the best of our knowledge, few studies have applied radiomic signatures of pleura around the tumor to predict LNM. Researchers have concluded that LNM depends on selected cancer cells (the “seeds”) and micro-environments (the “soil”), and metastases formed only when the seeds and soil were compatible (29, 30). We thus hypothesized the “seed and soil” theory for LNM prediction. Based on the “seed and soil” theory, interestingly, we found that LNM was associated with both the tumor and the phenotype of its nearby pleura. This finding might partly be explained by the rich subpleural lymph drainage and direct drainage route into the mediastinum, through which tumor cells may spread and metastasize easily (6, 31). We concluded that tumor invasion to the network of subpleural lymph vessel would lead to higher occurrence of LNM. Moreover, radiomic signatures of R1+R2, which contained 31 characteristics in total, showed good performance in predicting LNM in patients with PLADC, with AUCs of 0.878 and 0.870 in the training and validation cohorts, respectively.

Previous studies have confirmed that several CT features and clinical risk factors were closely related to LNM of lung adenocarcinoma (8, 32–39). Similarly, we found that sex, smoking history, and eight CT morphological features of tumors, including long diameter, tumor density, air bronchogram, spiculation, lobulation, necrosis, pleural effusion, and pleural involvement, were significantly associated with LNM in this study. Therefore, we further established a prediction model that combined radiomic signatures of R1 and R2, CT features, and clinical risk factors. The combined model is of great value in predicting LNM with AUCs of 0.897 and 0.883 in the training and validation cohorts, respectively. The decision curve showed that the combined model was of great help in clinical decision-making. We have also developed a radiomics nomogram and calibration curve of the combined model, both of which showed that the combined model had good predictive ability for LNM in patients with PLADC.

Asamura et al. (40) reported that the 5-year survival rates in patients with lung cancer according to the pathological N statuses were 75% (N0), 49% (N1), 36% (N2), and 20% (N3). Therefore, the survival differed significantly between all neighboring nodal categories, and it is very important to accurately evaluate the metastasis status of lymph nodes before operation. In the present study, the radiomics model was also used to distinguish N0, N1, N2, and N3, and the combined model revealed good diagnostic performance in estimating N stages for patients with PLADC.

The present study had several limitations. All data were collected within a single institution, but we are preparing to conduct a multicenter study to verify the reliability and general applicability of this model. Previous studies have shown the relationship between different pleural involvement and LNM or nodal staging. Radiomics was used only to further quantify the relevant features, and we believe that we can achieve good performance in external verification. Moreover, due to the lack of

MRI and PET images, there is scope for improving the performance of the model, especially under the condition wherein PET/CT can provide better reference for evaluating LNM. We chose only three slices instead of the whole tumor for image-feature extraction. Future work might benefit from automatic target area delineation software, and more auxiliary information around the tumor can be added to achieve an accurate assessment of tumor lymph nodes.

## CONCLUSION

This study showed that obtaining information about the primary tumor and pleura around the tumor provides complementary information that can be useful in clinical decision-making. The combined model, which incorporates radiomic signatures, CT features, and clinical factors, can be used as an auxiliary tool to predict LNM in patients with PLADC.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Review Committee of the First Affiliated Hospital of Chongqing Medical University. The ethics committee waived the requirement of written informed consent for participation.

## AUTHOR CONTRIBUTIONS

QL and X-qH have contributed equally to this work and share first authorship. J-wL and T-yL contributed equally to this work and share correspondence authorship. All authors contributed to the article and approved the submitted version.

## FUNDING

This research was supported by the Science and Technology Innovation Program of Social Undertakings and People's Livelihood Security of Chongqing Science and Technology Commission to T-yL: Technology Exploration and Application of Precision Radiotherapy for Disease. (cstc2016shms-ztxx10002).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2021.675877/full#supplementary-material>

## REFERENCES

- Siegel RL, Fedewa SA, Miller KD, Goding-Sauer A, Pinheiro PS, Martinez-Tyson D. Cancer Statistics for Hispanics/Latinos. *CA Cancer J Clin* (2015) 65 (6):457–30. doi: 10.3322/caac.21314
- Moon Y, Lee KY, Sung SW, Park JK. Differing Histopathology and Prognosis in Pulmonary Adenocarcinoma At Central and Peripheral Locations. *J Thorac Dis* (2016) 8(1):169–77. doi: 10.3978/j.issn.2072-1439.2016.01.15
- Kukhon FR, Lan X, Helgeson SA, Arunthari V, Fernandez-Bussy S, Patel NM. Occult Lymph Node Metastasis in Radiologic Stage I non-Small Cell Lung Cancer: The Role of Endobronchial Ultrasound. *Clin Respir J* (2021) 00:1–7. doi: 10.1111/crj.13344
- Chen YC, Lin YH, Chien HC, Hsu PK, Hung JJ, Huang CS. Preoperative Consolidation-to-Tumor Ratio is Effective in the Prediction of Lymph Node Metastasis in Patients With Pulmonary Ground-Glass Component Nodules. *Thorac Cancer* (2021) 12(8):1203–9. doi: 10.1111/1759-7714.13899
- Hsu JS, Han IT, Tsai TH, Lin SF, Jaw TS, Liu GC. Pleural Tags on CT Scans to Predict Visceral Pleural Invasion of Non-Small Cell Lung Cancer That Does Not Affect the Pleura. *Radiology* (2016) 279(2):590–6. doi: 10.1148/radiol.201511120
- Eriguchi T, Takeda A, Tsurugai Y, Sanuki N, Kibe Y, Hara Y. Pleural Contact Decreases Survival in Clinical T1N0M0 Lung Cancer Patients Undergoing SBRT. *Radiother Oncol* (2019) 134:191–8. doi: 10.1016/j.radonc.2019.02.005
- Chang YL, Lin MW, Shih JY, Wu CT, Lee YC. The Significance of Visceral Pleural Surface Invasion in 321 Cases of Non-Small Cell Lung Cancers With Pleural Retraction. *Ann Surg Oncol* (2012) 19(9):3057–64. doi: 10.1245/s10434-012-2354-y
- Wang L, Jiang W, Zhan C, Shi Y, Zhang Y, Lin Z. Lymph Node Metastasis in Clinical Stage IA Peripheral Lung Cancer. *Lung Cancer* (2015) 90(1):41–6. doi: 10.1016/j.lungcan.2015.07.003
- Eichhorn F, Klotz LV, Muley T, Kobinger S, Winter H, Eichhorn ME. Prognostic Relevance of Regional Lymph-Node Distribution in Patients With N1-positive non-Small Cell Lung Cancer: A Retrospective Single-Center Analysis. *Lung Cancer* (2019) 138:95–101. doi: 10.1016/j.lungcan.2019.10.018
- Katsumata S, Aokage K, Ishii G, Nakasone S, Sakai T, Okada S. Prognostic Impact of the Number of Metastatic Lymph Nodes on the Eighth Edition of the TNM Classification of NSCLC. *J Thorac Oncol* (2019) 14(8):1408–18. doi: 10.1016/j.jtho.2019.04.016
- Coroller TP, Agrawal V, Huynh E, Narayan V, Lee SW, Mak RH. Radiomic-Based Pathological Response Prediction From Primary Tumors and Lymph Nodes in NSCLC. *J Thorac Oncol* (2017) 12(3):467–76. doi: 10.1016/j.jtho.2016.11.2226
- Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) non-Small Cell Lung Cancer. *Radiology* (2016) 281(3):947–57. doi: 10.1148/radiol.2016152234
- Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrta A. Artificial Intelligence in Cancer Imaging: Clinical Challenges and Applications. *CA Cancer J Clin* (2019) 69(2):127–57. doi: 10.3322/caac.21552
- Shimada AY, Kudo Y, Furumoto H, Imai K, Maehara S, Tanaka T. Computed Tomography Histogram Approach to Predict Lymph Node Metastasis in Patients With Clinical Stage IA Lung Cancer. *Ann Thorac Surg* (2019) 108 (4):1021–8. doi: 10.1016/j.athoracsurg.2019.04.082
- Cong M, Feng H, Ren JL, Xu Q, Cong L, Hou Z. Development of a Predictive Radiomics Model for Lymph Node Metastases in Pre-Surgical CT-based Stage IA non-Small Cell Lung Cancer. *Lung Cancer* (2020) 139:73–9. doi: 10.1016/j.lungcan.2019.11.003
- Zhang C, Pang G, Ma C, Wu J, Wang P, Wang K. Preoperative Risk Assessment of Lymph Node Metastasis in Ctl Lung Cancer: A Retrospective Study From Eastern China. *J Immunol Res* (2019) 2019:6263249. doi: 10.1155/2019/6263249
- Wang X, Zhao X, Li Q, Xia W, Peng Z, Zhang R. Can Peritumoral Radiomics Increase the Efficiency of the Prediction for Lymph Node Metastasis in Clinical Stage T1 Lung Adenocarcinoma on CT? *Eur Radiol* (2019) 29 (11):6049–58. doi: 10.1007/s00330-019-06084-0
- De Leyn P, Dooms C, Kuzdzal J, Lardinois D, Passlick B, Rami-Porta R. Revised ESTS Guidelines for Preoperative Mediastinal Lymph Node Staging for non-Small-Cell Lung Cancer. *Eur J Cardiothorac Surg* (2014) 45(5):787–98. doi: 10.1093/ejcts/ezu028
- Thomas PA. Intraoperative Lymph-Node Assessment During NSCLC Surgery: The Need for Standardisation and Quality Evaluation. *Lancet Oncol* (2019) 20(1):23–5. doi: 10.1016/S1470-2045(18)30768-X
- Detterbeck FC, Chansky K, Groome P, Bolejack V, Crowley J, Shemanski L. The IASLC Lung Cancer Staging Project: Methodology and Validation Used in the Development of Proposals for Revision of the Stage Classification of NSCLC in the Forthcoming (Eighth) Edition of the TNM Classification of Lung Cancer. *J Thorac Oncol* (2016) 11(9):1433–46. doi: 10.1016/j.jtho.2016.06.028
- Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. *Chest* (2017) 151(1):193–203. doi: 10.1016/j.chest.2016.10.010
- Ohno Y, Koyama H, Lee HY, Yoshikawa T, Sugimura K. Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET)/MRI for Lung Cancer Staging. *J Thorac Imaging* (2016) 31(4):215–27. doi: 10.1097/RTI.0000000000000210
- Usuda K, Funazaki A, Maeda R, Sekimura A, Motono N, Matoba M. Economic Benefits and Diagnostic Quality of Diffusion-Weighted Magnetic Resonance Imaging for Primary Lung Cancer. *Ann Thorac Cardiovasc Surg* (2017) 23(6):275–80. doi: 10.5761/atcs.ra.17-00097
- Sun XY, Chen TX, Chang C, Teng HH, Xie C, Ruan MM. Suvmax of (18)FDG PET/CT Predicts Histological Grade of Lung Adenocarcinoma. *Acad Radiol* (2021) 28(1):49–57. doi: 10.1016/j.acra.2020.01.030
- Fréchet B, Kazakov J, Thiffault V, Ferraro P, Liberman M. Diagnostic Accuracy of Mediastinal Lymph Node Staging Techniques in the Preoperative Assessment of Nonsmall Cell Lung Cancer Patients. *J Bronchology Interv Pulmonol* (2018) 25(1):17–24. doi: 10.1097/LBR.0000000000000425
- Jett JR, Schild SE, Kesler KA, Kalemkerian GP. Treatment of Small Cell Lung Cancer: Diagnosis and Management of Lung Cancer, 3rd Ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* (2013) 143(5 Suppl):e400S–19S. doi: 10.1378/chest.12-2363
- Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT. Methods for Staging non-Small Cell Lung Cancer: Diagnosis and Management of Lung Cancer, 3rd Ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* (2013) 143(5 Suppl):e211S–50S. doi: 10.1378/chest.12-2355
- Yang X, Pan X, Liu H, Gao D, He J, Liang W, et al. A New Approach to Predict Lymph Node Metastasis in Solid Lung Adenocarcinoma: A Radiomics Nomogram. *J Thorac Dis* (2018) 10(Suppl 7):S807–19. doi: 10.21037/jtd.2018.03.126
- Paget S. The Distribution of Secondary Growths in Cancer of the Breast. 1889. *Cancer Metastasis Rev* (1989) 8(2):98–101.
- Fidler IJ. The Pathogenesis of Cancer Metastasis: The ‘Seed and Soil’ Hypothesis Revisited. *Nat Rev Cancer* (2003) 3(6):453–8. doi: 10.1038/nrc1098
- Yonemura Y, Bandou E, Kawamura T, Endou Y, Sasaki T. Quantitative Prognostic Indicators of Peritoneal Dissemination of Gastric Cancer. *Eur J Surg Oncol* (2006) 32(6):602–6. doi: 10.1016/j.ejso.2006.03.003
- Cohen JG, Raymond E, Jankowski A, Brambilla E, Arbib F, Lantuejoul S. Lung Adenocarcinomas: Correlation of Computed Tomography and Pathology Findings. *Diagn Interv Imaging* (2016) 97(10):955–63. doi: 10.1016/j.diii.2016.06.021
- Moon Y, Lee KY, Park JK. The Prognosis of Invasive Adenocarcinoma Presenting as Ground-Glass Opacity on Chest Computed Tomography After Sublobar Resection. *J Thorac Dis* (2017) 9(10):3782–92. doi: 10.21037/jtd.2017.09.40
- Lederlin M, Puderbach M, Muley T, Schnabel PA, Stenzinger A, Kauczor HU. Correlation of Radio- and Histomorphological Pattern of Pulmonary Adenocarcinoma. *Eur Respir J* (2013) 41(4):943–51. doi: 10.1183/09031936.00056612
- Maeyashiki T, Suzuki K, Hattori A, Matsunaga T, Takamochi K, Oh S. The Size of Consolidation on Thin-Section Computed Tomography is a Better Predictor of Survival Than the Maximum Tumour Dimension in Resectable Lung Cancer. *Eur J Cardiothorac Surg* (2013) 43(5):915–8. doi: 10.1093/ejcts/ezs516
- Hattori A, Suzuki K, Matsunaga T, Fukui M, Kitamura Y, Miyasaka Y. Is Limited Resection Appropriate for Radiologically “Solid” Tumors in Small



- Lung Cancers? *Ann Thorac Surg* (2012) 94(1):212–5. doi: 10.1016/j.athoracsur.2012.03.033
37. Yoshino I, Ichinose Y, Nagashima A, Takeo S, Motohiro A, Yano T. Clinical Characterization of Node-Negative Lung Adenocarcinoma: Results of a Prospective Investigation. *J Thorac Oncol* (2006) 1:825–31. doi: 10.1097/01243894-200610000-00011
  38. Cong M, Yao H, Liu H, Huang L, Shi G. Development and Evaluation of a Venous Computed Tomography Radiomics Model to Predict Lymph Node Metastasis From non-Small Cell Lung Cancer. *Med (Baltimore)* (2020) 99(18):e20074. doi: 10.1097/MD.00000000000020074
  39. Liu Y, Kim J, Balagurunathan Y, Hawkins S, Stringfield O, Schabath MB. Prediction of Pathological Nodal Involvement by CT-based Radiomic Features of the Primary Tumor in Patients With Clinically Node-Negative Peripheral Lung Adenocarcinomas. *Med Physics* (2018) 45(6):2518–26. doi: 10.1002/mp.12901
  40. Asamura H, Chansky K, Crowley J, Goldstraw P, Rusch VW, Vansteenkiste JF. The International Association for the Study of Lung Cancer Lung Cancer Staging Project: Proposals for the Revision of the N Descriptors in the Forthcoming 8th Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* (2015) 10(12):1675–84. doi: 10.1097/JTO.0000000000000678
- 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.
- Copyright © 2021 Li, He, Fan, Zhu, Lv and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Artificial Intelligence Can Cut Costs While Maintaining Accuracy in Colorectal Cancer Genotyping

Alec J. Kacew<sup>1</sup>, Garth W. Strohbehn<sup>2</sup>, Loren Saulsberry<sup>3</sup>, Neda Laiteerapong<sup>2</sup>, Nicole A. Cipriani<sup>4</sup>, Jakob N. Kather<sup>5\*</sup> and Alexander T. Pearson<sup>2\*</sup>

<sup>1</sup> Pritzker School of Medicine, University of Chicago, Chicago, IL, United States, <sup>2</sup> Department of Medicine, University of Chicago, Chicago, IL, United States, <sup>3</sup> Department of Public Health Sciences, University of Chicago, Chicago, IL, United States, <sup>4</sup> Department of Pathology, University of Chicago, Chicago, IL, United States, <sup>5</sup> Department of Medicine, University Hospital Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Aachen, Germany

## OPEN ACCESS

### Edited by:

George Bebis,  
University of Nevada, Reno,  
United States

### Reviewed by:

Shujiro Okuda,  
Niigata University, Japan  
Kun Wang,  
National Institutes of Health (NIH),  
United States

### \*Correspondence:

Jakob N. Kather  
jkather@ukaachen.de  
Alexander T. Pearson  
Apearson5@  
medicine.bsd.uchicago.edu

### Specialty section:

This article was submitted to  
Cancer Immunity and  
Immunotherapy,  
a section of the journal  
Frontiers in Oncology

**Received:** 18 November 2020

**Accepted:** 13 May 2021

**Published:** 08 June 2021

### Citation:

Kacew AJ, Strohbehn GW,  
Saulsberry L, Laiteerapong N,  
Cipriani NA, Kather JN and  
Pearson AT (2021) Artificial Intelligence  
Can Cut Costs While Maintaining  
Accuracy in Colorectal  
Cancer Genotyping.  
Front. Oncol. 11:630953.  
doi: 10.3389/fonc.2021.630953

Rising cancer care costs impose financial burdens on health systems. Applying artificial intelligence to diagnostic algorithms may reduce testing costs and avoid wasteful therapy-related expenditures. To evaluate the financial and clinical impact of incorporating artificial intelligence-based determination of mismatch repair/microsatellite instability status into the first-line metastatic colorectal carcinoma setting, we developed a deterministic model to compare eight testing strategies: A) next-generation sequencing alone, B) high-sensitivity polymerase chain reaction or immunohistochemistry panel alone, C) high-specificity panel alone, D) high-specificity artificial intelligence alone, E) high-sensitivity artificial intelligence followed by next generation sequencing, F) high-specificity artificial intelligence followed by next-generation sequencing, G) high-sensitivity artificial intelligence and high-sensitivity panel, and H) high-sensitivity artificial intelligence and high-specificity panel. We used a hypothetical, nationally representative, population-based sample of individuals receiving first-line treatment for *de novo* metastatic colorectal cancer (N = 32,549) in the United States. Model inputs were derived from secondary research (peer-reviewed literature and Medicare data). We estimated the population-level diagnostic costs and clinical implications for each testing strategy. The testing strategy that resulted in the greatest project cost savings (including testing and first-line drug cost) compared to next-generation sequencing alone in newly-diagnosed metastatic colorectal cancer was using high-sensitivity artificial intelligence followed by confirmatory high-specificity polymerase chain reaction or immunohistochemistry panel for patients testing negative by artificial intelligence (\$400 million, 12.9%). The high-specificity artificial intelligence-only strategy resulted in the most favorable clinical impact, with 97% diagnostic accuracy in guiding genotype-directed treatment and average time to treatment initiation of less than one day. Artificial intelligence has the potential to reduce both time to treatment initiation and costs in the metastatic colorectal cancer setting without meaningfully sacrificing diagnostic accuracy. We expect the artificial intelligence value proposition to improve in coming years, with increasing diagnostic accuracy and decreasing costs of processing power. To extract maximal value from the technology,

health systems should evaluate integrating diagnostic histopathologic artificial intelligence into institutional protocols, perhaps in place of other genotyping methodologies.

**Keywords:** deep learning, microsatellite instability (MSI), colorectal (colon) cancer, financial implication, digital biomarker, digital pathology, cost savings, artificial intelligence

## INTRODUCTION

Oncologic diagnostic algorithms, specifically those involving next-generation sequencing (NGS), financially burden healthcare systems. Just as the advent of NGS was an advancement over polymerase chain reaction (PCR) and immunohistochemistry (IHC) for some applications, artificial intelligence (AI) may be the next innovative oncologic diagnostic agent. From routine histopathology images, AI can recapitulate genetic information with area under the receiver-operator curve (ROC) approaching 0.9 (1, 2). AI may help overcome NGS-related challenges like cost, packing and shipping delays, and turnaround time. Due to massive scalability, AI costs, following initial investment, would be a fraction of other technologies' costs. Since tumors grow in the absence of treatment, AI's faster turnaround (and associated earlier treatment initiation) could impact clinical outcomes.

AI may be especially impactful in common malignancies. In the United States (U.S.), nearly 150,000 cases of colorectal cancer (CRC) are diagnosed annually (3). In the metastatic setting (22% of cases), deficient mismatch repair (dMMR) or high microsatellite instability (MSI-H) – genetic features seen in 5% of metastatic CRC (mCRC) cases – are predictive and prognostic (4). For individuals with dMMR/MSI-H mCRC, KEYNOTE-177 demonstrated superior outcomes for front-line pembrolizumab over cytotoxic chemotherapy (5). The high price of immunotherapies (like pembrolizumab) could portend a significant escalation in the total cost of mCRC care. Diagnostic strategies can limit immunotherapy use to only those patients who are most likely to benefit. Like NGS, PCR, and IHC, AI, although not currently part of routine clinical practice, can infer actionable genetic features like MMR/MSI, *KRAS*, and *BRAF* status from histopathology (1, 2). In the present study, we projected the financial and clinical impacts of incorporating AI into the diagnostic algorithm. We are unaware of any prior research in estimating the financial impact of implementing AI in a clinical context – this study is the first one in our knowledge to do so. Our results, along with future real-world confirmation across cancers, could inform policy and practice to optimize oncologic diagnostic pathways.

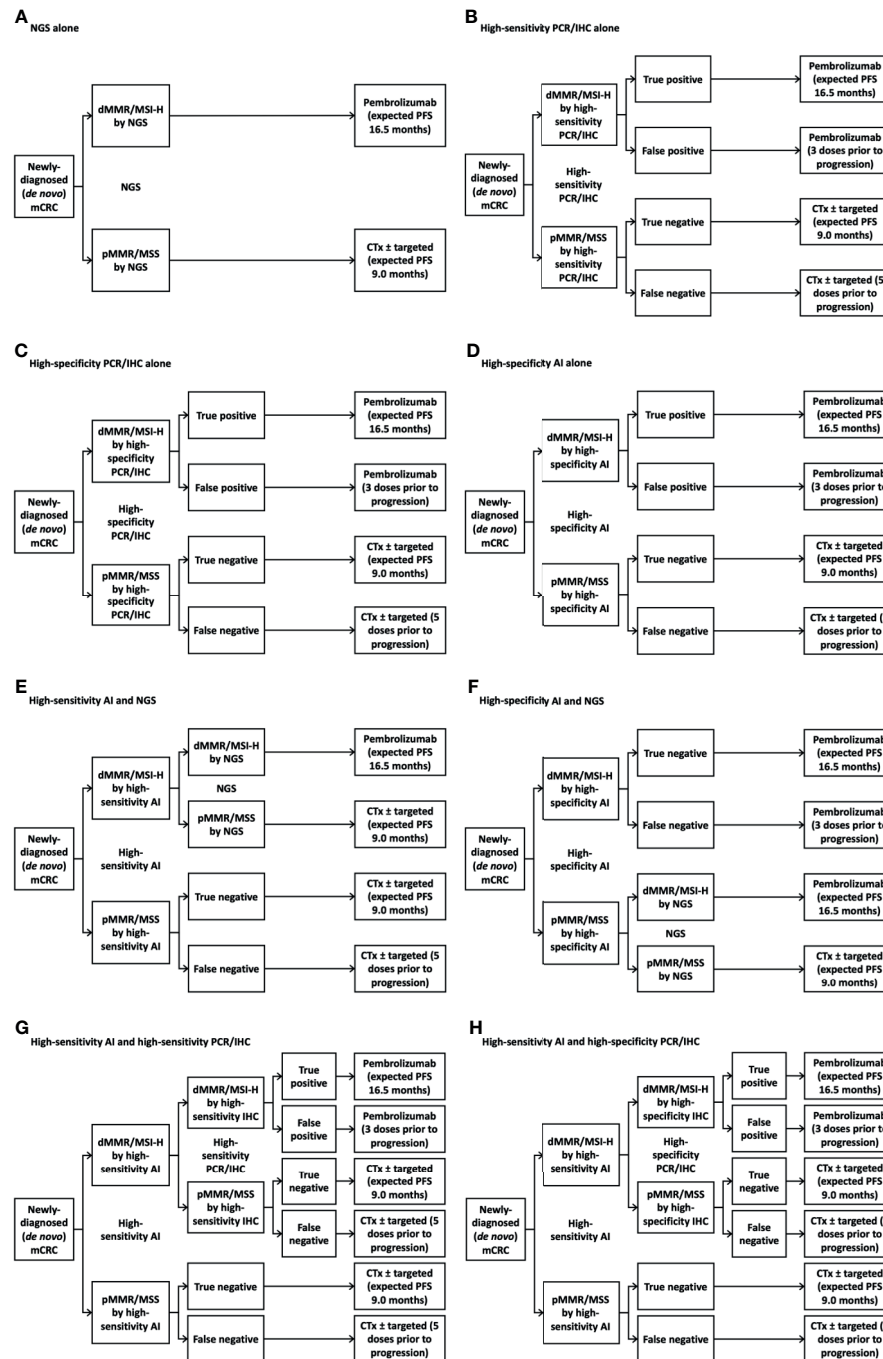
## MATERIALS AND METHODS

We generated eight potential diagnostic algorithms for determining MMR/MSI status in the U.S. first-line newly-diagnosed (*de novo*) mCRC population: NGS alone, high-sensitivity PCR or IHC panel ("panel" for short) alone, high-specificity panel alone, high-specificity AI alone, high-sensitivity AI with confirmatory NGS for patients testing negative by AI, high-specificity AI with confirmatory NGS for patients testing positive by AI, high-

sensitivity AI with confirmatory high-sensitivity panel for patients testing positive by AI, and high-sensitivity AI with confirmatory high-specificity panel for patients testing positive by AI (**Figure 1**). We chose these diagnostic scenarios based on current standard-of-care and, based on clinical and cost considerations, where AI might reasonably fit within the diagnostic paradigm. We took "NGS alone" as the reference approach, as it was expected to be the costliest, and chose other scenarios to include clinically reasonable permutations of using NGS, panel, and/or AI. We evaluated costs from the perspective of the U.S. healthcare system, agnostic of payer. We assessed costs over one year, as longer time horizons might not account for yet-unknown future changes in oncologic technology and practice over longer timeframes. Given the relatively short time horizon, our model did not incorporate a discount rate. Consideration of opportunity costs (e.g., potential use of cost-savings derived from new diagnostic approaches) was outside our scope.

We incorporated data from peer-reviewed literature and government sources into a financial and clinical model (**Table 1** and **Figure 2**) (6, 7, 10–14). All cost assumptions were based on values reported in 2017–2020. We aimed to use as few data sources as possible in the interest of minimizing heterogeneity of assumptions. We gathered nearly all dollar values from publicly available reimbursement schedules of the Centers for Medicare & Medicaid Services. Test characteristics of the AI platform is based on our group's previous work. Absolute population and incidence estimates were derived from the Surveillance, Epidemiology, and End Results Program (SEER) database, while proportions of patients falling into genetic and treatment subgroups was derived from a variety of peer-reviewed publications. The timing of restaging scans was based on the restaging cadence in the KEYNOTE-177 trial. We chose the AI sensitivity/specificity cutoffs based on two points along our previously-developed ROC (**Figure 1**) (2). Briefly, we developed our AI algorithm using hematoxylin and eosin-stained slides for samples that had previously been analyzed for MSI-H/dMMR status by either IHC or PCR. Pathologists who had been blinded to clinical data and MSI-H/dMMR status determined sample quality and area of tumor tissue. Images were saved digitally, color-normalized, then subjected to our deep learning system. We did not incorporate the cost of developing the deep learning model into our financial estimate, as we have already developed this approach.

We grouped PCR and IHC and assessed two sets of test characteristics (high-specificity and high-sensitivity) for these platforms, as characteristics vary across studies (8, 9). Our primary objective was to compare total costs of testing and first-line therapy across the scenarios. To assess clinical impact of each diagnostic strategy, we estimated time to treatment initiation, proportion of patients receiving results within guideline-



**FIGURE 1 |** Treatment decision tree with various testing strategies. **(A)** Next-generation sequencing alone; **(B)** High-sensitivity immunohistochemistry panel alone; **(C)** High-specificity immunohistochemistry panel alone; **(D)** High-specificity artificial intelligence alone; **(E)** High-sensitivity artificial intelligence followed by next-generation sequencing for individuals with deficient mismatch repair/microsatellite instability-high tumors by artificial intelligence; **(F)** High-specificity artificial intelligence followed by next-generation sequencing for individuals with intact mismatch repair/microsatellite stable tumors by artificial intelligence; **(G)** High-sensitivity artificial intelligence followed by high-sensitivity immunohistochemistry panel for individuals with deficient mismatch repair/microsatellite instability-high tumors by artificial intelligence; **(H)** High-sensitivity artificial intelligence followed by high-specificity immunohistochemistry panel for individuals with deficient mismatch repair/microsatellite instability-high tumors by artificial intelligence. AI, artificial intelligence; CTx, chemotherapy; dMMR, deficient mismatch repair; FU, fluorouracil; IHC, immunohistochemistry; mCRC, metastatic colorectal cancer; MSI-H, microsatellite instability-high; MSS, microsatellite-stable; NGS, next-generation sequencing; PCR, polymerase chain reaction; PFS, progression-free survival; pMMR, proficient mismatch repair.

**TABLE 1 |** Model assumptions and inputs.

Model input	Assumed value (reference)
<b>Population characteristics</b>	
# newly diagnosed colorectal cancer per year in the U.S.	147,950 (3)
% metastatic	22% (3)
# newly diagnosed ( <i>de novo</i> ) metastatic colorectal cancer per year in the U.S.	32,549
% dMMR/MSI-H	5% (4)
% pMMR/MSS	95% (4)
<b>Diagnostic characteristics</b>	
Cost per patient of next-generation sequencing	\$3,500.00 (6)
Cost per patient of PCR or IHC panel	\$1,206.25
KRAS/NRAS	\$682.29 (6)
BRAF	\$175.40 (6)
dMMR/MSI-H	\$348.56 (6)
Cost per patient of artificial intelligence (digital image scanning)	\$6.07 <sup>a</sup>
Time for next-generation sequencing (days)	12 (7)
Time for PCR or IHC panel (days)	4 (7)
Time for artificial intelligence (months) – assumed nominal value	-
Next generation sequencing sensitivity – conservative assumption	100%
Next generation sequencing specificity – conservative assumption	100%
PCR or IHC dMMR/MSI-H panel sensitivity (high sensitivity cutoff):	100% (8)
PCR or IHC dMMR/MSI-H panel specificity (high sensitivity cutoff):	81% (8)
PCR or IHC dMMR/MSI-H panel sensitivity (high specificity cutoff):	67% (9)
PCR or IHC dMMR/MSI-H panel specificity (high specificity cutoff):	93% (9)
Artificial intelligence dMMR/MSI-H sensitivity (high sensitivity cutoff)	98% (2)
Artificial intelligence dMMR/MSI-H specificity (high sensitivity cutoff)	79% (2)
Artificial intelligence dMMR/MSI-H sensitivity (high specificity cutoff)	70% (2)
Artificial intelligence dMMR/MSI-H specificity (high specificity cutoff)	98% (2)
<b>Therapeutic characteristics</b>	
Cost per patient per month for dMMR/MSI-H therapy	\$23,021.13 (5)
Weighted average cost per patient per month of 5-fluorouracil-based therapy <sup>b</sup>	\$7,625.88
% receiving FOLFOX + bevacizumab	35% (10)
Cost per patient per month for FOLFOX + bevacizumab	\$6,316.70 (11)
% receiving FOLFOX + cetuximab	45% (10)
Cost per patient per month for FOLFOX + cetuximab	\$11,945.73 (11)
% receiving 5-fluorouracil + leucovorin	20% (10)
Cost per patient per month for 5-fluorouracil + leucovorin	\$179.76 (11)
Weighted average cost per patient per dose of 5-fluorouracil-based therapy	\$3,807.68
% receiving FOLFOX + bevacizumab	35% (10)
Cost per patient per dose for FOLFOX + bevacizumab	\$3,158.35 (11)
% receiving FOLFOX + cetuximab	45% (10)
Cost per patient per dose for FOLFOX + cetuximab	\$5,972.86 (11)
% receiving 5-fluorouracil + leucovorin	20% (12)
Cost per patient per dose for 5-fluorouracil + leucovorin	\$63.63 (11)
Weighted average median time on of 5-fluorouracil-based therapy (months)	9.0
% receiving FOLFOX + bevacizumab	35% (10)
Median time on therapy for FOLFOX + bevacizumab	10.3 (13)
% receiving FOLFOX + cetuximab	45% (10)

(Continued)

**TABLE 1 |** Continued

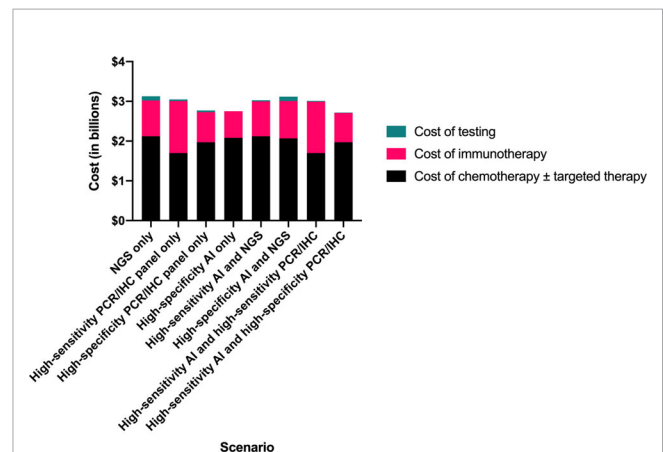
Model input	Assumed value (reference)
Median time on therapy for FOLFOX + cetuximab	10 (13)
% receiving 5-fluorouracil + leucovorin	20% (12)
Median time on therapy for 5-fluorouracil + leucovorin	4.4 (14)
Time between scans (months)	2.07 (5)
Number of pembrolizumab doses before first restaging scans	3
Time between pembrolizumab doses (months)	0.69 (5)
Number of chemotherapy ± targeted therapy doses before first restaging scans	5
Time between chemotherapy ± targeted therapy doses (months)	0.46 (5)

Superscript numbers represent references. Values without references are calculated from other values in the table unless otherwise noted.

<sup>a</sup>Internal, documentation available upon request.

<sup>b</sup>Among patients with pMMR/MSS disease, we assumed that patients ineligible for intensive therapy would receive 5-fluorouracil (5-FU) and leucovorin (LV). Among the remaining patients with pMMR/MSS disease, we assumed that all patients with RAS wild type disease would receive 5-FU, LV, oxaliplatin (FOLFOX), and cetuximab, while all patients with RAS mutant disease would receive FOLFOX and bevacizumab.

dMMR, deficient mismatch repair; FOLFOX, 5-fluorouracil, leucovorin, oxaliplatin; IHC, immunohistochemistry; MSI-H, microsatellite instability-high; MSS, microsatellite-stable; PCR, polymerase chain reaction; pMMR, proficient mismatch repair; U.S., United States.

**FIGURE 2 |** Comparison of total testing and treatment-related costs by clinical scenario. AI, artificial intelligence; IHC, immunohistochemistry; NGS, next-generation sequencing; PCR, polymerase chain reaction.

recommended 10 working days from laboratory sample receipt (15), and proportion of patients receiving first-line therapy supported by KEYNOTE-177. Since we did not have a direct way of linking these clinical consequences with clinical outcomes, we did not compare incremental cost-effectiveness ratios.

## RESULTS

We projected the high-sensitivity AI followed by high-specificity panel strategy to result in the lowest total testing and first-line drug therapy cost, \$2.72 billion, compared to \$3.13 billion for



NGS alone, representing savings of \$400 million (12.9%) (**Table 2**). The high-specificity panel-only and the high-specificity AI-only scenarios resulted in nearly as much cost savings (\$360 million and \$370 million, respectively).

The high-specificity AI-only scenario was associated with the shortest time to treatment initiation (<1 day) (versus 12 days for NGS), with 100% of patients receiving results within the guideline-recommended ten working days (versus 0% for NGS). Compared with the NGS-only scenario, in which all 32,549 (100%) patients received KEYNOTE-177-supported therapy, 31,442 of 32,549 (97%) patients received KEYNOTE-177-supported therapy in the high-specificity AI-only scenario.

We estimate that the accuracy of AI is similar to the accuracies of PCR and IHC in determining MSI/MMR status. For the high-sensitivity context (i.e., as screening tests), we estimate 98% sensitivity and 79% specificity for AI compared to 100% sensitivity and 81% specificity for PCR/IHC. In the high-specificity context (i.e., as confirmatory tests), we estimate 70% sensitivity and 98% specificity for AI compared to 67% sensitivity and 93% specificity for PCR/IHC.

## DISCUSSION

The \$400 million (12.9%) difference between the most and least expensive scenarios highlights that testing approach can significantly impact costs in the setting of first line mCRC. The least costly scenario, high-sensitivity AI with confirmatory high-specificity panel, comes with the tradeoff of 9% of patients (2,815) receiving a first-line therapy not supported by KEYNOTE-177 data (versus 0% with NGS-only). The second-least costly scenario, using high-specificity AI alone, results in only 3% of patients (1,107) receiving a non-supported therapy. It is our view that the ability to start therapy earlier due to elimination of treatment initiation delay (e.g., time for packing and shipping of tissue samples to outside facilities, time to conduct tests) may compensate, to some degree, for any reduction in median progression-free survival (PFS) resulting from that 3%. Moreover, the Kaplan-Meier PFS curves from KEYNOTE-177 suggest that PFS for pembrolizumab and chemotherapies are similar for the first eight months of therapy. Only after this timepoint do the curves separate, disease tends to progress (PFS 8.2 months), and patients will likely switch therapy. We could draw the conclusion, then, that chemotherapy offers similar benefit to pembrolizumab in dMMR/MSI-H disease for several months. If we accept this premise, at least in part, then perhaps treating 1,107 patients with a non-KEYNOTE-177-supported therapy, and avoiding additional immunotherapy cost, becomes more reasonable. If we consider where else in the health system the \$400 million in savings could be spent, the prospect becomes more palatable still.

It is important to acknowledge that established tests are only as powerful as the biomarkers that they assess. While 43.8% of dMMR/MSI-H patients respond to pembrolizumab, health systems would benefit from diagnostic tools that could help avoid using costly immunotherapy in the dMMR/MSI-H patients who will not respond (i.e., the majority of these patients). With potential to consider tumor characteristics beyond genetics (e.g., intratumoral

**TABLE 2** | Cost of therapy and clinical impact by diagnostic strategy.

	NGS only	High-sensitivity PCR/IHC only	High-specificity PCR/IHC only	High-sensitivity AI and NGS	High-specificity AI and NGS	High-sensitivity AI and high-sensitivity PCR/IHC	High-specificity PCR/IHC and high-specificity AI and NGS
Total cost of diagnostic testing and first-line therapy	\$3.13	\$3.05	\$2.76	\$2.75	\$3.03	\$3.12	\$2.72
Cost of chemotherapy ± targeted therapy	\$2.12	\$1.70	\$1.97	\$2.08	\$2.12	\$2.07	\$1.97
Cost of immunotherapy	\$0.90	\$1.31	\$0.76	\$0.67	\$0.88	\$0.94	\$0.74
Cost of testing	\$0.11	\$0.04	\$0.04	\$0.00	\$0.03	\$0.11	\$0.01
Cost savings compared to reference scenario (NGS only) (absolute)	Reference	\$0.07	\$0.36	\$0.37	\$0.10	\$0.01	\$0.40
Cost savings compared to reference scenario (NGS only) (relative)	Reference	2.3%	11.6%	11.9%	3.2%	0.2%	12.9%
Weighted average time to treatment initiation	12	4	4	0	3.0	11.4	1.2
Percent of patients receiving results within guideline-recommended 10 working days (15)	0%	100%	100%	100%	75%	5%	100%
Percent of patients receiving first-line therapy supported by KEYNOTE-177	100%	81%	91%	97%	80%	97%	91%

Dollar values presented in billions. AI, artificial intelligence; IHC, immunohistochemistry; NGS, next-generation sequencing; PCR, polymerase chain reaction.

heterogeneity, three-dimensional structure), AI could prove to be even more predictive than NGS. In other words, the promise of AI is not to be a cost-effective approximating of existing technologies, but rather an improvement upon them, both in terms of clinical utility and cost.

It is important to recognize that applying artificial intelligence to digital histopathology is only one cog in a much broader wheel of strategies to curb healthcare spending. For example, screening for early detection of colorectal cancer is another vital component of a greater program to curb costs, as screening is estimated to be associated with \$1.50 to \$2.00 in returns for each dollar spent (16). Uptake of cost-cutting measures like AI relies on appropriate financial incentives presented to hospitals and clinics. Whereas classic buy-and-bill outpatient reimbursement actually encourages overspending (as the reimbursement is pegged to the cost of the purchase), structures like the oncology care model encourage providers to make choices that curb costs. Health systems must seek to target multiple levers (e.g., at the levels of screening, diagnosis, and treatment) to achieve financial sustainability in oncologic health. Besides, any dollar saved from one sector within the field of oncologic health can be routed towards spending on those areas with the highest value (e.g., investment in screening).

The main limitation of our study was the use of a theoretical model, which will require real-world validation. The integrity of our estimates is dependent on the validity of the sources that we used to develop input assumptions. Although we aimed to use as few sources as possible to allow for some standardization among or assumptions, our assumptions are derived from a diverse range of sources. We aimed, too, to use data from high-quality, prospective clinical trials, where possible, but there were numerous cases in which the required data was only available in the form of retrospective analyses. Since the application of AI to histopathology diagnostics has not been widely used in clinical contexts, we do not yet have access to real-world cost and outcomes data. Our model did not consider important aspects like heterogeneity in the population and varying costs by setting and payer, instead assuming a monolithic U.S. healthcare system for demonstrative purposes. Each individual institution's initial fixed costs associated with implementing digital histopathology are also outside of the scope of our study. These costs might include purchasing or renting hardware (e.g., slide scanners) and software (e.g., cloud data storage) from digital histopathology vendors. On an ongoing basis, additional pathology personnel would likely be required to perform new tasks like internal validation, maintaining hardware and software, and scanning slides. However, multiple previous analyses have suggested that gains in efficiency and productivity associated with implementing digital histopathology more than pay for these upfront and ongoing costs (17, 18). There may be additional costs of which we are not currently aware, as potential costs may arise in the real world that have not been encountered before, given the novelty of this platform. It is important to be conscientious, too, of more abstract implementation hurdles like earning clinicians' confidence in new technologies. NGS, IHC, and PCR are trusted tools on which clinicians have long relied for guiding treatment decisions. Encouraging the adoption

of a technology unlike any of the current diagnostic tools may be an uphill battle in some contexts. Finally, any new tool of this kind must undergo rigorous validation to ensure that it offers equal benefit across demographic groups (e.g., by race, ethnicity, socio-economic status). Our study did not account for any such heterogeneity. Our conclusions would benefit greatly from validation with future independent cohorts.

While we used first-line therapy for mCRC as an example, we view these findings as relevant across cancers whose diagnostic algorithm involves genetic evaluation – with savings far beyond this sliver of total spending on cancer care. Not only would the initial investment in AI eventually pay for itself, but, because of the nature of the technology, AI improves as the platform “learns” from each sample. In this way, every dollar spent on AI is an investment in a better technology. This valuable characteristic of AI differentiates it and positions it as a vehicle for improving the quality and cost of cancer care.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

## AUTHOR CONTRIBUTIONS

AK - Performed analysis and wrote manuscript. GS, LS, NL, and NC - Provided analytical advice and edited manuscript. JK - Formulated research question, designed analysis, supervised research, and edited manuscript. AP - Formulated research question, designed analysis, supervised research, edited manuscript, and provided resources and funding. All authors contributed to the article and approved the submitted version.

## FUNDING

JK is funded by the Max-Eder-Programme of the German Cancer Aid (Bonn, Germany, grant # 70113864) and the START Programme of the Medical Faculty Aachen (Aachen, Germany, grant #691906). AP is funded by NIH/NIDCR K08-DE026500 and NIH/NCI U01-CA243075, and has research funding from the Adenoid Cystic Carcinoma Research Foundation, Cancer Research Foundation, and University of Chicago Comprehensive Cancer Center. NL is funded by NIMHD R01MD013420 and NIDDK P30 DK092949.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2021.630953/full#supplementary-material>

## REFERENCES

- Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-Cancer Image-Based Detection of Clinically Actionable Genetic Alterations. *Nat Cancer* (2020) 1(8):789–99. doi: 10.1101/833756
- Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology* (2020) 159(4):14067–16.E11. doi: 10.1053/j.gastro.2020.06.021
- Howlander NNA, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA eds. *Seer Cancer Statistics Review, 1975-2017 Bethesda, Md.* Bethesda, MD: National Cancer Institute (2020). Available at: [https://seer.cancer.gov/csr/1975\\_2017/](https://seer.cancer.gov/csr/1975_2017/). updated based on November 2019 SEER data submission, posted to the SEER web site, April 2020.
- Venderbosch S, Nagtegaal ID, Maughan TS, Smith CG, Cheadle JP, Fisher D, et al. Mismatch Repair Status and BRAF Mutation Status in Metastatic Colorectal Cancer Patients: A Pooled Analysis of the CAIRO, Cairo2, COIN, and FOCUS Studies. *Clin Cancer Res* (2014) 20(20):5322–30. doi: 10.1158/1078-0432.CCR-14-0332
- Andre T, Shiu K-K, Kim TW, Jensen BV, Jensen LH, Punt CJA, et al. Pembrolizumab Versus Chemotherapy for Microsatellite Instability-High/Mismatch Repair Deficient Metastatic Colorectal Cancer: The Phase 3 KEYNOTE-177 Study. *J Clin Oncol* (2020) 38(18\_suppl):LBA4–LBA. doi: 10.1200/JCO.2020.38.18\_suppl.LBA4
- Centers for Medicare & Medicaid Services. *Clinical Laboratory Fee Schedule Files 2020*.
- Gregg JP, Li T, Yoneda KY. Molecular Testing Strategies in non-Small Cell Lung Cancer: Optimizing the Diagnostic Journey. *Transl Lung Cancer Res* (2019) 8(3):286–301. doi: 10.21037/tlcr.2019.04.14
- Southey MC, Jenkins MA, Mead L, Whitty J, Trivett M, Tesoriero AA, et al. Use of Molecular Tumor Characteristics to Prioritize Mismatch Repair Gene Testing in Early-Onset Colorectal Cancer. *J Clin Oncol* (2005) 23(27):6524–32. doi: 10.1200/JCO.2005.04.671
- Barnetson RA, Tenesa A, Farrington SM, Nicholl ID, Cetnarskyj R, Porteous ME, et al. Identification and Survival of Carriers of Mutations in DNA Mismatch-Repair Genes in Colon Cancer. *N Engl J Med* (2006) 354(26):2751–63. doi: 10.1056/NEJMoa053493
- Kafatos G, Niepel D, Lowe K, Jenkins-Anderson S, Westhead H, Garawin T, et al. RAS Mutation Prevalence Among Patients With Metastatic Colorectal Cancer: A Meta-Analysis of Real-World Data. *Biomark Med* (2017) 11(9):751–60. doi: 10.2217/bmm-2016-0358
- Centers for Medicare & Medicaid Services. *2020 ASP Drug Pricing Files 2020*.
- Parikh RC, Du XL, Morgan RO, Lairson DR. Patterns of Treatment Sequences in Chemotherapy and Targeted Biologics for Metastatic Colorectal Cancer: Findings From a Large Community-Based Cohort of Elderly Patients. *Drugs Real World Outcomes* (2016) 3(1):69–82. doi: 10.1007/s40801-015-0059-9
- Heinemann V, von Weikersthal LF, Decker T, Kiani A, Vehling-Kaiser U, Al-Batran S-E, et al. FOLFIRI Plus Cetuximab Versus FOLFIRI Plus Bevacizumab as First-Line Treatment for Patients With Metastatic Colorectal Cancer (FIRE-3): A Randomised, Open-Label, Phase 3 Trial. *Lancet Oncol* (2014) 15(10):1065–75. doi: 10.1016/S1470-2045(14)70330-4
- Saltz LB, Douillard JY, Pirootta N, Alakl M, Gruia G, Awad L, et al. Irinotecan Plus Fluorouracil/Leucovorin for Metastatic Colorectal Cancer: A New Survival Standard. *Oncologist* (2001) 6(1):81–91. doi: 10.1634/theoncologist.6-1-81
- Sepulveda AR, Hamilton SR, Allegra CJ, Grody W, Cushman-Vokoun AM, Funkhouser WK, et al. Molecular Biomarkers for the Evaluation of Colorectal Cancer. *Am J Clin Pathol* (2017) 147(3):221–60. doi: 10.1093/ajcp/aqw209
- Orsak G, Miller A, Allen CM, Singh KP, McGaha P. Return on Investment of Free Colorectal Cancer Screening Tests in a Primarily Rural Uninsured or Underinsured Population in Northeast Texas. *Pharmacoecon Open* (2020) 4(1):71–7. doi: 10.1007/s41669-019-0147-y
- Ho J, Ahlers SM, Stratman C, Aridor O, Pantanowitz L, Fine JL, et al. Can Digital Pathology Result in Cost Savings? A Financial Projection for Digital Pathology Implementation at a Large Integrated Health Care Organization. *J Pathol Inform* (2014) 5(1):33. doi: 10.4103/2153-3539.139714
- Hanna MG, Reuter VE, Samboy J, England C, Corsale L, Fine SW, et al. Implementation of Digital Pathology Offers Clinical and Operational Increase in Efficiency and Cost Savings. *Arch Pathol Lab Med* (2019) 143(12):1545–55. doi: 10.5858/arpa.2018-0514-OA

**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.

Copyright © 2021 Kacew, Strohbehn, Saulsberry, Laiteerapong, Cipriani, Kather and Pearson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# The Immune Subtypes and Landscape of Gastric Cancer and to Predict Based on the Whole-Slide Images Using Deep Learning

Yan Chen<sup>1,2†</sup>, Zepang Sun<sup>3,4†</sup>, Wanlan Chen<sup>5†</sup>, Changyan Liu<sup>1</sup>, Ruoyang Chai<sup>1</sup>, Jingjing Ding<sup>1</sup>, Wen Liu<sup>1,2</sup>, Xianzhen Feng<sup>1</sup>, Jun Zhou<sup>1</sup>, Xiaoyi Shen<sup>1</sup>, Shan Huang<sup>2\*</sup> and Zhongqing Xu<sup>1\*</sup>

## OPEN ACCESS

### Edited by:

Heyrim Cho,  
University of California, Riverside,  
United States

### Reviewed by:

Esther Giehl,  
University Hospital Carl Gustav Carus,  
Germany  
Benjamin Alexander Kansy,  
Essen University Hospital, Germany

### \*Correspondence:

Zhongqing Xu  
Zhongqing\_xu@126.com  
Shan Huang  
hs1147@126.com

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Cancer Immunity and  
Immunotherapy,  
a section of the journal  
Frontiers in Immunology

**Received:** 26 March 2021

**Accepted:** 24 May 2021

**Published:** 28 June 2021

### Citation:

Chen Y, Sun Z, Chen W, Liu C,  
Chai R, Ding J, Liu W, Feng X,  
Zhou J, Shen X, Huang S and Xu Z  
(2021) The Immune Subtypes and  
Landscape of Gastric Cancer and to  
Predict Based on the Whole-Slide  
Images Using Deep Learning.  
*Front. Immunol.* 12:685992.  
doi: 10.3389/fimmu.2021.685992

<sup>1</sup> Department of General Practice, Tongren Hospital, Shanghai Jiao Tong University, School of Medicine, Shanghai, China,

<sup>2</sup> Department of Endocrinology, Tongren Hospital, Shanghai Jiao Tong University, School of Medicine, Shanghai, China,

<sup>3</sup> Department of General Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, China, <sup>4</sup> Guangdong Provincial Key Laboratory of Precision Medicine for Gastrointestinal Tumor, Guangzhou, China, <sup>5</sup> Department of Cardiology, Tongren Hospital, Shanghai Jiao Tong University, School of Medicine, Shanghai, China

**Background:** Gastric cancer (GC) is a highly heterogeneous tumor with different responses to immunotherapy. Identifying immune subtypes and landscape of GC could improve immunotherapeutic strategies.

**Methods:** Based on the abundance of tumor-infiltrating immune cells in GC patients from The Cancer Genome Atlas, we used unsupervised consensus clustering algorithm to identify robust clusters of patients, and assessed their reproducibility in an independent cohort from Gene Expression Omnibus. We further confirmed the feasibility of our immune subtypes in five independent pan-cancer cohorts. Finally, functional enrichment analyses were provided, and a deep learning model studying the pathological images was constructed to identify the immune subtypes.

**Results:** We identified and validated three reproducible immune subtypes presented with diverse components of tumor-infiltrating immune cells, molecular features, and clinical characteristics. An immune-inflamed subtype 3, with better prognosis and the highest immune score, had the highest abundance of CD8+ T cells, CD4+ T-activated cells, follicular helper T cells, M1 macrophages, and NK cells among three subtypes. By contrast, an immune-excluded subtype 1, with the worst prognosis and the highest stromal score, demonstrated the highest infiltration of CD4+ T resting cells, regulatory T cells, B cells, and dendritic cells, while an immune-desert subtype 2, with an intermediate prognosis and the lowest immune score, demonstrated the highest infiltration of M2 macrophages and mast cells, and the lowest infiltration of M1 macrophages. Besides, higher proportion of EVB and MSI of TCGA molecular subtyping, over expression of CTLA4, PD1, PDL1, and TP53, and low expression of JAK1 were observed in immune subtype 3, which consisted with the results from Gene Set Enrichment Analysis.



These subtypes may suggest different immunotherapy strategies. Finally, deep learning can predict the immune subtypes well.

**Conclusion:** This study offers a conceptual frame to better understand the tumor immune microenvironment of GC. Future work is required to estimate its reference value for the design of immune-related studies and immunotherapy selection.

**Keywords:** tumor-infiltrating immune cells, immune subtypes, immunotherapy, deep learning, gastric cancer

## INTRODUCTION

Gastric cancer (GC) is the fifth most common malignant tumor and third leading cause of cancer-related death worldwide (1). Despite major advancements in therapies, the 5-year overall survival (OS) rate for patients in advanced stage remains 20% (2). Even patients with locally advanced disease underwent radical resection and perioperative chemotherapy, the 5-year OS rate is still less than 40% (3–7). Thus, more effective systemic treatments are obviously urgent.

Immunotherapy is catching attention in multiple solid tumors recently, including gastric cancer. Specifically, immune checkpoint inhibitors, such as cytotoxic T-lymphocyte associated protein 4 (CTLA4) antibodies and programmed cell death protein 1 (PD1) antibodies, presented unprecedented clinical benefit in a variety of tumors (8–18). However, for patients with advanced gastric cancer, only a small subset (10–20%) responded to anti-CTLA4 (ipilimumab) and anti-PD1 (nivolumab, pembrolizumab) (8–12). A randomized controlled phase 3 trial ONO-4538-12/ATTRACTION-2 indicates an improvement of objective response rate (ORR) of 11% for patients with advanced gastric cancer receiving nivolumab versus placebo (10). Also, the ORR remains similar for other clinical trials, including the phase 1b KEYNOTE-012 (ORR 22%) and phase II KEYNOTE-059 (ORR 12%) trials (9, 11). Therefore, researches to identify mechanisms of response and resistance to immune checkpoint inhibition and to screen underlying patients who may benefit are required. However, our understanding of the role of tumor microenvironment (TME) in immune response remains incomplete because of its complexity.

The tumor microenvironment is a complex system composed of extracellular matrix, cytokines, chemokines, and non-tumor cells (19). As an important component of non-tumor cells in TME, tumor infiltrating immune cells (TIICs) is associated with the promotion or inhibition of tumor growth (20–22). In particular, the presence of tumor-associated CD8<sup>+</sup> T cells, CD4<sup>+</sup> T cells, T follicular helper cells (T<sub>fh</sub>s), and natural killer (NK) cells in TME, suggesting activated immune response, is associated with good prognosis, while regulatory T cells (T<sub>regs</sub>), B cells, macrophages, mast cells and plasma cells inhibiting immune response indicate poor prognosis (22–31). Conventional detection techniques for TIICs, such as flow cytometry and immunohistochemistry, are generally confined to evaluate limited types of immune cells, due to inability to measure numbers of markers simultaneously (29, 32). However, the interactions among tumor-infiltrating immune cells are

extremely complicated. Thus, a systematic assessment of all immune cells in the TME offers better clinical value.

Immune subtypes have presented with meaningful clinical value in multiple tumors, including melanoma, esophageal cancer, lung cancer, and breast cancer (33, 34). Although the relationship between tumor infiltrating immune cells and gastric cancer has been described, the overall function of TME is ignored (35). Therefore, our understanding of the immune subtypes based on TIICs in gastric cancer is far from complete. From this perspective, our study is of great significance.

Deep learning performs excellently as a powerful technique for reading pathological images (36, 37). The emergence of pathological scanning copy for the whole slide images (WSIs) provides a platform for deep learning (34, 37). It is generally acknowledged that the histopathology images contain valuable information of TME (38). Therefore, deep learning could extract high dimensional data from standard medical images for clinical applications, such as distinguishing immune subtypes. Besides, convincing performance for deep learning has been observed in prediction of microsatellite instability status, immune cell types and prognosis in a variety of tumors (39–42), which provides reference for our study.

In the present study, we identified three robust immune subtypes of gastric cancer based on unsupervised consensus clustering of TIICs, and their reproducibility was further validated in an independent cohort. We observed that each of the three immune subtypes presented distinct immune cells proportion, molecular features, and clinical characteristics, which could provide reference for the design of immune-related studies and the choice of immunotherapy. Moreover, we verified the feasibility and prognostic value of this classification system in five pan-cancer data sets, including breast cancer, esophageal cancer, colorectal cancer, liver cancer, and pancreatic cancer. Finally, we developed and validated a deep learning model based on pathological images to predict the immune subtypes for easy-use in clinical practice.

## MATERIALS AND METHODS

### Patients and Data Sets

The discovery cohort to identify the immune subtypes consisted of 375 patients with gastric cancer obtained from The Cancer Genome Atlas (TCGA) database (<https://cancergenome.nih.gov>). Another cohort including 433 patients with gastric cancer in GSE84437 downloaded from the Gene Expression Omnibus



(GEO) database was used to validate the immune subtypes (<https://www.ncbi.nlm.nih.gov/geo/>). Besides, five independent cohorts (total  $n = 2230$ ), including breast cancer ( $n=1108$ ), esophageal cancer ( $n=185$ ), colorectal cancer ( $n=383$ ), liver cancer ( $n=375$ ), and pancreatic cancer ( $n=179$ ), acquired from UCSC Xena (<https://xenabrowser.net/>) were applied to further elucidate feasibility of the immune subtypes. For details about study design and data preprocessing, please refer to supplementary methods and **Figure S1**.

## Data Processing and Quantification of Immune Cells

Based on the gene expression profiles, the CIBERSORT algorithm was employed to quantify the proportions of 22 Tumor-infiltrating immune cells using the LM22 signature and 1,000 permutations (43). Cases with  $P < 0.05$  in CIBERSORT, which indicated that the deconvolution results were accurate, would be retained for further analysis. In this study, a total of 194 GC samples from discovery cohort and 299 GC samples from validation cohort were screened out (**Figure S2**). Finally, we obtained 22 types of immune cells, including B cells (naive B cells and memory B cells), CD8+ T cells, naive CD4+ T cells, resting memory CD4+ T cells, activated memory CD4+ T cells, T follicular helper cells (Tfh), regulatory T cells (Tregs), natural killer cell (resting NK cells, activated NK cells), macrophages (M0, M1 and M2), dendritic cells (resting DC and activated DC), mast cells (resting mast cells and activated mast cells), plasma cells, gamma delta T cells, monocytes, neutrophils, and eosinophils (**Figure S3**).

## Discovery and Validation of the Immune Subtypes

To dissect inter-tumor heterogeneity defined by TIICs, we applied unsupervised consensus clustering to define the robust subgroup of patients, i.e., immune subtypes. Specifically, the K-Means clustering algorithm with the Euclidean distance metric and performed 10,000 bootstraps, with 80% resampling of the immune cells. The consensus clustering algorithm was implemented with the ConsensusClusterPlus package (44). The number of clusters was determined by the optimal consensus matrix and explicit cluster allocation across permuted runs. Besides, in order to evaluate the reproducibility of the clusters, the same clustering procedure was performed independently in the validation cohort. We then calculated the in-group proportion (IGP) index with “clusterRepro” R package to quantitatively measure the similarity of clusters produced from the two data sets (45).

## Assessing the Clinical, Molecular, Cellular Characteristics Associated With the Immune Subtypes

We first evaluated the association of immune-related cellular features with immune subtypes using Kruskal-Wallis statistic. TIICs (naive CD4+ T cells, gamma delta T cells, monocytes, neutrophils, and eosinophils) with zero value in more than 40% of all samples were excluded from the analysis. Next, we described the distribution of demographic, clinicopathological characteristics, and molecular

feature of the immune subtypes, including age, sex, Lauren’s classification, pathological differentiation status, tumor location, stage, TCGA molecular subtyping, and stromal-immune score based on ESTIMATE algorithm (46, 47). Finally, log-rank test and multivariable Cox regression were used to measure the prognostic value of the immune subtypes with OS as the endpoint. For details about identification of TCGA molecular subtyping, please refer to **Supplementary Methods**.

## Validation Using Pan-Cancer Data Set

Tumor-infiltrating immune cells data were extracted based on the CIBERSORT method described above from the pan-cancer data sets (breast cancer, esophageal cancer, colorectal cancer, liver cancer, and pancreatic cancer). However, for those cohorts with too few samples, we chose  $P < 0.1$  as the cutoff point. Then, the consensus clustering algorithm and Kaplan-Meier analysis were performed to illustrate the feasibility of our immune subtypes.

## Functional Enrichment Analyses for Immune Subtypes

Differentially expressed genes (DEGs) were identified between any two immune subtypes (IS1 vs IS2, IS1 vs IS3, IS2 vs IS3) using an R package “limma”. An absolute value of  $\log_2$  (fold change)  $> 1$  combined with the false discovery rate (FDR) adjusted p-value  $< 0.05$  was selected as the threshold for DEG identification. The intersection of the DEGs in TCGA-GC cohort and GSE84437 cohort was applied to Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Set Enrichment Analysis (GSEA). For the enrichment analysis, we focused on the immune related gene sets and cancer hallmark gene sets. Besides, several classic immune checkpoints (PD1, PDL1, and CTLA4) and cancer related genes (TP53, JAK1) were evaluated among the immune subtypes.

## Deep Learning to Identify Immune Subtypes

Deep learning can identify the macroscopic contents of pathological images, including tumor cells and TIICs nuclear size, nuclear location, nuclear morphology, etc. It can even identify high-dimensional data, such as color matrix, histogram matrix, and high-order matrix, which cannot be distinguished by naked eye. Thus, we trained a convolutional neural network with deep residual learning (based on ResNet-18) model to detect the immune subtype by transfer learning using patches segmented from the whole slide images (WSIs). First, high-quality WSIs without obvious interfering factors, including bleeding, creases, necrosis, and blurred areas, were screened and divided into training, validation and test sets at a 5:3:2 ratio for further processing. Next, tumor regions of interest (ROIs) on WSIs were manually delineated by expert pathologists. All WSIs were digitalized at 20× objective lens. Then, ROIs were subsequently separated into 512 pixels  $\times$  512 pixels patches. Finally, after preprocessed with random cutting, random horizontal flipping, and random affine transformation, center cropping (224 pixels  $\times$  224 pixels), and normalization, patches were put into the deep learning model based on ResNet-18. For details about data preprocessing, please refer to **Supplementary Methods**.

## Statistical Analysis

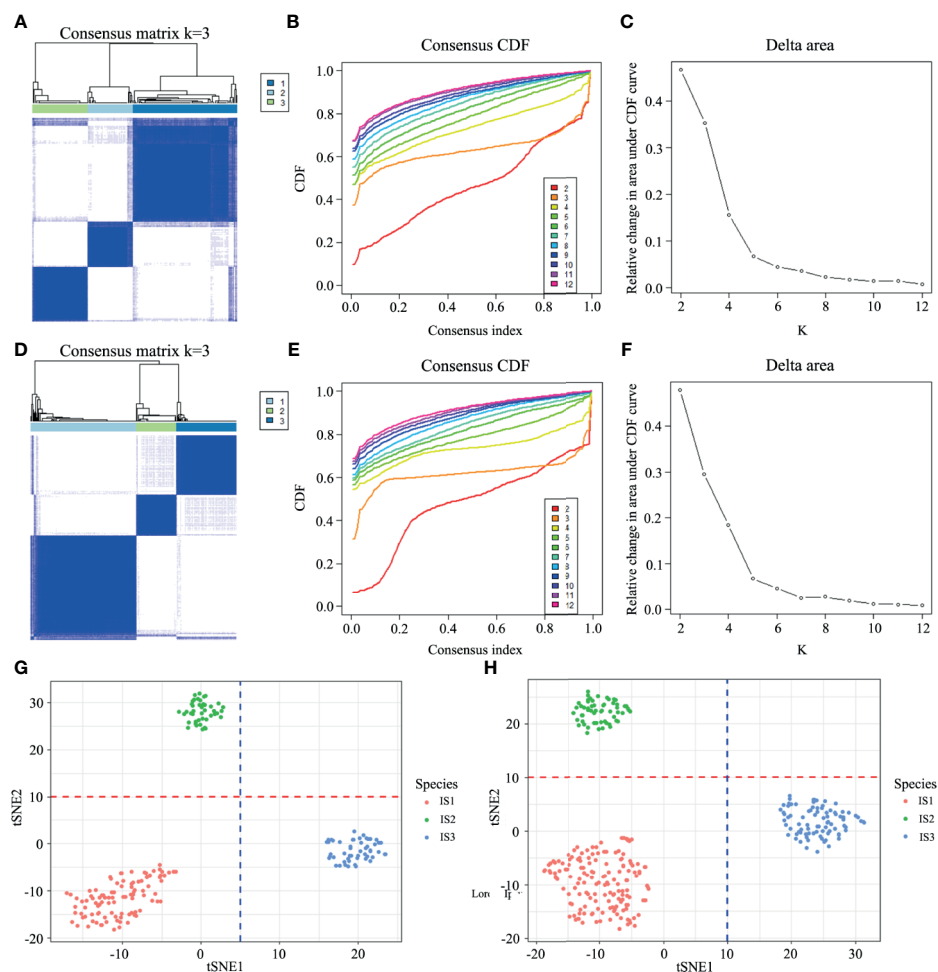
All the statistical significance values were set as two-tail and  $P < 0.05$  was considered statistically significant. Statistical analyses were performed using SPSS statistical software (version 22.0), GraphPad Prism software (version 7.00), Perl 5 software (version 5.28.1) and R software (version 3.5.3). Deep learning was implemented with torch library in Python software (version 3.6.7).

## RESULTS

### Immune Subtypes Discovery and Validation

By performing the unsupervised consensus clustering on the 194 GC cases from TCGA based on the 22 TIICs, the optimal number of

clusters was found to be three with maximal consensus within clusters and minimal ambiguity among clusters (**Figures 1A–C**). Based on this, we identified three robust immune subtypes—immune subtype 1 (IS1), immune subtype 2 (IS2) and immune subtype 3 (IS3). To evaluate the reproducibility of the immune subtypes, we performed the same algorithm in the 299 GC cases from GSE84437. Interestingly, we found that the optimal number of clusters was three, too (**Figures 1D–F**). The tSNE analysis well represented the discrete distribution of three clusters and the consistency of the discovery and validation cohorts (**Figures 1G, H**). Furthermore, we calculated the in-group proportion (IGP) statistic to quantify the similarity of the immune subtypes between the discovery and validation cohort. And immune subtypes showed good consistency between the two cohorts, with corresponding IGP value at 79%, 81%, and 86% in IS1, IS2, and IS3, respectively.

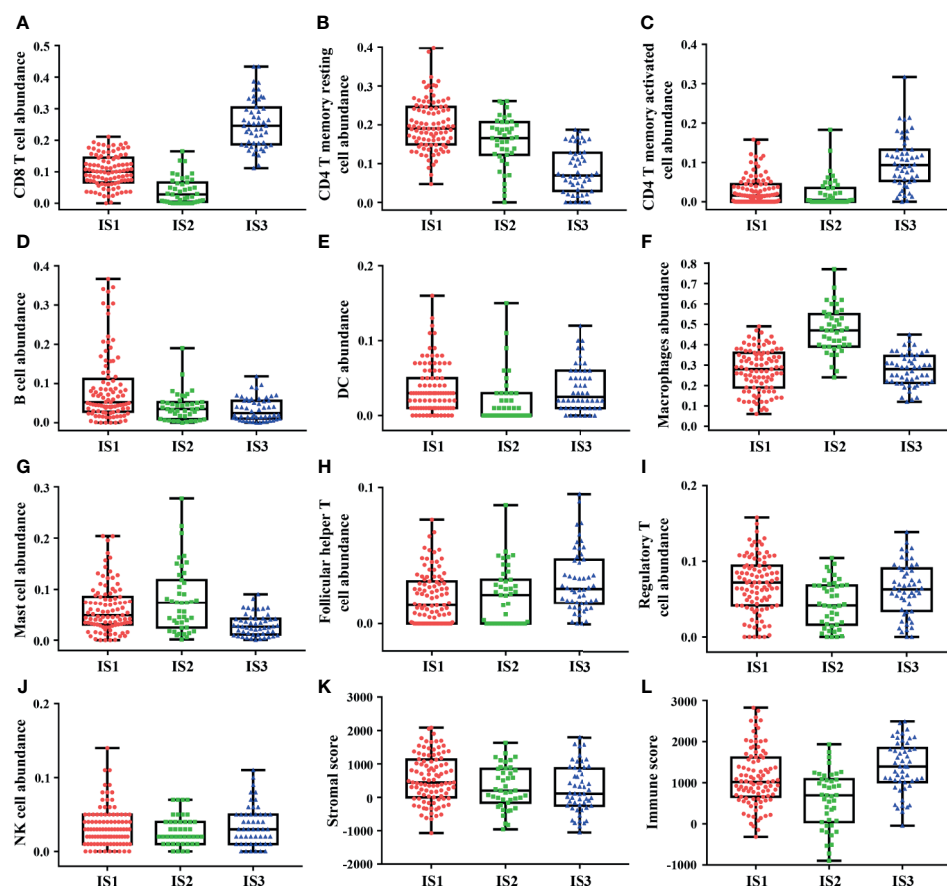


**FIGURE 1** | Discovery and validation of the immune subtypes in TCGA (**A**) and GEO (**D**). Patient samples are both in rows and columns, and consensus values range from 0 (never clustered together) to 1 (always clustered together). The optimal cluster number ( $K = 3$ ) is determined by the area under the cumulative distribution function (CDF) curve in the discovery (**B, C**) and validation cohort (**E, F**), which corresponds to the largest number of clusters that induced the smallest incremental change in the area under the CDF curves. The tSNE well represents the discrete distribution of three clusters (**G, H**).

## Profound Differences in Immune Infiltration Among Immune Subtypes

Each of the three immune subtypes represented distinct immune cells expression patterns in the discovery cohort, which was found to be highly consistent with the validation cohort, surprisingly. Highest CD8+ T cells and M1 macrophages abundance was confirmed in IS3 (Figures 2A and S5B). IS3 was also characterized by the highest abundance of activated CD4+ T memory cells (Figure 2C). However, the least abundance of resting CD4+ T memory cells and M2 macrophages presented in IS3 (Figures 2B and S5C). Moreover, IS3 was also associated with highest abundance of Tfh and NK cells (Figures 2H, J) and lowest abundance of B cells and mast cells (Figures 2D, G). Besides, the expression of DC and Tregs of IS3 was in the middle among three immune subtypes (Figures 2E, I), whereas the expression of plasma cells in IS3 was high in the discovery cohort and low in the validation cohort (Figure S4I, K). In comparison, IS1 exhibited the highest density of CD4+ T memory resting cells, B cells, DC cells, and Tregs (Figures 2B, D, E, I), and the lowest density of macrophages and Tfh (Figures 2F, H). The

expression of CD8+ T cells of IS1 was in the middle among three immune subtypes (Figure 2A). Furthermore, compared with IS1 and IS3, the highest abundance of macrophages in IS2 was confirmed (Figure 2F), accompanied with the lowest abundance of CD8+ T cells, DC and Tregs (Figures 2A, E, I). Besides, the highest abundance of M0 and M2 macrophages was observed in IS2 (Figures S5A, C, D, F), while the lowest abundance of M1 macrophages was observed in IS2 compared with IS1 and IS3 (Figure S5). And the expression of CD4+ T memory resting cells, B cells, Tfh of IS2 was in the middle among three immune subtypes (Figures 2B, D, H). The expression of activated CD4+ T memory cells, mast cells, NK cells was inconsistent in IS1 or IS2 between the discovery cohort and validation cohort (Figures 2C, S4C, 2G, S4G, 2J and S4J). Additionally, comparison of TIICs between any two immune subtypes (IS1 vs IS2, IS1 vs IS3, IS2 vs IS3) was provided in supplementary results. Lastly, we identified that IS3 exhibited the lowest stromal score and highest immune score (Figures 2K, L), while IS1 exhibited the highest stromal score (Figures 2K) and IS2 exhibited the lowest immune score



**FIGURE 2** | The discovery cohort shows heterogeneity of immune infiltration among immune subtypes. Highest abundance of CD8+ T cells, CD4+ T memory activated cells, follicular helper T cells, and NK cells was observed in IS3 (A, C, H, J), while highest abundance of CD4+ T memory resting cells, B cells, macrophages cells, and mast cells was observed in IS3 (B, D, F, G). DC cells and regulatory T cells abundance showed one highest and one lowest in IS1 and IS2 (E, I). Besides, IS3 exhibited the lowest stromal score and highest immune score (K, L). The plot of patient immune cells abundance shows the median, 25th and 75th percentile values (horizontal bar, bottom, and top bounds of the box), and the highest and lowest values (top and bottom whiskers, respectively).

(Figures 2L). The results of these findings were consistent with the validation cohort (Figures S4 and S5). Specific values of TIICs abundance and their P values are presented in Table 1. These results suggested that IS3 had the strongest immune activity accompanied with a weaker immune-suppression (immune-inflamed phenotype), while IS1 had a moderate immune response accompanied by a stronger immune-suppression (immune excluded phenotype), while IS3 was characterized by immune deficiency (immune-desert phenotype).

## Clinical Characteristics, Molecular Features, and Prognoses of the Immune Subtypes

The TCGA cohort containing GC patients with available clinicopathologic information and molecular features, stratified by immune subtypes, was analyzed and listed in Table 2. Compared to IS1 and IS3, the median age of IS2 is slightly higher (Figure 3B). In addition, IS2 was associated with highest proportion of men and intestinal type tumor (Figures 3A, F). Furthermore, IS3 was associated with a lower incidence of cardia/fundus cancer, while presented with worse pathological differentiation (Figures 3D, E). Besides, there was no significant difference in the proportion of TNM stages among the three immune subtypes (Figure 3C). In terms of TCGA molecular subtyping, IS3 revealed more EVB and MSI, and less CIN and GS than that in 1 and 2 (Figure 3G). The clinicopathological information available in the validation cohort is listed in Table S1 and Figure S6. Lastly, we observed that the immune subtypes revealed significantly prognostic impact in TCGA-GC and GEO cohort (Figures 3H, I). Overall, the immune-hot subtype IS3 was associated with the best prognosis for OS among all subtypes. By contrast, the immune-cold subtype IS1 and IS2 was associated with poor outcomes. This survival difference was confirmed after excluding confounding factors of age, gender, tumor location, Lauren's classification, pathological differentiation and stage and was showed in Tables 3 and S2.

## Validation Using Pan-Cancer Data Set

The consensus clustering algorithm was conducted using the 22 TIICs based on patients in the pan-cancer data set (breast cancer, esophageal cancer, colorectal cancer, liver cancer, and pancreatic cancer). We observed that the optimal number of clusters was four in liver cancer, two in colorectal cancer, four in breast cancer, two in esophageal cancer, and two in pancreatic cancer. And survival difference was found in liver cancer, breast cancer, and pancreatic cancer. However, significant statistical differences were found only in liver cancer and pancreatic cancer. The total results were visualized in Figure S8.

## Functional Enrichment Analyses

To investigate the underlying functional differences among immune subtypes, we conducted GO and GSEA analyses on the differentially expressed genes. Through the abovementioned analysis, GC cases in TCGA and GEO database were divided into three immune subtypes —IS1, IS2, and IS3. Thus, the functional enrichment analyses were performed between any two immune subtypes (IS1 vs IS2, IS1 vs IS3, IS2 vs IS3). First, 1639 DEGs, including 737 up-regulated expression (UE) and 902 down-regulated expression

TABLE 1 | The proportion of tumor-infiltrating immune cells in patients with gastric cancer from TCGA and GEO.

Variables	TCGA				GEO			
	IS1 (n=99) median (IQR)	IS2 (n=43) median (IQR)	IS3 (n=52) median (IQR)	P value	IS1 (n=153) median (IQR)	IS2 (n=59) median (IQR)	IS3 (n=87) median (IQR)	P value
B cell	0.053 (0.029-0.117)	0.035 (0.010-0.053)	0.025 (0.011-0.056)	<0.001	0.055 (0.029-0.092)	0.042 (0.017-0.067)	0.024 (0.011-0.053)	<0.001
Plasma cells	0.039 (0.009-0.066)	0.018 (0.00-0.018)	0.028 (0.015-0.068)	0.089	0.015 (0.00-0.073)	0.002 (0.00-0.035)	0.024 (0.00-0.063)	0.0348
CD8+ T cell	0.100 (0.066-0.145)	0.028 (0.005-0.028)	0.246 (0.187-0.304)	<0.001	0.064 (0.025-0.117)	0.030 (0.008-0.061)	0.159 (0.111-0.222)	<0.001
CD4+ T memory resting cells	0.190 (0.149-0.246)	0.166 (0.122-0.207)	0.069 (0.029-0.128)	<0.001	0.208 (0.168-0.285)	0.130 (0.086-0.189)	0.050 (0.003-0.085)	<0.001
CD4+ T memory activated cells	0.016 (0.00-0.045)	0.004 (0.00-0.035)	0.093 (0.053-0.132)	<0.001	0.028 (0.003-0.063)	0.045 (0.00-0.098)	0.173 (0.121-0.224)	<0.001
T follicular helper cells	0.013 (0.00-0.031)	0.021 (0.00-0.033)	0.026 (0.015-0.047)	0.002	0.024 (0.00-0.046)	0.023 (0.001-0.051)	0.041 (0.018-0.070)	<0.001
T regulatory cells	0.072 (0.042-0.094)	0.042 (0.016-0.068)	0.063 (0.034-0.091)	0.001	0.002 (0.00-0.016)	0.006 (0.00-0.028)	0.00 (0.00-0.008)	<0.001
NK cells	0.029 (0.014-0.047)	0.021 (0.011-0.022)	0.030 (0.012-0.051)	0.515	0.026 (0.014-0.048)	0.037 (0.028-0.051)	0.046 (0.028-0.070)	<0.001
Macrophages	0.279 (0.189-0.361)	0.468 (0.393-0.551)	0.280 (0.214-0.345)	<0.001	0.223 (0.160-0.291)	0.474 (0.394-0.527)	0.243 (0.205-0.305)	<0.001
DC cell	0.028 (0.011-0.053)	0.004 (0.00-0.027)	0.026 (0.014-0.061)	<0.001	0.022 (0.012-0.035)	0.011 (0.00-0.029)	0.020 (0.011-0.045)	0.004
Mast cell	0.049 (0.031-0.086)	0.074 (0.025-0.118)	0.026 (0.011-0.041)	<0.001	0.159 (0.099-0.214)	0.089 (0.061-0.121)	0.080 (0.048-0.126)	<0.001

IS1, immune subtype 1; IS2, immune subtype 2; IS3, immune subtype 3.



**TABLE 2 |** Clinicopathological characteristics of patients with gastric cancer in TCGA.

Variables	TCGA					
	IS1 (n=99)		IS2 (n=43)		IS3 (n=52)	
	N	%	N	%	N	%
<b>Age (median, IQR, Y)</b>	65 (57-72)		69 (58-75)		65 (56-75)	
<b>Gender</b>						
Male	58	58.6	34	79.1	30	57.7
Female	41	41.4	9	20.9	22	42.3
<b>Lauren's type</b>						
Intestinal	29	47.5	22	84.6	17	54.8
Diffuse	32	52.5	4	15.4	14	45.2
Unknown	38	NA	17	NA	21	NA
<b>Differentiation</b>						
Well	26	26.3	24	55.8	6	11.5
Poor	73	73.7	19	44.2	46	88.5
<b>Location</b>						
Cardia/Fundus	41	43.2	20	47.6	15	29.4
Body	22	23.2	8	19.1	16	31.4
Antrum/Pylorus	32	33.6	14	33.3	20	39.2
Unknown	4	NA	1	NA	1	NA
<b>Stage</b>						
I	8	8.1	6	14.0	6	11.5
II	39	39.4	18	41.9	21	40.4
III	41	41.4	18	41.9	23	44.2
IV	11	11.1	1	2.3	2	3.8
<b>Stromal score (median, IQR)</b>	445.4 (13.8-1104.2)		202.7 (-161-853.8)		113.9 (-248-853.2)	
<b>Immune score (median, IQR)</b>	1017.9 (659.1-1611.7)		694.7 (45.9-1086.7)		1394.3 (1016.3-1848.1)	
<b>Molecular characterization</b>						
EVB	3	3.0	0	0	19	37.3
MSI	14	14.1	7	16.7	17	33.3
CIN	53	53.5	31	73.8	11	21.6
GS	29	29.3	4	9.5	4	7.8
Unknown	0	NA	1	NA	1	NA

IS1, immune subtype 1; IS2, immune subtype 2; IS3, immune subtype 3; NA represents the meaningless values.

(DE), were filtered out from TCGA cohort and 208 DEGs (106 UE and 167 DE) were filtered out from GEO cohort in IS1 vs IS2 (Figures 4A, D, and S7). Next, a total of 115 DEGs (35 UE and 80 DE) were observed in the intersection between them (Table S4 and Figure 4G). In IS1 vs IS3, 1312 DEGs (363 UE and 949 DE) were filtered out from TCGA and 272 DEGs (55 UE and 217 DE) were filtered out from GEO (Figures 4B, E, and S7). Next, a total of 124 DEGs (31 UE and 93 DE) were observed in the intersection between them (Table S5 and Figure 4H). In IS2 vs IS3, 1685 DEGs (637 UE and 1048 DE) were filtered out from TCGA and 293 DEGs (111 UE and 182 DE) were filtered out from GEO (Figures 4C, F, and S7). Next, a total of 136 DEGs (65 UE and 71 DE) were observed in the intersection between them (Table S6 and Figure 4I). Furthermore, GO, KEGG, and GSEA analyses were performed based on the DEGs separately (Figures 4D, E).

From the above, we found significant difference of chemokine pathway between IS1 and IS2. IS presented with more active chemokine responsiveness and interactions (Figures 5A–C). Additionally, compared to IS3, TGF- $\beta$  signaling was significantly enriched in IS1 and IS2, which suggested immunosuppression (Figures 5F, I). Also, IS3 was associated with significantly upregulated T cell receptor signaling, antigen processing, and presentation signaling, suggesting that active inflammation and immune infiltration (Figures 5D, E, G, H). In addition, the classic

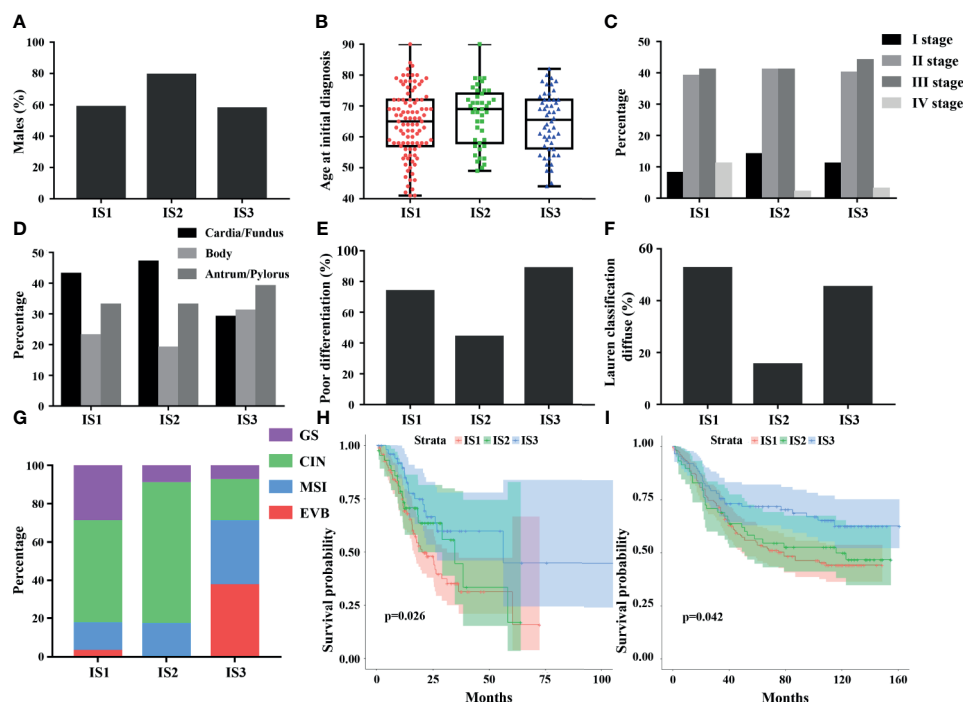
tumor suppressor signaling P53 was observed to enrich in IS3 and the typical carcinogenic signaling JAK-STAT enriched in IS1-2 (Figures 5F, I). The results consisted with the profound differences in immune infiltration among immune subtypes. And these may explain why IS3 has a better prognosis than IS1-2. Based on the above, we studied the relationship among several immune checkpoints (PD1, PDL1, and CTLA4), cancer related genes (TP53, JAK1) and the immune subtypes (Table S3). Interestingly, we found that compared with that in IS1-2, expression of PD1, PDL1, CTLA4, and TP53 was higher in IS3 and expression of JAK1 was lower, which was consistent with the functional enrichment analyses (Figure S9).

## Deep Learning Can Identify Immune Subtypes

After removing low quality pathological images, 169 samples with WSIs were divided into training (84 cases), validation (51 cases), and test cohorts (34 cases), and then tumor ROI was separated into  $512 \times 512$  patches. Finally, the training cohort contained 12,986 normalized tiles marked as IS1, 3,399 normalized tiles marked as IS2 and 6,323 normalized tiles marked as IS3. The validation cohort contained 11,070 normalized tiles marked as IS1, 3,003 normalized tiles marked as IS2, 5,114 normalized tiles marked as IS3. And the test cohort contained 5,344 normalized tiles marked as IS1, 1,790 normalized tiles marked as IS2, and 2,508 normalized

**TABLE 3** | Univariable and multivariable analyses for overall survival in patients with gastric cancer.

Variables	Univariable analysis (N=194)		Multivariable analysis (N=194)	
	OR (95%CI)	P	OR (95%CI)	P
Age (years)	1.018 (0.999-1.038)	0.061	NA	NA
Gender (female vs. male)	1.549 (0.984-2.437)	0.059	NA	NA
Lauren type (intestinal vs diffuse)	1.100 (0.859-1.409)	0.453	NA	NA
Differentiation (well vs. poor)	1.075 (0.671-1.723)	0.764	NA	NA
Location (cardia/fundus vs. body vs. antrum/pylorus)	0.865 (0.685-1.093)	0.224	NA	NA
Stage	1.399 (1.055-1.854)	0.020	1.371 (1.037-1.810)	0.026
Stromal score	1.020 (1.001-1.006)	0.086	NA	NA
Immune score	1.032 (1.000-1.144)	0.800	NA	NA
Immune subtype				
IS1	1	NA	1	NA
IS2	0.729 (0.425-2.148)	0.249	0.747 (0.435-1.282)	0.289
IS3	0.480 (0.277-0.834)	0.009	0.491 (0.282-0.853)	0.012

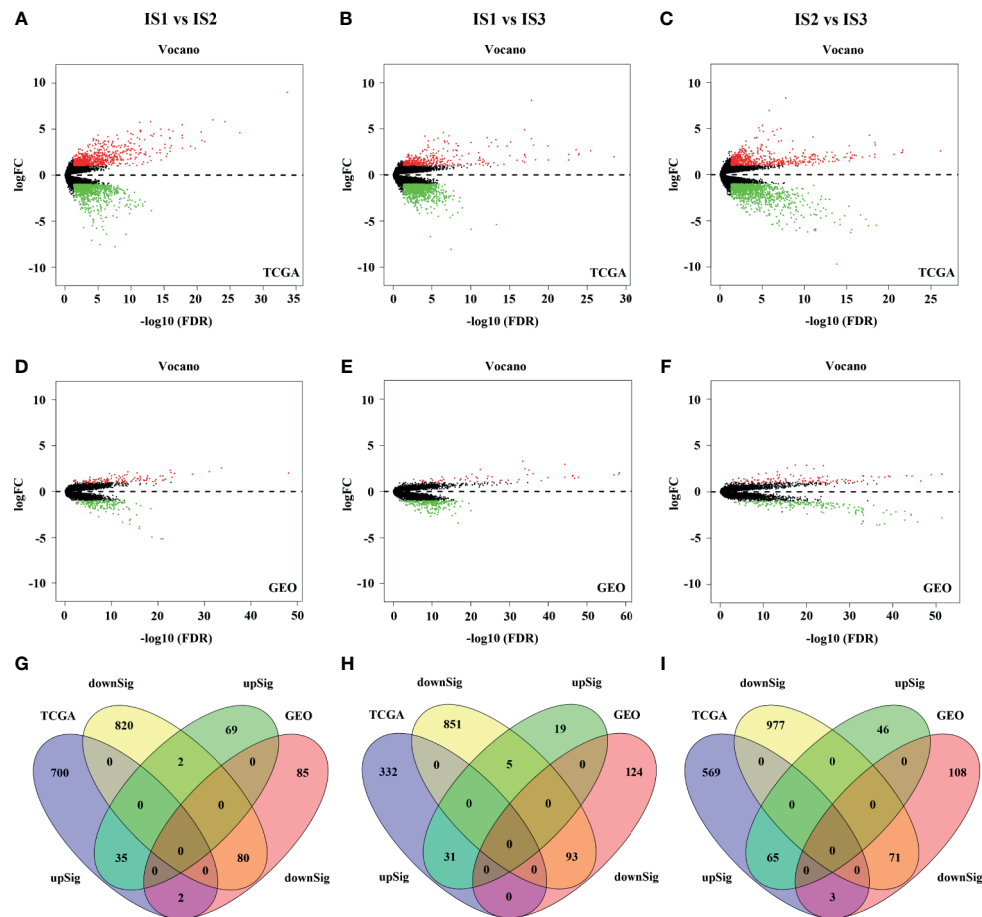
**FIGURE 3** | Differences in clinical and histological characteristics among immune subtypes, including age, sex, stage, tumor location, pathological differentiation, and Lauren classification (A–F). The plot of patient age at initial diagnosis shows the median, 25th and 75th percentile values (horizontal bar, bottom, and top bounds of the box), and the highest and lowest values (top and bottom whiskers, respectively). The distribution of TCGA molecular subtyping among immune subtypes (G). The prognostic value of the immune subtypes in TCGA (H) and GEO (I), indicating best prognosis of IS3.

tiles marked as IS3. Next, we developed a ResNet-18 deep learning model to predict the immune subtypes based on training and validation, and measured the performance in the test cohort. The model first predicted the probability of immune subtypes for each patch. We found that the accuracy of IS prediction for each patch in the training, validation, and test cohort was 80.23%, 74.45%, and 68.89%, respectively. Then GC cases would be designated as one of the three subtypes (IS1 or IS2 or IS3) according to the accumulated number of patches in tumor ROI (Figure 6). We observed that the accuracy of IS prediction ResNet-18 model for GC cases was about 85.71%, 80.39%, 76.47% in the training, validation, and test cohorts,

separately (Figures S10A, C). Additionally, we observed that the accuracy of IS3 prediction would increase to about 90% when IS1 and the two were combine as IS1-2 (Figures S10D–F). More details refer to supplementary results.

## DISCUSSIONS

Immunotherapy is increasingly being recognized for its potential therapeutic effect on a variety of tumors. However, only a subset of patients has response or survival benefit to immunotherapy. This may



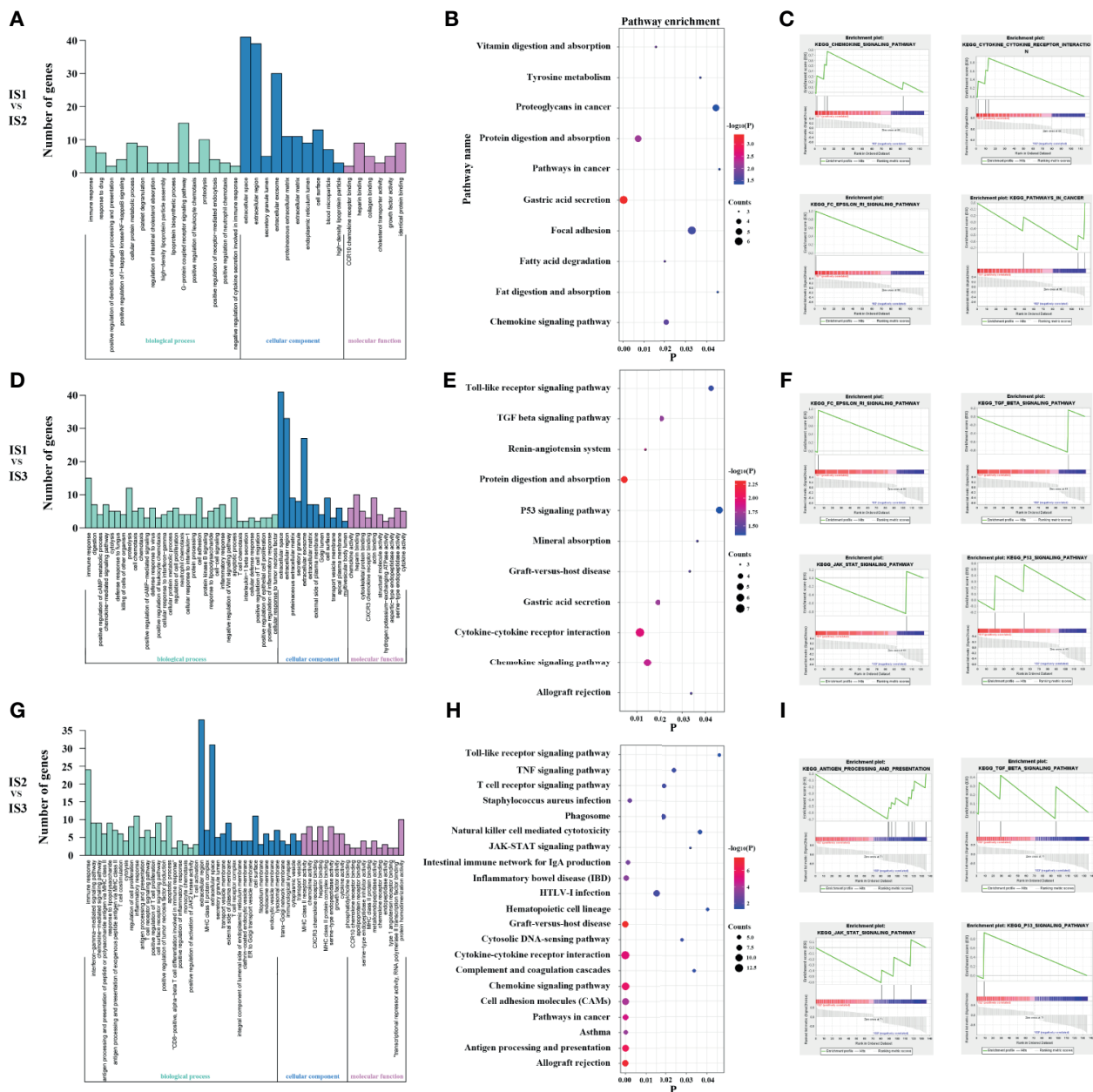
**FIGURE 4** | The results of differential expression analysis in the TCGA and GEO cohort (A–F). 115 DEGs was found in IS1 vs IS2 (G). 124 DEGs was found in IS1 vs IS3 (H). 136 DEGs was found in IS2 vs IS3 (I).

be caused by our incomplete understanding of the tumor immune microenvironment. Thus, to better understand the tumor immune microenvironment and further to filter out patients suitable for immunotherapy are particularly important. Here, we present the identification and validation of three reproducible immune subtypes of GC in a retrospective study with multiple cohorts. We observe that each of the immune subtypes presented with distinct composition of tumor infiltrating immune cells, and hence demonstrated widely different modes in gene expression profiles, functional orientation, molecular feature and clinical characteristics. Moreover, validation of a pan-cancer cohort can reinforce the credibility of our results. Lastly, a deep learning model with good performance to predict the status of immune subtypes in gastric cancer based on the whole-slide images is presented. This study provides a concept of immune subtypes to understand the immune microenvironment of GC and make it easy-use in clinical implications, which may have benefit for personalized immunotherapy and prognosis evaluation.

Immune microenvironment has been confirmed to be associated with prognosis in gastric cancer. However, traditional methods simply describe the relationship between the cell composition of the immune microenvironment and prognosis

according to the known outcome. Our method is ‘unsupervised’, which can better represent the complex and obscure information within the immune microenvironment. Significant survival differences are observed among the immune subtypes in this study, which can be a supplement to traditional TNM staging system. Specifically, an immune-hot subtype 3 presents with better prognosis, and by contrast, the immune-cold subtype 1 to 2 demonstrated a poor prognosis. Furthermore, the proportion EVB and MSI of the TCGA molecular subtyping in the IS3 are significantly higher than that in IS1-2, which is consistent with the previous report (48). EVB and MSI subtyping always present a more active immune response.

Appropriate classification for GC is essential to individual treatment. Several subtyping systems have been proposed in the past few decades, including the World Health Organization (WHO) classification, the Lauren’s classification, intrinsic Subtypes, Lei subtypes, The Cancer Genome Atlas (TCGA) subtypes, Asian Cancer Research Group (ACRG) subtypes, and some other additional classifications (49). Some are based on morphology or pathology, and some are based on the molecular and genetic features. However, classification based on immune



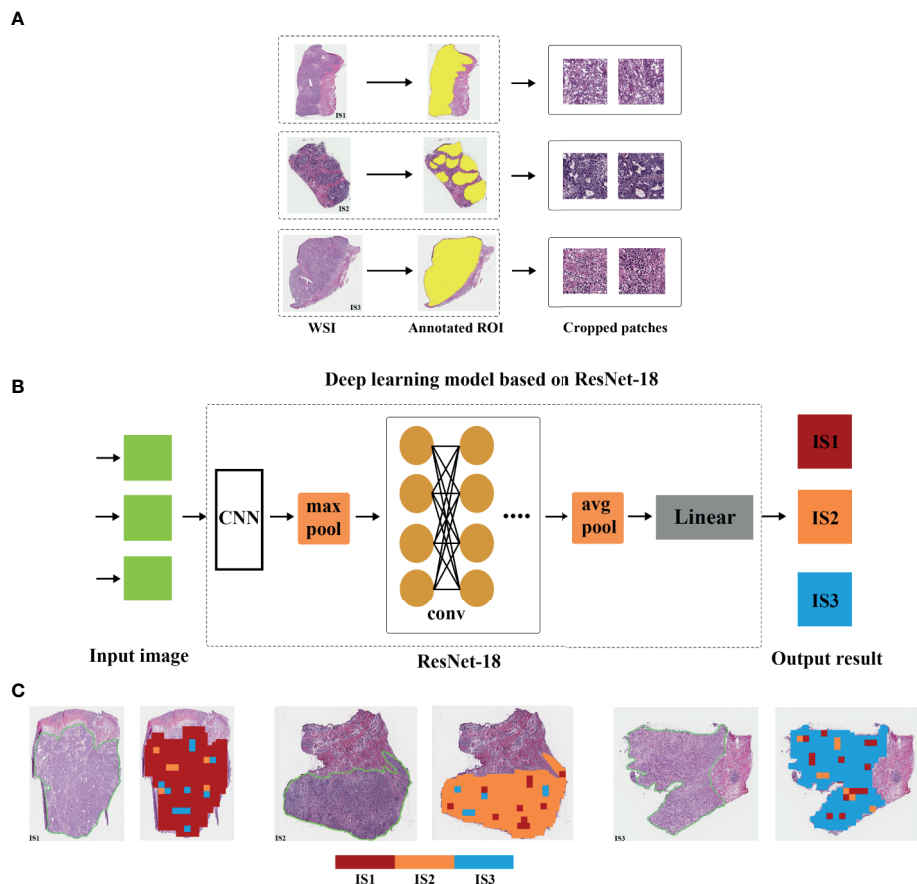
**FIGURE 5** | Lots of cytokine secretion and immune regulation pathways were found in the GO, KEGG and GSEA analysis for IS1 vs IS2 (**A–C**). Lots of classical tumor and immune-related pathways were found in the GO, KEGG and GSEA analysis for IS1 vs IS3 (**D–F**). Lots of classical tumor and immune-related pathways were found in the GO, KEGG, and GSEA analyses for IS2 vs IS3 (**G–I**). These suggest more active immune respond and antitumor reaction in IS3.

data for GC is not well elaborated. Tumor-related immune information plays an important role in the development of tumors. The immune subtypes proposed in this study based on TIICs are independent of existing classifications. Interestingly, we also find that the proportion of different Lauren's classification (intestinal and diffuse) and TCGA classification (EVB, MSI, CIN, GS) is different among three immune subtypes, which suggests an interaction between them. This underlying implication is worth further study. Besides, different from the previous classification,

our immune subtypes have the underlying value to guide the immunotherapy and to predict prognosis.

The relationship between various types of immune cells as immune-suppressive and immune-promoting elements and tumor has been widely explored. The high abundance of tumor-associated lymphocytes, including CD8+ T cell, CD4+ T cell, and NK cell, plays a positive impact on prognosis of gastric cancer by dissolving tumor cells directly (22–25). Also, Tfh cells promote tumor-associated lymphocytes to play an anti-tumor





**FIGURE 6** | Overview of the deep learning model. The whole slide image (WSI) of each patient was obtained and annotated with regions of carcinoma (ROI) **(A)**. Then, tumor of ROI was segmented into patches, and the immune subtypes likelihood of each patch was predicted by deep learning model based on ResNet-18 **(B)**. Finally, multiple patch-level IS likelihoods were integrated into a WSI-level IS prediction **(C)**.

role in gastric cancer by producing diverse antibodies and cytokines (23). By contrast, tumor-associated macrophages (TAMs), Tregs, B cell, and mast cells play the central role in the antitumor immune responses, such as negatively regulating T cell immunity (22–26). Besides, DC, as the key role in antigen presenting cells, had many subtypes. Some could induce the generation of CD8+ effector T cells through presenting the MHC class I molecules to T cells and some may inhibit immune response. In this study, we find that IS3 with high abundance of CD8+ T cells, NK cells, and Tfh cells have a better prognosis, and IS1-2 with high abundance of DC, Tregs, B cell, and mast cell have a poor prognosis. Interestingly, the two subtyping of CD4+ T cells show an opposite trend of aggregation in immune subtypes, which may play different immune functions. Furthermore, different research directions of immunotherapy could be suggested according to the immune subtypes. For IS3, it may be sufficient to mobilize the antitumor function of tumor-associated lymphocytes alone; whereas IS1-2, inhibition of anti-tumor immune response, and promoting the formation of tumor-associated lymphocytes are equally important in immunotherapy.

A series of classical tumor and immune-related pathways are found in the GO and GSEA analyses. For example, in our study, gastric cancer of IS3 is demonstrated with the highest enrichment of T cell receptor signaling and P53 signaling. In comparison, tumors of IS1-2 are confirmed with the highest enrichment of TGF-BETA signaling and JAK-STAT signaling. This reflects the difference in the composition of immune microenvironment and partly explains the difference in prognosis between them. More interestingly, high expression of PD1, PDL1, CTLA4, and TP53, and low expression of JAK1 are found in IS3. Currently, the most well-studied immune checkpoint inhibitors, such as ipilimumab and pembrolizumab, target at CTLA4 and PD1, then releases effector T cells from negative feedback pathway. Therefore, immune-hot IS3 tumor with high expression of CTLA4 and PD1 may respond better to current immunotherapy which should be fully considered in immunotherapy.

T cell infiltration and immune checkpoint (PD-1, PD-L1, and CTLA-4) are known as predictors to immunotherapy (22). Tumor immune microenvironment involves the interaction of multiple immune cells, which contains a more complex relationship and is closely related to immunotherapy. Relationship between T cell

infiltration and immune response is not clear. The location of T cell and other immune cells (e.g. Tregs and DC) also play an important role (25, 50). Meanwhile, not all patients with positive immune checkpoints respond well to the immunotherapy (15, 16). In this study, we find that the IS1 was also infiltrated with abundant T cells, but the expression of immune checkpoint is not high, and the prognosis is poor. This may be related to its strong immunosuppression, such as high abundance of Tregs and low abundance of Tfh cells (31, 51). Meanwhile, the expression level of almost all immune-infiltrating cells, except for macrophages, is low in IS2. And M2 macrophages abundance is the highest in IS2, while M1 macrophages' abundance is the lowest in IS2. This indicates a status of immunologic deficiency and immunosuppression (52). IS3 shows an immune-hot status with high T cell infiltration. These findings suggest that it is more valuable to study the tumor immune cell microenvironment as a whole and suggest the possibility of different immunotherapeutic strategies for different immune subtypes. For IS1, appropriate treatment targeting regulatory cells (e.g., Tregs and DC) is also important (31, 51). For IS2, in addition to enhancing immune activity, it can also be considered to promote M1 polarization of macrophages and inhibit M2 polarization to promote the immune response (52). For IS3, enhancing the function of existing T cells may be enough.

The whole transcriptome sequencing data are difficult to obtain due to its high cost. Besides, flow cytometry to detect all immune cells in the immune microenvironment is difficult and requires complex protocol and high quality of GC tissue. Thus, we hope to get information about the immune subtypes in a more convenient way. Therefore, considering the extensive and easy application of HE pathological sections in clinical practice, a deep learning model based on ResNet-18 is developed and validated for our immune subtypes based on the whole-slide image. We put forward such a conceptual framework that the immune subtype could be predicted based on the whole-slide pathological image. With limited cases, we find that deep learning can predict the immune subtypes well. In the future, if enough cases and a perfect deep learning model are available, the immune subtypes can be easily used in clinical practice.

There are several limitations to this study. First, our analysis is only focused on tumor-infiltrating immune cells, while other components in tumor microenvironment might also play important role. Second, immune infiltration cells were generated from gene expression profiles, which means the location information of immune cells could not be further analyzed. Third, the possibility of selection bias in this retrospective study could not be excluded. Fourth, gastric cancer is a highly heterogenous cancer. Three subtypes to predict the response to immunotherapy may not be enough. In the future, we will focus on the discovery of new immune subtypes for GC. Fifth, the exactly parameters used by deep learning to distinguish subtypes cannot be acquired. Finally, a small sample size should not be ignored.

In conclusion, we confirm three reproducible immune subtypes of gastric cancer. Each of the three immune subtypes possess distinct compositions of tumor immune-infiltrating cells, molecular features, and clinical characteristics. We then

develop and validate a deep learning model based on pathological images to predict the immune subtypes. Our study puts forward a conceptual framework of immune subtypes to understand the immune microenvironment of gastric cancer better, which may provide references for the future design of immune-related studies and immunotherapy selection.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

## ETHICS STATEMENT

This study was deemed exempt from institutional review board approval by Tongren Hospital, Shanghai Jiao Tong University, School of Medicine (Shanghai, China).

## AUTHOR CONTRIBUTIONS

All authors listed had made a substantial contribution to the work. ZQX and SH put forward the conception and designed the study. CYL, RYC, WL and JJD collected and collated the data and do the language editing. YC, ZPS and WLC analyzed data and wrote the manuscript together. XZF, JZ and XYS made contribution to proofread the article. Finally, All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by grants from: Health Commission of Changning District, Shanghai (YXMZK009) and Tongren Hospital, Shanghai Jiao Tong University, School of Medicine (TR2020xk28).

## ACKNOWLEDGMENTS

We acknowledge TCGA and GEO database for providing their platforms and contributors for uploading their meaningful data sets.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2021.685992/full#supplementary-material>

## REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J Clin* (2018) 68 (6):394–424. doi: 10.3322/caac.21492
- Sarela AI, Yelluri S. Leeds Upper Gastrointestinal Cancer Multidisciplinary T. Gastric Adenocarcinoma With Distant Metastasis: Is Gastrectomy Necessary? *Arch Surg* (2007) 142(2):143–9. doi: 10.1001/archsurg.142.2.143
- Cunningham D, Allum WH, Stenning SP, Thompson JN, Van de Velde CJ, Nicolson M, et al. Perioperative Chemotherapy Versus Surgery Alone for Resectable Gastroesophageal Cancer. *N Engl J Med* (2006) 355(1):11–20. doi: 10.1056/NEJMoa055531
- Lee JH, Chang KK, Yoon C, Tang LH, Strong VE, Yoon SS. Lauren Histologic Type Is the Most Important Factor Associated With Pattern of Recurrence Following Resection of Gastric Adenocarcinoma. *Ann Surg* (2018) 267 (1):105–13. doi: 10.1097/SLA.0000000000002040
- Sun Z, Liu H, Yu J, Huang W, Han Z, Lin T, et al. Frequency and Prognosis of Pulmonary Metastases in Newly Diagnosed Gastric Cancer. *Front Oncol* (2019) 9:671. doi: 10.3389/fonc.2019.00671
- Sun Z, Zheng H, Yu J, Huang W, Li T, Chen H, et al. Liver Metastases in Newly Diagnosed Gastric Cancer: A Population-Based Study From SEER. *J Cancer* (2019) 10(13):2991–3005. doi: 10.7150/jca.30821
- Sun Z, Chen H, Han Z, Huang W, Hu Y, Zhao M, et al. Genomics Score Based on Genome-Wide Network Analysis for Prediction of Survival in Gastric Cancer: A Novel Prognostic Signature. *Front Genet* (2020) 11:835. doi: 10.3389/fgene.2020.00835
- Cohen NA, Strong VE, Janjigian YY. Checkpoint Blockade in Esophagogastric Cancer. *J Surg Oncol* (2018) 118(1):77–85. doi: 10.1002/jso.25116
- Fuchs CS, Doi T, Jang RW, Muro K, Satoh T, Machado M, et al. Safety and Efficacy of Pembrolizumab Monotherapy in Patients With Previously Treated Advanced Gastric and Gastroesophageal Junction Cancer: Phase 2 Clinical Keynote-059 Trial. *JAMA Oncol* (2018) 4(5):e180013. doi: 10.1001/jamaoncol.2018.0013
- Kang YK, Boku N, Satoh T, Ryu MH, Chao Y, Kato K, et al. Nivolumab in Patients With Advanced Gastric or Gastro-Oesophageal Junction Cancer Refractory to, or Intolerant of, At Least Two Previous Chemotherapy Regimens (ONO-4538-12, ATTRACTION-2): A Randomised, Double-Blind, Placebo-Controlled, Phase 3 Trial. *Lancet* (2017) 390(10111):2461–71. doi: 10.1016/S0140-6736(17)31827-5
- Muro K, Chung HC, Shankaran V, Geva R, Catenacci D, Gupta S, et al. Pembrolizumab for Patients With PD-L1-Positive Advanced Gastric Cancer (KEYNOTE-012): A Multicentre, Open-Label, Phase 1b Trial. *Lancet Oncol* (2016) 17(6):717–26. doi: 10.1016/S1470-2045(16)00175-3
- Janjigian YY, Bendell J, Calvo E, Kim JW, Ascierto PA, Sharma P, et al. Checkmate-032 Study: Efficacy and Safety of Nivolumab and Nivolumab Plus Ipilimumab in Patients With Metastatic Esophagogastric Cancer. *J Clin Oncol* (2018) 36(28):2836–44. doi: 10.1200/JCO.2017.76.6212
- Pardoll DM. The Blockade of Immune Checkpoints in Cancer Immunotherapy. *Nat Rev Cancer* (2012) 12(4):252–64. doi: 10.1038/nrc3239
- Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, Activity, and Immune Correlates of Anti-PD-1 Antibody in Cancer. *N Engl J Med* (2012) 366(26):2443–54. doi: 10.1056/NEJMoa1200690
- Brahmer JR, Tykodi SS, Chow LQ, Hwu WJ, Topalian SL, Hwu P, et al. Safety and Activity of Anti-PD-L1 Antibody in Patients With Advanced Cancer. *N Engl J Med* (2012) 366(26):2455–65. doi: 10.1056/NEJMoa1200694
- Brahmer J, Reckamp KL, Baas P, Crino L, Eberhardt WE, Poddubskaya E, et al. Nivolumab Versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N Engl J Med* (2015) 373(2):123–35. doi: 10.1056/NEJMoa1504627
- Garon EB, Rizvi NA, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for The Treatment of Non-Small-Cell Lung Cancer. *N Engl J Med* (2015) 372(21):2018–28. doi: 10.1056/NEJMoa1501824
- Wolchok JD, Chiarion-Sileni V, Gonzalez R, Rutkowski P, Grob JJ, Cowey CL, et al. Overall Survival With Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N Engl J Med* (2017) 377(14):1345–56. doi: 10.1056/NEJMoa1709684
- Yang N, Zhu S, Lv X, Qiao Y, Liu YJ, Chen J. MicroRNAs: Pleiotropic Regulators in The Tumor Microenvironment. *Front Immunol* (2018) 9:2491. doi: 10.3389/fimmu.2018.02491
- Xiong Y, Wang K, Zhou H, Peng L, You W, Fu Z. Profiles of Immune Infiltration in Colorectal Cancer and Their Clinical Significance: A Gene Expression-Based Study. *Cancer Med* (2018) 7(9):4496–508. doi: 10.1002/cam4.1745
- Liu Z, Zhu Y, Xu L, Zhang J, Xie H, Fu H, et al. Tumor Stroma-Infiltrating Mast Cells Predict Prognosis and Adjuvant Chemotherapeutic Benefits in Patients With Muscle Invasive Bladder Cancer. *Oncoimmunology* (2018) 7(9):e1474317. doi: 10.1080/2162402X.2018.1474317
- Oya Y, Hayakawa Y, Koike K. Tumor Microenvironment in Gastric Cancers. *Cancer Sci* (2020) 111(8):2696–707. doi: 10.1111/cas.14521
- Wei M, Shen D, Mulmi Shrestha S, Liu J, Zhang J, Yin Y. The Progress of T Cell Immunity Related to Prognosis in Gastric Cancer. *BioMed Res Int* (2018) 2018:3201940. doi: 10.1155/2018/3201940
- Thompson ED, Zahurak M, Murphy A, Cornish T, Cuka N, Abdelfatah E, et al. Patterns of PD-L1 Expression and CD8 T Cell Infiltration in Gastric Adenocarcinomas and Associated Immune Stroma. *Gut* (2017) 66(5):794–801. doi: 10.1136/gutjnl-2015-310839
- Chen LJ, Zheng X, Shen YP, Zhu YB, Li Q, Chen J, et al. Higher Numbers of T-bet(+) Intratumoral Lymphoid Cells Correlate With Better Survival in Gastric Cancer. *Cancer Immunol Immunother* (2013) 62(3):553–61. doi: 10.1007/s00262-012-1358-6
- Sarvaria A, Madrigal JA, Saudemont A. B Cell Regulation in Cancer and Anti-Tumor Immunity. *Cell Mol Immunol* (2017) 14(8):662–74. doi: 10.1038/cmi.2017.35
- Lee K, Hwang H, Nam KT. Immune Response and the Tumor Microenvironment: How They Communicate to Regulate Gastric Cancer. *Gut Liver* (2014) 8(2):131–9. doi: 10.5009/gnl.2014.8.2.131
- Du Y, Wei Y. Therapeutic Potential of Natural Killer Cells in Gastric Cancer. *Front Immunol* (2018) 9:3095. doi: 10.3389/fimmu.2018.03095
- Sharonov GV, Serebrovskaya EO, Yuzhakova DV, Britanova OV, Chudakov DM. B Cells, Plasma Cells and Antibody Repertoires in the Tumour Microenvironment. *Nat Rev Immunol* (2020) 20(5):294–307. doi: 10.1038/s41577-019-0257-x
- Couillaud C, Germain C, Dubois B, Kaplon H. Identification of Tertiary Lymphoid Structure-Associated Follicular Helper T Cells in Human Tumors and Tissues. *Methods Mol Biol* (2018) 1845:205–22. doi: 10.1007/978-1-4939-8709-2\_12
- Joshi NS, Akama-Garren EH, Lu Y, Lee DY, Chang GP, Li A, et al. Regulatory T Cells in Tumor-Associated Tertiary Lymphoid Structures Suppress Anti-Tumor T Cell Responses. *Immunity* (2015) 43(3):579–90. doi: 10.1016/j.immuni.2015.08.006
- Zhou R, Zhang J, Zeng D, Sun H, Rong X, Shi M, et al. Immune Cell Infiltration as a Biomarker for the Diagnosis and Prognosis of Stage I-III Colon Cancer. *Cancer Immunol Immunother* (2019) 68(3):433–42. doi: 10.1007/s00262-018-2289-7
- Li B, Cui Y, Nambiar DK, Sunwoo JB, Li R. The Immune Subtypes and Landscape of Squamous Cell Carcinoma. *Clin Cancer Res* (2019) 25(12):3528–37. doi: 10.1158/1078-0432.CCR-18-4085
- Ming W, Xie H, Hu Z, Chen Y, Zhu Y, Bai Y, et al. Two Distinct Subtypes Revealed in Blood Transcriptome of Breast Cancer Patients With an Unsupervised Analysis. *Front Oncol* (2019) 9:985. doi: 10.3389/fonc.2019.00985
- Wang M, Li Z, Peng Y, Fang J, Fang T, Wu J, et al. Identification of Immune Cells and mRNA Associated With Prognosis of Gastric Cancer. *BMC Cancer* (2020) 20(1):206. doi: 10.1186/s12885-020-6702-1
- Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial Intelligence in Digital Pathology - New Tools for Diagnosis and Precision Oncology. *Nat Rev Clin Oncol* (2019) 16(11):703–15. doi: 10.1038/s41571-019-0252-y
- Niazi MKK, Parwani AV, Gurcan MN. Digital Pathology and Artificial Intelligence. *Lancet Oncol* (2019) 20(5):e253–e61. doi: 10.1016/S1470-2045(19)30154-8
- Bhargava R, Madabhushi A. Emerging Themes in Image Informatics and Molecular Analysis for Digital Pathology. *Annu Rev BioMed Eng* (2016) 18:387–412. doi: 10.1146/annurev-bioeng-112415-114722

39. Wang H, Jiang Y, Li B, Cui Y, Li D, Li R. Single-Cell Spatial Analysis of Tumor and Immune Microenvironment on Whole-Slide Image Reveals Hepatocellular Carcinoma Subtypes. *Cancers (Basel)* (2020) 12(12):3562–77. doi: 10.3390/cancers12123562
40. Cao R, Yang F, Ma SC, Liu L, Zhao Y, Li Y, et al. Development and Interpretation of a Pathomics-Based Model for the Prediction of Microsatellite Instability in Colorectal Cancer. *Theranostics* (2020) 10(24):11080–91. doi: 10.7150/thno.49864
41. Kather JN, Pearson AT, Halama N, Jger D, Luedde T. Deep Learning Can Predict Microsatellite Instability Directly From Histology in Gastrointestinal Cancer. *Nat Med* (2019) 25(7):1054–63. doi: 10.1038/s41591-019-0462-y
42. Jiang Y, Jin C, Yu H, Wu J, Chen C, Yuan Q, et al. Development and Validation of a Deep Learning CT Signature to Predict Survival and Chemotherapy Benefit in Gastric Cancer: A Multicenter, Retrospective Study. *Ann Surg* (2020). doi: 10.1097/SLA.0000000000003778
43. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust Enumeration of Cell Subsets From Tissue Expression Profiles. *Nat Methods* (2015) 12(5):453–7. doi: 10.1038/nmeth.3337
44. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A Class Discovery Tool With Confidence Assessments and Item Tracking. *Bioinformatics* (2010) 26(12):1572–3. doi: 10.1093/bioinformatics/btq170
45. Kapp AV. Are Clusters Found in One Dataset Present in Another Dataset? *Biostatistics* (2007) 8(1):9–31. doi: 10.1093/biostatistics/kxj029
46. Cancer Genome Atlas Research N. Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* (2014) 513(7517):202–9. doi: 10.1038/nature13480
47. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612
48. Derks S, de Klerk LK, Xu X, Fleitas T, Liu KX, Liu Y, et al. Characterizing Diversity in the Tumor-Immune Microenvironment of Distinct Subclasses of Gastroesophageal Adenocarcinomas. *Ann Oncol* (2020) 31(8):1011–20. doi: 10.1016/j.annonc.2020.04.011
49. Wang Q, Liu G, Hu C. Molecular Classification of Gastric Adenocarcinoma. *Gastroenterol Res* (2019) 12(6):275–82. doi: 10.14740/gr1187
50. Farhood B, Najafi M, Mortezaee K. CD8(+) Cytotoxic T Lymphocytes in Cancer Immunotherapy: A Review. *J Cell Physiol* (2019) 234(6):8509–21. doi: 10.1002/jcp.27782
51. Kondelkova K, Vokurkova D, Krejsek J, Borska L, Fiala Z, Ctirad A. Regulatory T Cells (TREG) and Their Roles in Immune System With Respect to Immunopathological Disorders. *Acta Med (Hradec Kralove)* (2010) 53(2):73–7. doi: 10.14712/18059694.2016.63
52. Najafi M, Hashemi Goradel N, Farhood B, Salehi E, Nashtaei MS, Khanlarkhani N, et al. Macrophage Polarity in Cancer: A Review. *J Cell Biochem* (2019) 120(3):2756–65. doi: 10.1002/jcb.27646

**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.

Copyright © 2021 Chen, Sun, Chen, Liu, Chai, Ding, Liu, Feng, Zhou, Shen, Huang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## OPEN ACCESS

### Edited by:

George Bebis,  
University of Nevada, Reno,  
United States

### Reviewed by:

Long Liu,  
First Affiliated Hospital of Zhengzhou  
University, China  
Xiaocan Jia,  
Zhengzhou University, China  
Gregory Riedlinger,  
Rutgers Cancer Institute of  
New Jersey, United States

### \*Correspondence:

Yong Luo  
luoyonganzhen@163.com  
Yongguang Jiang  
yongguangjiangazzy@126.com

<sup>†</sup>These authors share first authorship

### Specialty section:

This article was submitted to  
Cancer Immunity  
and Immunotherapy,  
a section of the journal  
Frontiers in Immunology

**Received:** 21 August 2021

**Accepted:** 27 September 2021

**Published:** 12 October 2021

### Citation:

Feng T, Zhao J, Wei D, Guo P, Yang X,  
Li Q, Fang Z, Wei Z, Li M, Jiang Y and  
Luo Y (2021) Immunogenomic  
Analyses of the Prognostic Predictive  
Model for Patients With Renal Cancer.  
Front. Immunol. 12:762120.  
doi: 10.3389/fimmu.2021.762120

# Immunogenomic Analyses of the Prognostic Predictive Model for Patients With Renal Cancer

Tao Feng<sup>1†</sup>, Jiahui Zhao<sup>1†</sup>, Dechao Wei<sup>1</sup>, Pengju Guo<sup>1</sup>, Xiaobing Yang<sup>1</sup>, Qiankun Li<sup>2</sup>,  
Zhou Fang<sup>3</sup>, Ziheng Wei<sup>3</sup>, Mingchuan Li<sup>1</sup>, Yongguang Jiang<sup>1\*</sup> and Yong Luo<sup>1\*</sup>

<sup>1</sup> Department of Urology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China, <sup>2</sup> Department of Urology, Beijing Huairou Hospital, Beijing, China, <sup>3</sup> Department of Cardiovascular Surgery, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

**Background:** Renal cell carcinoma (RCC) is associated with poor prognostic outcomes. The current stratifying system does not predict prognostic outcomes and therapeutic benefits precisely for RCC patients. Here, we aim to construct an immune prognostic predictive model to assist clinician to predict RCC prognosis.

**Methods:** Herein, an immune prognostic signature was developed, and its predictive ability was confirmed in the kidney renal clear cell carcinoma (KIRC) cohorts based on The Cancer Genome Atlas (TCGA) dataset. Several immunogenomic analyses were conducted to investigate the correlations between immune risk scores and immune cell infiltrations, immune checkpoints, cancer genotypes, tumor mutational burden, and responses to chemotherapy and immunotherapy.

**Results:** The immune prognostic signature contained 14 immune-associated genes and was found to be an independent prognostic factor for KIRC. Furthermore, the immune risk score was established as a novel marker for predicting the overall survival outcomes for RCC. The risk score was correlated with some significant immunophenotypic factors, including T cell infiltration, antitumor immunity, antitumor response, oncogenic pathways, and immunotherapeutic and chemotherapeutic response.

**Conclusions:** The immune prognostic, predictive model can be effectively and efficiently used in the prediction of survival outcomes and immunotherapeutic responses of RCC patients.

**Keywords:** renal cell carcinoma, tumor immune microenvironment, prognostic model, risk score, immunotherapy



## BACKGROUND

The prevalence of renal cell carcinoma (RCC), a lethal urogenital cancer, ranks third after prostate and bladder cancers (1–3). In 2020, about 73,750 new RCC cases were diagnosed, with approximately 14,830 deaths in the USA (3). Nowadays, a range of treatments, such as surgery accompanied with or without postoperative adjuvant therapy, chemotherapy, immunotherapy, and target therapy, have been developed for RCC. Although these options have certain therapeutic effects, the overall prognosis of RCC patients remains dismal, especially in the late-stage RCC (4).

Over recent decades, the development of immunotherapy has revolutionized cancer treatment paradigms and has been recognized as a promising therapeutic frontier (5–7). For example, immune checkpoint blockade (ICB) is a new therapeutic strategy for several cancer types, such as breast cancer (8, 9), melanoma (10, 11), and lung cancer (12, 13). ICB has also evolved in RCC and showed certain practical application value through the years based on the phase III CheckMate 025 study, whether or not patients have been previously treated (14, 15). In addition, accumulating evidence has also proven that the tumor immune microenvironment (TIME), which encompasses immune cells, fibroblasts, extracellular matrix, endothelial cells, and various cytokines, is associated with tumor progression and metastasis (16–19). In 2017, Chevrier et al. depicted an in-depth Immune Atlas of Clear Cell Renal Cell Carcinoma by applying mass cytometry for the high-dimensional single-cell analysis of kidney primary Tumors (20). In addition, an increased number of studies have proved that multiple immune cells, including CD8<sup>+</sup> T cells, CD4<sup>+</sup> T cells and NK cells et al, have been associated with ccRCC tumor (21, 22). An in-depth understanding of TIME is critical to identifying potential immunotherapeutic targets for RCC. However, the majority of the studies have only evaluated gene expressions in the prediction of survival rates for RCC patients, and most of these biomarkers only reveal the status of TIME in some aspects (23, 24). Hence, a comprehensive immune-based model might provide an in-depth insight into the association between prognosis and TIME in RCC.

**Abbreviations:** RCC, renal cell carcinoma; KIRC, kidney renal clear cell carcinoma; DEGs, differentially expressed genes; ccRCC, clear cell renal cell carcinoma; CMAP, connectivity map; ssGSEA, single sample gene set enrichment analysis; GSEA, gene set enrichment analysis; NES, normalized enrichment score; IS, immune cells; TS, tumor cells; ICB, immune checkpoint blockade; TIME, tumor immune microenvironment; TME, tumor microenvironment; TMB, tumor mutation burden; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; TCIA, the cancer immune group atlas; IRGs, immune-related genes; ImmPort, Immunology Database and Analysis Portal; DE-IRGs, differentially expressed immune-related genes; FDR, false discovery rate; LASSO, least absolute shrinkage and selection operator; ROC, receiver operating characteristic curves; C-index, concordance index; DCA, decision curve analysis; ES, ESTIMATE scores, TP, tumor purity, SS, stromal scores; IS, immune scores; CYT, cytolytic activity; PD, progressive disease; SD, stable disease; PR, partial response; CR, complete response; IPS, immunophenoscore; GDSC, Genomics of Drug Sensitivity in Cancer; PCA, principal component analysis; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; CC, cellular component; MF, molecular function; PAC, antigen-presenting cell; OS, overall survival.

In the current study, we established an immune prognostic signature model for RCC using the training cohort and further confirmed the effectiveness of the prognosis model in the testing and the entire cohort. Additionally, the associations between the risk score subtypes and immune checkpoints, antitumor immunity, antitumor response, oncogenic pathways, immune cell infiltration, and tumor mutation burden (TMB) were explored. Also, the models' ability for the prediction of chemotherapeutic and immunotherapeutic responses was evaluated. Finally, we screened out two compounds that could improve the prognosis of RCC.

## MATERIALS AND METHODS

### Data Acquisition as Well as Preprocessing

Transcriptional expression profiles, mutation patterns, and related clinical data for KIRC patients were retrieved from the Cancer Genome Atlas (TCGA) cohort (<https://cancergenome.nih.gov/>). Immune-associated genes (IRGs) were derived from the Immunology Database as well as Analysis Portal (ImmPort) database (25). The immunophenoscore (IPS) for RCC patients were retrieved from the cancer immune group atlas (TCIA) (<https://tcia.at/home>). In addition, the advanced urothelial cancer database of administered anti-PD-L1 therapy was downloaded using the R package “IMvigor210CoreBiologies” (version 1.0.0) (26). The malignant melanoma dataset that received anti-PD-1 and antiCTLA4 therapy were obtained from the GSE91061 cohort. All data were subjected to background correction and logarithmic conversion using R software.

### Differentially Expressed Immune-Related Genes (DE-IRGs) and Functional Enrichment Analyses

Differential gene expression analysis between tumor and corresponding normal tissues in KIRC were screened based on the count data for TCGA kidney cancer cohort using the R package “DESeq2” (27), according to the screening criteria ( $\log_2$  fold change  $> 2$ , P-value  $< 0.05$ ). The IRGs involved in oncogenesis were provide by IMMPORT website. Then, DE-IRGs were identified by the intersection between DEGs and IRGs.

The R package “clusterProfiler” was used for Gene Ontology (GO) as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of these significant DE-IRGs and their visualization (28). Next, we defined the pathways and terms using false discovery rate (FDR)  $\leq 0.05$  as statistically significant.

### Establishment of the Immune-Related Risk Score

Among 538 KIRC with mRNA expression data, 517 patients with the overall survival (OS) data were retained for further analyses. First, 70% of samples were randomly drawn and grouped as training cohort to develop a prognostic risk model, and the other

30% of samples comprised the validation set, which was used in evaluating the model's predictive ability and robustness in the entire cohort. Then, DE-IRGs were screened out by univariate Cox proportional hazard regression through the “coxph” R-function from the “survival” package (29). Subsequently, the least absolute shrinkage and selection operator (LASSO) Cox regression analysis was carried out to select the prognostic genes using the R package “glmnet” (30). Finally, the immune-associated risk score was calculated using LASSO Cox regression hazard regression – retrieved regression coefficients to multiply expression levels of genes (the risk score = mRNA expression levels of gene a  $\times$  coefficient a + mRNA expression levels of gene b  $\times$  coefficient b + ..... + mRNA expression levels of gene n  $\times$  coefficient n).

In addition, by setting the median of risk score as the cutoff value, the patients were classified into a high-risk group and a low-risk group. To establish the prognostic accuracy of the established model, we used Kaplan–Meier survival curve analysis, concordance (C)-index, log-rank test in addition to time-dependent receiver operating characteristic curves (ROC) and XGBoost algorithm.

### Independent Prognostic Value of the Immune-Associated Prognostic Signature

Multivariate Cox regression analysis with the forward stepwise procedure was performed to investigate if the risk score is an independent prognostic factor. The immune-associated risk score and other clinical variables with  $P < 0.05$  were identified as independent prognostic risk factors.

### Establishment and Validation of the Nomogram

To develop a prognostic signature for 1-, 3-, and 5-year survival rates, a nomogram was constructed using the identified independent prognostic variables, such as stage, age, and risk score (31). Moreover, the C-index, calibration curve, decision curve analysis (DCA), and ROC analysis were performed to determine its predictive accuracy and discriminatory capacity (32). The C-index was evaluated using a bootstrap method involving 1000 resamples (33). The C-index values, dependent on the nomogram's predictive ability, ranged from 0.5 (no discrimination) to 1 (perfect discrimination). The consistency between the predictive survival rate and the actual survival rate in unknown samples was assessed using calibration curves. Additionally, DCA (34) was used to evaluate the clinical utility and the net benefits of the nomogram as it takes both discrimination and calibration into consideration. Finally, the area under the receiver operating characteristic (ROC) curve (AUC) was also determined for each variable to evaluate the discriminative performance of the nomograms.

### Immune Cell Proportion Analyses and Immune Related Features

To explore immune cell abundance in KIRC tissues, CIBERSORT (35) was employed to evaluate the proportions of 22 immune cell types using a deconvolution algorithm by the R

package with default parameters. In addition, the ESTIMATE scores (ES), tumor purity (TP), stromal scores (SS), and immune scores (IS) for each KIRC sample were evaluated using the ESTIMATE algorithm (19) of the “estimate” package. The cytolytic activity (CYT) index is a geometric mean of mRNA expression levels of GZMA and PRF1, and was utilized to assess the intratumoral immune cytolytic T-cell activities (36).

### Immunotherapy and Chemotherapeutic Response in Risk Score Subtype

As immune checkpoint molecules are widely explored in the immunotherapeutic studies of multiple cancers, programmed cell death 1 (PDCD1, also referred to as PD-1), CD274 molecule (also referred to as PD-L1), and cytotoxic T-lymphocyte protein 4 (CTLA4) were used to evaluate the associations between risk scores and immunotherapeutic efficacies. The urothelial cancer dataset (IMvigor210) comprising of administered anti-PD-L1 therapy was used to establish the therapeutic benefits between high- and low-risk score subtypes using four treatment categories: progressive disease (PD), stable disease (SD), complete response (CR), and partial response (PR).

IPS is a machine learning-based scoring system applied for the prediction of patients' responses to immune checkpoint inhibitor (ICI) treatment based on the weight average Z scores representing immune-related genes expression in cell types (37). High IPS scores reflect increased immunogenicity.

As chemotherapy and targeted therapy are widely used to treat clear cell renal cell carcinoma (ccRCC), risk scores were used to predict the drug sensitivity based on half-maximal inhibitory concentrations (IC50) for each KIRC patient from the Genomics of Drug Sensitivity in Cancer (GDSC) website (38) using the R package “pRRophetic” (39–44). The common target drugs, such as Cisplatin, Gefitinib, Gemcitabine, Sorafenib, Sunitinib, Vinblastine, Vinorelbine, and Vorinostat, were selected for ccRCC.

### Tumor Mutational Burden (TMB), Connectivity Map (CMAP) and Molecular Docking Analysis

KIRC patients' somatic variants data were analyzed and visualized by “maftool” R package (45) to identify the mutation burden of KIRC in the high- and low-risk scores. Then, the TMB of each patient was calculated as follows: mutations/million bases.

Next, to identify the potentially small molecules related to this signature, genes in the model were assessed *via* CMAP analysis. Thus, the positive mean represented that these selected drugs may share similar functions with the model, while the negative mean indicated that these drugs could improve the prognosis of RCC. Herein, we screened compounds by setting the criteria as  $P < 0.05$ .

Moreover, the crystal structure of the protein was obtained from RCSB Protein Data Bank. The three-dimensional structures for all compounds were downloaded from PubChem database using MOL2 format. The molecular docking calculations were conducted using Schrodinger and Pymol 2.1 software.

## Statistical Analysis

The differences between variables were determined by chi-square as well as Student's t-tests. For baseline clinical data, the Wilcoxon test and the Kruskal-Wallis were utilized to evaluate the significant differences between two or multiple groups, respectively. The Kaplan-Meier survival curves were compared using the log-rank test.  $P < 0.05$  indicated statistical significance. R 4.0.3 and SPSS 26.0 software were used for all analyses.

## RESULTS

### Identification and Functional Analyses of DE-IRGs

All 517 KIRC samples with OS information were split into training (367 patients) and test groups (151 patients). Between the training and validation cohorts, no significant differences were detected among most of the clinical characteristics (Table 1).

With the cutoff value  $|\log_2 \text{fold change (logFC)}| > 2$  and adjusted  $P < 0.05$ , 953 DEGs were filtered, of which 539 genes were significantly elevated, while 414 genes were significantly suppressed in tumor samples compared to normal samples (Figures 1A, B). Moreover, principal component analysis (PCA) results (Figure 1C) revealed that KIRC samples clustered separately from normal samples. Subsequently, the intersection between DEGs and immune-associated genes retrieved from the ImmPort database was determined, and 98 DE-IRGs were selected and visualized on a Venn diagram (Figure 1D).

These 98 DE-IRGs were further utilized in functional enrichment analyses, including KEGG and GO analyses. Based

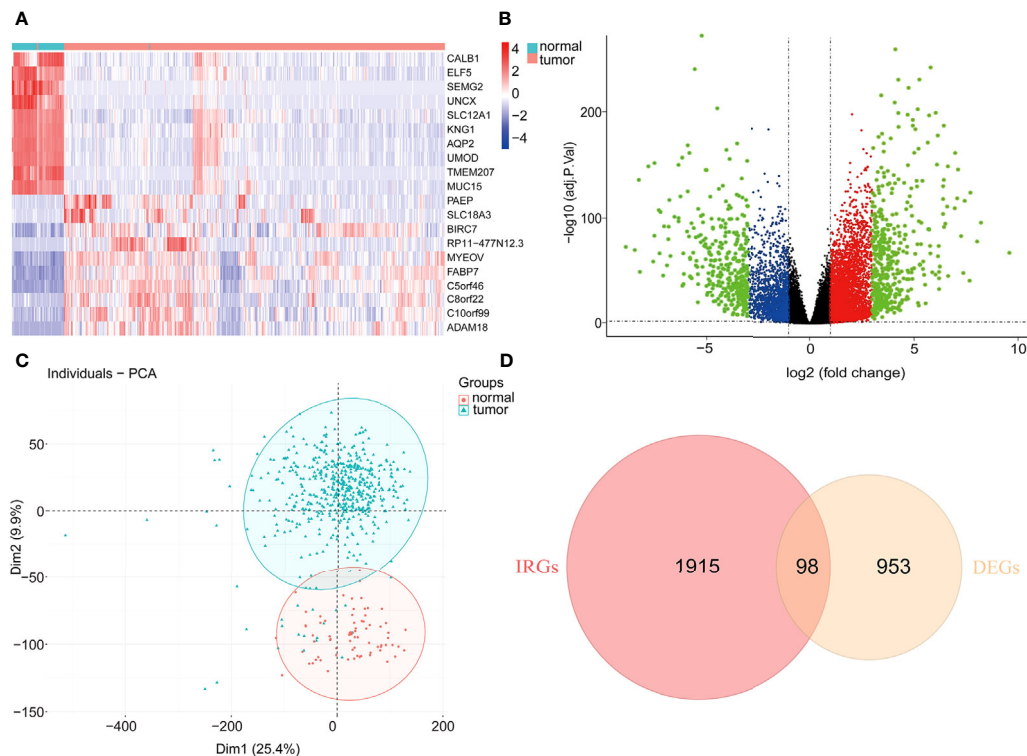
on GO analysis, in the biological process (BP), these DE-IRGs were enriched in cell chemotaxis, leukocyte chemotaxis, lymphocyte chemotaxis, positive regulation of cell adhesion, and T cell activation (Figure 2A). In the cellular component (CC) category, the DE-IRGs were mainly enriched in the cytoplasmic vesicle lumen, plasma membrane's external side, platelet alpha granule, platelet alpha granule lumen, and vesicle lumen (Figure 2B). Regarding molecular function (MF), these DE-IRGs were enriched in cytokine receptor binding, growth factor activity, receptor ligand activity, cytokine activity, and signaling receptor activator activity (Figure 2C). Regarding KEGG pathways analysis, these DE-IRGs were mainly involved in the calcium signaling pathway, chemokine signaling pathways, cytokine-cytokine receptor interactions, Ras signaling pathways, and viral protein interactions with cytokine receptors and cytokines (Figure 2D).

### Establishment and Validation of Prognostic Immune Score Model

All 367 KIRC samples in the training cohort were utilized in a prognostic model establishment. First, univariate Cox regression analysis was carried out to explore the association between DE-IRGs and the OS outcomes for KIRC samples. Among 98 DE-IRGs, 47 genes were selected. To avoid overfitting, we further conducted the LASSO Cox regression analysis with minimized lambda (Figures 3A, B). A total of 14/47 genes were used to establish the prognostic immune score model using the following formula: risk score =  $(\text{SAA1} \times 0.08215) + (\text{IL20RB} \times 0.07643) + (\text{TNFSF14} \times 0.09743) + (\text{ESRRG} \times -0.09743) + (\text{FGF21} \times 0.23324) + (\text{IFNG} \times 0.05956) + (\text{CTLA4} \times 0.01439) + (\text{KLRK1} \times 0.00717) + (\text{IL11} \times 0.01639) + (\text{GDF6} \times -$

**TABLE 1** | The clinical characteristics of KIRC patients.

Variables	Group	Total set (n = 517)	Training set (n = 367)	Testing set (n = 151)	P value
Vital status	Alive	361	249	111	0.136
	Dead	156	118	38	
Survival time		1054.797	1046.125	1070.073	0.815
Clinical Stage	I	257	189	68	0.233
	II	54	36	19	
	III	123	80	43	
	IV	83	62	21	
T stage	T1	262	192	70	0.637
	T2	65	46	20	
	T3	179	122	57	
	T4	11	7	4	
N stage	N0	233	166	67	0.348
	N1	15	8	7	
	NX	269	193	77	
M stage	M0	413	286	128	0.144
	M1	78	59	19	
	MX	26	22	4	
Grade	G1	13	11	2	0.478
	G2	225	161	65	
	G3	204	146	58	
	G4	75	49	26	
Gender	Male	339	236	104	0.32
	Female	178	131	47	
Age	<65	326	233	94	0.791
	≥65	191	134	57	



**FIGURE 1** | Differentially expressed immune-associated genes. **(A)** Heatmap of top 10 up- and down-regulated genes between normal and tumor tissues. **(B)** Volcano plot for DEGs between normal and tumor tissues. **(C)** PCA plot of the data. **(D)** Venn diagram for intersections between DEGs and IRGs.

$0.02484) + (BMP7 \times 0.05433) + (GNLY \times 0.11780) + (AVPR1B \times -0.06460) + (CXCL11 \times 0.03907)$  (**Figure 3C**).

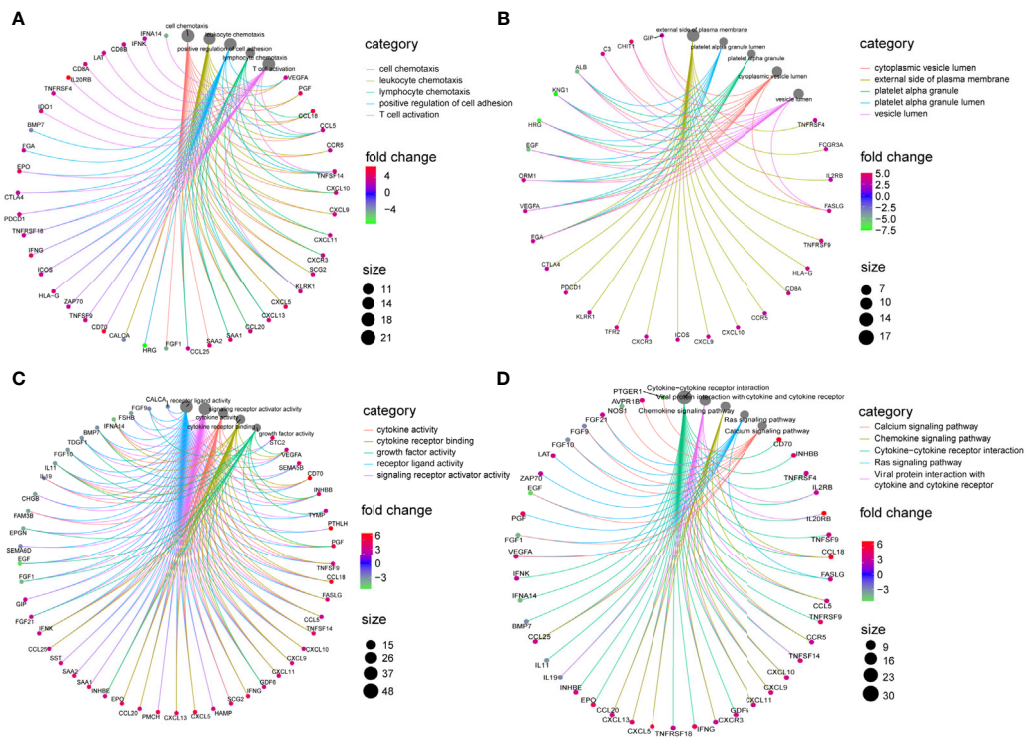
Based on the above calculated formula, the risk scores for every patient in the training set were computed. Then, with the median risk score as the basis, patients were allocated into the high- and low-risk groups. The high-risk patients exhibited significantly poor OS outcomes compared to low-risk patients ( $P=1.398E-10$ ) (**Figure 4A**). As illustrated in **Figure 4B**, the AUCs of risk scores for 1-, 3-, and 5-years were 0.725, 0.723, and 0.745, respectively, in the training group. The distribution of the risk scores, survival time, survival status, and the expression of 14 OS-associated DE-IRGs for KIRC patients in the training cohort are displayed in **Figures 4C–E**. The model's C-index was 0.698 (95% confidence interval (CI): 0.647–0.750,  $P=6.849E-14$ ). To further explore whether the prognostic model was independent of other clinical elements, such as grade, age, T stage, and clinical stage, univariate and multivariate Cox regression analyses were conducted (**Table 2**). The risk score was confirmed as an independent prognostic factor (HR=2.699, 95% CI: 1.716–4.243,  $P<0.0001$ ).

In addition, a quantitative strategy for the prediction of the prognostic outcomes of patients was established by constructing a nomogram that integrated the risk scores as well as other independent clinical prognostic factors for OS (**Figure 4F**). Then, the nomogram's performance was determined using the ROC curve, C-index, calibration curve, and decision curve

analyses. The AUCs of the nomogram were 0.828, 0.783, and 0.774 for 1-, 3-, and 5-year survival times, respectively (**Figure 4G**). The C-index was 0.762 (95% CI: 0.720–0.804,  $P=1.800E-34$ ). Based on the calibration curve, the training cohort predicted that 1-, 3-, and 5-year survival probabilities were good (**Figure 4H**). For the decision curve, the nomogram exhibited a higher net benefit than other schemes to predict the OS (**Figure 4I**).

To delineate the robustness and versatility of the immune score model, the risk score in the training cohort was validated in the testing and entire cohorts. The participants in the testing and entire cohorts were grouped into high- and low-risk score subtypes using the same formula. The findings in the testing and entire datasets were similar. The Kaplan–Meier survival curves revealed poor survival rates for the high-risk group in the testing ( $P=3.58E-8$ ) (**Figure 5A**) and the entire cohorts ( $P=3.616E-12$ ) (**Figure 6A**). The AUC for 1-, 3-, and 5-years are 0.858, 0.842, and 0.857 in the testing group (**Figure 5B**) and 0.736, 0.727, and 0.746 in the entire group (**Figure 6B**). The survival data, risk score, scatterplots, and gene expression pattern distributions in the testing and entire cohorts are shown in **Figure 5C–E** and **Figures 6C–E**. The C-indices of the model were 0.835 (95% CI: 0.782–0.888,  $P=1.034E-35$ ) and 0.709 (95% CI: 0.666–0.752,  $P=7.520E-22$ ) in the testing and entire cohorts, respectively. Univariate and multivariate Cox regression analyses for clinicopathological parameters were carried out in the testing





**FIGURE 2 |** Enrichment analysis of DE-IRGs. **(A)** Visualization of top 5 enriched GO analysis in BP. **(B)** Visualization of top 5 enriched GO analysis in CC. **(C)** Visualization of top 5 enriched GO analysis in MF. **(D)** Visualization of top 5 enriched KEGG pathways.

and entire cohorts. Also, the risk score was an independent prognostic indicator of OS in KIRC patients (Table 2). To improve the prognostic immune score model, the nomogram system was established based on testing and entire cohorts (Figure 5F and Figure 6F). The AUC of our nomogram for predicting 1-, 3-, and 5-year OS was 0.9, 0.875, and 0.891, respectively, in the testing cohort and 0.858, 0.808, and 0.787, respectively, in the entire cohort (Figure 5G and Figure 6G). The C-indices of the nomogram in the testing and entire cohorts were 0.859 (95% CI: 0.809–0.909,  $P=4.980E-45$ ) and 0.786 (95% CI: 0.752–0.821,  $P=4.201E-59$ ), respectively. Finally, the calibration curves and decision curves for 1-, 3-, and 5-year survival probabilities were established (Figures 5H, I and Figures 6H, I). These findings indicated that the nomogram has excellent predictive performance in all cohorts.

### Associations Between DE-IRGs Signature and Clinical Characteristics of KIRC Patients

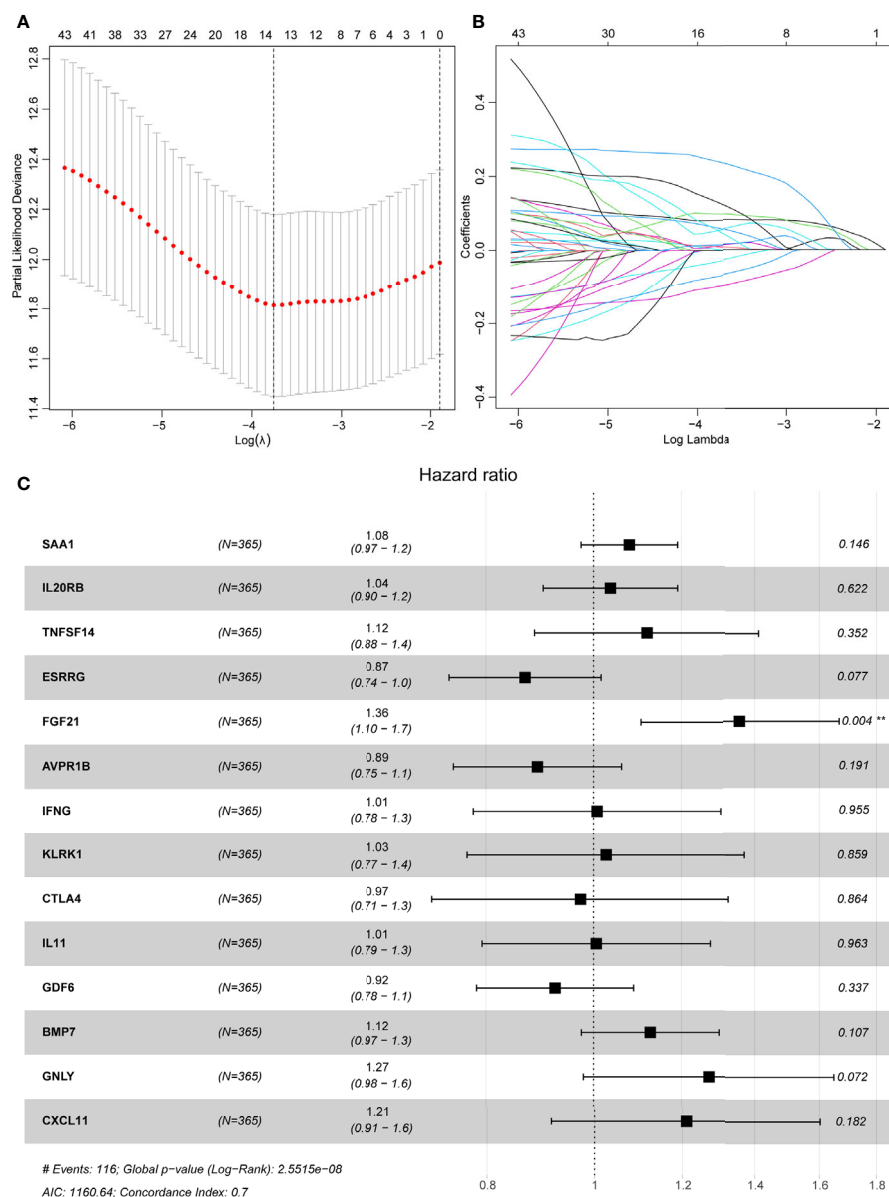
Next, we further investigate the association between clinical characteristics, including tumor burden, age at diagnosis, gender, grade, clinical stage, T stage, and the prognostic risk signature. A significant correlation was established between high-risk score and a high tumor burden ( $P=9.87E-08$ ), male gender ( $P=0.03$ ), advanced grade ( $P=2.54E-09$ ), higher stage ( $P=8.85E-11$ ), and T stage ( $P=5.6E-08$ ) (Figure S1). Additionally, no statistical

significance was observed between < 65-year-old group and >65-year-old group ( $P=0.1$ ). Subsequently, we also assessed whether the model could assess the survival probability in subgroups exhibiting varying clinical patterns. The prognostic model could be utilized for the prediction of survival probabilities for various clinicopathological parameters ( $P<0.05$ ) (Figure S2).

### Immune Cell Proportions Between High- and Low-Risk Score Patients

Using the CIBERSORT algorithm, 22 immune cell types were determined in each KIRC sample between high- and low-risk score subtypes. The proportions of 22 immune cells and their distribution in tumor samples are illustrated in Figure 7A and Figures 7B, C, respectively. Compared to the low-risk group, the high-risk score group exhibited significantly elevated proportions of plasma cells, T cells CD8<sup>+</sup>, T cells follicular helper, T regulatory cells (Tregs), and M0 macrophages ( $P<0.05$ ) (Figures 7C, D). Conversely, the proportions of macrophages M1, activated natural killer (NK) cells, naïve B cells, macrophages M2, resting NK cells, monocytes, T cells CD4<sup>+</sup> memory resting, and resting mast cells in the high-risk score subtype were remarkably elevated compared to those in the low-risk score subtypes ( $P<0.05$ ) (Figures 7C, D). In addition, in 22 immune cell types, high plasma cells, Tregs, follicular helper T cells, and monocytes M0 level were remarkably correlated with





**FIGURE 3 |** LASSO regression analyses and a forest plot describing Cox regression model findings of 14 immune-associated genes. **(A)** Partial likelihood deviance with changing of log ( $\lambda$ ) plotted by LASSO regression in 10-fold cross-validations. Vertical dotted lines were described at the optimal values using minimum criteria and the 1-SE criteria. **(B)** The LASSO coefficient profiles for 14 DE-IRGs in the 10-fold cross-validation. **(C)** Forest plot representing correlations between the expression levels of 14 DE-IRGs and overall survival outcomes in the training dataset. HR, 95% CI, and P-values were evaluated by LASSO regression analyses.

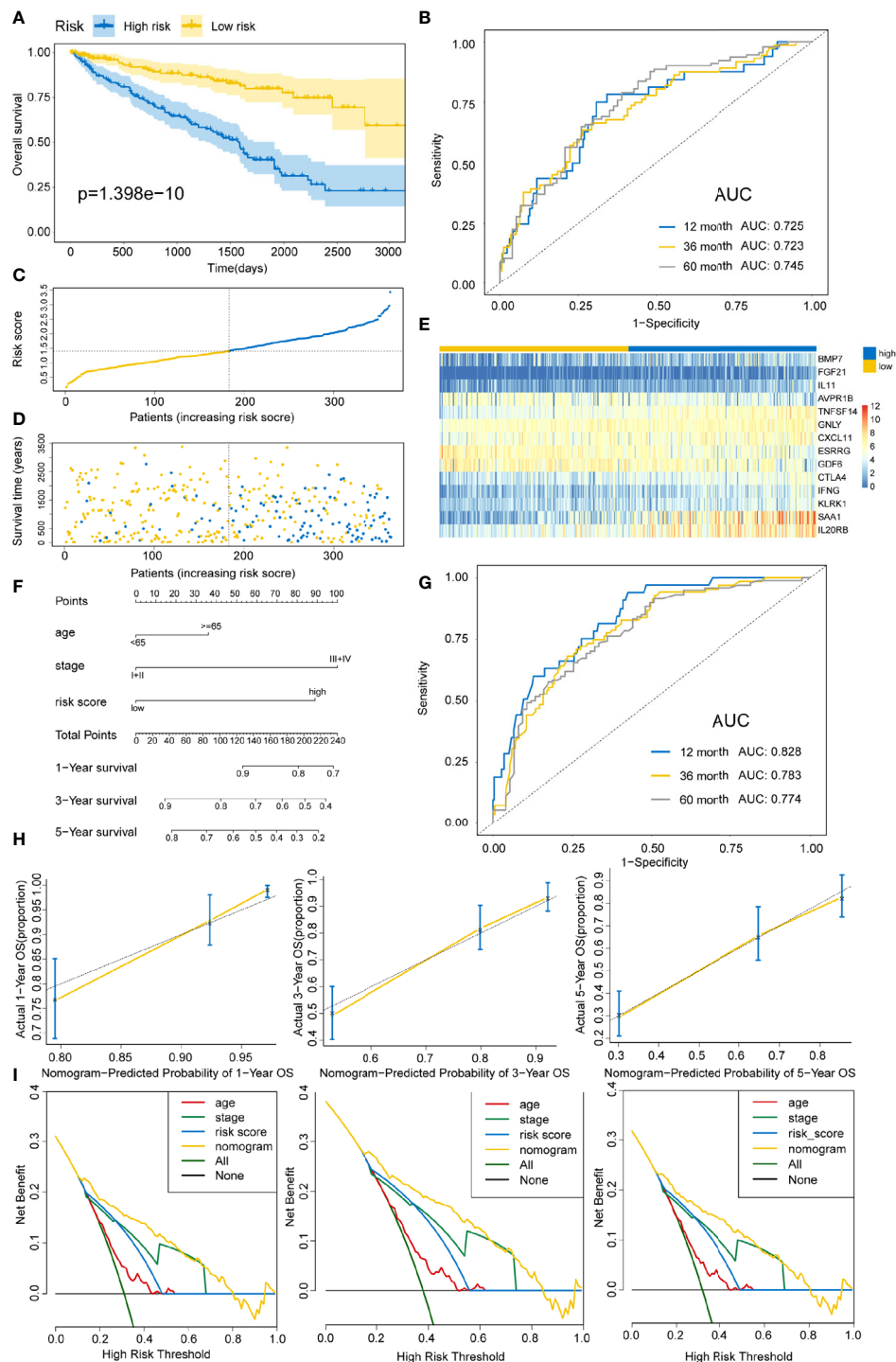
\*\* means  $P < 0.01$ .

poor OS outcomes ( $P=0.01$ ,  $0.0019$ ,  $<0.0001$ , and  $0.031$ , respectively), while the increase in activated dendritic cells was related to better OS ( $P=0.0079$ ) (Figure S3). Figure S4 displayed a weak or moderate correlation between the levels of various tumor-infiltrating immune cells and the risk score.

## Immune Landscape in KIRC Patients

Subsequently, the associations between risk score and some immune-associated features were assessed. The cGAS-STING

pathway has been shown to be a key signaling pathway in antitumor immunity and cancer therapeutics (46–48). Thus, four key genes (*TBK1*, *IRF3*, *MB21D1*, and *TMEM173*) in the cGAS-STING signaling pathway, three immune checkpoint molecules (PD-L1, CTLA-4, and PD-1), CYT, and the results of ESTIMATE algorithm (SS, IS, ES, and TP) and risk score were investigated. Figure 8A shows that the risk score values are correlated with the immune score, tumor purity, *TBK1*, ESTIMATE score, *IRF3*, stromal score, *MB21D1*, PD-1, and



**FIGURE 4 |** Constructing an immune risk score predictive model using the training set. **(A)** Kaplan-Meier curves for OS outcomes in the training cohort grouped into high- and low-risk score groups. **(B)** Time-dependent ROC curves for prediction of 1-, 3-, and 5-year survival outcomes. **(C)** Distribution of risk scores of the training cohort. **(D)** Vital statuses for patients in high- and low-risk patients. **(E)** Expression patterns for 14 immune-associated genes in high- and low-risk score cohorts. **(F)** A nomogram for the estimation of 1-, 3-, and 5-year OS probabilities in the training cohort. Risk scores and other independent prognostic factors are incorporated in the model. **(G)** Time-dependent ROC curves for the prediction of 1-, 3-, and 5-year survival rates using the nomogram. **(H)** Calibration plot of nomogram in the training cohort according to the agreement between predicted and observed 1-, 3-, and 5-year outcomes. The model's ideal performance is shown by dashed lines. **(I)** Decision curve analysis for 1-, 3-, and 5-year risk using the nomogram. Black line represents the hypothesis that no patient died at 1-, 3-, and 5-years.

**TABLE 2 |** Univariate and multivariate Cox regression analysis.

Variables	Univariate analysis		Multivariate analysis	
	HR (95%CI)	P value	HR (95%CI)	P value
Training set				
Age ( $\geq 65$ vs $<65$ )	1.595 (1.106–2.301)	0.012	1.518 (1.044–2.29)	0.0291
Grade (G3+4 vs G1+2)	2.427 (1.601–3.679)	$<0.001$	1.264 (0.803–1.990)	0.3109
T stage (T3+4 vs T1+2)	3.108 (2.143–4.508)	$<0.001$	1.986 (1.028–3.915)	0.0514
Stage (III+IV vs I+II)	4.066 (2.749–6.015)	$<0.001$	5.384 (2.551–11.369)	$<0.001$
Risk score (high vs low)	3.678 (2.372–5.703)	$<0.001$	2.699 (1.716–4.243)	$<0.001$
Testing set				
Age ( $\geq 65$ vs $<65$ )	2.015 (1.061–3.826)	0.032	2.737 (1.412–5.305)	0.003
Grade (G3+4 vs G1+2)	5.956 (2.311–15.350)	$<0.001$	2.318 (0.788–6.821)	0.127
T stage (T3+4 vs T1+2)	5.279 (2.543–10.957)	$<0.001$	2.443 (0.321–18.575)	0.127
Stage (III+IV vs I+II)	5.265 (2.480–11.77)	$<0.001$	1.510 (0.199–11.461)	0.69
Risk score (high vs low)	10.371 (3.656–29.418)	$<0.001$	7.991 (2.684–23.786)	$<0.001$
Total set				
Age ( $\geq 65$ vs $<65$ )	1.674 (1.220–2.298)	0.001	1.734 (1.262–2.383)	0.001
Grade (G3+4 vs G1+2)	2.887 (1.987–4.196)	$<0.001$	1.766 (1.189–2.622)	0.005
T stage (T3+4 vs T1+2)	3.468 (2.500–4.811)	$<0.001$	1.352 (0.722–2.530)	0.346
Stage (III+IV vs I+II)	4.248 (3.012–5.993)	$<0.001$	3.903 (2.018–7.549)	$<0.001$
Risk score (high vs low)	3.358 (2.316–4.868)	$<0.001$	2.424 (1.648–3.563)	$<0.001$

CTLA-4. **Figure S5** showed significant differences in the CYT, immune score, ESTIMATE score, stromal score, and tumor purity based on the Wilcoxon test between the two risk score subtypes ( $P<0.0001$ ). Importantly, the expression of IRF3, MB21D1, TMEM173, PD-1, and CTLA-4 was elevated in the high-risk than in the low-risk score subtype.

To further characterize immune cell infiltration, 28 immune cell signatures (25, 49–53) from diverse resources were investigated based on the single sample gene set enrichment analysis (ssGSEA) algorithm. As shown in **Figure 8B**, 23 immune subpopulations (multiple T cell signatures, including T helper cells, central memory CD8<sup>+</sup> T cells, and activated CD T cells) were enriched in high-risk patient cohort, whereas only two subpopulations (immature dendritic cells and neutrophils) were enriched in the low-risk patient group. Furthermore, DEGs between low- and high-risk groups were determined by gene set enrichment analysis (GSEA) using two MeSH terms (gene2pubmed and gendoo) to explore their immune-related functions. The DEGs were enriched in multiple immune-associated terms, including CD4-CD8 ratio, immune tolerance, lymphocyte cooperation, lymphocyte count, immunologic memory, and T-cell antigen receptor specificity in gendoo and gene2pubmed (**Figures 8C, D**).

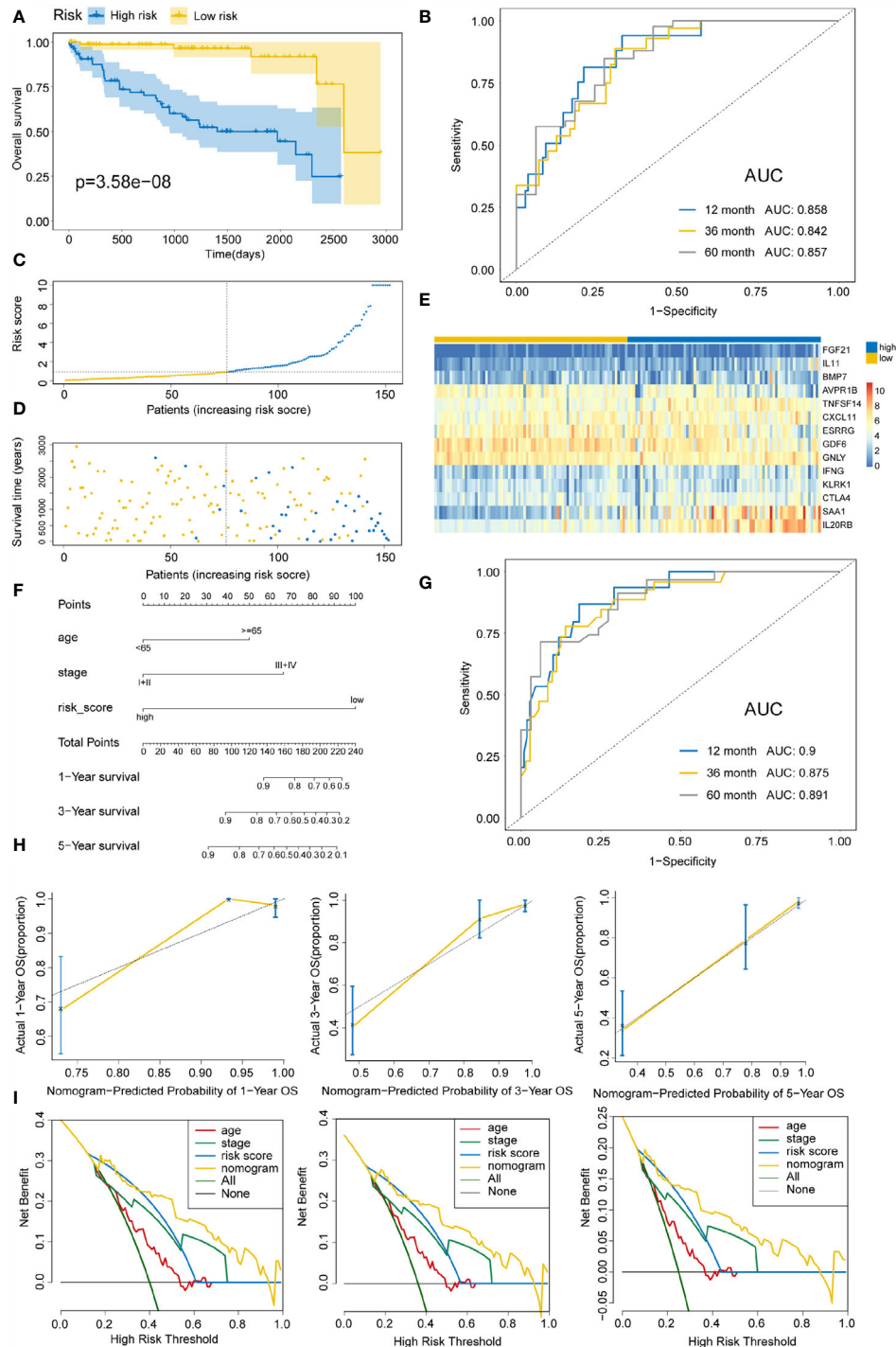
## Correlation Between Risk Score Model and T Cell Infiltrations, Antitumor Immunity, Antitumor Responses, and Oncogenic Pathways

Several studies have shown that cDC1 cells play a central role in the initiation of antitumor CD8<sup>+</sup> T cells and driving tumor-specific CD8<sup>+</sup> T cells by activating CXCL10 (54–57). Some studies (57–59) also clarified that the two key chemokines (CCL4 and CCL5) are the key modulators of cDC1 recruitment into tumors *via* activating CCR5 expression. Moreover, chemokines CXCR3, CXCL9, and CXCL10 have

been documented on T cell infiltration and NK cell recruitment (60). Thus, we investigated the expression level of CCL4, CXCR3, CXCL9, CCL5, and CXCL10 between high- and low-risk subtypes and the correlations between these genes and the risk score. The high-risk group patients exhibited higher expression levels compared to low-risk patients ( $P<0.05$ ) (**Figures S6A–E**). Moreover, strong positive correlations were established between risk scores and CXCR3, CCL5, CXCL9, CCL4, and CXCL10 ( $P<0.05$ ) (**Figures S6F–J**).

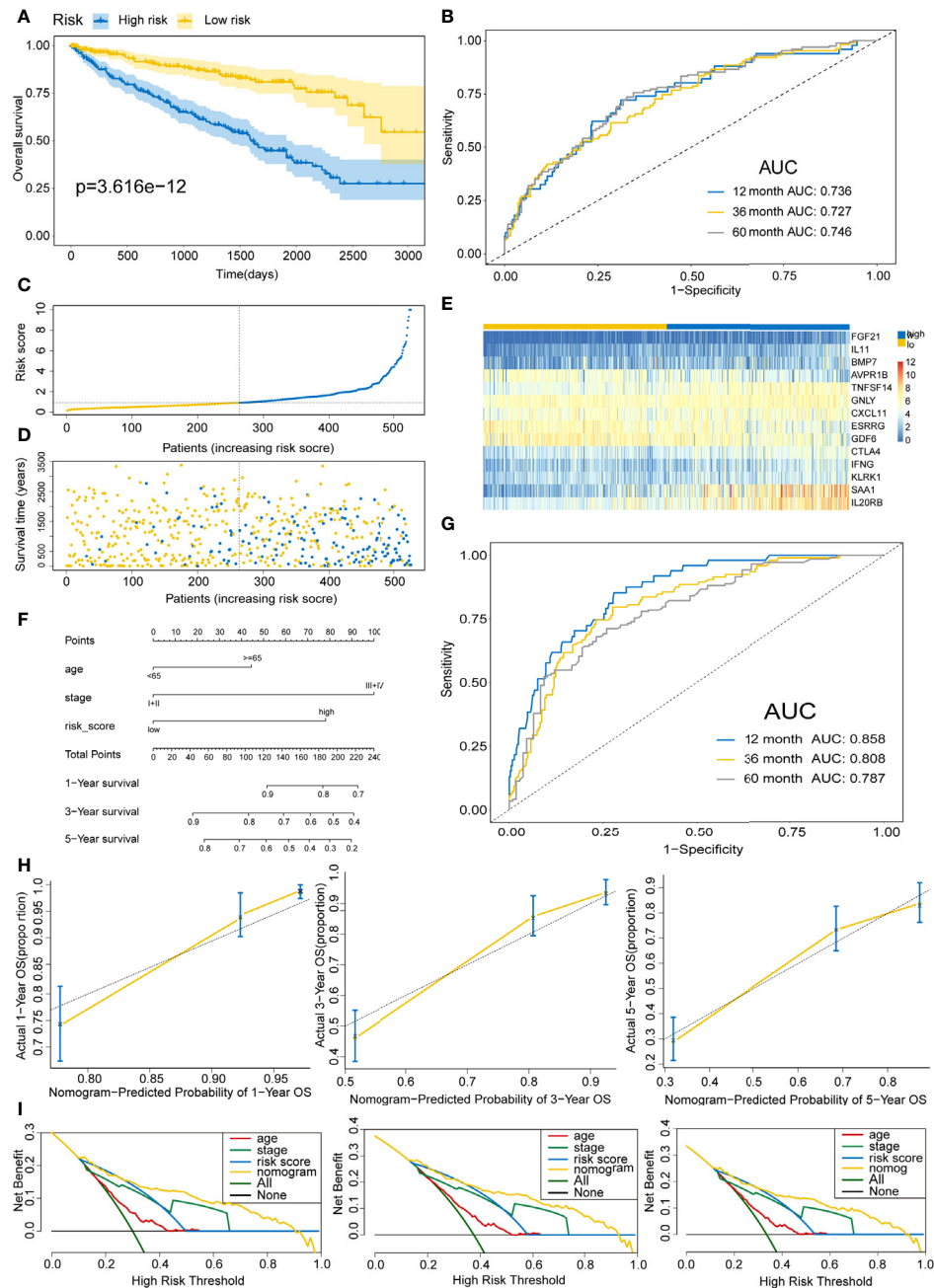
Moreover, we explored the association between risk scores, T cell infiltrations, and antitumor response scores (BATF3\_DC, IFNA, IFNG, IL\_1\_speed, T\_cell\_infiltration\_1, T\_cell\_infiltration\_2, TFNA, and TNFa\_speed) determined by ssGSEA from the corresponding TME gene signatures (57, 61). For the high-risk group, the ssGSEA scores for T cell infiltrations and antitumor responses were significantly elevated compared to the low-risk group, as determined by the Wilcoxon test ( $P<0.05$ ) (**Figure S7A**). A strong positive correlation was established between risk scores and ssGSEA scores save to BATF3\_DC ( $P<0.05$ ) (**Figures S7B–I**). Conclusively, high-risk score patients exhibited elevated T cell infiltration levels.

The differences in the normalized enrichment score (NES) value of 10 oncogenic pathways between low- and high-risk groups were calculated using ssGSEA algorithm; also, the correlation between the NES value and the risk score was evaluated. Compared to the low-risk group, cell cycle and TP53-related pathways exhibited significantly elevated NES values in the high-risk patient group, whereas the Hippo-, NRF2-, PI3K-, RAS-, and TGF- $\beta$ -related pathways in the high-risk patient group had lower NES value ( $P<0.05$ ) (**Figure S8A**). The correlations between the risk score and the NES value in the cell cycle ( $P=1.44\text{e-}12$ ) and TP53-related ( $P=0.024$ ) pathways were found to be positive (**Figure S8B**). Nevertheless, we also observed that the NES value of the Hippo-, NRF2-, PI3K-, RAS-, and TGF- $\beta$ -related pathways had a negative correlation with the risk score (**Figure S8B**).

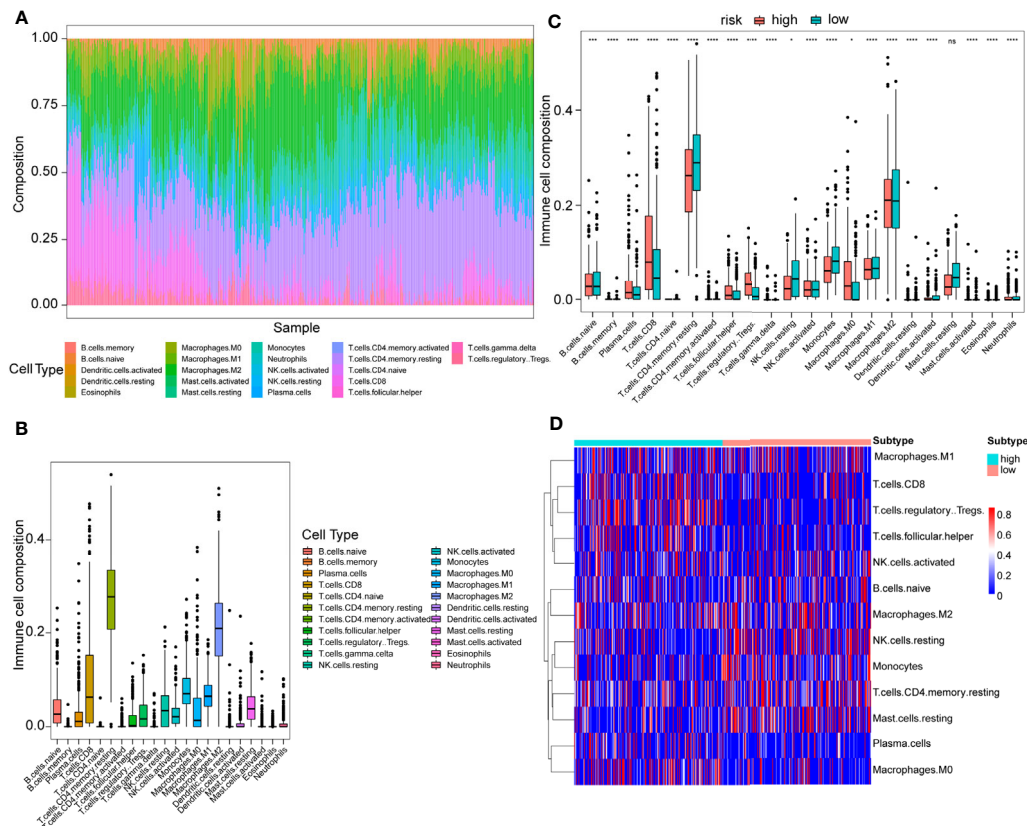


**FIGURE 5 |** Validating immune risk score prognostic predictive model in the testing set. **(A)** Kaplan–Meier curves for OS outcomes in the testing cohort divided by high- and low-risk score groups. **(B)** The time-dependent ROC curves for predicting 1-, 3-, and 5-year survival outcomes using this signature. **(C)** Risk score distribution in the testing cohort. **(D)** Vital statuses of patients in high- and low-risk patients. **(E)** Expression patterns for 14 immune-associated genes in the high- and low-risk score cohorts. **(F)** Nomogram developed for the prediction of probabilities for 1-, 3-, and 5-year OS outcomes in the testing cohort. Risk scores and other independent prognostic factors were incorporated in the nomogram. **(G)** Time-dependent ROC curves for prediction of 1-, 3-, and 5-year survival outcomes using the nomogram. **(H)** Calibration plot of nomogram in the training cohort according to the agreement between estimated and observed 1-, 3-, and 5-year outcomes. Dashed lines represent the nomograms' ideal performance. **(I)** Decision curve analysis for 1-, 3-, and 5-year risk using the nomogram. Black line represents the hypothesis that no patient died after 1-, 3-, and 5-years.





**FIGURE 6 |** Validating the immune risk score prognostic predictive model for the entire set. **(A)** Kaplan–Meier curves of the OS outcomes in the entire cohort divided as high- and low-risk score groups. **(B)** Time-dependent ROC curves for prediction of 1-, 3-, and 5-year survival outcomes using this signature. **(C)** Risk score distributions for the entire cohort. **(D)** Vital statuses for high- and low-risk group patients. **(E)** Expression patterns for 14 immune-associated genes in the high- and low-risk score cohorts. **(F)** Nomogram for the prediction of the probability of 1-, 3-, and 5-year OS outcomes in the entire cohort. Risk scores and other independent prognostic factors were incorporated into the model. **(G)** Time-dependent ROC curves for prediction of 1-, 3-, and 5-year survival outcomes using the nomogram. **(H)** Calibration plot of nomogram in the training cohort according to the agreement between observed and predicted 1-, 3-, and 5-year outcomes. The models' ideal performance is shown by the dashed lines. **(I)** Decision curve analysis for 1-, 3-, and 5-year risks using the nomogram. Black line represents the hypothesis that no patient died after 1-, 3-, and 5-years.



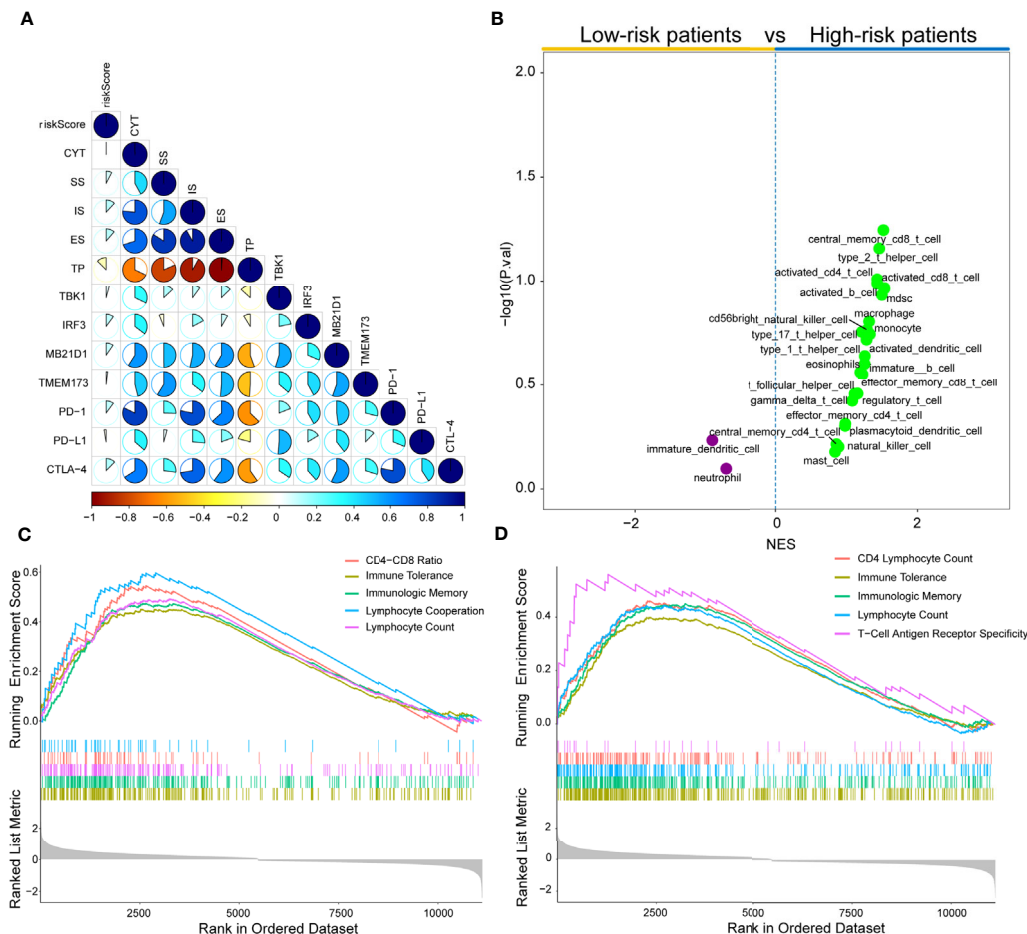
**FIGURE 7 |** Immune cell proportion analyses in the TCGA cohort between high- and low-risk score patients. **(A)** Overall view of relative proportions of immune cell infiltrations for 22 immune signatures. **(B)** Boxplots for 22 immune cell proportions in the TCGA cohort. **(C)** Boxplots for different immune cell infiltrations in the high- and low-risk score patients. Significance: ns≥0.05, \* < 0.05, \*\* < 0.01, and \*\*\* < 0.001, and \*\*\*\* < 0.0001. **(D)** Immune cell heatmap for patients in the high- and low-risk score subtypes. Only immune cells whose non-zero proportions exceeded half in all samples were plotted.

## Therapeutic Benefit of the Risk Score

Recently, ICB therapies have exhibited striking clinical benefits. However, the main challenge faced by ICB therapies is the limitation of effective predictive markers with only a few patients showing therapeutic response. Herein, the urothelial cancer database (IMvigor210) consisting of anti-PD-L1 therapy and the malignant melanoma database (GSE91061) administered with anti-PD-1 and-CTLA-4 therapy were used to investigate the association between risk score and immunotherapeutic benefits. **Figures 9A–F** and **Figures S9A–C** showed the distribution of clinical and molecular characteristics (immunotherapy response, binary response, immune phenotype, immune cells (IC) level, and tumor cells (TC) level and correlation with risk scores between high- and low-risk groups in the IMvigor210 cohort and GSE91061 cohort separately. For the immunotherapy response, the risk score of RCC with CR/PR were significantly lower than those of RCC with SD/PD, as assessed by the chi-squared test (IMvigor210 dataset:  $P < 0.001$ , GSE91061 dataset:  $P = 0.036$ ) (**Figure 9A** and **Figure S9C**). The violin plot further revealed that the risk scores in CR/PR were lower than those in SD/PD, as assessed by the Wilcoxon test (IMvigor210 cohort:

$P = 1.3e-08$ , GSE91061 cohort:  $P = 0.0075$ ) (**Figure 9C** and **Figure S9B**). Strikingly, Kaplan–Meier curves showed that high-risk score patients exhibited worse prognosis compared to the low-risk score patients in IMvigor210 ( $P < 0.0001$ ) (**Figure 9G**) and GSE91061 cohort ( $P = 0.00016$ ) (**Figure S8D**). In addition, IMvigor210 and GSE91061 were used to plot a time-dependent ROC. The current results displayed that the AUCs of our model for OS were 0.61 at 6 months, 0.673 at 12 months, and 0.729 at 18 months in the IMvigor210 cohort (**Figure 9H**) and 0.746 at 12 months, 0.712 at 18 months, and 0.753 at 24 months in the GSE91061 cohort (**Figure S9A**).

To further expand this study, the machine learning-based score (IPS) was determined to predict patients' response to ICI treatment. Four subtypes of IPS values (CTLA4\_neg\_PD1\_neg, CTLA4\_pos\_PD1\_neg, CTLA4\_neg\_PD1\_pos, and CTLA4\_pos\_PD1\_pos) were carried out to predict the KIRC patients' responses to anti-CTLA4 and anti-PD1 treatment. We found that relative probabilities to response to anti-PD1 were elevated in high-risk score patients ( $P = 0.023$ ), and the similar results were obvious in the combination treatment of anti-PD1 and anti-CTLA4 ( $P = 2.24e-04$ ) (**Figure 10A**). In addition, CTLA-



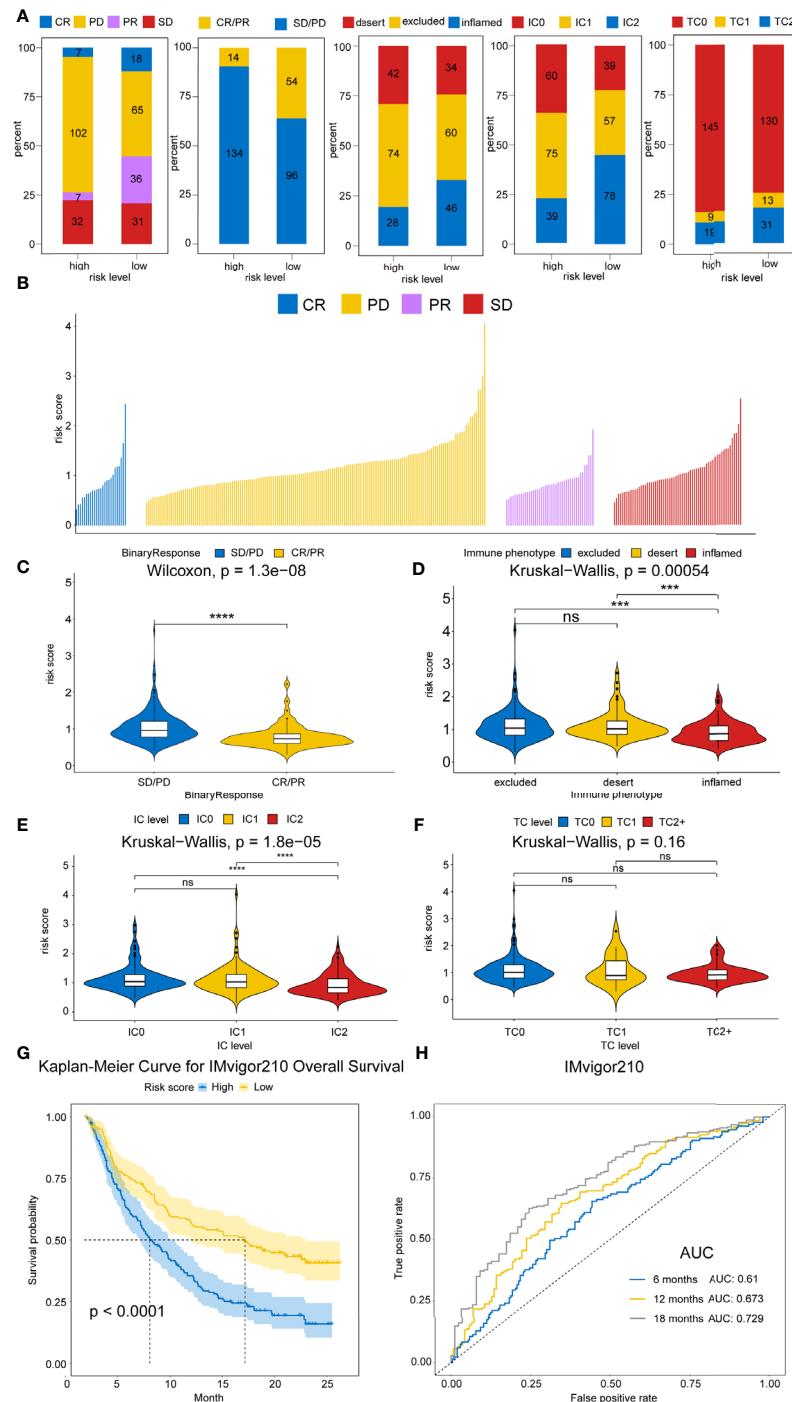
**FIGURE 8 |** Immune landscape of risk score in the TCGA cohort. **(A)** Correlations between risk score, levels of expression of PD-L1, CYT, TBK1, IRF3, MB21D1, CTLA-4, PD-1, and TMEM173, immune score, stromal score, ESTIMATE score, and tumor purity in the TCGA cohort. **(B)** Volcano plots for immune cell sub-population enrichment in high- and low-risk patients according to NES scores from ssGSEA. **(C)** Gene set enrichment analyses described the MeSH terms correlated with risk score using gendoo term in the TCGA cohort. **(D)** Gene set enrichment analysis described the MeSH terms correlated with the risk score using gene2pubmed term in the TCGA cohort.

4 and *PD-1* mRNA expression levels in the high-risk score group were significantly elevated compared to the low-risk score patients ( $P=1.07 \times 10^{-14}$  and  $P=2.02 \times 10^{-15}$ ), whereas no obvious difference was detected in the *PD-L1* mRNA expression level between high- and low-risk patients ( $P=0.603$ ) (Figure 10B). This phenomenon was consistent with the concept that high expression of ICI genes had a poor prognosis. Owing to the complex environment between immune infiltration and ICI genes, we further examined whether immune infiltration had consequences on the clinical prognosis in ICI genes. Figure 10C shows that low-risk score patients with high PD-1 exhibited better clinical outcomes compared to high-risk score and high PD-1, and the outcomes of low-risk score patients with low PD-1 were superior to those of high-risk score patients and low PD-1 levels ( $P<0.0001$ ). Also, patient groups showed similar findings, and survival patterns were yielded using risk score and PD-L1 or CTLA4 ( $P<0.0001$ ) (Figures 10C).

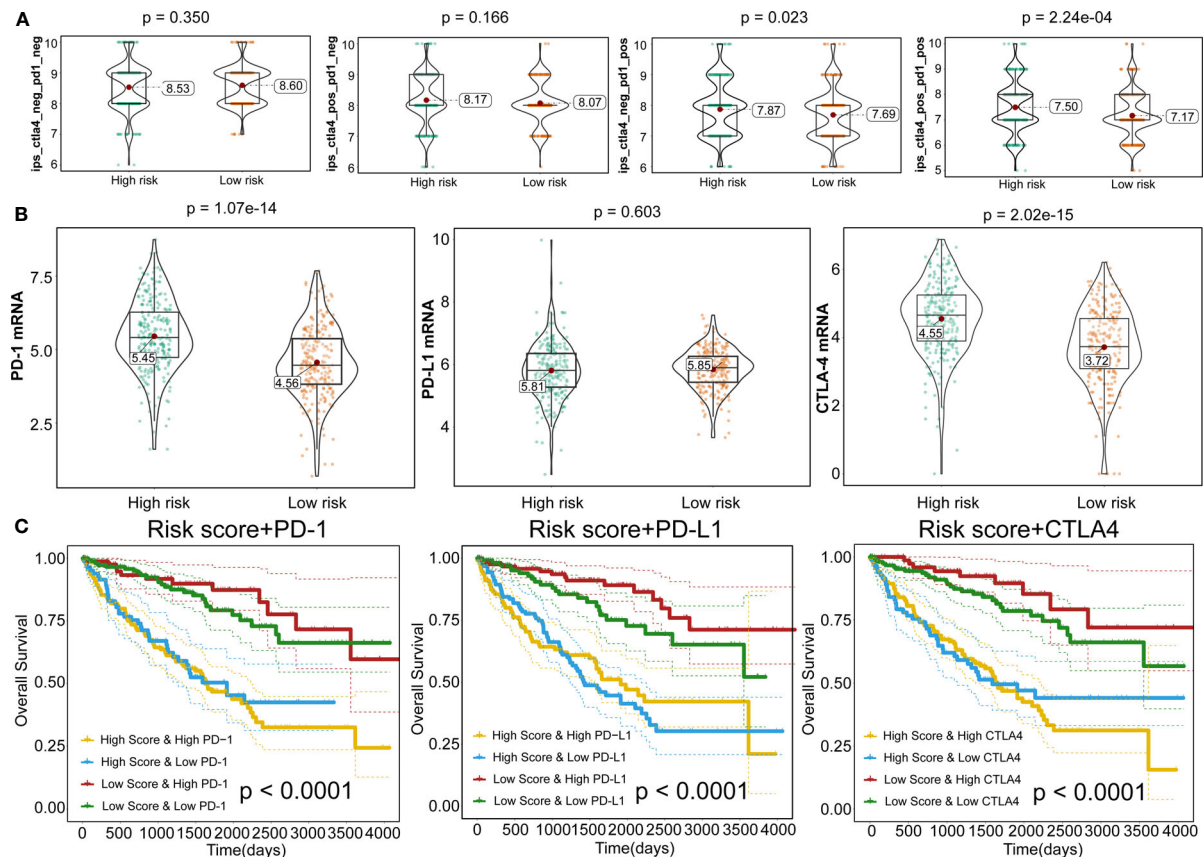
The responsive predictive values of the risk score to chemotherapy and target-therapy were also investigated by the IC50 of eight drugs. The estimated IC50 values of Cisplatin, Gemcitabine, Sorafenib, and Vinorelbine in high-risk patients were significantly elevated compared to low-risk patients, which indicating the high-risk patients showed a stronger drug resistance ( $P<0.05$ ) (Figures 11A, B). Similarity, patients with high-risk group were associated with increased sensitivity to Gefitinib, Vinblastine, and Sunitinib relative to low-risk patients ( $P<0.05$ ) (Figures 11A, B).

## Risk Score and TMB

Next, we analyzed the gene mutations of each KIRC patient. The waterfall chart showed the top 20 genes with the highest mutation frequencies: *VHL*, *PBRM1*, *SETD2*, *MTOR*, *TTN*, *MUC16*, *KDM5C*, *BAP1*, *HMCN1*, *DNAH9*, *LRP2*, *ATM*, *ARID1A*, *CSMD3*, *DST*, *KMT2C*, *ERBB4*, *SMARCA4*, *USH2A*,



**FIGURE 9** | Therapeutic benefits of risk scores calculated by our model. **(A)** Bar graphs illustrate the distribution of the clinicopathological parameters for IMvigor210 dataset in high- and low-risk patients based on chi-square test. ( $P=4.8008E-08$ ,  $P=4.8008E-08$ ,  $P=0.3305$ ,  $P=6.0E-6$ , and  $P=1.6023E-59$ , respectively). **(B)** Waterfall plot illustrates the risk score distributions for patients exhibiting different immunotherapeutic responses in the IMvigor210 dataset. **(C)** Violin plot illustrates the risk score distributions for patients exhibiting different anti-PD-L1 immunotherapies in IMvigor210 dataset. **(D)** Violin plot illustrates the risk score distributions for patients exhibiting different immune phenotypes in the IMvigor210 dataset. **(E)** Violin plot illustrates the risk score distributions for patients with varying IC levels in the IMvigor210 dataset. **(F)** Violin plot illustrates the risk score distributions for patients with varying TC levels in the IMvigor210 dataset. **(G)** Kaplan-Meier curves for OS outcomes in the IMvigor210 cohort assigned into high- and low-risk score groups. **(H)** Time-dependence ROC curves of anti-PD-L1 immunotherapy response prediction at 0.5-, 1-, and 1.5-year survival rate in the IMvigor210 dataset. Significance: ns $\geq 0.05$ , \*\*\* $<0.001$ , and \*\*\*\* $<0.0001$ .



**FIGURE 10 |** Responses to immune checkpoint inhibitors. **(A)** Violin plots illustrate the relative probabilities for anti-PD-1 and anti-CTLA-4 treatment responses between high- and low-risk groups. **(B)** Violin plots for expression levels of PD-1, CTLA-4, and PD-L1 between high- and low-risk patients. **(C)** Kaplan-Meier curves for OS outcomes among four groups, according to risk score and PD-1, CTLA-4, and PD-L1.

and PCLO (Figure 12A). Subsequently, the TMB for each sample was determined and was found to be higher in the high-risk patients ( $P=0.037$ ) (Figure 12B) and related to shorter OS ( $P=0.023$ ) than in low-risk patients (Figure 12C).

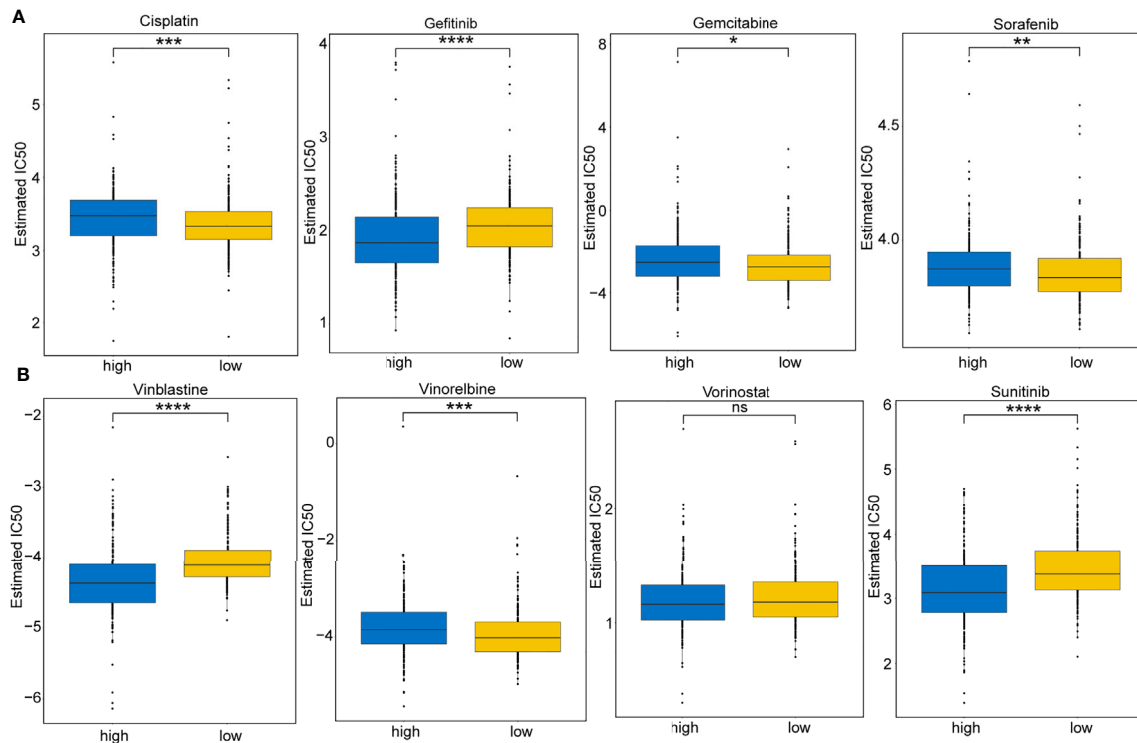
## Prediction of High- and Low-Risk Scores by XGBoost Algorithm

XGBoost is an efficient and reliable machine learning classifier based on gradient boosting, designed to solve data science challenges accurately and rapidly in bioinformatics (62, 63). Using this approach, a classifier that could predict high- and low-risk score groups for KIRC patients based on expression levels of 14 selected genes was constructed for the training cohort. SHAP dependency plot and the importance of 14 features were visualized in Figures 13A, B to evaluate the contribution of each feature towards prediction. Figure 13C showed that the AUC of the training cohort was 100%. Then, classification model performance was assessed using the testing and entire total cohorts (Figures S10A, B and Figures 13C). Taken together, the middle cutoff value might be suitable to classify KIRC patients.

## Identification of Potential Small Molecule Drugs

According to CMAP analysis, 10 small molecule drugs with highly significant correlations are listed in Table 3. Among these, Finasteride, Biperiden, Merbromin, Cefamandole, Fludrocortisone, and Vincamine displayed a high negative correlation and potential to improve the prognosis of RCC. Subsequently, the SAA1 gene contributing to the model according to the feature importance was docked with these 10 compounds (Table 4). Next, we identified the compounds except for Orphenadrine that showed a high binding affinity against the target protein due to their binding energy  $< -5$  kcal/mol. Moreover, the three-dimensional structure of top two high-affinity compounds combined with SAA1 is shown in Figures S11A, B. In SAA1-merbromin complex, due to multiple phenylene rings and active groups, merbromin forms hydrogen bonds with activity groups of amino acids, such as GLN-66, ARG-25, and TRP-53, indicating that merbromin could match well with SAA1 protein. Similarly, the SAA1-Cefamandole complex can be formed by multiple interactions, such as the cooperation of hydrogen bonding and multiple  $\pi$ - $\pi$  stacking interactions. Hence, these





**FIGURE 11 |** Immunotherapeutic and chemotherapeutic responses for high- and low-risk patients. **(A)** Boxplots illustrate the immunotherapeutic and chemotherapeutic responses of Cisplatin, Gefitinib, Gemcitabine, and Sorafenib in the high- and low-risk patients. **(B)** Boxplots illustrate the immunotherapeutic and chemotherapeutic responses of Vinblastine, Vinorelbine, Vorinostat, and Sunitinib in the high- and low-risk patients. Significance: ns $\geq$ 0.05, \* $<$ 0.05, \*\* $<$ 0.01, \*\*\* $<$ 0.001, and \*\*\*\* $<$ 0.0001.

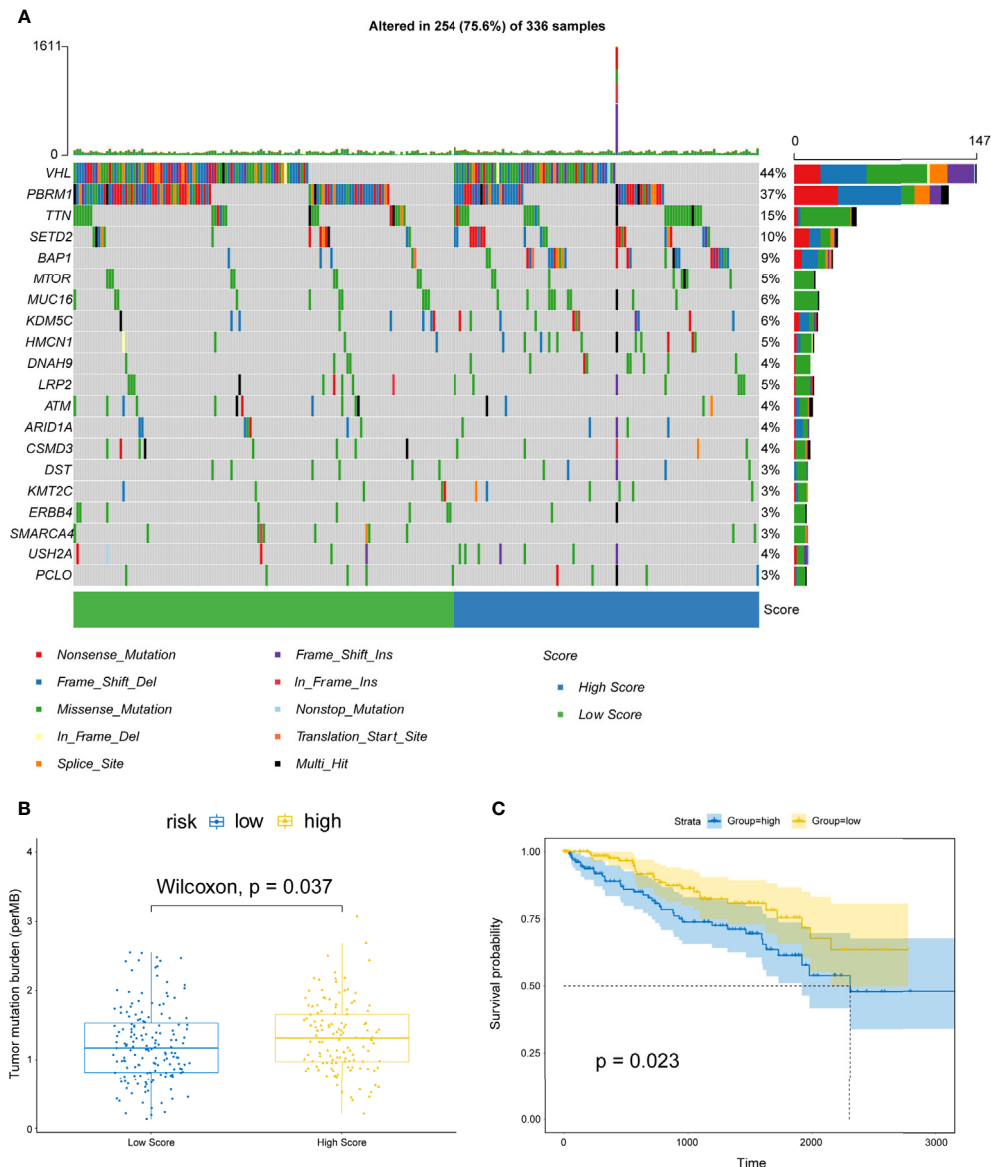
two compounds were both regarded as potential SAA1 inhibitors that could improve the prognosis of RCC.

## DISCUSSION

Epidemiological evidence indicated that the incidence of RCC had a continually increasing trend with high mortality (64, 65). Clinical decision-making tools were effective prognostic biomarkers to predict the survival outcomes of RCC patients, rendering them a viable choice for clinicians. To date, the prognostic prediction of RCC patients relies on the TNM staging system according to the clinical practice guidelines (66). However, this system failed to taken the influence of gene level of RCC into consideration and made it not always able to predict the patients accurately. In recent years, IRGs have gradually gained attention with in-depth studies on immune-escape and immunotherapeutic mechanisms. Hence, an immune-related prognostic system is an urgent requirement for a supplementary TNM staging system.

Next, we screened for immune-associated DEGs in RCC. To minimize the potential for overfitting, 14 genes established the prognostic immune signature and were validated in TCGA through the univariate Cox proportional hazard regression and LASSO Cox analysis. Subsequently, we confirmed the

independent predictive roles of this signature. Then, a personalized, predictive nomogram with a risk score was developed, which served as a predictive indicator; the signature encompassed a total of 14 IRGs. Among these, SAA1, TNFSF14, FGF21, IFNG, BMP7, and IL11 are biomarkers for predicting RCC outcomes (67–72). For example, as a member of the serum amyloid A family of apolipoproteins, SAA1 can increase the invasive capacity of tumor cells in RCC by inducing MMP-9 expression (73), which make it serve as a biomarker for the diagnosis and prognosis of advanced and metastatic renal cell carcinoma. In addition, as a member of the IL-6 family of cytokines, IL-11 exerts pleiotropic oncogenic activities may by stimulating angiogenesis and metastasis, which make it become an independent indicator of poor prognosis in RCC (71). The other IRGs, such as IL20RB, ESRRG, GDF6, were reported to be involved in the regulation of carcinogenesis (74–76) but not yet investigated in RCC. Moreover, some IRGs were also involved in TIME. For example, NKG2D receptor, KLRK1, is expressed in NK cells and activated CD8<sup>+</sup> T cells, involved in innate immune responses (77). In some studies also identified GNLY as the first lymphocyte-derived alarmin protein to promote antigen-presenting cell (APC) recruitment, activation, and antigen-specific immune responses (78). CTLA-4 is a negative regulator and modulates T cell activation, and induces tolerance (79). CXCL11 is activated by IFN- $\gamma$  and IFN- $\beta$  and

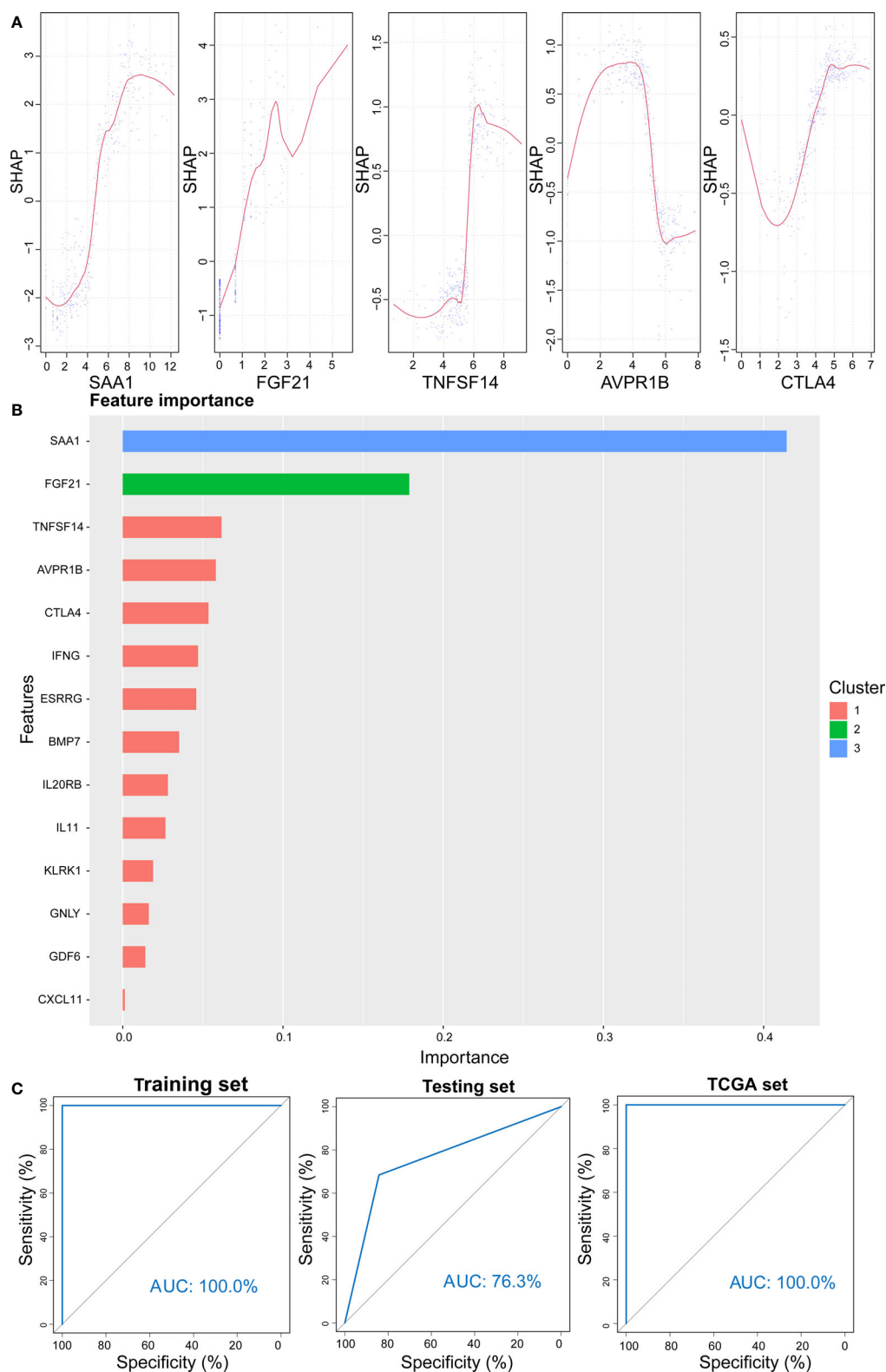


**FIGURE 12 |** Correlations between risk scores and TMB. **(A)** OncoPrint displays the mutation profile of top 20 frequently mutated genes. Each column represents individual patients and mutated genes arranged by mutation rates. The right shows the mutation percentage, and color-coding indicates the mutation type. **(B)** Boxplot shows the difference of TMB between high- and low-risk patients. **(C)** Kaplan-Meier curves for OS divided by the high TMB group and the low TMB group.

can stimulate immune cells by promoting Th1 polarization and enhancing the antitumor immunity (80). To sum up, these IRGs may affected the prognosis and treatment of RCC by influencing TIME.

Herein, some self-validation processes, including the associations between risk scores and immune cell proportions, T cell infiltrations, antitumor immunity, antitumor response, GSEA analysis, and oncogenic pathways, were conducted to identify the risk score effectiveness in characterizing the immune landscape features of RCC patients. For immunotherapeutic development, anti-PD-1, anti-CTLA-4,

and anti-PD-L1 treatment have been under intensive focus in solid tumors. Nevertheless, a small number of patients respond to such treatment, and some studies (81–83) pointed out that PD-L1 and PD-1 expression levels are not reliable biomarkers to predict ICI treatment. Hence, it is necessary for clinicians to develop a reliable tool for appropriate patient selection in immunotherapy. Based on these findings, we established that the risk score is a robust immune classifier for classifying RCC patients in different subtypes. Moreover, we also demonstrated that high-score patients were more immunotherapeutically suitable compared to patients in the low-risk score group.

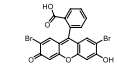
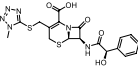
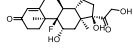
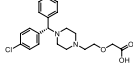
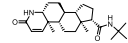
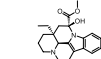
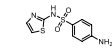
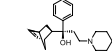
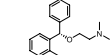
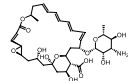


**FIGURE 13** | Prediction results from the XGBoost algorithm. **(A)** SHAP contribution dependency plots for the training cohort. **(B)** Importance of 14 features of the training cohort. **(C)** ROC curve for XGBoost algorithm for the prediction of high- and low-risk patients in training, testing, and entire cohorts.

**TABLE 3 |** The results of CMAP analysis.

rank	Cmap name	mean	n	enrichment	p	specificity
1	cetirizine	0.62	4	0.902	0.0001	0
2	finasteride	-0.385	6	-0.791	0.00016	0
3	orphenadrine	0.499	6	0.779	0.00028	0
4	biperiden	-0.516	5	-0.83	0.00034	0.0204
5	merbromin	-0.524	5	-0.807	0.00062	0.0081
6	natamycin	0.572	4	0.849	0.00074	0
7	sulfathiazole	0.515	5	0.785	0.00104	0
8	cefamandole	-0.478	4	-0.834	0.00137	0
9	fludrocortisone	-0.308	8	-0.63	0.00144	0.0704
10	vincamine	-0.539	6	-0.699	0.00171	0.0177

**TABLE 4 |** The selected compounds of docking results.

Name	Compound Structure	Target	Binding Energy (kcal/mol)	Combination Type
merbromin		SAA1	-7.85	Hydrogen bonds, Hydrophobic interactive, $\pi$ -stacking
cefamandole		SAA1	-7.43	Hydrogen bonds, Hydrophobic interactive, $\pi$ -stacking
fludrocortisone		SAA1	-7.35	Hydrogen bonds, Hydrophobic interactive
cetirizine		SAA1	-7.26	Hydrogen bonds, Hydrophobic interactive, $\pi$ -stacking
finasteride		SAA1	-7.09	Hydrogen bonds, Hydrophobic interactive
vincamine		SAA1	-6.96	Hydrogen bonds, Hydrophobic interactive
sulfathiazole		SAA1	-6.01	Hydrogen bonds, Hydrophobic interactive
biperiden		SAA1	-5.61	Hydrophobic interactive, $\pi$ -stacking
natamycin		SAA1	-5.35	Hydrophobic interactive, $\pi$ -stacking
orphenadrine		SAA1	0	0

Targeted therapy is currently the main treatment strategy for metastatic RCC. Thus, it is necessary to identify patients with the potential to benefit from targeted therapy for RCC. Interestingly, our data showed that high-risk patients had a high sensitivity to Gefitinib, Vinblastine, and Sunitinib compared to low-risk score patients, who exhibited high sensitivity to Cisplatin, Sorafenib, Gemcitabine, and Vinorelbine. These responses could be attributed to the differences in the drug target. In addition, the TMB values of the high-risk score patients were elevated compared to those of the low-risk score patients. This finding was consistent with the concept that elevated TMB values are associated with a high probability of satisfactory immunotherapeutic outcomes (84, 85).

Nevertheless, the present study had some limitations. First, although our model exhibited precise predictive capability to predict the survival of RCC patients, multiple large external

cohorts of patients with RCC are also needed to further validate. Secondly, only the median risk score was used to classify the RCC patients into high- and low-risk score subtypes. An optimal cutoff of the risk score is essential for the stratification of RCC patients. Although our model had been correlated with immune cells, the mechanism underlying poor prognosis is unclear, requiring additional experimental and theoretical studies on immune cells in RCC to further understand their functional role.

## CONCLUSIONS

Taken together, our proposed immune prognostic, predictive model could be used as a robust classifier for the prediction of survival outcomes and individual treatment guidance of adjuvant chemotherapy and anticancer immunotherapy for RCC.



## DATA AVAILABILITY STATEMENT

The RNA-seq data and corresponding clinical information were observed from the TCGA (<https://portal.gdc.cancer.gov/>). The immune-related gene list was got from the IMMPORT website (<https://www.immport.org/>).

## AUTHOR CONTRIBUTIONS

All authors participated in the design, interpretation of the studies, analysis of the data, and review of the manuscript. TF and JZ conceived and designed the whole project and wrote the manuscript. TF, DW, ZF, and ZW analyzed and visualized the data. TF, QL, PG, and XY interpreted the data and partook in the discussion. ML, YJ, and YL revised the final version of the manuscript. All authors contributed to the article and approved the submitted version.

## REFERENCES

- Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2019. *CA Cancer J Clin* (2019) 69:7–34. doi: 10.3322/caac.21551
- Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2020. *CA Cancer J Clin* (2020) 70:7–30. doi: 10.3322/caac.21590
- Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA: A Cancer J Clin* (2021) 71:7–33. doi: 10.3322/caac.21654
- Makhov P, Joshi S, Ghatalia P, Kutikov A, Uzzo RG, Kolenko VM. Resistance to Systemic Therapies in Clear Cell Renal Cell Carcinoma: Mechanisms and Management Strategies. *Mol Cancer Ther* (2018) 17:1355–64. doi: 10.1158/1535-7163.MCT-17-1299
- Aoun F, Rassy EE, Assi T, Albisinni S, Katan J. Advances in Urothelial Bladder Cancer Immunotherapy, Dawn of a New Age of Treatment. *Immunotherapy* (2017) 9:451–60. doi: 10.2217/imt-2017-0007
- Hoos A. Development of Immuno-Oncology Drugs — From CTLA4 to PD1 to the Next Generations. *Nat Rev Drug Discov* (2016) 15:235–47. doi: 10.1038/nrd.2015.35
- Kamal Y, Cheng C, Frost HR, Amos CI. Predictors of Disease Aggressiveness Influence Outcome From Immunotherapy Treatment in Renal Clear Cell Carcinoma. *Oncoimmunology* (2019) 8:e1500106. doi: 10.1080/2162402X.2018.1500106
- Bu X, Yao Y, Li X. Immune Checkpoint Blockade in Breast Cancer Therapy. *Adv Exp Med Biol* (2017) 1026:383–402. doi: 10.1007/978-981-10-6020-5\_18
- Hu ZL, Ho AY, McArthur HL. Combined Radiation Therapy and Immune Checkpoint Blockade Therapy for Breast Cancer. *Int J Radiat Oncol Biol Phys* (2017) 99:153–64. doi: 10.1016/j.ijrobp.2017.05.029
- Mahoney KM, Freeman GJ, McDermott DF. The Next Immune-Checkpoint Inhibitors: PD-1/PD-L1 Blockade in Melanoma. *Clin Ther* (2015) 37:764–82. doi: 10.1016/j.clinthera.2015.02.018
- Van Allen D, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic Correlates of Response to CTLA-4 Blockade in Metastatic Melanoma. *Science* (2015) 350:207–11. doi: 10.1126/science.aad0095
- Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, White J, et al. Evolution of Neoantigen Landscape During Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. *Cancer Discov* (2017) 7:264–76. doi: 10.1158/2159-8290.CD-16-0828
- Xia L, Liu Y, Wang Y. PD-1/PD-L1 Blockade Therapy in Advanced Non-Small-Cell Lung Cancer: Current Status and Future Directions. *Oncologist* (2019) 24:S31–41. doi: 10.1634/theoncologist.2019-IO-S1-s05
- Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, et al. Nivolumab Versus Everolimus in Advanced Renal-Cell Carcinoma. *N Engl J Med* (2015) 373(19):1803–13. doi: 10.1056/NEJMoa1510665
- Motzer RJ, Tannir NM, McDermott DF, Arén Frontera O, Melichar B, Choueiri TK, et al. Nivolumab Plus Ipilimumab Versus Sunitinib in Advanced Renal-Cell Carcinoma. *N Engl J Med* (2018) 378(14):1277–90. doi: 10.1056/NEJMoa1712126
- Cooper LA, Gutman DA, Chisolm C, Appin C, Kong J, Rong Y, et al. The Tumor Microenvironment Strongly Impacts Master Transcriptional Regulators and Gene Expression Class of Glioblastoma. *Am J Pathol* (2012) 180:2108–19. doi: 10.1016/j.ajpath.2012.01.040
- Curry JM, Sprandio J, Cognetti D, Luginbuhl A, Bar-ad V, Pribitkin E, et al. Tumor Microenvironment in Head and Neck Squamous Cell Carcinoma. *Semin Oncol* (2014) 41:217–34. doi: 10.1053/j.seminoncol.2014.03.003
- Senbabaoglu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G, et al. Tumor Immune Microenvironment Characterization in Clear Cell Renal Cell Carcinoma Identifies Prognostic and Immunotherapeutically Relevant Messenger RNA Signatures. *Genome Biol* (2016) 17:231. doi: 10.1186/s13059-016-1092-z
- Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612
- Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, Bacac M, et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* (2017) 169(4):736–749.e18. doi: 10.1016/j.cell.2017.04.016
- Kopecký O, Lukesová S, Vroblová V, Vokurková D, Morávek P, Safránek H, et al. Phenotype Analysis of Tumour-Infiltrating Lymphocytes and Lymphocytes in Peripheral Blood in Patients With Renal Carcinoma. *Acta Med (Hradec Kralove)* (2007) 50(3):207–12. doi: 10.14712/18059694.2017.84
- Komohara Y, Hasita H, Ohnishi K, Fujiwara Y, Suzu S, Eto M, et al. Macrophage Infiltration and Its Prognostic Relevance in Clear Cell Renal Cell Carcinoma. *Cancer Sci* (2011) 102(7):1424–31. doi: 10.1111/j.1349-7006.2011.01945.x
- Wang B, Liu B, Yu G, Huang Y, Lv C. Differentially Expressed Autophagy-Related Genes Are Potential Prognostic and Diagnostic Biomarkers in Clear-Cell Renal Cell Carcinoma. *Aging* (2019) 11(20):9025–42. doi: 10.18632/aging.102368
- Wang Y, Chen L, Wang G, Cheng S, Qian K, Liu X, et al. Fifteen Hub Genes Associated With Progression and Prognosis of Clear Cell Renal Cell Carcinoma Identified by Coexpression Analysis. *J Cell Physiol* (2019) 234:10225–37. doi: 10.1002/jcp.27692
- Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: Disseminating Data to the Public for the Future of Immunology. *Immunol Res* (2014) 58:234–9. doi: 10.1007/s12026-014-8516-1
- Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. Tgfb Attenuates Tumour Response to PD-L1 Blockade by Contributing to Exclusion of T Cells. *Nature* (2018) 554:544–8. doi: 10.1038/nature25501

## FUNDING

This study was supported by the National Science Foundation of Beijing (7172068 and 7192053).

## ACKNOWLEDGMENTS

We are grateful to the TCGA, TCIA, ImmPort and GEO database for the availability of the data.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2021.762120/full#supplementary-material>

27. Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With DESeq2. *Genome Biol* (2014) 15:550. doi: 10.1186/s13059-014-0550-8
28. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* (2012) 16:284–7. doi: 10.1089/omi.2011.0118
29. Groeneveld CS, Chagas VS, Jones SJM, Robertson AG, Ponder BAJ, Meyer KB, et al. RTNsurvival: An R/Bioconductor Package for Regulatory Network Survival Analysis. *Bioinformatics* (2019) 35:4488–9. doi: 10.1093/bioinformatics/btz229
30. R T. The Lasso Method for Variable Selection in the Cox Model. *Stat Med* (1997) 16:385–95. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3
31. Iasonos A, Schrag D, Raj GV, Panageas KS. How to Build and Interpret a Nomogram for Cancer Prognosis. *J Clin Oncol* (2008) 26:1364–70. doi: 10.1200/JCO.2007.12.9791
32. Duan J, Xie Y, Qu L, Wang L, Zhou S, Wang Y, et al. A Nomogram-Based Immunoprofile Predicts Overall Survival for Previously Untreated Patients With Esophageal Squamous Cell Carcinoma After Esophagectomy. *J Immunother Cancer* (2018) 6:100. doi: 10.1186/s40425-018-0418-7
33. Kiran M, Chatrath A, Tang X, Keenan DM, Dutta A. A Prognostic Signature for Lower Grade Gliomas Based on Expression of Long Non-Coding RNAs. *Mol Neurobiol* (2019) 56:4786–98. doi: 10.1007/s12035-018-1416-y
34. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making* (2006) 26:565–74. doi: 10.1177/0272989X06295361
35. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust Enumeration of Cell Subsets From Tissue Expression Profiles. *Nat Methods* (2015) 12:453–7. doi: 10.1038/nmeth.3337
36. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and Genetic Properties of Tumors Associated With Local Immune Cytolytic Activity. *Cell* (2015) 160:48–61. doi: 10.1016/j.cell.2014.12.033
37. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* (2017) 18:248–62. doi: 10.1016/j.celrep.2016.12.019
38. Geleher P CN, Huang RS. Clinical Drug Response can be Predicted Using Baseline Gene Expression Levels and In Vitro Drug Sensitivity in Cell Lines. *Genome Biol* (2014) 15(3):R47. doi: 10.1186/gb-2014-15-3-r47
39. Liu Z, Zhang Y, Shi C, Zhou X, Xu K, Jiao D, et al. A Novel Immune Classification Reveals Distinct Immune Escape Mechanism and Genomic Alterations: Implications for Immunotherapy in Hepatocellular Carcinoma. *J Transl Med* (2021) 19(1):5. doi: 10.1186/s12967-020-02697-y
40. Liu Z, Wang L, Guo C, Liu L, Jiao D, Sun Z, et al. TTN/OBSCN 'Double-Hit' Predicts Favourable Prognosis, 'Immune-Hot' Subtype and Potentially Better Immunotherapeutic Efficacy in Colorectal Cancer. *J Cell Mol Med* (2021) 25(7):3239–51. doi: 10.1111/jcmm.16393
41. Liu Z, Liu L, Lu T, Wang L, Li Z, Jiao D, et al. Hypoxia Molecular Characterization in Hepatocellular Carcinoma Identifies One Risk Signature and Two Nomograms for Clinical Management. *J Oncol* (2021) 2021:6664386. doi: 10.1155/2021/6664386
42. Liu Z, Liu L, Jiao D, Guo C, Wang L, Li Z, et al. Association of RYR2 Mutation With Tumor Mutation Burden, Prognosis, and Antitumor Immunity in Patients With Esophageal Adenocarcinoma. *Front Genet* (2021) 12:669694. doi: 10.3389/fgene.2021.669694
43. Liu Z, Wang L, Liu L, Lu T, Jiao D, Sun Y, et al. The Identification and Validation of Two Heterogenous Subtypes and a Risk Signature Based on Ferroptosis in Hepatocellular Carcinoma. *Front Oncol* (2021) 11:619242. doi: 10.3389/fonc.2021.619242
44. Liu Z, Lu T, Wang L, Liu L, Li L, Han X. Comprehensive Molecular Analyses of a Novel Mutational Signature Classification System With Regard to Prognosis, Genomic Alterations, and Immune Landscape in Glioma. *Front Mol Biosci* (2021) 8:682084. doi: 10.3389/fmolb.2021.682084
45. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer. *Genome Res* (2018) 28:1747–56. doi: 10.1101/gr.239244.118
46. An X, Zhu Y, Zheng T, Wang G, Zhang M, Li J, et al. An Analysis of the Expression and Association With Immune Cell Infiltration of the cGAS/STING Pathway in Pan-Cancer. *Mol Ther - Nucleic Acids* (2019) 14:80–9. doi: 10.1016/j.omtn.2018.11.003
47. Li T, Cheng H, Yuan H, Xu Q, Shu C, Zhang Y, et al. Antitumor Activity of cGAMP via Stimulation of cGAS-cGAMP-STING-IRF3 Mediated Innate Immune Response. *Sci Rep* (2016) 6:19049. doi: 10.1038/srep19049
48. Ramanjulu JM, Pesiridis GS, Yang J, Concha N, Singhaus R, Zhang S-Y, et al. Design of Amidobenzimidazole STING Receptor Agonists With Systemic Activity. *Nature* (2018) 564:439–43. doi: 10.1038/s41586-018-0705-y
49. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression. *Genome Biol* (2016) 17:218. doi: 10.1186/s13059-016-1070-5
50. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf Anna C, et al. Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* (2013) 39:782–95. doi: 10.1016/j.immuni.2013.10.003
51. Miao YR, Zhang Q, Lei Q, Luo M, Xie GY, Wang H, et al. ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and Its Application in Cancer Immunotherapy. *Adv Sci (Weinh)* (2020) 7:1902880. doi: 10.1002/advs.201902880
52. Nirmal AJ, Regan T, Shih BB, Hume DA, Sims AH, Freeman TC. Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. *Cancer Immunol Res* (2018) 6:1388–400. doi: 10.1158/2326-6066.CIR-18-0342
53. Wolf DM, Lenburg ME, Yau C, Boudreau A, van 't Veer LJ. Gene Co-Expression Modules as Clinically Relevant Hallmarks of Breast Cancer Diversity. *PloS One* (2014) 9:e88309. doi: 10.1371/journal.pone.0088309
54. Hildner K EB, Purtha WE. Batf3 Deficiency Reveals a Critical Role for CD8alpha+ Dendritic Cells in Cytotoxic T Cell Immunity. *Science* (2008) 322:1097. doi: 10.1126/science.1164206
55. Fuentes MB, Kacha AK, Kline J, Woo S-R, Kranz DM, Murphy KM, et al. Host Type I IFN Signals Are Required for Antitumor CD8+ T Cell Responses Through CD8alpha+ Dendritic Cells. *J Exp Med* (2011) 208:2005–16. doi: 10.1084/jem.20101159
56. Roberts EW, Broz ML, Binnewies M, Headley MB, Nelson AE, Wolf DM, et al. Critical Role for CD103(+)/CD141(+) Dendritic Cells Bearing CCR7 for Tumor Antigen Trafficking and Priming of T Cell Immunity in Melanoma. *Cancer Cell* (2016) 30:324–36. doi: 10.1016/j.ccell.2016.06.003
57. Spranger S, Dai D, Horton B, Gajewski TF. Tumor-Residing Batf3 Dendritic Cells Are Required for Effector T Cell Trafficking and Adoptive T Cell Therapy. *Cancer Cell* (2017) 31:711–23.e4. doi: 10.1016/j.ccell.2017.04.003
58. Böttcher JP, Bonavita E, Chakravarty P, Blees H, Cabeza-Cabrerizo M, Sammicheli S, et al. NK Cells Stimulate Recruitment of Cdc1 Into the Tumor Microenvironment Promoting Cancer Immune Control. *Cell* (2018) 172:1022–37.e14. doi: 10.1016/j.cell.2018.01.004
59. Spranger S, Bao R, Gajewski TF. Melanoma-Intrinsic  $\beta$ -Catenin Signalling Prevents Anti-Tumour Immunity. *Nature* (2015) 523:231–5. doi: 10.1038/nature14404
60. Harlin H, Meng Y, Peterson AC, Zha Y, Tretiakova M, Slingluff C, et al. Chemokine Expression in Melanoma Metastases Associated With CD8+ T-Cell Recruitment. *Cancer Res* (2009) 69:3077–85. doi: 10.1158/0008-5472.CAN-08-2281
61. Cheng WC, Tsui YC, Ragusa S, Koelzer VH, Mina M, Franco F, et al. Uncoupling Protein 2 Reprograms the Tumor Microenvironment to Support the Anti-Tumor Immune Cycle. *Nat Immunol* (2019) 20:206–17. doi: 10.1038/s41590-018-0290-0
62. Ogunleye AA WQ. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans Comput Biol Bioinform* (2019) 17(6):2131–40. doi: 10.1109/TCBB.2019.2911071
63. Zhong J, Sun Y, Peng W, Xie M, Yang J, Tang X. XGBFEMF: An XGBoost-Based Framework for Essential Protein Prediction. *IEEE Trans Nanobioscience* (2018) 17:243–50. doi: 10.1109/TNB.2018.2842219
64. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer Incidence and Mortality Patterns in Europe: Estimates for 40 Countries and 25 Major Cancers in 2018. *Eur J Cancer* (2018) 103:356–87. doi: 10.1016/j.ejca.2018.07.005
65. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Pineros M, et al. Estimating the Global Cancer Incidence and Mortality in 2018:

- GLOBOCAN Sources and Methods. *Int J Cancer* (2019) 144:1941–53. doi: 10.1002/ijc.31937
66. Motzer RJ, Jonasch E, Michaelson MD, Nandagopal L, Gore JL, George S, et al. NCCN Guidelines Insights: Kidney Cancer, Version 2.2020. *J Natl Compr Cancer Network* (2019) 17:1278–85. doi: 10.6004/jnccn.2019.0054
  67. Berglund A, Amankwah EK, Kim YC, Spiess PE, Sexton WJ, Manley B, et al. Influence of Gene Expression on Survival of Clear Cell Renal Cell Carcinoma. *Cancer Med* (2020) 9:8662–75. doi: 10.1002/cam4.3475
  68. Knott ME, Minatta JN, Roulet L, Gueglio G, Pasik L, Ranuncolo SM, et al. Circulating Fibroblast Growth Factor 21 (Fgf21) as Diagnostic and Prognostic Biomarker in Renal Cancer. *J Mol biomark Diagn* (2016) 1(Suppl 2):015. doi: 10.1158/1538-7445.AM2016-431
  69. Lin J YM, Xu X, Wang Y, Xing H, An J, Yang J, et al. Identification of Biomarkers Related to CD8 T Cell Infiltration With Gene Co-Expression Network in Clear Cell Renal Cell Carcinoma. *Aging (Albany NY)* (2020) 12(4):3694–712. doi: 10.18632/aging.102841
  70. Markic D, Celic T, Grskovic A, Spanjol J, Fuckar Z, Grahovac B, et al. mRNA Expression of Bone Morphogenetic Proteins and Their Receptors in Human Renal Cell Carcinoma. *Urol Int* (2011) 87:353–8. doi: 10.1159/000330797
  71. Pan D, Xu L, Liu H, Zhang W, Liu W, Liu Y, et al. High Expression of Interleukin-11 Is an Independent Indicator of Poor Prognosis in Clear-Cell Renal Cell Carcinoma. *Cancer Sci* (2015) 106:592–7. doi: 10.1111/cas.12638
  72. Xu F, Guan Y, Zhang P, Xue L, Yang X, Gao K, et al. The Impact of TNFSF14 on Prognosis and Immune Microenvironment in Clear Cell Renal Cell Carcinoma. *Genes Genomics* (2020) 42:1055–66. doi: 10.1007/s13258-020-00974-0
  73. Paret C, Schön Z, Szponar A, Kovacs G. Inflammatory Protein Serum Amyloid A1 Marks a Subset of Conventional Renal Cell Carcinomas With Fatal Outcome. *Eur Urol* (2010) 57(5):859–66. doi: 10.1016/j.eururo.2009.08.014
  74. Cui XF, Cui XG, Leng N. Overexpression of Interleukin-20 Receptor Subunit Beta (IL20RB) Correlates With Cell Proliferation, Invasion and Migration Enhancement and Poor Prognosis in Papillary Renal Cell Carcinoma. *J Toxicol Pathol* (2019) 32:245–51. doi: 10.1293/tox.2019-0017
  75. Kang MH, Choi H, Oshima M, Cheong JH, Kim S, Lee JH, et al. Estrogen-Related Receptor Gamma Functions as a Tumor Suppressor in Gastric Cancer. *Nat Commun* (2018) 9:1920. doi: 10.1038/s41467-018-04244-2
  76. Venkatesan AM, Vyas R, Gramann AK, Dresser K, Gujja S, Bhatnagar S, et al. Ligand-Activated BMP Signaling Inhibits Cell Differentiation and Death to Promote Melanoma. *J Clin Invest* (2017) 128:294–308. doi: 10.1172/JCI92513
  77. Raulet DH. Roles of the NKG2D Immunoreceptor and Its Ligands. *Nat Rev Immunol* (2003) 3:781–90. doi: 10.1038/nri1199
  78. Tewary P, Yang D, de la Rosa G, Li Y, Finn MW, Krensky AM, et al. Granulysin Activates Antigen-Presenting Cells Through TLR4 and Acts as an Immune Alarmin. *Blood* (2010) 116:3465–74. doi: 10.1182/blood-2010-03-273953
  79. Chambers CAK, Michael S, Egen, Jackson G, Allison, James P. CTLA-4-Mediated Inhibition in Regulation of T Cell Responses Mechanisms and Manipulation in Tumor Immunotherapy. *Annu Rev Immunol* (2001) 19(1):565–94. doi: 10.1146/annurev.immunol.19.1.565
  80. Tokunaga R, Zhang W, Naseem M, Puccini A, Berger MD, Soni S, et al. CXCL9, CXCL10, CXCL11/CXCR3 Axis for Immune Activation – A Target for Novel Cancer Therapy. *Cancer Treat Rev* (2018) 63:40–7. doi: 10.1016/j.ctrv.2017.11.007
  81. Fuchs CS, Doi T, Jang RW, Muro K, Satoh T, Machado M, et al. Safety and Efficacy of Pembrolizumab Monotherapy in Patients With Previously Treated Advanced Gastric and Gastroesophageal Junction Cancer: Phase 2 Clinical KEYNOTE-059 Trial. *JAMA Oncol* (2018) 4:e180013. doi: 10.1001/jamaoncol.2018.0013
  82. Panda A, Mehnert JM, Hirshfield KM, Riedlinger G, Damare S, Saunders T, et al. Immune Activation and Benefit From Avelumab in EBV-Positive Gastric Cancer. *J Natl Cancer Inst* (2018) 110:316–20. doi: 10.1093/jnci/djx213
  83. Roh W, Chen PL, Reuben A, Spencer CN, Prieto PA, Miller JP, et al. Integrated Molecular Analysis of Tumor Biopsies on Sequential CTLA-4 and PD-1 Blockade Reveals Markers of Response and Resistance. *Sci Transl Med* (2017) 9(379):eaah3560. doi: 10.1126/scitranslmed.aah3560
  84. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* (2017) 16:2598–608. doi: 10.1158/1535-7163.MCT-17-0386
  85. Korbakis D, Soosaipillai A, Diamandis EP. Study of Kallikrein-Related Peptidase 6 (KLK6) and Its Complex With Alpha1-Antitrypsin in Biological Fluids. *Clin Chem Lab Med* (2017) 55:1385–96. doi: 10.1515/cclm-2017-0017

**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.

Copyright © 2021 Feng, Zhao, Wei, Guo, Yang, Li, Fang, Wei, Li, Jiang and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Mathematical Modeling of Locoregional Recurrence Caused by Premalignant Lesions Formed Before Initial Treatment

Mitsuaki Takaki and Hiroshi Haeno \*

Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

## OPEN ACCESS

### Edited by:

Maria Rodriguez Martinez,  
IBM Research–Zurich, Switzerland

### Reviewed by:

Jeffrey West,  
Moffitt Cancer Center, United States  
Kevin Leder,  
University of Minnesota Twin Cities,  
United States

### \*Correspondence:

Hiroshi Haeno,  
haeno@edu.k.u-tokyo.ac.jp

### Specialty section:

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

**Received:** 18 July 2021

**Accepted:** 20 September 2021

**Published:** 13 October 2021

### Citation:

Takaki M and Haeno H (2021)  
Mathematical Modeling of  
Locoregional Recurrence Caused  
by Premalignant Lesions Formed  
Before Initial Treatment.  
Front. Oncol. 11:743328.  
doi: 10.3389/fonc.2021.743328

Locoregional recurrence after surgery is a major unresolved issue in cancer treatment. Premalignant lesions are considered a cause of cancer recurrence. A study showed that premalignant lesions surrounding the primary tumor drove a high local cancer recurrence rate after surgery in head and neck cancer. Based on the multistage theory of carcinogenesis, cells harboring an intermediate number of mutations are not cancer cells yet but have a higher risk of becoming cancer than normal cells. This study constructed a mathematical model for cancer initiation and recurrence by combining the Moran and branching processes in which cells require two specific mutations to become malignant. There are three populations in this model: (i) normal cells with no mutation, (ii) premalignant cells with one mutation, and (iii) cancer cells with two mutations. The total number of healthy tissue is kept constant to represent homeostasis, and there is a rare chance of mutation every time a cell divides. If a cancer cell with two mutations arises, the cancer population proliferates, violating the homeostatic balance of the tissue. Once the number of cancer cells reaches a certain size, we conduct computational resection and remove the cancer cell population, keeping the ratio of normal and premalignant cells in the tissue unchanged. After surgery, we considered tissue dynamics and eventually observed the second appearance of cancer cells as recurrence. Consequently, we computationally revealed the conditions where the time to recurrence became short by parameter sensitivity analysis. Particularly, when the premalignant cells' fitness is higher than normal cells, the proportion of premalignant cells becomes large after the surgical resection. Moreover, the mathematical model was fitted to clinical data on disease-free survival of 1,087 patients in 23 cancer types from the TCGA database. Finally, parameter values of tissue dynamics are estimated for each cancer type, where the likelihood of recurrence can be elucidated. Thus, our approach provides insights into the concept to identify the patients likely to experience recurrence as early as possible.

**Keywords:** mathematical modeling, tumor recurrence, premalignant lesions, stochastic processes, field cancerization



## INTRODUCTION

Locoregional recurrence after surgery appears in many cancer types. About 8% of invasive breast cancer patients exhibited local recurrence after surgical resection with free resection margins (1). In non-small-cell lung cancer, about 25% of patients showed locoregional recurrence after wedge resection (2). In colorectal cancer, over 4% of patients developed locoregional recurrence after surgery (3). To prevent the emergence of recurrent tumors, treatment strategies, such as adjuvant chemotherapy has been examined and improved (4). However, tumor recurrence remains a problem.

A major cause of local recurrence is field cancerization (5–7). Field cancerization was initially defined as the presence of histologically abnormal tissue surrounding primary cancer, but currently, the concept includes the spread of histologically normal but genetically altered cells (5, 8). These cells are prone to be hotbeds for recurrent tumors because they have already accumulated specific cancer-related mutations, and a small number of additional ones is necessary to trigger cancer initiation there. Molecular evidence of field cancerization has been investigated in each tissue (6, 8–10). For example, in breast cancer, microsatellite markers, epigenetic aberrations, and hTERT overexpression have been detected in histologically normal mammary tissues (8). In head and neck cancer, loss of heterozygosity of chromosome 9p was commonly observed in benign squamous hyperplasia (9). In colon cancer patients with Crohn's ileocolitis, the same mutations of *KRAS*, *CDKN2A*, and *TP53* were observed within neoplasia and non-tumor epithelium (10). Interestingly, locoregional recurrence rates and field cancerization molecular mechanism vary among cancer types. Therefore, understanding field cancerization formation process will contribute to the estimation of the risk of locoregional recurrence and the development of optimal treatment in each tissue.

Theoretical studies have investigated field cancerization impacts on the emergence of recurrent tumors (11–15). Jeon et al. examined the multistage clonal expansion model by employing the Poisson process to consider the effects of premalignant cells on cancer initiation (11). The model was applied to the clinical practice of neoplasia in Barrett's esophagus. In this study, they succeeded in demonstrating the clinical utility of the model by predicting the long-term impact of ablative treatments on reducing esophageal adenocarcinoma incidence (13). Foo et al. developed a spatial evolutionary framework to study the cancer field effect. They analytically showed the size distribution of histologically undetectable premalignant fields during diagnosis (12). The model was applied to the head and neck cancer and revealed that the patient's age was a critical predictor of the size and multiplicity of precancerous lesions (14). Although theoretical studies have shed light on field cancerization effects on the emergence of primary and recurrent cancers, the relationship between tissue kinetic parameters and the incidence of recurrent cancers is unclear.

This study developed a novel mathematical model of recurrent tumor evolution. We employed a stochastic process of a multistage model to represent the accumulation of mutations in a tissue, leading to cancer relapse after surgical resection of the first tumor. Particularly, we focused on the

relationship between the tissue compositions at the time of surgery and the time until the emergence of recurrent tumors. Our approach provided insights on how to predict the time of recurrence from the tissue dynamics at the time of surgery and how to intervene patients to prevent the recurrence.

## MATERIAL AND METHODS

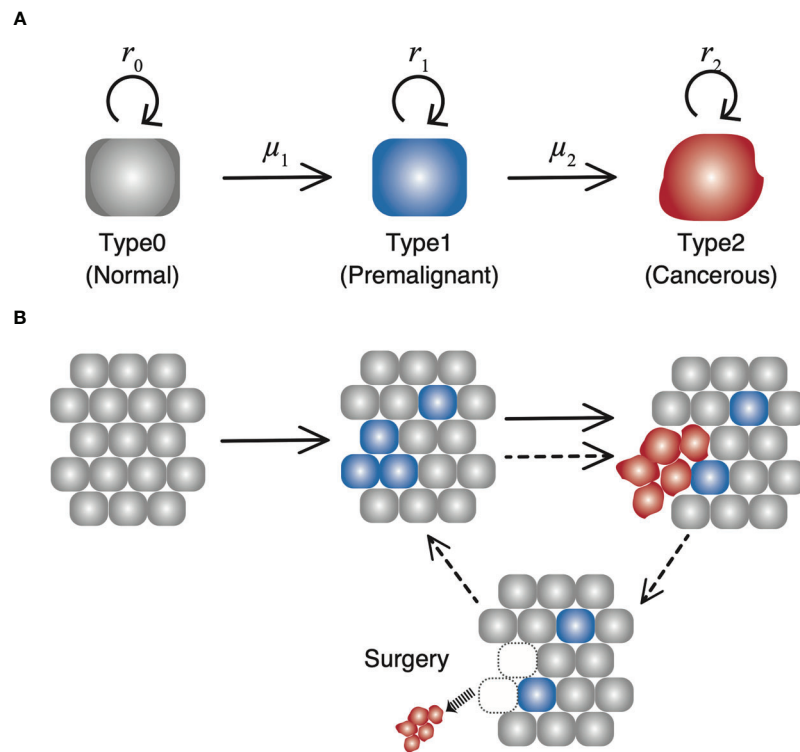
### Mathematical Model

Let us consider the dynamics of three types of cells in a tissue (**Figure 1**). “Type0,” “Type1,” and “Type2” represent normal healthy cells with no mutation, premalignant cells with one cancer-related mutation, and cancer cells with two cancer-related mutations, respectively. We assume that a normal healthy tissue consists of Type0 and Type1 cells performing a turnover of cells with a small probability of a mutation. Moran process is employed to consider the tissue turnover dynamics, where the total number of Type0 and Type1 cells is kept constant as  $N$  (16). The average turnover time of a whole tissue is defined by  $\delta$  days. Type2 cells are considered as uncontrolled cancer cells proliferating. The branching process is employed to consider the process of Type2 proliferation (17).

Initially,  $N$  Type0 cells occupy the tissue. There is a rare chance of a mutation every time a cell divides, and a daughter cell may change into a Type1 cell with a mutation rate,  $\mu_1$ . A cell to be divided in a tissue is selected depending on the fitness of Type0 cells ( $r_0$ ) and that of Type1 cells ( $r_1$ ) weighted by the proportion of Type0 and Type1 cells in a tissue. When a Type1 cell divides, a daughter cell may change into a Type2 cell with a mutation rate,  $\mu_2$ . Once a Type2 cell appears, the cells proliferate indefinitely based on the growth rate of Type2 cells,  $r_2$ , ignoring a number restrictions of a tissue unless they go extinct stochastically. In other words, the net growth of Type0 and Type1 cells is zero (equal frequency of cell division and death), while that of Type2 cells is positive. Type0 and Type1 cells consist of a healthy tissue based on the Moran process, so  $r_0$  and  $r_1$  are just parameters to determine which to choose as a dividing cell at the time of a cell turnover. Alternatively,  $r_2$  is the growth rate, which determines the average number of increases in Type2 cells during a unit time. When the number of Type2 cells reaches  $10^9$  at the first time, all the Type2 cells are discarded to represent surgical resection, whereas the number of Type1 cells in a tissue is preserved so that the time until the emergence of the recurrent tumor is influenced by the frequency of residual Type1 cells. Since the conversion from the number of cells to the tumor volume is frequently done using the following relationship as  $10^9$  cells in a  $1\text{ cm}^3$  tumor, the time of surgery in this model is conducted when the size of the tumor becomes  $1\text{ cm}^3$ . After the first treatment, the simulation continues until the next Type2 cell appears from the tissue and number reach  $10^9$  again, representing the recurrence of the tumor after surgery.

### Simulation Framework

To integrate the Moran process and branching process, we adopted stochastic simulations based on Gillespie's algorithm (18) as follows: We firstly considered three events: (i) cell



**FIGURE 1** | The schematic diagram of our model. **(A)** There are three types of cells with each own mutation rate and fitness in the model. **(B)** In a normal tissue, composed of Type0 and Type1 cells, cell turnover is conducted according to the Moran process, and the number of cells is kept constant. If a Type2 cell emerges, it proliferates unlimitedly over the tissue, and grows up to  $10^9$ . Once the number reaches  $10^9$ , all the Type2 cells are resected while the number of Type1 cells in a tissue are preserved. Then the time until the next Type2 population reaches  $10^9$  is measured as time of recurrence.

turnover in a healthy tissue, (ii) death of a Type2 cell, and (iii) birth of a Type2 cell. The rates of each event at time  $t$  is given by (i)  $\frac{1}{8}N$  (ii)  $d_2X_2(t)$ , and (iii)  $r_2X_2(t)$ , respectively. Here  $d_2$ ,  $r_2$ , and  $X_2(t)$  were a death rate, a proliferation rate, and the number of Type2 cells, respectively. Then an average time until one of the three events happens,  $\Delta T$ , is given by

$$\Delta T = \frac{1}{\frac{1}{8}N + d_2X_2(t) + r_2X_2(t)} \quad (1)$$

When the event of cell turnover in a healthy tissue occurs, one of  $N$  cells is selected as a cell to die, and another cell divides within the time step to complete cell turnover. In detail, there are three possibilities of state transitions in the tissue dynamics: the number of Type1 cells (i) increases by one, (ii) decreases by one, and (iii) does not change. Let us denote the number of Type1 cells by  $i$ .

First of all, the case (i) occurs through two ways: (a) A Type0 cell dies, and a Type1 cell divides without a mutation; and (b) a Type0 cell dies, and another Type0 cell divides with a mutation to be a Type1 cell. Exceptionally, when a Type0 dies, and a Type1 cell divides with a mutation to be a Type2 cell, an additional selection of a cell to divide is done because a Type2 cell cannot reside in a normal tissue under the assumption of the model. In this situation, if a Type1 cell is selected to divide

without a mutation, the number of Type1 cells increases by one. The probabilities of these three events are given by  $\frac{N-i}{N} \cdot \frac{r_1 i (1-\mu_2)}{F}$ ,  $\frac{N-i}{N} \cdot \frac{r_0 (N-i) \mu_1}{F}$ , and  $\frac{N-i}{N} \cdot \frac{r_1 i \mu_2 c_1}{F}$  respectively. Here  $F = r_0 (N-i) + r_1 i$  is a scaling factor for the probability to be chosen for a dividing cell and  $c_1 = \frac{r_1 i}{r_0 (N-i) + r_1 i}$  is the probability that a Type1 cell is selected to divide in an additional round after a mutation of a Type1 cell to be a Type2 cell. The probability that a Type0 cell is selected to die is given by  $\frac{N-i}{N}$ . Taken together, the transition probability that the number of Type1 cells increases by one is given by

$$\Pr[i \rightarrow i+1] = \frac{r_0 (N-i) \mu_1 + r_1 i (1-\mu_2 + \mu_2 c_1)}{F} \cdot \frac{N-i}{N} \quad (2)$$

Secondly, the case (ii) occurs in such a way that a Type1 cell dies and a Type0 cell divides without a mutation. Exceptionally, when a Type1 cell dies, and another Type1 cell divides with a mutation to be a Type2 cell, an additional selection for a cell division is done. In this case, if a Type0 cell is selected for the additional cell division, the number of Type1 cells decreases by one. The probabilities of the two events are given by  $\frac{1}{N} \cdot \frac{r_0 (N-i) (1-\mu_1)}{F}$  and  $\frac{i}{N} \cdot \frac{r_1 i \mu_2 c_0}{F}$  and  $c_0 = \frac{r_0 (N-i)}{r_0 (N-i) + r_1 i}$  is the probability that a Type0 cell is selected to divide in an additional round after a mutation of a Type1 cell to be a Type2 cell. The probability that a Type1 cell is selected to die is given by  $\frac{i}{N}$ . Taken together, the

transition probability that the number of Type1 cells decrease by one is given by

$$\Pr[i \rightarrow i-1] = \frac{r_0(N-i)(1-\mu_1) + r_1 i \mu_2 c_0}{F} \cdot \frac{i}{N} \quad (3)$$

Finally, the probability that the number of Type1 does not change [case (iii)] is given by

$$\Pr[i \rightarrow i] = 1 - \Pr[i \rightarrow i+1] - \Pr[i \rightarrow i-1] \quad (4)$$

In summary, the time of one step in simulations is calculated using Eq. (1), and in one step, one of the following three processes occurs: (i) cell turnover in a tissue, (ii) the death of a Type2 cell, or (iii) the birth of a Type2 cell. When case (i) happens, there are three possibilities in tissue dynamics. The number of type1 cells increases by one, decreases by one, or does not change. Initially, all the cells are Type0. Once the number of Type2 cells reaches  $10^9$ , computational surgical resection to set the number of Type2 cells to be 0 again will be conducted. After that, the time until the number of Type2 cells reaches  $10^9$  again is measured as recurrence time.

## Deterministic Approximation of Type2 Growth

As for the calculation of the Type2 growth, we assumed that when the number of cells is small, the stochastic effect should be considered. When the number of Type2 cells exceed twice as large as the size of the normal tissue,  $2N$ , growth can be regarded as a deterministic process. Then the time duration from when the number of Type2 cells is  $2N$  to  $10^9$ ,  $\Delta t_s$ , is given by

$$\Delta t_s = (r_2 - d_2) \ln \left( \frac{10^9}{2N} \right) \quad (5)$$

During  $\Delta t_s$ , tissue dynamics to reflect the cell turnover is conducted.

## Clinical Data

The data used in our analysis were downloaded from TCGA Pan-Cancer Clinical Data Resource provided in the previous

publication (19). We adopted the data of disease-free intervals from 23 cancer types. Data processing was performed on Excel.

## Survival Time Analysis

Disease-free survival of clinical data were calculated using the Kaplan–Meier method from disease-free intervals mentioned in *Clinical Data* section. In this study, disease-free interval is defined as the survival time without cancer recurrence of each patient, which corresponds to the time to recurrence of each simulation trial. Disease-free survivals *in silico* were then calculated from that.

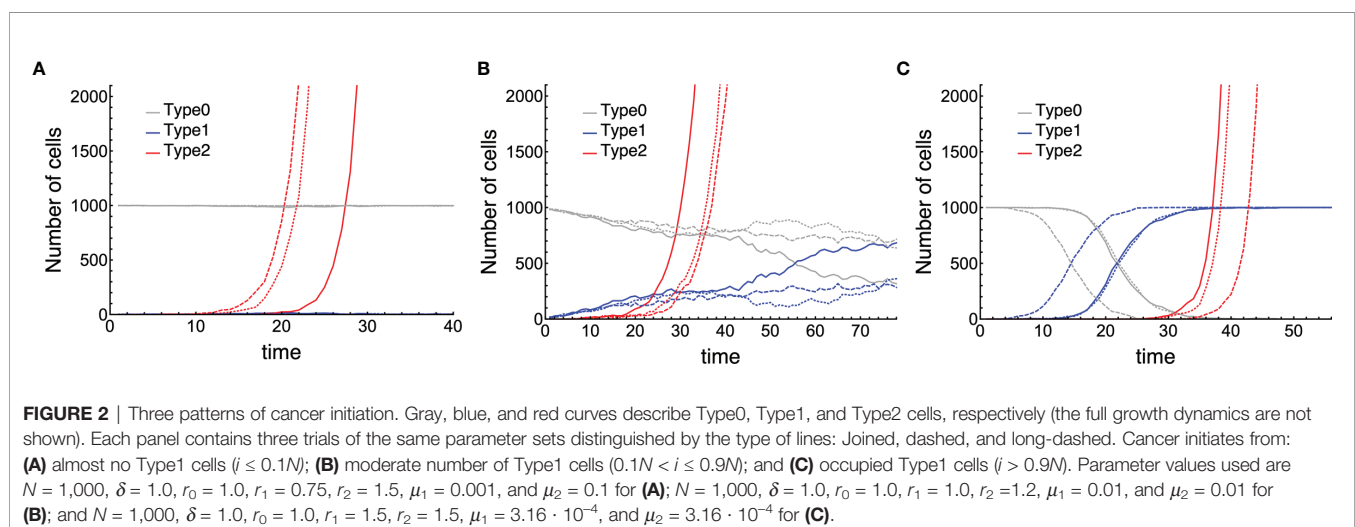
## Simulation and Statistical Analysis

The whole process of our model was conducted on C++. Parameter optimization was conducted using the Nelder–Mead method on R (version 3.6.2). The survival time analysis was conducted on Prism (version 8.4.3).

## RESULTS

### Three Patterns of Cancer Initiation

First of all, we conducted stochastic simulations of the model for the initial cancer progression, and the time courses of three populations: Type0, Type1, and Type2 were shown (**Figure 2**). We classified the tissue dynamics until the emergence of Type2 cells into three patterns. When Type1 cells had less fitness than Type0 cells, sporadic cancer initiation from a tissue dominated by Type0 cells could be observed (**Figure 2A**). In this case, Type1 cells could not spread in a normal tissue, and cancer initiation depended on two sequential mutations in one Type1 cell. After surgical resection of the first Type2 lineage, the time to recurrence would be almost the same as that of the first cancer initiation because the frequency of Type1 cells in a tissue was almost the same as the initial condition. When the fitness of Type1 cells was as high as that of Type0 cells, cancer initiation in a moderate frequency of Type1 cells could be observed (**Figure 2B**). In this case, the time to recurrence could be faster



than that of the first cancer initiation because the proportion of Type1 cells in a tissue was larger than that in the initial condition. When Type1 cells had much higher fitness than Type0 cells, multiple cancer initiations from a Type2-dominated tissue could be observed (**Figure 2C**). In this case, the recurrence of tumors happened easily. From these results, we found that different situations of Type1 cells at the time of cancer initiation were considered to influence the difficulty of recurrence, and they could be classified by parameter regions.

## Parameter Dependency

Next, we examined the time to recurrence after surgical resection and the proportion of premalignant (Type1) lesions at the time of surgery in a vast parameter range (**Figure 3**). The mean recurrence time became shorter as the fitness of Type1 cells increased because higher fitness enabled Type1 cells to dominate the normal tissue, which facilitated the emergence of recurrent cancer (Type2) (**Figures 3A, B**). When the size of the normal tissue is small, the effect of fitness advantage on the proportion of Type1 cells in a tissue became large (**Figures 3A, B**). **Figures 3C, D** showed that recurrence time became shorter when the growth rate of Type2 cells was large. Compared to the case where the fitness of Type1 was large, the early recurrence occurred from the small proportion of Type1 cells in a tissue (**Figures 3C, D**). High mutation rates accelerated the time of recurrence (**Figures 3E–H**). A higher mutation rate from Type0 to Type1 made the proportion of Type1 cells larger (**Figures 3E, F**), while a higher mutation rate from Type1 to Type2 made proportion smaller (**Figures 3G, H**). Furthermore, when the size of normal tissues became large, the time to recurrence became short, and the variation became small (**Figures 3B, D, F, H**).

## Relationship Between the Proportion of Type1 Cells and Time to Recurrence

To investigate the relationship between the proportion of Type1 cells during initial treatment and time to recurrence comprehensively, we conducted computational simulations with parameter sets randomly picked (**Figure 4A**). Additionally, we did 1,000 runs of stochastic simulations with the same parameter set to obtain each point. A total of 1,200 parameter combinations were examined.

We confirmed that recurrence time was significantly different among the proportion of Type1 cells during the first treatment (**Figure 4B**). It would be intuitive that the time to recurrence became long when the proportion of Type1 cells was very small (between 0 and 0.2 of a tissue). Interestingly, the proportion of Type1 cells that minimize recurrence time was not the largest group (between 0.8 and 1.0 of a tissue), but the moderate group (**Figure 4B**). This result showed that patients with a moderate number of premalignant cells (Type1) have a risk of shorter recurrence time in many cases. When we investigated the characteristics of parameter values in each category (**Figures 4C–F**), we found that the fitness of Type1 cells was lower, and their mutation rate was higher in areas a and b than those in areas e and f (**Figures 4C, F**). These results suggested that Type1 cells could occupy the normal tissue before the first

treatment when Type1 cells could spread rapidly and were hardly mutated to be Type2 cells. The mutation rate of Type0 cells did not affect the proportion of Type1 cells at the first treatment (**Figure 4E**). Points with time to recurrence more than  $10^3$  only resided in area b, indicating that there was no parameter set that could realize both conditions of a large proportion of Type1 cells at the time of first treatment and a long recurrence time (**Figure 4A**). In area a, time to recurrence was short despite small premalignant cells (Type1). In that case, the fitness of Type1 cells was almost neutral, and the mutation rate of Type1 cells and the growth rate of Type2 cells were relatively high (**Figures 4C, D, F**). In area f, recurrence was relatively long, although the normal tissue was occupied by premalignant cells (Type1). In that case, the growth rate of Type 2 cells was extremely small (**Figure 4D**). Mutation rates of areas d, e, and f were almost the same, and their difference was generated by the fitness of Type1 cells and the growth rate of Type2 cells (**Figures 4C, D**).

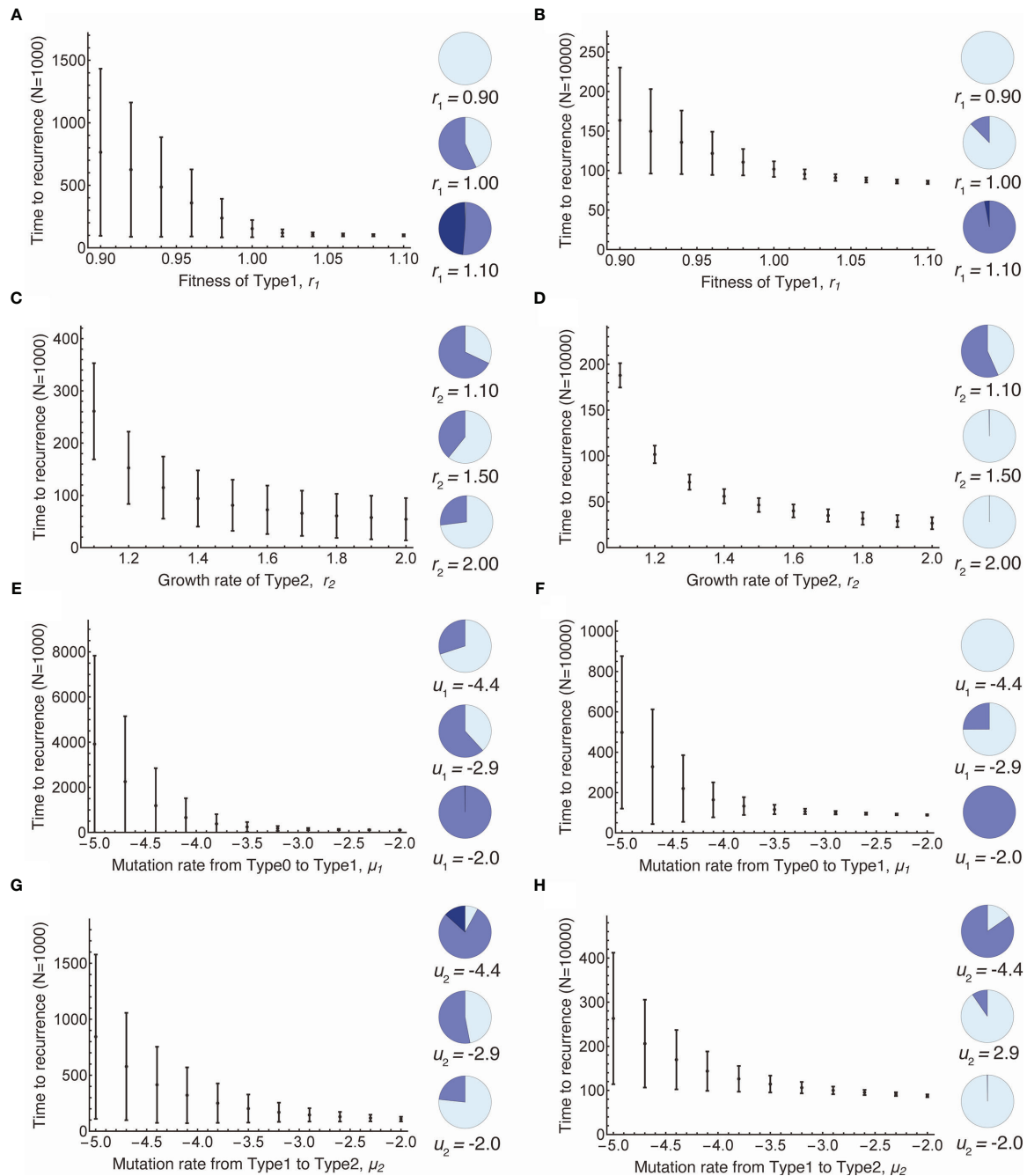
## Fitting to Clinical Data of Time to Recurrence

Results of recurrence time *in silico* were fitted to published clinical data of disease-free survivals in 23 cancer types (**Figure 5** and **Table 1**) (19). A thousand runs of stochastic simulations with a single parameter combination for each cancer type were conducted. The sum of squared logarithmic residuals (log-SSR) between outputs *in silico* and five data points extracted from clinical data was calculated. A set of the five data points was when 20, 40, 60, 80, and 100% of patients experienced a recurrence. We then investigated the parameter sets that could minimize log-SSR for each cancer type (**Table 1**), and depicted the survival curves with the estimated parameters (**Figure 5**). We also conducted a log-rank test between the curves of clinical and simulated data (**Table 1**). In most clinical data, we could find the optimal parameter sets, and with these parameters, significant differences were not observed between simulation results and clinical outcomes. However, in some cancer types (BRCA, CHOL, LUAD, OV, SARC, and THCA), significant deviations were observed ( $p < 0.05$ ). Notably, the fitness of Type1 cells was lower than that of Type0 cells, 1.0, among most cancer types, indicating a cancer-related mutation tends to be disadvantageous before the emergence of cancer cells (Type2). Mutation rates were distributed around  $10^{-3.6}$  for almost all cancer types. Alternatively, the growth rates of Type2 were widely distributed.

## DISCUSSION

In this study, we constructed a mathematical model that could describe cell population dynamics in both normal tissue and cancer tissues. We revealed the relationship between the proportion of premalignant cells and recurrence time (**Figures 3** and **4**). Importantly, we found that recurrence time became shorter when the mutation rate or growth rate of cancer cells was large, while the time became longer when the fitness of premalignant cells or growth rate of cancer cells was





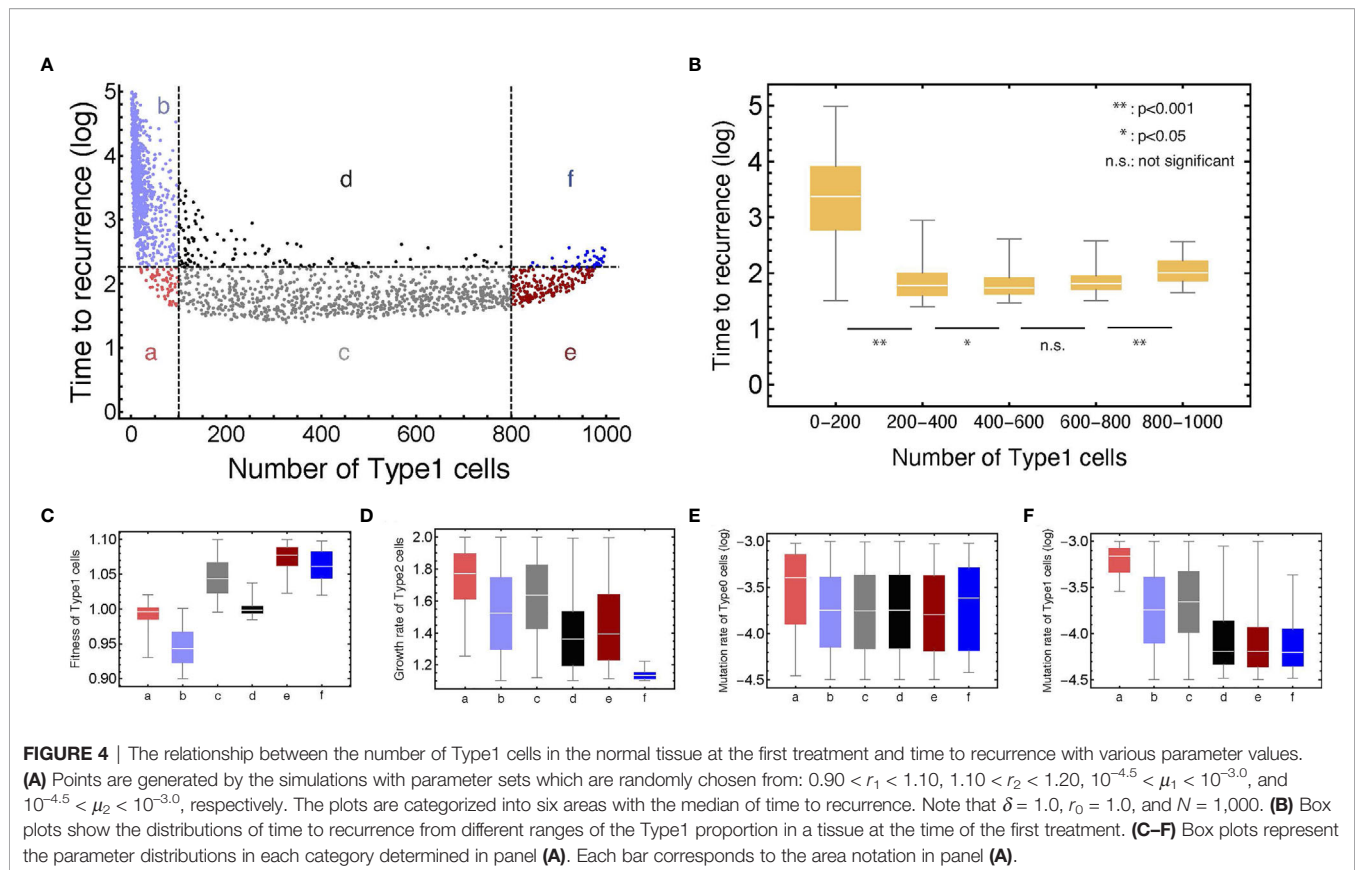
**FIGURE 3 |** Parameter dependence on recurrence time. Mean values obtained from the simulations are shown by dots, and standard deviations are indicated by bars. Pie charts in the panels indicate the proportion of Type1 cells in normal tissue at the first treatment. Light blue, blue, dark blue represent small ( $i \leq 0.1N$ ), intermediate ( $0.1N < i \leq 0.9N$ ), and large ( $i > 0.9N$ ) proportion of Type1 cells, respectively. Standard parameter values used in (A–H) are  $\delta = 1.0$ ,  $r_0 = 1.0$ ,  $r_1 = 1.0$ ,  $r_2 = 1.2$ ,  $\mu_1 = 0.001$ ,  $\mu_2 = 0.001$ ; and  $N = 1,000$  in (A, C, E, G); and  $N = 10,000$  in (B, D, F, H).

low (Figure 4). Moreover, we successfully estimated the characteristic parameter sets of the computational model by fitting the model results to the clinical data of disease-free survival in each cancer type (Figure 5 and Table 1). This study is the first attempt to quantitatively predict recurrence time after the first treatment in various cancer types with a

mathematical model by considering the effect of premalignant cells in a healthy tissue.

This model successfully reproduced the disease-free survivals in 17 out of 23 cancer types (Figure 5 and Table 1). Notably, the estimated fitness values of premalignant cells ( $r_1$ ) were less than those of normal cells in many cancer types (Table 1). According

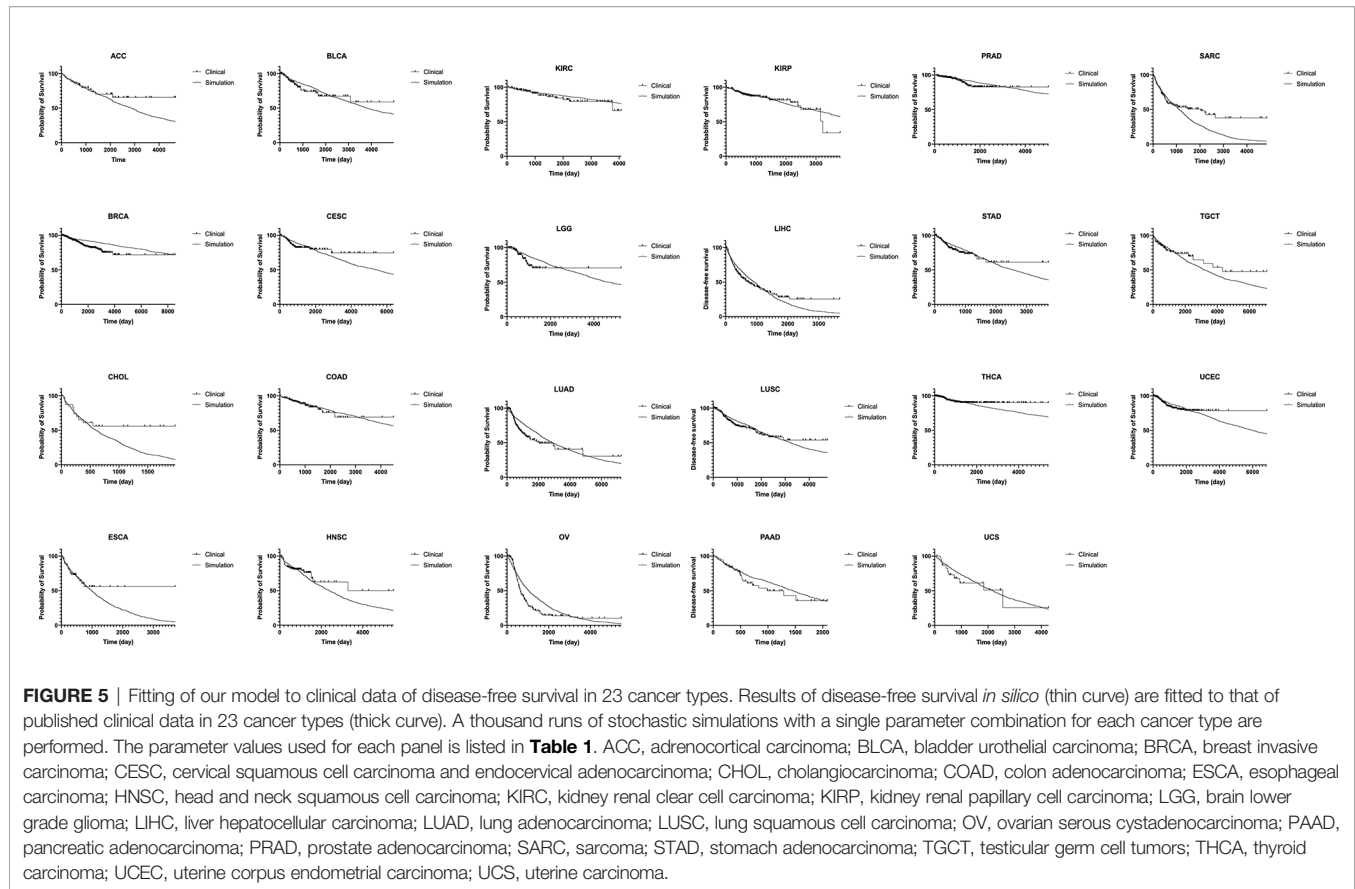




to the analysis on how the proportion of premalignant cells depended on their fitness (**Figure 4C**), the characteristics of those cancers residing in area b in **Figure 4A** suggest the small abundance of premalignant cells during the first treatment. Therefore, the efforts to find and eradicate the residual premalignant lesions in a normal tissue after the first treatment may be inefficient; rather, the suppression of the emergence of new premalignant cells from the normal cells by adjuvant therapy should be recommended. In most cancer types, the fitting tends to work for the early reduction of the disease-free survivals and not for the long tail of the survivals (**Figure 5**). Because the estimated parameters of the low fitness of premalignant cells ( $r_1$ ) indicate that recurrence arises from the almost non-mutated tissue, it implies that the deviance recurrence time in the same cancer type is caused by variations of mutation rates or efficiency of adjuvant therapy among patients, not incorporated into the model. It suggests the importance of identifying a biomarker to classify recurrence-prone patients (20).

For the model's simplicity, we prepared only one population for intermediate cell type as premalignant cells. However, the multistage theory suggested more than two steps to generate a cancer cell from a normal cell (21). This restriction resulted in the simple tendency of the survival curves from the model and failure to fit the long tail of clinical survival curves (**Figure 5**). With multiple stages of

pre-malignant cells in the model, the premalignant cells after the first treatment have several mutational distances to recurrence, which may generate multiple inclinations of the survival curves. In contrast, the number of mutations required to be a cancer cell varies in each patient, even in the same cancer type, so that it was difficult to determine it accurately for each cancer type. This simple model structure had the abovementioned weakness but still could imply that the single-intermediate population might be enough to reproduce the data of well-fitted cancer types, while more populations would be required for the others. We also adopted a spatially homogeneous process, though a spatial process can contain detailed information, such as molecular mechanisms of field cancerization and cell competition. Note that this study focused on constructing the basic mathematical model extensible for various types of cancer to quantitatively predict recurrence time after the first treatment by considering the effect of premalignant cells. Molecular mechanisms vary among cancer types, and cell competition can be regarded as dynamics based on the fitness and the number of the cells. The simple model structure enabled us to analyze the various types of cancer by uniform parameters, fitness, and mutation rate. This was the first attempt, and even at the current stage, we obtained many new insights. A spatial structure and additional intermediate populations optimized for each cancer type would be a possible future extension of the model.



**FIGURE 5 |** Fitting of our model to clinical data of disease-free survival in 23 cancer types. Results of disease-free survival *in silico* (thin curve) are fitted to that of published clinical data in 23 cancer types (thick curve). A thousand runs of stochastic simulations with a single parameter combination for each cancer type are performed. The parameter values used for each panel is listed in **Table 1**. ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; SARC, sarcoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinoma.

**TABLE 1 |** Estimated parameters and p-values by fitting the outputs from our simulations to clinical data.

Cancer type	$r_1$	$r_2$	$\text{Log}_{10}\mu$	log-SSR	p-value
ACC	0.916	1.62	-3.61	0.272	0.2008
BLCA	0.908	1.43	-3.60	0.717	0.4658
BRCA	0.922	1.52	-4.02	0.294	<0.0001
CEC	0.905	1.36	-3.63	0.815	0.8958
CHOL	0.964	1.52	-3.43	0.113	0.0272
COAD	0.924	1.40	-3.71	0.564	0.5966
ESCA	0.934	1.60	-3.42	0.128	0.3458
HNSC	0.914	1.58	-3.56	0.926	0.3446
KIRC	0.920	1.27	-3.77	0.314	0.3945
KIRP	0.908	1.38	-3.62	0.981	0.6651
LGG	0.905	1.35	-3.62	0.0312	0.0803
LIHC	0.962	1.72	-3.54	0.647	0.8949
LUAD	0.920	1.62	-3.62	1.52	0.0039
LUSC	0.904	1.43	-3.55	0.588	0.4501
OV	0.905	1.56	-3.37	0.604	<0.0001
PAAD	0.918	1.54	-3.44	0.139	0.3649
PRAD	0.913	1.34	-3.80	0.165	0.1207
SARC	0.930	1.51	-3.42	1.05	0.0036
STAD	0.917	1.59	-3.59	0.432	0.3859
TGCT	0.916	1.61	-3.63	1.68	0.4146
THCA	1.04	1.10	-3.31	7.74	<0.0001
UCEC	0.909	1.33	-3.65	0.168	0.4777
UCS	0.904	1.53	-3.49	0.633	0.6923

ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; SARC, sarcoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinoma.

Conclusively, this model suggests special care of recurrence in the clinic when the fitness of premalignant cells and the growth rate of recurrent tumors is high. Furthermore, this approach can be extended to explore the deviance of recurrence rates among cancer types by introducing the variations of mutational stages and standard adjuvant therapies in each cancer according to growing knowledge.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

## REFERENCES

- Holleczer B, Stegmaier C, Radosa JC, Solomayer EF, Brenner H. Risk of Locoregional Recurrence and Distant Metastases of Patients With Invasive Breast Cancer Up to Ten Years After Diagnosis – Results From a Registry-Based Study From Germany. *BMC Cancer* (2019) 19:520. doi: 10.1186/s12885-019-5710-5
- Maurizi G, D'Andrilli A, Ciccone AM, Ibrahim M, Andreotti C, Tierso S, et al. Margin Distance Does Not Influence Recurrence and Survival After Wedge Resection for Lung Cancer. *Ann Thorac Surg* (2015) 100:918–24. doi: 10.1016/j.athoracsurg.2015.04.064
- Pugh SA, Shinkins B, Fuller A, Mellor J, Mant D, Primrose JN. Site and Stage of Colorectal Cancer Influence the Likelihood and Distribution of Disease Recurrence and Postrecurrence Survival: Data From the FACS Randomized Controlled Trial. *Ann Surg* (2016) 263:1143–7. doi: 10.1097/SLA.0000000000001351
- Aebi S, Gelber S, Anderson SJ, Lang I, Robidoux A, Martin M, et al. Chemotherapy for Isolated Locoregional Recurrence of Breast Cancer (CALOR): A Randomised Trial. *Lancet Oncol* (2014) 15:156–63. doi: 10.1016/S1470-2045(13)70589-8
- Slaughter DP, Southwick HW, Smejkal W. Field Cancerization in Oral Stratified Squamous Epithelium: Clinical Implications of Multicentric Origin. *Cancer* (1953) 6:963–8. doi: 10.1002/1097-0142(195309)6:5<963::aid-cnrcr2820060515>3.0.co;2-q
- Tabor MP, Brakenhoff RH, van Houten M, Kummer JA, Snel MH, Snijders PJ, et al. Persistence of Genetically Altered Fields in Head and Neck Cancer Patients: Biological and Clinical Implications. *Clin Cancer Res* (2001) 7:1523–32.
- Braakhuis BJM, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH. A Genetic Explanation of Slaughter's Concept of Field Cancerization: Evidence and Clinical Implications. *Cancer Res* (2003) 63:1727–30.
- Heaphy CM, Griffith JK, Bisoffi M. Mammary Field Cancerization: Molecular Evidence and Clinical Importance. *Breast Cancer Res Treat* (2009) 118:229–39. doi: 10.1007/s10549-009-0504-0
- Califano J, Ahrendt SA, Meiningner G, Westra WH, Koch WM, Sidransky D. Detection of Telomerase Activity in Oral Rinses From Head and Neck Squamous Cell Carcinoma Patients. *Cancer Res* (1996) 56:5720–2.
- Galandiuk S, Rodriguez-Justo M, Jeffery R, Nicholson AM, Cheng Y, Oukrif D, et al. Field Cancerization in the Intestinal Epithelium of Patients With Crohn's Ileocolitis. *Gastroenterology* (2012) 142:855–64.e8. doi: 10.1053/j.gastro.2011.12.004
- Jeon J, Meza R, Moolgavkar SH, Luebeck EG. Evaluation of Screening Strategies for Pre-Malignant Lesions Using a Biomathematical Approach. *Math Biosci* (2008) 213:56–70. doi: 10.1016/j.mbs.2008.02.006

## AUTHOR CONTRIBUTIONS

HH supervised the work. MT performed theoretical analysis. MT and HH wrote manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

The work is supported by National Cancer Center Research and Development Fund (2021A-7), a research grant from SRL, H.U Group Research Institute, and JSPS KAKENHI Grant Number 20J22335. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

- Foo J, Leder K, Ryser MD. Multifocality and Recurrence Risk: A Quantitative Model of Field Cancerization. *J Theor Biol* (2014) 355:170–84. doi: 10.1016/j.jtbi.2014.02.042
- Curtius K, Hazelton WD, Jeon J, Luebeck EG. A Multiscale Model Evaluates Screening for Neoplasia in Barrett's Esophagus. *PLoS Comput Biol* (2015) 11: e1004272. doi: 10.1371/journal.pcbi.1004272
- Ryser MD, Lee WT, Ready NE, Leder KZ, Foo J. Quantifying the Dynamics of Field Cancerization in Tobacco-Related Head and Neck Cancer: A Multiscale Modeling Approach. *Cancer Res* (2016) 76:7078–88. doi: 10.1158/0008-5472.CAN-16-1054
- Curtius K, Wright NA, Graham TA. An Evolutionary Perspective on Field Cancerization. *Nat Rev Cancer* (2018) 18:19–32. doi: 10.1038/nrc.2017.102
- Moran PAP. *The Statistical Processes of Evolutionary Theory*. Oxford: Clarendon Press (1962).
- Athreya KB, Ney PE. *Branching Processes*. New York: Dover Publications (2004).
- Gillespie DT. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J Comput Phys* (1976) 22:403–34. doi: 10.1016/0021-9991(76)90041-3
- Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* (2018) 173:400–16.e11. doi: 10.1016/j.cell.2018.02.052
- Lee CW, Simin K, Liu Q, Plescia J, Guha M, Khan A, et al. A Functional Notch-Survivin Gene Signature in Basal Breast Cancer. *Breast Cancer Res* (2008) 20:R97. doi: 10.1186/bcr2200
- Armitage P, Doll R. The Age Distribution of Cancer and a Multi-Stage Theory of Carcinogenesis. *Br J Cancer* (1954) 8:1–12. doi: 10.1038/bjc.1954.1

**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.

**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.

Copyright © 2021 Takaki and Haeno. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Re-Identification of Patient Subgroups in Uveal Melanoma

## OPEN ACCESS

### Edited by:

Max A. Alekseyev,  
George Washington University,  
United States

### Reviewed by:

Rosario Caltabiano,  
University of Catania, Italy  
Valluru Manoj Kumar,  
The University of Sheffield,  
United Kingdom  
Justin Moser,  
HonorHealth, United States  
Jiang Qian,  
Fudan University, China

### \*Correspondence:

Quang-Huy Nguyen  
huynguyen96.dnu@gmail.com  
Duc-Hau Le  
hauldhut@gmail.com

<sup>†</sup>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 June 2021

**Accepted:** 29 September 2021

**Published:** 20 October 2021

### Citation:

Nguyen THY, Nguyen T,  
Nguyen Q-H and Le D-H (2021)  
Re-Identification of Patient  
Subgroups in Uveal Melanoma.  
Front. Oncol. 11:731548.  
doi: 10.3389/fonc.2021.731548

Thi Hai Yen Nguyen<sup>1†</sup>, Tin Nguyen<sup>2</sup>, Quang-Huy Nguyen<sup>1\*†</sup> and Duc-Hau Le<sup>1,3\*</sup>

<sup>1</sup> Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam, <sup>2</sup> Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, United States, <sup>3</sup> College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam

Uveal melanoma (UM) is a comparatively rare cancer but requires serious consideration since patients with developing metastatic UM survive only for about 6–12 months. Fortunately, increasingly large multi-omics databases allow us to further understand cancer initiation and development. Moreover, previous studies have observed that associations between copy number aberrations (CNA) or methylation (MET) versus messenger RNA (mRNA) expression have affected these processes. From that, we decide to explore the effect of these associations on a case study of UM. Also, the current subtypes of UM display its weak association with biological phenotypes and its lack of therapy suggestions. Therefore, the re-identification of molecular subtypes is a pressing need. In this study, we recruit three omics profiles, including CNA, MET, and mRNA, in a UM cohort from The Cancer Genome Atlas (TCGA). Firstly, we identify two sets of genes, CNAexp and METexp, whose CNA and MET significantly correlated with their corresponding mRNA, respectively. Then, single and integrative analyses of the three data types are performed using the PINSPlus tool. As a result, we discover two novel integrative subgroups, IntSub1 and IntSub2, which could be a useful alternative classification for UM patients in the future. To further explore molecular events behind each subgroup, we identify their subgroup-specific genes computationally. Accordingly, the highest expressed genes among IntSub1-specific genes are mostly enriched with immune-related processes. On the other hand, IntSub2-specific genes are highly associated with cellular cation homeostasis, which responds effectively to chemotherapy using ion channel inhibitor drugs. In addition, we detect that the two integrative subgroups show different age-related risks and survival rates. These discoveries can influence the frequency of metastatic surveillance and support medical practitioners to choose an appropriate treatment regime.

**Keywords:** uveal melanoma, clustering, multi-omics, molecular subtypes, biomolecular markers



# 1 INTRODUCTION

Uveal melanoma (UM) is a comparatively rare cancer formed from melanocytes within the uveal tract of the eye involving either in the iris, ciliary body, or mostly choroid (1) and responsible for about five cases per million per year (2). Although current first-line treatment approaches receive good results for this malignancy, specifically, UM patients can live longer, but we want to improve early diagnosis more with the hope of raising overall patient survival as smaller tumors are treated, resulting in achieving local disease control and vision preservation with the possibility to prevent metastases (3). However, it has still remained challenging. Indeed, UM patients with the metastatic disease only lived for approximately 6–12 months (4). This emphasizes a pressing need of improving the diagnosis, prevention, and treatment of UM patients.

Besides, several recent large-scale and multi-omics databases have enabled us to see associations between the genetic or epigenetic alterations versus the tumorigenesis and progression of UM. For example, the importance of different types of RNA such as mRNA, microRNA (miRNA), and long non-coding RNA (lncRNA) was investigated in UM (5, 6). Based on an *in silico* and experimental biology, lncRNA LINC00518 was identified to be a oncogene in UM and could be used in RNA-based therapeutic approaches as a promising target (6). Additionally, UM has frequently had copy number aberrations (CNA) gain regions of chromosomes 6p and 8q as well as loss regions of chromosomes 1p, 3, 6q, 8p, and 16q (7, 8). Particularly, *BAP1* mutations related to chromosome 3 monosomy and *SF3B1* and *SRSF2* alterations related to chromosome 3 disomy contributed to high risk of metastasis. Meanwhile, mutations on *EIF1AX* related to chromosome 3 disomy were associated with low metastatic risk (9). In addition, Yang et al. (4) have made a comprehensive review of the role of DNA methylation in the development and metastasis of UM. They highlighted that several tumor suppressor genes comprising *RASSF1A* and *p16INK4a* have been altered by DNA methylation (MET) and contributed to controlling cell migration and invasion in UM. Moreover, *p16INK4a* expression was reported in all UM liver metastatic cases and may have potential in discriminating UM and cutaneous melanoma (10). Besides, the autophagy has been hypothesized to have a role in inhibiting tumor growth when investigating this process-related protein, Beclin-1. The high level of immunohistochemistry in Beclin-1 was found to be a positive prognosis of UM patients (11).

Moreover, multiple prior studies have been conducted to stratify UM patients using various kinds of -omic data. Among them, the most popular work proposed by Robertson et al. (5) has conducted a multiplatform analysis of 80 UM patients using only one single data type of omics data, including mRNA expression, miRNA, long non-coding RNA, MET, and CNA, and successfully identified four different subtypes: two associated with poor-prognosis monosomy 3 (M3) and the others with better-prognosis disomy 3 (D3). However, we claim that not a single data alone but instead integrated omics data are powerful enough to explain the interplay of molecules and the biological phenotypes of cancer holistically (12–14). This motivates us to

do this study in order to discover novel subgroups of UM patients that adopt an integrative approach.

In this study, we aimed to analyze three omics profiles, namely, CNA, MET, and mRNA, in a UM cohort from The Cancer Genome Atlas (TCGA). To this purpose, we identified the significant correlation between CNA and MET versus their own corresponding expression levels (**Figure 1**). It was of importance to note that the omics experiments were conducted with thousands of simultaneous hypothesis tests (15). Therefore, the adjusted *P*-value using the Benjamini–Hochberg procedure (16) as a measure of significant tests controlling the number of false discoveries was necessarily considered in this work. Then, single and joint analyses of the three data types were performed using the tool PINSplus (17, 18). As a result, we discovered two novel integrative subgroups, IntSub1 and IntSub2, which could be potentially a future classification system for UM patients. These discoveries could influence the frequency of metastatic surveillance and support medical practitioners to choose an appropriate treatment regime.

# 2 MATERIALS AND METHODS

## 2.1 Materials

The three datasets, namely, CNA, MET, and mRNA expression, were collected from the TCGA project (TCGA, Firehose Legacy) (5) and downloaded from the cBioPortal website (19, 20). The UM cohort is described in **Table 1**.

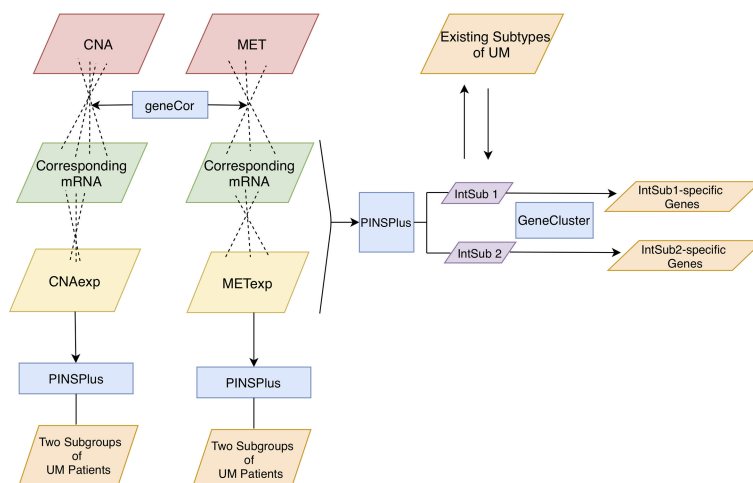
## 2.2 Data Preprocessing

There were two preprocessing steps applied to the three profiles (i.e., mRNA, CNA, and MET) from the data. We first checked if the 80 patients from each of the three profiles and clinical data were matched. Then, we detected genes whose missing values were more than 50% using the *k*-nearest neighbor algorithm (21) from the CancerSubtypes package (version 1.14.0) (22).

## 2.3 Identification and Examination of the Relationship of CNAexp and METexp Genes

Here, we kept only genes shared between CNA and mRNA, as well as between MET and mRNA. To identify and examine the relationship of CNAexp and METexp genes, we used the R tool geneCor (14). Roughly, the tool first computed the correlation coefficients (*r*) between MET and mRNA, as well as between CNA and mRNA based on Spearman's rank method, and then, the conversion of significant *r* (i.e., adjusted *P*-value  $\leq 0.05$ ; Benjamini–Hochberg (16); two-sided) into *Z* values by Fisher's *Z*-transformation following the equation:  $Z = 0.5 \ln[(1 + r)/(1 - r)]$ . Secondly, the overall distributions of calculating *Z* values were pictured automatically. Thirdly, geneCor computed the skewness of the *Z*-score distributions using the D'Agostino test. The overall skewness illustrated whether CNA or MET was correlated positively or negatively with their own corresponding mRNA. Parallely, geneCor also issued two sets of genes, CNAexp and METexp, whose CNA and MET significantly correlated with their





**FIGURE 1 |** Analysis pipeline. Firstly, we inputted CNA and MET datasets with their corresponding mRNA data to the function geneCor to identify a list of CNAexp and METexp genes, respectively. Then, PINSPPlus was used to extract different patient subgroups for individual CNAexp and METexp datasets and integration of CNAexp + METexp + mRNA data through single and integrated analyses, respectively. Finally, we discovered subtype-specific genes within each identified integrated subgroup, IntSub1 and IntSub2, using the R package GeneCluster. UM, uveal melanoma.

**TABLE 1 |** Description of a cohort of UM patients used in the study.

Omics data	Platform	Description
mRNA	mRNA sequencing	A continuous matrix whose columns (the number of samples) are 80 samples and rows (the number of genes) are 20,440 genes
CNA	Affymetrix SNP6 Whole-exome sequencing	A discrete matrix whose columns (the number of samples) are 80 samples and rows (the number of genes) are 24,776 genes. There are four copy-number levels indicated for each gene, namely, -2, -1, 1, and 2. Two levels presented with minus value (i.e., -2, -1) show the loss level of copy-number compared with the two positive values (i.e., 1, 2) expressing the additional copies degree. For the 0 level, the gene is located in the diploid chromosomal region.
MET	Illumina Infinium HumanMethylation 450 platform	A continuous matrix whose columns (the number of samples) are 80 samples and rows (the number of genes) are 15,477 genes
Clinical data		Samples: 80 Overall survival (OS) status was defined as vital status (dead or alive), whereas OS time was identified as the time to UM death or last follow-up (unit: day). The follow-up time OS was truncated to 2,600 days.

corresponding mRNA expression levels, respectively. Further analysis was performed using FSbyCOX in the package CancerSubtypes (version 1.14.0) (22) to only retain a small number of genes associated significantly with a prognostic value ( $P$ -value  $\leq 0.0005$ ; log-rank test; two-tailed) in the two gene sets (i.e., CNAexp and METexp).

## 2.4 Single and Integrated Subtyping

The related study proposed by Robertson et al. (5) found the four different molecular groups based on highly expressed genes, CNA and MET, separately. We hypothesized that an integrative clustering analysis, comprising the three profiles above, would be a more powerful approach. Moreover, our clustering tool, PINSPPlus (version 2.0.5) (17, 18), demonstrated its great ability in cancer subtyping, in general, using multi-omics data. Especially, it classified breast cancer patients into two subgroups that have possessed biologically and clinically meaningful properties (14). We, therefore, continued applying this tool to seeking

the optimal group number of UV patients. In this study, we kept all the parameters of PINSPPlus as default (i.e., clustering method was  $k$ -means); except for the number of candidate groups,  $k$  was set to a range from 2 to 10. The area under the receiver operating characteristic (AUC) value allowed us to choose the optimal  $k$ .

## 2.5 Subgroup-Specific Gene Determination and Enrichment Analysis

To observe the biological differences between identified UM subgroups, we sought to discover the subtype-specific genes using the package GeneCluster (version 0.1.0) (14). Given the lists of genes (i.e., METexp and CNAexp), this tool computed the mean expression level of each gene in each identified patient subgroup across all samples. Then, the gene whose mean expression value was the highest will be allotted to a cluster if the  $P$ -value  $\leq 0.05$  (one-way ANOVA test; two-sided). Finally, the gene will be recognized officially as belonging to that subtype if

the adjusted  $P$ -value  $\leq 0.05$  (Benjamini–Hochberg procedure (16); two-tailed).

Subsequently, in order to investigate further the biological themes from the gained subgroup-specific genes, we implemented the enrichment analysis using the DAVID tool (version 6.8) (23, 24). Also, the output was concentrated into functional-related gene groups or different meaningful terms that were convenient to translate into the clinic. The significance levels of these terms were assessed based on  $P$ -value (Fisher's exact test). In other words, a list of genes with a smaller  $P$ -value was more overrepresented and had a stronger association to the subtype phenotypes.

## 3 RESULTS

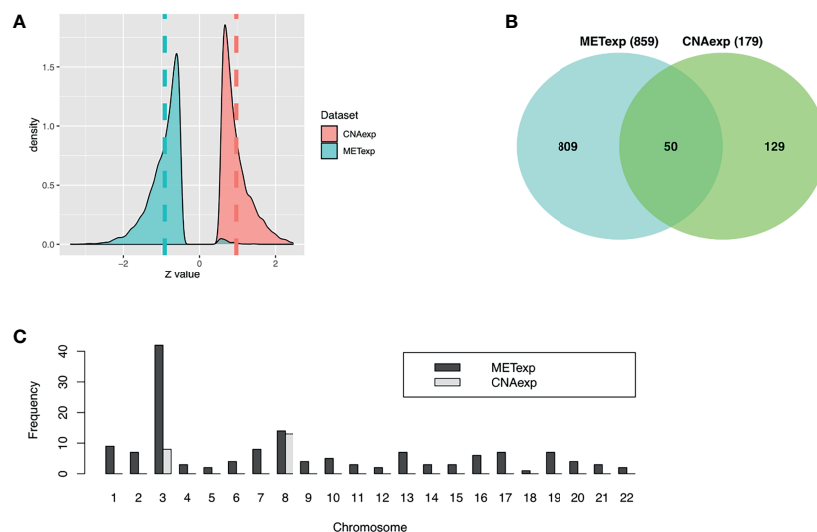
### 3.1 Identification and Examination of the Relationship of CNAexp and METexp Genes

Our tool geneCor provided us with the two sets comprising 4,139 CNAexp genes and 8,157 METexp genes (see **Supplementary Table S1**). As pictured in **Figure 2A**, the CNAexp genes were significantly skewed to the right (skewness = 1.3511,  $P$ -value  $< 2.2 \times 10^{-16}$ ; D'Agostino test; two-sided) consistent with the results reported in (25), while the METexp genes were significantly skewed to the left (skewness =  $-0.3419$ ,  $P$ -value  $< 2.2 \times 10^{-16}$ ; D'Agostino test; two-sided) consistent with the results reported in (26). This indicated that there was a consistently converse relation of mRNA with CNA and MET genes. As mentioned, we truncated genes per the gene set above (i.e., CNAexp and METexp) based on the association with the OS of patients. Particularly, due to an overwhelming number of genes in each set, we only preserved genes per set if  $P$ -value  $< 0.0005$ . Finally, 179 CNAexp genes and 859 METexp genes were obtained. It was a weak intersection (50 genes) between CNAexp and METexp, indicating that the CNAexp and

METexp were two poorly non-disjoint events (**Figure 2B**). **Figure 2C** shows the frequency of the CNAexp or METexp genes against the total count of genes in each chromosome arm. Of particular interest, CNAexp only distributed in two chromosomes 3 and 8, especially almost in chromosome 8, implying not only a poor prognosis but also a considerably reduced survival (27–30). Also, we could observe that the METexp genes displayed a regional genomic preference for MET, particularly on chromosome 3, involving in high metastatic risk (26).

### 3.2 Single and Integrated Subtyping

As described in the *Materials and Methods* section, we implemented the single clustering analyses for CNAexp and METexp, separately. For METexp, the  $k$  of two with the AUC of 1.0000 was optimal (**Figure 3A**). Similarly, for CNAexp, the same  $k$  and AUC were also optimal again (**Figure 3A**). Notably, the number of patients assigned to either of the two CNAexp subgroups significantly overlapped with that of the two METexp subgroups ( $P$ -value =  $3.6714 \times 10^{-15}$ ;  $\chi^2$  test; two-sided; **Figure 3B**). The heatmap shows the expression patterns of CNAexp subgroups and METexp subgroups from integrated analysis by PINSPPlus (**Supplementary Figure S1**). Moreover, the association between our integrated subgroups, IntSub1 and IntSub2, versus patient subtypes in (5) using mRNA data is also shown in **Supplementary Figure S2**. Interestingly, IntSub1 was divided almost into subgroups 1 to 3, whereas most patients in IntSub2 belonged previously to subtype 4. We then employed the survival analysis for the acquired subgroups of CNAexp and METexp. The two CNAexp subgroups were revealed to be statistically meaningful to the OS ( $P$ -value =  $1.7844 \times 10^{-5}$ ; two-sided; **Figure 3C**). Also, with the Cox  $P$ -value =  $1.1006 \times 10^{-6}$ , the two METexp subgroups were significantly correlated with the OS (**Figure 3C**). These results told us that the data single clustering strategy seemed to be effective



**FIGURE 2** | Characteristics of CNAexp and METexp in UM. **(A)** Two Z-score distributions showed two associations of MET or CNA with their respective mRNA. **(B)** Intersection between 859 METexp genes and 179 CNAexp genes. **(C)** Side-by-side bar chart showed the frequency of the CNAexp or METexp genes against the total count of genes in each chromosome arm. CNA, DNA copy number aberrations; MET, epigenetic DNA methylation.

in this case. However, the given single analyses might only show the results that reflected the solitary aberration in UM pathology.

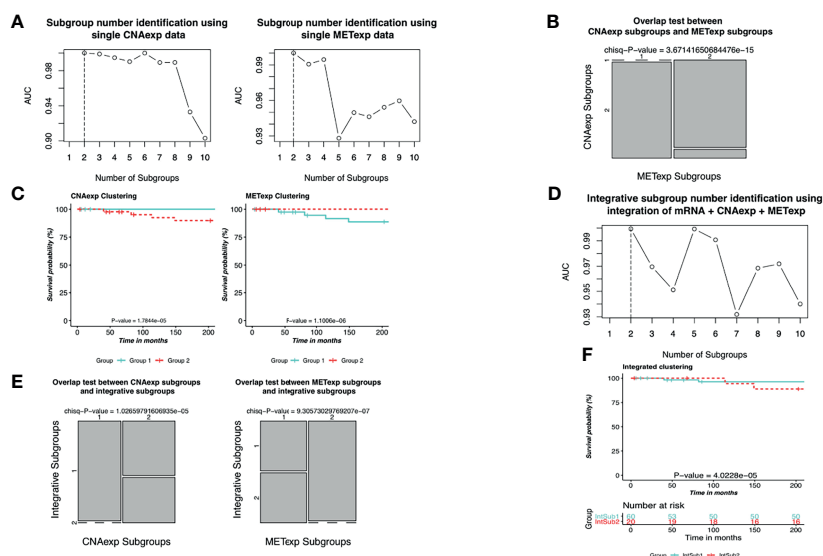
Next, the integrative clustering analysis was leveraged for a combination of CNAexp, METexp, and mRNA gene sets in a similar manner with the single clustering analysis above. Interestingly, PINSPPlus classified UM patients into two integrative subgroups called IntSub1 ( $n = 60$ ) and IntSub2 ( $n = 20$ ) (Figure 3D). Specially, they were consistent significantly with the single subgroups of the CNAexp dataset ( $P$ -value =  $1.0266 \times 10^{-5}$ ;  $\chi^2$  test; two-sided; Figure 3E) and the METexp dataset ( $P$ -value =  $9.3057 \times 10^{-7}$ ;  $\chi^2$  test; Figure 3E). On top of that, we then investigated the survival analysis which revealed that the two integrated subgroups possessed statistically different factors for the survival of UM patients ( $P$ -value =  $4.0228 \times 10^{-5}$ ; log-rank test; Figure 3F). Also, in Figure 3F, the patients in IntSub2 were significantly worse than those in IntSub1 (hazard ratio of 6.1204 and 95% confidence interval between 2.5970 and 14.4200; IntSub1 was reference; log-rank test). Also, we reviewed the statistical descriptions for UM patients, containing age, gender, tumor stages, and metastasis status, between the IntSub1 and IntSub2 provided in Supplementary Table S2. These results bolstered our confidence in the effectiveness of our previous strategy (14) in discovering the novel UM patient subgroups under the perspective of integration.

### 3.3 Molecular Characteristics of Integrated Subgroups

#### 3.3.1 Determination of Subgroup-Specific Genes

As mentioned earlier, the GeneCluster tool was leveraged to exploit subtype-specific gene lists. Accordingly, we extracted

three subgroup-specific gene lists for the two integrative subgroups using three kinds of profiles: mRNA, CNAexp, and METexp. Specifically, these lists were established on average mRNA expression levels (IntSub1: 347 genes and IntSub2: 431 genes; Supplementary Table S3), average CNA aberrations (IntSub1: 108 genes and IntSub2: 71 genes; Supplementary Table S4), and average MET aberrations (IntSub1: 492 and IntSub2: 345 genes; Supplementary Table S5). Notably, we checked the intersection of the subgroup-specific genes from mRNA with UM immune single-cell gene signature from Durante et al. (31) and revealed that 46 overlapped genes (13.26%) in IntSub1 belonged to B-cell cluster, CD4 T follicular helper cluster, M2 macrophage cluster, Mitotic CD8 T-cell cluster, etc. (Supplementary Table S6). Meanwhile, 107 overlapped genes (24.82%) in IntSub2 were associated with immune cells such as B cells, CD4 T follicular helper, CD8, gamma delta T cells, and mitotic CD8 T cells (Supplementary Table S6). This indicated that the UM pathology had a strong connection to the abnormally expressed genes related to immune cells. Interestingly, we found that that the highest expressed gene based on copy number aberrations, *SLCO5A1*, was identified to associate with poor outcome (32), which could be a prospective interpretation for the worse prognosis of IntSub2 patients compared with those in IntSub1. Notably, *SLCO5A1* was considered as a prognosis gene correlated with the immune infiltrates. The immune cell infiltration level was noted to be a crucial factor in predicting the UM prognosis (33). Supplementally, we sought out that *BAP1* was associated with abnormal DNA methylation within IntSub2 samples rather than other subtypes. It was reported that about 22% of familial UM



**FIGURE 3** | Identification of UM molecular subgroups using individual CNAexp and METexp genes for single clustering and mRNA + CNAexp + METexp for joint clustering. (A, D) AUC values obtained for each value of  $k$ . The optimal  $k$  has the highest AUC value, in which (A—left, A—right, and D) the results are of CNAexp alone, METexp alone, and integration of mRNA + CNAexp + METexp, respectively. (B) Overlap test between subgroups of CNAexp and METexp. (E) Overlap test between integrative subgroups versus CNAexp subgroups (left) and versus METexp subgroups (right). (C, F) Kaplan–Meier survival curves for the CNAexp subgroups (C—left), METexp subgroups (C—right), and (F) integrated subgroups.

cases found the muted *BAP1*. *BAP1* mutations raised not only a large tumor diameter percentage but also the metastasis risk in UM patients. This indicated that *BAP1* testing is a reasonable recommendation for hereditary melanoma (34). Additionally, *PTP4A3*, the most overexpressed gene ranked by mean expression value among specific genes of IntSub2, was defined as a marker of poor prognosis involved in cell migration and metastatic progression (35). Furthermore, metastasis is a confident signal of the poor outcome, resulting in death in most UM cases (36).

### 3.3.2 Enrichment Analysis Using the DAVID Tool

We next performed the enrichment analysis as described above with the given subgroup-specific genes. Remarkably, the top biological processes for IntSub1-specific CNAexp genes included endonuclease activity and interleukin-17 receptor activity and transcription factor binding (**Supplementary Table S7** and **Supplementary Figure S3**); IntSub1-specific METexp genes were associated with the positive regulation of cell migration, immune effector process, and positive regulation of hydrolase activity (**Supplementary Table S8** and **Supplementary Figure S4**). Conversely, the IntSub2 was characterized most in cellular cation homeostasis embracing, especially, the regulation of pH and the regulation of calcium ion in the CNAexp profile (**Supplementary Table S7** and **Supplementary Figure S5**). Also, the IntSub2 was distinguished by common abnormalities of METexp genes related to the regulation of gene expression and cellular macromolecule biosynthetic process (**Supplementary Table S8** and **Supplementary Figure S6**).

In this study, we also compared the subgroup-specific genes from the two lists: mRNA (**Supplementary Table S3**) and CNAexp (**Supplementary Table S4**) with the FoundationOne CDx (updated on June 15, 2020) that included 321 genes relating closely to cancer and participating in the process of tumorigenesis. Consequently, we revealed 22 subgroup-specific mRNA expression genes (bold red gene names in **Supplementary Table S3**) and eight subgroup-specific CNAexp genes (bold red gene names in **Supplementary Table S4**) included in the database above. Collectively, our results reinforced the clinical association between the obtained subgroup-specific genes and melanoma formation.

### 3.3.3. Prognostic Factor Identification

We then sought to conduct the age at diagnosis and survival time analyses in order to define the prognosis factor of two UM subtypes. The results are shown in **Table 2**. It is worth mentioning that 60-year-old or older patients were highly risky to have UM. In addition, there was a distinct difference in the average survival day between IntSub1 and IntSub2 patients: 885.2667 and 617.0000 days, respectively. This indicated that

the OS of UM patients could be foreknown dependent partly on which subgroup a patient is assigned to, to some extent. Besides, the patients in the IntSub1 were characterized by the average age of 60.3333 as well as the average OS of 885.2667 days, whereas those numbers in the IntSub2 were 65.6000 years old and 617.0000 days. Obviously, although the average age of the patients in the IntSub1 was only 5 years younger than that of their counterparts in IntSub2, they could live about 9 months longer than the patients in IntSub2. These results should be understood that age-related risks and survival rates might be separate in these integrative subgroups. For a better understanding, we took into account the risk of the two age groups in each subgroup comprising the mid-adults (21–65 years) and the older adults (>65 years) from the 80 UM patients (22–86 years old) in the clinical data. The reason we chose the threshold of 65 years old was because **Figure 4A** illustrates a bimodal age distribution, implying that we had two groups naturally.

The two age groups, the non-old group and the old group, in each subgroup were interrogated by the survival analyses. Observing the results reported in **Figure 4B**, we revealed a significant survival difference between the two age groups in the IntSub2, whereas no statistical significance in patient outcome between the two age groups was seen in the IntSub1, indicating that age factor could be a risk factor to predict the survival time.

## 4 DISCUSSION AND CONCLUSION

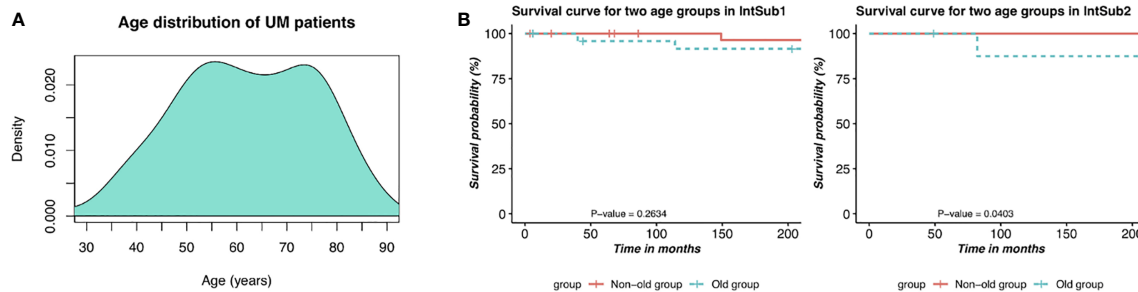
Recently, genomic profiling at multiple levels (e.g., genomics, epigenomics, transcriptomics) has been boomed (37). The abundant omics type of data has been easily accessed from public databases like TCGA facilitating a better understanding of molecular events behind cancer progression. Additionally, based on the associations between the three types of omics data (mRNA, MET, and CNA), we successfully classified breast cancer into two patient subsets which improved the weak manifestations of the intrinsic subtypes, especially in association with the biological phenotype in a prior work. With these concerns in mind, we have decided to apply this successful framework to a rare cancer like UM.

Here, we defined the two lists of CNAexp and METexp based on the correlations of CNA and MET with their mRNA at first. The resulting lists are leveraged to stratify not only individually but also integratively the 80 UM patients using the PINSPlus tool. We revealed the two molecular subgroups (IntSub1 and IntSub2) along with their subtype-specific genes that help to uncover significantly different clinical characteristics as well as

**TABLE 2 |** Average diagnosis ages and survival time of the UM patients in the two integrated subgroups.

	IntSub1	IntSub2
Average age (years)	60.3333	65.6000
Average survival time (days)	885.2667	617.0000





**FIGURE 4** | Description of prognostic risks of UM patients in each integrative subtype. **(A)** Bimodal age distribution of UM patients. **(B)** Kaplan–Meier survival curves of patients in the two age groups in the two identified subgroups, IntSub1 and IntSub2.

patient outcomes. Importantly, there existed several poorly prognostic genes (*SLCO5A1*, *BAP1*, and *PTP4A3*) which could lead to shorter OS of IntSub2 patients. We next recruited the DAVID tool to perform the enrichment analysis in each integrated clustering. Notably, the IntSub1 showed the overexpression of genes enriched significantly in the immune system process (**Supplementary Table S4**). Besides, the IntSub2 displayed CNAexp genes known to be key factors in cellular cation homeostasis and regulation of calcium. These findings are likely to help oncologists and physicists find out distinct treatment strategies for the two subgroups.

The finding of these subgroups could be a suggestion in clinical application for UM treatment. For example, in the IntSub1, the IL-17 (*IL17RE*, *IL17RD*, and *IL17RC*) played vital roles in immune responses which stimulated the tumor growth and repressed the antitumor activity (38). Fabre et al. (39) affirmed in their study that the IL-17/IL-17R axis could be a novel immunotherapeutic target relevant to the antitumor purpose. Besides, the dense appearance of mutated genes is enriched in the cellular cation homeostasis group (i.e.,  $K^+$ ,  $Ca^{2+}$ ,  $Na^+$ , and  $H^+$ ). Cell proliferation and apoptosis were regulated by various cation channels. For instance,  $K^+$  channels participated in the stimulation of the cell end, thus declining the cell number. Therefore, the changeable potassium channels contributed to the malignant expression of cancer (40). In the cellular cation homeostasis gene group, *SGK3* played an activation role of potassium channels (41). Moreover, several prior studies showed the promising therapy of  $K^+$  channel blocking in cancer treatment. This enhanced the consideration of using drugs inhibiting the potassium channels as chemotherapy for UM patients. As an example, astemizole was repositioned in its use by blocking the *EAG1* channel which was one of the major potassium channels and brought remarkable efficacy for cancer cell growth (42). Alternatively, the small molecule which was able to block, inhibit, or regulate the calcium ion transport was reported to be a potential anticancer drug, such as brilliant blue G, oxidized ATP for melanoma cases (43). Taken together, targeted therapies may be efficient for the IntSub1 subgroup, while the combination of the cation channel blocker and chemotherapeutic drugs has the potential for IntSub2 patients.

In addition, we saw that the baselines of both IntSub1 and IntSub2 subgroups varied depending potentially on several clinical

features being vital factors for prognosis. Thus, the survival comparison between the two subgroups was further interrogated by utilizing a multivariate Cox regression model in terms of age groups, tumor stages, gender, and histology cell type comparisons. The analysis results are shown in **Supplementary Table S8**. As a consequence, old age groups, tumor stage IV, and histology cell type comparison between spindle cell and predominant mixed spindle cell were considered as significantly independent prognostic factors.

Furthermore, some powerful predictive genes (except *BAP1*) for prognosis used in clinical routine in UM are not identified by our strategy. This can be regarded as a potential restriction of our work when deliberately leveraging the power of integration of multi-omics data. The following are several factors giving rise to the poor performance of our strategy. The first factor can be the “curse of dimensionality” being a typical problem when using multimodal data. Another factor can be possibly due to the different nature of data types. Most of the statistical tools only work well on continuous data, whereas the minority of them do well on discrete data. In this study, we have combined the two types.

In conclusion, multi-omics data integration contributes to dealing with the bottleneck in getting insights into complex multi-mechanism diseases like cancer in general and UM in particular. We determined the two clinically and molecularly distinct integrative subgroups, IntSub1 and IntSub2, which not only can be a potential alternative classification system in the future but also give more effective suggestions for UM treatment.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

## AUTHOR CONTRIBUTIONS

THYN drafted the manuscript, which was edited by all co-authors. Q-HN conceived and designed the approach, coded the geneCor and GeneCluster tools, and wrote the R codes for analyses. THYN ran the codes. THYN and Q-HN analyzed



output data. TN coded the PINSPlus algorithm. Q-HN and D-HL jointly directed and supervised the work. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was partially supported by NIH NIGMS under grant number GM103440 and by NSF under grant numbers 2001385 and

2019609. This research is also supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA18.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2021.731548/full#supplementary-material>

## REFERENCES

- Shields CL, Furuta M, Thangappan A, Nagori S, Mashayekhi A, Lally DR, et al. Metastasis of Uveal Melanoma Millimeter-by-Millimeter in 8033 Consecutive Eyes. *Arch Ophthalmol* (2009) 127(8):989–98. doi: 10.1001/archophthalmol.2009.208
- Anbunathan H, Verstraten R, Singh AD, Harbour JW, Bowcock AM. Integrative Copy Number Analysis of Uveal Melanoma Reveals Novel Candidate Genes Involved in Tumorigenesis Including a Tumor Suppressor Role for PHF10/BAF45a. *Clin Cancer Res* (2019) 25(16):5156. doi: 10.1158/1078-0432.CCR-18-3052
- Kaliki S, Shields CL. Uveal Melanoma: Relatively Rare But Deadly Cancer. *Eye (London England)* (2017) 31(2):241–57. doi: 10.1038/eye.2016.275
- Yang Z-K, Yang J-Y, Xu Z-Z, Yu W-H. DNA Methylation and Uveal Melanoma. *Chin Med J* (2018) 131(7):845–51. doi: 10.4103/0366-6999.228229
- Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell* (2017) 32(2):204–20.e15. doi: 10.1016/j.ccell.2017.07.003
- Barbagallo C, Caltabiano R, Broggi G, Russo A, Puzzo L, Avitabile T, et al. LncRNA LINC00518 Acts as an Oncogene in Uveal Melanoma by Regulating an RNA-Based Network. *Cancers (Basel)* (2020) 12(12):3867. doi: 10.3390/cancers12123867
- Damato B, Dopierala J, Klaasen A, van Dijk M, Sibbring J, Coupland SE. Multiplex Ligation-Dependent Probe Amplification of Uveal Melanoma: Correlation With Metastatic Death. *Invest Ophthalmol Visual Sci* (2009) 50(7):3048–55. doi: 10.1167/iov.08-3165
- Kilic E, van Gils W, Lodder E, Beverloo HB, van Til ME, Mooy CM, et al. Clinical and Cytogenetic Analyses in Uveal Melanoma. *Invest Ophthalmol Visual Sci* (2006) 47(9):3703–7. doi: 10.1167/iov.06-0101
- Vivet-Noguer R, Tarin M, Roman-Roman S, Alsafadi S. Emerging Therapeutic Opportunities Based on Current Knowledge of Uveal Melanoma Biology. *Cancers (Basel)* (2019) 11(7):1019. doi: 10.3390/cancers11071019
- Russo D, Di Crescenzo RM, Broggi G, Merolla F, Martino F, Varricchio S, et al. Expression of P16INK4a in Uveal Melanoma: New Perspectives. *Front Oncol* (2020) 10:562074. doi: 10.3389/fonc.2020.562074
- Broggi G, Ieni A, Russo D, Varricchio S, Puzzo L, Russo A, et al. The Macro-Autophagy-Related Protein Beclin-1 Immunohistochemical Expression Correlates With Tumor Cell Type and Clinical Behavior of Uveal Melanoma. *Front Oncol* (2020) 10:589849. doi: 10.3389/fonc.2020.589849
- Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-Omics Data Integration, Interpretation, and Its Application. *Bioinf Biol Insights* (2020) 14:1177932219899051. doi: 10.1177/1177932219899051
- 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(4):e35236. doi: 10.1371/journal.pone.0035236
- Nguyen Q-H, Nguyen H, Nguyen T, Le D-H. Multi-Omics Analysis Detects Novel Prognostic Subgroups of Breast Cancer. *Front Genet* (2020) 11:574661. doi: 10.3389/fgene.2020.574661
- Dunkler D, Sanchez-Cabo F, Heinze G. Statistical Analysis Principles for Omics Data. *Methods Mol Biol (Clifton NJ)* (2011) 719:113–31. doi: 10.1007/978-1-61779-027-0\_5
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B (Methodological)* (1995) 57(1):289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: A Tool for Tumor Subtype Discovery in Integrated Genomic Data. *Bioinf (Oxford England)* (2018) 35(16):2843–6. doi: 10.1093/bioinformatics/bty1049
- Nguyen T, Tagett R, Diaz D, Draghici S. A Novel Approach for Data Integration and Disease Subtyping. *Genome Res* (2017) 27(12):2025–39. doi: 10.1101/gr.215129.116
- 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 Discov* (2012) 2(5):401–4. doi: 10.1158/2159-8290.CD-12-0095
- Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the Cbioportal. *Sci Signal* (2013) 6(269):p1. doi: 10.1126/scisignal.2004088
- Batista GEAPA, Monard MC. A Study of K-Nearest Neighbour as an Imputation Method. *HIS* (2002) 87:251–60.
- Xu T, Le TD, Liu L, Su N, Wang R, Sun B, et al. CancerSubtypes: An R/Bioconductor Package for Molecular Cancer Subtype Identification, Validation and Visualization. *Bioinf (Oxford England)* (2017) 33(19):3131–3. doi: 10.1093/bioinformatics/btx378
- Huang da W, Sherman BT, Lempicki RA. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat Protoc* (2009) 4(1):44–57. doi: 10.1038/nprot.2008.211
- Huang da W, Sherman BT, Lempicki RA. Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res* (2009) 37(1):1–13. doi: 10.1093/nar/gkn923
- Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy Number Variation is Highly Correlated With Differential Gene Expression: A Pan-Cancer Study. *BMC Med Genet* (2019) 20(1):175. doi: 10.1186/s12881-019-0909-5
- Field MG, Kuznetsov JN, Bussies PL, Cai LZ, Alawa KA, Decatur CL, et al. BAP1 Loss Is Associated With DNA Methylation Repatterning in Highly Aggressive Class 2 Uveal Melanomas. *Clin Cancer Res* (2019) 25(18):5663. doi: 10.1158/1078-0432.CCR-19-0366
- Hammond DW, Al-Shammari NS, Danson S, Jacques R, Rennie IG, Sisley K. High-Resolution Array CGH Analysis Identifies Regional Deletions and Amplifications of Chromosome 8 in Uveal Melanoma. *Invest Ophthalmol Vis Sci* (2015) 56(6):3460–6. doi: 10.1167/iov.14-16215
- Sisley K, Rennie IG, Parsons MA, Jacques R, Hammond DW, Bell SM, et al. Abnormalities of Chromosomes 3 and 8 in Posterior Uveal Melanoma Correlate With Prognosis. *Genes Chromosomes Cancer* (1997) 19(1):22–8. doi: 10.1002/(SICI)1098-2264(199705)19:1<22::AID-GCC4>3.0.CO;2-2
- Versluis M, de Lange MJ, van Pelt SI, Ruivenkamp CA, Kroes WG, Cao J, et al. Digital PCR Validates 8q Dosage as Prognostic Tool in Uveal Melanoma. *PLoS One* (2015) 10(3):e0116371. doi: 10.1371/journal.pone.0116371
- van den Bosch T, van Beek JG, Vaarwater J, Verdijk RM, Naus NC, Paridaens D, et al. Higher Percentage of FISH-Determined Monosomy 3 and 8q Amplification in Uveal Melanoma Cells Relate to Poor Patient Prognosis. *Invest Ophthalmol Vis Sci* (2012) 53(6):2668–74. doi: 10.1167/iov.11-8697
- Durante MA, Rodriguez DA, Kurtenbach S, Kuznetsov JN, Sanchez MI, Decatur CL, et al. Single-Cell Analysis Reveals New Evolutionary Complexity in Uveal Melanoma. *Nat Commun* (2020) 11(1):496. doi: 10.1038/s41467-019-14256-1

32. Luo H, Ma C. Identification of Prognostic Genes in Uveal Melanoma Microenvironment. *PLoS One* (2020) 15(11):e0242263–e. doi: 10.1371/journal.pone.0242263
33. Narasimhaiah D, Legrand C, Damotte D, Remark R, Munda M, De Potter P, et al. DNA Alteration-Based Classification of Uveal Melanoma Gives Better Prognostic Stratification Than Immune Infiltration, Which has a Neutral Effect in High-Risk Group. *Cancer Med* (2019) 8(6):3036–46. doi: 10.1002/cam4.2122
34. Rai K, Pilarski R, Boru G, Rehman M, Saqr AH, Massengill JB, et al. Germline BAP1 Alterations in Familial Uveal Melanoma. *Genes Chromosomes Cancer* (2017) 56(2):168–74. doi: 10.1002/gcc.22424
35. Duciel L, Anezo O, Mandal K, Laurent C, Planque N, Coquelle FM, et al. Protein Tyrosine Phosphatase 4A3 (PTP4A3/PRL-3) Promotes the Aggressiveness of Human Uveal Melanoma Through Dephosphorylation of CRMP2. *Sci Rep* (2019) 9(1):2990. doi: 10.1038/s41598-019-39643-y
36. Lane AM, Kim IK, Gragoudas ES. Survival Rates in Patients After Treatment for Metastasis From Uveal Melanoma. *JAMA Ophthalmol* (2018) 136(9):981–6. doi: 10.1001/jamaophthalmol.2018.2466
37. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *BioMed Res Int* (2018) 2018:9836256. doi: 10.1155/2018/9836256
38. Yang B, Kang H, Fung A, Zhao H, Wang T, Ma D. The Role of Interleukin 17 in Tumour Proliferation, Angiogenesis, and Metastasis. *Mediators Inflamm* (2014) 2014:623759–. doi: 10.1155/2014/623759
39. Fabre J, Giustiniani J, Garbar C, Antonicelli F, Merrouche Y, Bensussan A, et al. Targeting the Tumor Microenvironment: The Protumor Effects of IL-17 Related to Cancer Type. *Int J Mol Sci* (2016) 17(9):1433. doi: 10.3390/ijms17091433
40. Comes N, Serrano-Albarrás A, Capera J, Serrano-Novillo C, Condom E, Ramón YCS, et al. Involvement of Potassium Channels in the Progression of Cancer to a More Malignant Phenotype. *Biochim Biophys Acta* (2015) 1848(10 Pt B):2477–92. doi: 10.1016/j.bbame.2014.12.008
41. Gamper N, Feng Y, Friedrich B, Lang PA, Henke G, Huber SM, et al. K<sup>+</sup> Channel Activation by All Three Isoforms of Serum- and Glucocorticoid-Dependent Protein Kinase SGK. *Pflugers Arch* (2002) 445(1):60–6. doi: 10.1007/s00424-002-0873-2
42. Downie BR, Sánchez A, Knötgen H, Contreras-Jurado C, Gymnopoulos M, Weber C, et al. Egl1 Expression Interferes With Hypoxia Homeostasis and Induces Angiogenesis in Tumors. *J Biol Chem* (2008) 283(52):36234–40. doi: 10.1074/jbc.M801830200
43. Babcock JJ, Du F, Xu K, Wheelan SJ, Li M. Integrated Analysis of Drug-Induced Gene Expression Profiles Predicts Novel hERG Inhibitors. *PLoS One* (2013) 8(7):e69513–e. doi: 10.1371/journal.pone.0069513

**Conflict of Interest:** Authors THYN, Q-HN and D-HL were employed by Vingroup Big Data Institute.

The remaining 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.

Copyright © 2021 Nguyen, Nguyen, Nguyen and Le. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# SMRT: Randomized Data Transformation for Cancer Subtyping and Big Data Analysis

Hung Nguyen<sup>1</sup>, Duc Tran<sup>1</sup>, Bang Tran<sup>1</sup>, Monikrishna Roy<sup>1</sup>, Adam Cassell<sup>1</sup>, Sergiu Dascalu<sup>1</sup>, Sorin Draghici<sup>2</sup> and Tin Nguyen<sup>1\*</sup>

<sup>1</sup> Department of Computer Science and Engineering, University of Nevada Reno, Reno, NV, United States,

<sup>2</sup> Department of Computer Science, Wayne State University, Detroit, MI, United States

## OPEN ACCESS

### Edited by:

David A. Hornuth, II,  
The University of Texas at Austin,  
United States

### Reviewed by:

Soumita Ghosh,  
National University of Singapore,  
Singapore  
Jiaqi Liu,  
National Cancer Center of China,  
China

### \*Correspondence:

Tin Nguyen  
tinn@unr.edu

### Specialty section:

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

**Received:** 14 June 2021

**Accepted:** 28 September 2021

**Published:** 20 October 2021

### Citation:

Nguyen H, Tran D, Tran B, Roy M, Cassell A, Dascalu S, Draghici S and Nguyen T (2021) SMRT: Randomized Data Transformation for Cancer Subtyping and Big Data Analysis. *Front. Oncol.* 11:725133. doi: 10.3389/fonc.2021.725133

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. The treatment options, as well as treatment success, are highly dependent on the correct subtyping of individual patients. With the advancement of high-throughput platforms, we have the opportunity to differentiate among cancer subtypes from a holistic perspective that takes into consideration phenomena at different molecular levels (mRNA, methylation, etc.). This demands powerful integrative methods to leverage large multi-omics datasets for a better subtyping. Here we introduce Subtyping Multi-omics using a Randomized Transformation (SMRT), a new method for multi-omics integration and cancer subtyping. SMRT offers the following advantages over existing approaches: (i) the scalable analysis pipeline allows researchers to integrate multi-omics data and analyze hundreds of thousands of samples in minutes, (ii) the ability to integrate data types with different numbers of patients, (iii) the ability to analyze unmatched data of different types, and (iv) the ability to offer users a convenient data analysis pipeline through a web application. We also improve the efficiency of our ensemble-based, perturbation clustering to support analysis on machines with memory constraints. In an extensive analysis, we compare SMRT with eight state-of-the-art subtyping methods using 37 TCGA and two METABRIC datasets comprising a total of almost 12,000 patient samples from 28 different types of cancer. We also performed a number of simulation studies. We demonstrate that SMRT outperforms other methods in identifying subtypes with significantly different survival profiles. In addition, SMRT is extremely fast, being able to analyze hundreds of thousands of samples in minutes. The web application is available at <http://SMRT.tinnguyen-lab.com>. The R package will be deposited to CRAN as part of our PINSPlus software suite.

**Keywords:** cancer subtyping, multi-omics integration, web application, CRAN package, survival analysis

# 1 INTRODUCTION

Since cancer is a heterogeneous disease, the correct identification of cancer subtypes is essential for accurate prognosis and improved treatment. With the advancement of high-throughput platforms, subtyping methods have shifted toward multi-omics integration in order to differentiate between subtypes from a holistic perspective that takes into consideration phenomena at different molecular levels (mRNA, methylation, etc.). Vast amounts of molecular data have accumulated in public repositories, including The Cancer Genome Atlas datasets (TCGA) (1), Genomic Data Commons Data Portal (GDC) (2), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (3), and UK Biobank (4). This demands powerful yet fast analysis methods to leverage large multi-omics datasets for a more accurate subtype discovery.

Current approaches for multi-omics integration and cancer subtyping can be categorized into four categories based on their integration strategy. The first strategy is to concatenate different types of data into a single matrix and then partition the patients using the concatenated data. For example, users can normalize and concatenate multiple data types (e.g., mRNA, methylation, miRNA, etc.) into one single matrix and then apply well-known methods developed for single-omics analysis, such as ConsensusClusterPlus (5), to determine the subtypes. Such approaches are simple and computationally efficient. However, they do not account for data heterogeneity, e.g., different data types might have different scales, dimensions and might require different normalization procedures.

The second strategy is to model the multi-omics data as a mixture of statistical models. Methods in this category include LRACluster (6), rMKL-LPP (7), iClusterPlus (8), iClusterBayes (9), OTRIMLE (10), SBC (11), BCC (12), MID (13), JIVE (14), MCIA (15), moCluster (16), and sMBPLS (17). These methods typically maximize a joint likelihood function to determine the model parameters and the subtypes. Though statistically sound, these methods need to estimate a large number of parameters that often lead to overfitting and high computational complexity. Therefore, an added step of gene filtering or data transformation is often applied before the statistical analysis.

The third strategy is to project all data types into a joint latent space. A common technique used for this strategy is non-negative matrix factorization. Methods in this category include MvNMF (18), MultiNMF (19), IntNMF (20), iNMF (21), jointNMF (22). Another method is MCCA (23) that performs correlation analysis and then concatenates the correlation matrices into one single matrix. After projecting the data onto a joint space, cluster analysis is performed to determine the final subtypes. Similar to the second strategy, methods in this category often have excessive computational complexity and cannot be applied on the whole genome-scale. Therefore, gene filtering is a necessary step in the data processing.

The fourth strategy is also called similarity-based strategy. Methods in this category include SNF (24), PSDF (25), PFA (26), IS-Kmeans (27), NEMO (28), PINS (29, 30), SCFA (31), and CIMLR (32). These methods first compute a pair-wise

connectivity matrix for each data type, that represents the similarity/connectivity between patients. The connectivity matrices are then fused onto a single similarity matrix that can be used for the final clustering. Although powerful, the similarity matrix requires a quadratic memory space. This is problematic when the number of samples increases. As we will demonstrate in our analysis, these methods cannot analyze data with tens of thousands of samples.

Here we introduce Subtyping Multi-omics using a Randomized Transformation (SMRT), a new method for cancer subtyping and big data analysis. This method offers important advantages over existing software: (i) it allows researchers to analyze hundreds of thousands of samples in minutes, (ii) it can integrate data types with different numbers of patients, (iii) the ability to integrate and analyze un-matched data of different types, and (iv) the web application offers a convenient data analysis pipeline. We also improve the efficiency of our ensemble-based, perturbation clustering to support analysis on machines with memory constraints. Our extensive analysis on 37 TCGA and two METABRIC datasets shows that SMRT is more accurate than state-of-the-art subtyping methods in identifying subtypes with significantly different survival profiles. In addition, our simulations with big data show that SMRT is fast and many-fold more scalable than existing methods. Specifically, SMRT is able to analyze hundreds of thousands of samples in minutes.

# 2 MATERIALS AND METHODS

## 2.1 The SMRT Pipeline

The overall workflow of SMRT is presented in **Figure 1**. This workflow offers two different analysis pipelines for big data and data with a moderate size. In the first case, given a multi-omics dataset with a moderate size (e.g., less than 2,000 samples), SMRT performs subtyping as follows. It first projects each data type onto a lower-dimensional space using randomized singular value decomposition (RSVD) and then performs a perturbation clustering (PINS) (29, 30) to determine the subtypes within each data level. It also builds a pair-wise connectivity matrix for each data type that represents the connectivity between patients red (See **Supplementary Section 5** for the differences between SMRT and PINS). Next, the method combines the connectivity matrices into a single similarity matrix and then determines the final subtypes using an ensemble of multiple similarity-based methods. In the second case, when the data has more than 2,000 samples, SMRT splits the data into two different sets of patients: a sampled set and a propagated set. It then performs the subtyping on the sampled set and then assigns the patients from the propagated set to the identified subtypes. Note that the number 2,000 is chosen to balance between the accuracy and time complexity of the method. This moderate number of samples allows SMRT to perform a fast and accurate analysis in limited memory (see **Supplementary Section 3**). Our simulation studies show that the results do not change when



we vary this number. However, users are free to change this parameter when using the R package. Below is the description of each of these analysis modules.

## 2.2 Dimension Reduction Using Randomized Singular Value Decomposition

The goal of this step is to project the multi-omics data into a lower-dimensional space using randomized singular value decomposition (RSVD). For data with hundreds of thousands of dimensions (e.g., Illumina 450k), this step substantially reduces the required computational power while maintaining the clustering accuracy. Let us denote  $X \in \mathbb{R}^{n \times m}$  as the input matrix, where  $n$  is the number of samples/patients, and  $m$  is the number of genes/features. Briefly, the RSVD method starts by generating a random projection matrix  $P \in \mathbb{R}^{m \times r}$  from a standard normal distribution where  $r \ll m$ . It then projects  $X \in \mathbb{R}^{n \times m}$  to the column space of  $P$  to get a matrix  $Z$  such that  $Z = XP$ . Due to the random projection,  $Z$  and  $X$  will have approximately the same dominant columns (features). Now, we can obtain the orthogonalized matrix  $Q$  of  $Z$  by using QR decomposition, where  $Q$  has the same size as  $Z$  of  $n \times r$ . In the next step, the method projects  $X$  into a smaller space to get a matrix  $Y \in \mathbb{R}^{n \times r}$  such that  $Y = Q^{\wedge T} * X$  and then computes singular value decomposition (SVD) of  $Y$  as  $Y = U \Sigma V^*$  using the traditional SVD method (33).  $U$  and  $V$  matrices only keep at most  $r$  eigenvectors so the size of  $U$  is  $r \times r$  and the size of  $V^*$  is  $m \times r$ . Finally, the low rank rotated data of the original matrix  $X$  can be computed using:  $X' = XV^*$ .

In practice, RSVD is faster and requires less memory than the traditional SVD. To further speed up our approach, we implement a parallel version of RSVD that can efficiently utilize multiple cores available in modern processors. Note that when the input data is large (e.g., more than 2,000 samples), we do not perform RSVD on the whole input. Instead, we split the data into two sets of patients: a sampled set and a propagated set. We first perform RSVD on the sampled set, and then project the original data matrix (both sampled and propagated set) to the subspace of the *sampled set* by multiplying it with the rotation matrix obtained from the RSVD for the sampled set. This implementation allows us to perform SVD in at most a few seconds, even for datasets with hundreds of thousands of samples and features.

The output of this module is multiple matrices – one per data type. In each matrix, the rows represent patients while the columns represent the principal components (PCA). These matrices will serve as input of the next module: perturbation clustering that will be described in the next section. This will compute the perturbed connectivity matrices and determine the subtypes.

## 2.3 Subtype Discovery Using One Data Type

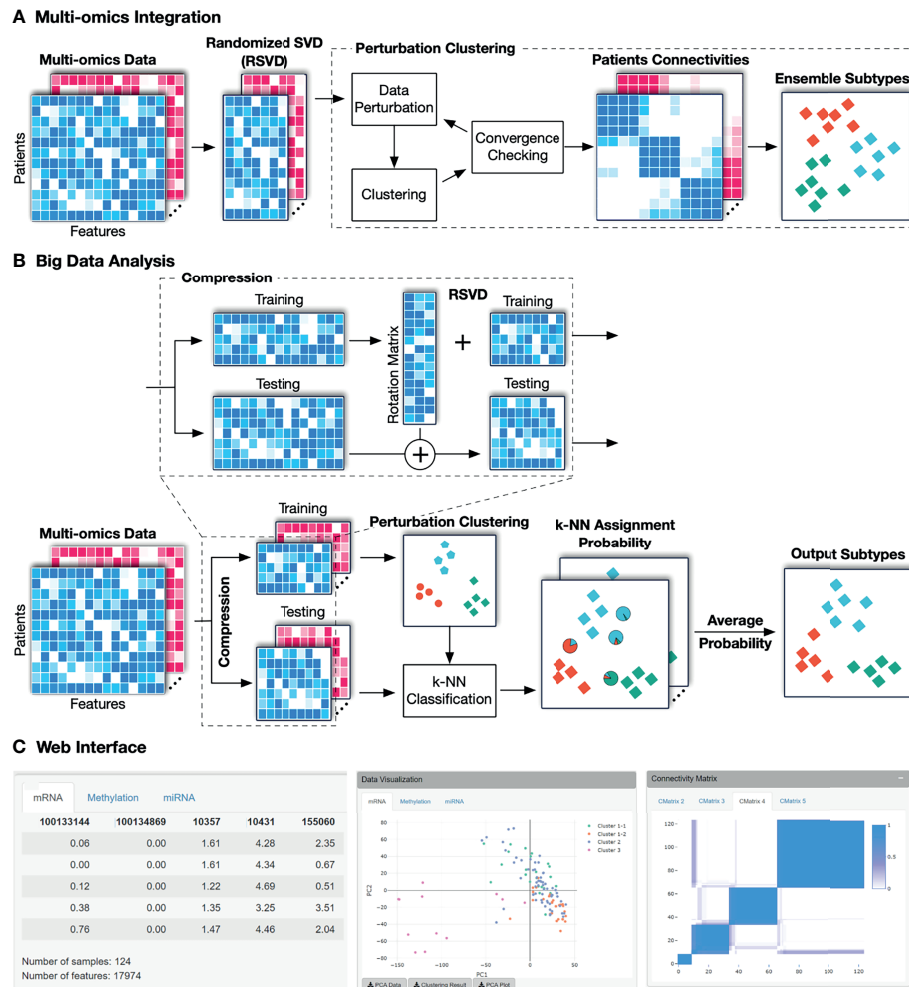
Given a single data type, SMRT utilizes our previously developed perturbation clustering (PINS) (29, 30) to partition the data.

Briefly, we perturb the data (by adding Gaussian noise) and repeatedly partition the patients (using k-means by default). For each partitioning, we build a pair-wise connectivity matrix of 0's and 1's in which 1 means that the two patients belong to the same cluster, and 0 otherwise. By perturbing and clustering the data multiple times, we obtain multiple connectivity matrices that represent how stable the connectivity between patient pairs. Finally, we choose the partitioning that is the most stable to data perturbation. This algorithm automatically determines the number of clusters and patient subgroups.

When the number of samples is large, the perturbation clustering becomes slow and memory-inefficient. The perturbation clustering algorithm relies on the pair-wise connectivity of size  $n \times n$  for clustering ( $n$  is the number of patients). The time and space complexity (running time and memory usage) of this method increase quadratically when the number of samples increases. Therefore, when the number of samples is large (by default setting, when  $n > 2,000$ ), we perform a sub-sampling process over the original data to obtain a subset of 2,000 patients/samples. Next, we transform the data into a lower-dimensional space, and use the perturbation clustering to partition these patients. After this step, each of the 2,000 patients has a subtype. Let us refer to this selected set of 2,000 patients as the *sampled set*. The next step is to determine the subtypes for the rest of the patients, called the *propagated set*. For this purpose, we use the fast k-nearest neighbor searching algorithms (FKNN) algorithm (34, 35) to assign each patient from the propagated set to one of the subtypes in the sampled set. Briefly, the FKNN method calculates the distance between the new patient to the  $k$  nearest patients in the sampled set. Next, the FKNN method classifies the new patient using vote counting (i.e., it chooses the subtype with the most patients among the  $k$  neighbors). By default,  $k$  is determined using the Elbow method on the sampled set using 5-fold cross-validation. The sampled set is divided randomly into 5 equally smaller sets. In each round, the combination of 4 sets is used as the training set, and the other is used as the validation set for the KNN algorithm with  $k$  ranges from 5 to a maximum of 50. The  $k$  that yields the lowest average classification error rate will be used as the optimal  $k$ . However, users are also free to modify the value of this parameter. **Supplementary Section 6** provides more details on the performance of using the Elbow method *versus* using a fixed number of  $k$ .

One note of caution is that the number of dimensions of the data can be high, thus slowing the process of distance calculation and neighbor finding. Therefore, instead of calculating the distance between patients in the original space, we calculate the distance between patients in the principal component (PC) space of the sampled set. As described above, we project the original data matrix (both sampled and propagated set) to the subspace of the *sampled set* by multiplying it with the projection matrix obtained from the RSVD for the sampled set. After this transformation, the pair-wise distance between patients will be calculated in the new space with a much lower number of dimensions.





**FIGURE 1 |** The overall workflow of SMRT. **(A)** Analysis pipeline for data with moderate size. First, SMRT projects each data type to a lower-dimensional space using randomized singular value decomposition (RSVD). Next, it performs a perturbation clustering to determine the subtypes, and to build a pair-wise patient connectivity for each data type. Finally, it merges the connectivity matrices onto a single similarity matrix and then determines the final subtypes using a cluster ensemble. The output is the clustering results for each data type, as well as the results after the multi-omics data integration. **(B)** Analysis pipeline for big data. SMRT first splits the data into two different sets: a sampled and a propagated set. The method first determines the subtypes using the sampled set and then assigns the patients from the propagated set to subtypes identified using the sampled set. The sampled data is partitioned using the pipeline described assignments for samples in the propagated set are determined by averaging the probabilities from all k-NN models. **(C)** An example of the subtypes discovered by the SMRT web service for the KIRC dataset. The left panel shows a preview of the uploaded data. The middle panel shows the visualization of the discovered subtypes and export functions. The right panel shows patient connectivity matrices for each data type.

## 2.4 Subtype Discovery Using Multi-Omics Data

When the number of samples is small (by default, when  $n \leq 2,000$ ), we utilize an ensemble strategy to partition the patients. The method first clusters each data type (using the algorithm described in Section 2.3) and constructs the perturbed connectivity matrices. It then merges the connectivity matrices of all data types to a single similarity matrix that represents the similarity between patients across all data types by averaging the connectivity values for each pair of samples. Next, to cluster the similarity matrix, it uses several similarity-based algorithms, including hierarchical clustering, partitioning around medoids (36), and dynamic tree

cut (37) and then chooses the partitioning that agrees the most with the partitioning of individual data types. This ensemble strategy ensures that the identified subtypes are consistent across all data types and are robust against the choice of clustering algorithms.

When the number of samples is large (by default, when  $n > 2,000$ ), we perform a sub-sampling and classifying procedure that is similar to the algorithm described in the Section 2.3. The difference here is that multiple data types are involved. First, we randomly select 2,000 samples/patients and then apply the multi-omics algorithm described above to partition the selected samples. We refer to this selected set of 2,000 patients as the

*sampled set* and the remaining patients as the *propagated set*. The next task is to determine the subtypes of patients in the propagated set. Given a patient in the propagated set, we perform the FKNN procedure for each data type to obtain the probability that it belongs to each subtype using the labels obtained from the nearest neighbors. The final probabilities are calculated by averaging the probabilities across all data types. Finally, we classify the patient to the subtype that has the highest probability. This strategy is also applied when integrating multi-omics data whose each data type has different number of samples. Here the sampled set will be the set of patients (by default, maximum 2,000 patients) that have data in all data types, and the remaining patients will be in the propagated set.

## 2.5 The SMRT Web Interface

The web application is publicly available at <http://SMRT.tinnguyen-lab.com>. The website is built using the R Shiny framework (38). Shiny is an R package that allows developers to directly build an interactive web interface using the R programming language. We use the web interface to forward data and requests from users to the new SMRT method to perform data integration and clustering. Because of the efficiency of the SMRT method, the website is able to return the results in minutes even for datasets with hundreds of thousands of samples.

Analysis using the web application is simple and straightforward. Users can either upload expression data in .csv files or a single .rds file using the upload function on the left panel. Each data type is presented as a matrix in which rows represent samples and columns represent genes/features. SMRT can automatically determine the number of subtypes. It does not require any extra configuration or parameters to perform the analysis. See **Supplementary Section 4** and **Figures S6, S7** for a more detailed description of the web application.

## 3 RESULTS

To assess the performance of SMRT, we perform an extensive analysis using 39 cancer datasets and simulated data. First, we demonstrate that SMRT is able to identify cancer subtypes with significantly different survival profiles. Second, we provide an in-depth analysis for a Glioma dataset. Finally, we illustrate the scalability of SMRT by analyzing simulated datasets with hundreds of thousands of samples. We also provide a comparative analysis between subtypes discovered by SMRT and those of PAM50 classifier on three Breast cancer datasets (TCGA-BRCA, METABRIC\_Discovery, and METABRIC\_Validation) in **Supplementary Section 7**.

### 3.1 Experimental Studies Using 39 Cancer Datasets

In this article, we analyze 37 TCGA and 2 METABRIC datasets. For TCGA datasets, we downloaded the matched mRNA, DNA methylation, and miRNA expression data from the TCGA data portal. For the METABRIC datasets, we were able to obtain

matched mRNA and copy number variation data from the European Genome-Phenome Archive. We also downloaded clinical data and survival information of each patient, which will be used to assess the performance of the subtyping methods. **Supplementary Tables 1, 2** provide more details of the datasets.

We compare SMRT with eight state-of-the-art subtyping algorithms: SNF (24), CIMLR (32), NEMO (28), moCluster (16), iClusterBayes (9), LRACluster (6), MCCA (23), and IntNMF (20). The following packages were used in our comparison: SNFtool v2.3.0 on CRAN for SNF, CIMLR v1.0.0 at <https://github.com/danro9685/CIMLR> for CIMLR, NEMO v0.1.0 at <https://github.com/Shamir-Lab/NEMO> for NEMO, moga v1.16.0 on Bioconductor for moCluster, iClusterPlus on Bioconductor v1.18.0 for iClusterBayes, LRACluster v1.18.0 at <http://bioinfo.au.tsinghua.edu.cn/member/jgu/lracluster/> for LRACluster, PMA v1.2.1 on CRAN for MCCA, and IntNMF on CRAN v1.2.0 for IntNMF. When the number of dimensions exceeded 2,000, we used only the top 2,000 variables with the largest variance for iClusterBayes, IntNMF, and MCCA, because these methods cannot analyze the data on the whole-genome scale. For all methods, we used default parameters and let all methods automatically determine the optimal number of clusters. For MCCA, which is not a clustering method itself, we follow the implementation at <https://github.com/Shamir-Lab/Multi-Omics-Cancer-Benchmark> for cluster analysis.

Using each method, we partition the patients in each dataset, and then assess the survival difference of the discovered patient groups using Cox regression (39). Overall survival data is used for TCGA datasets and Disease-free survival data is used for METABRIC datasets. **Table 1** shows the Cox p-values obtained from each dataset and method (See **Supplementary Section 9, Figures S10–S17** for the Kaplan-Meier survival curves for each dataset). There are seven datasets in which no method is able to identify subtypes with significant Cox p-values. For the remaining 32 datasets, SMRT has significant p-values in 28 datasets, whereas NEMO has significant p-values in 19 datasets and all other methods have significant p-values in 15 datasets or less. SMRT has the most significant p-values in 12 datasets out of those 28 datasets, while SNF, CIMLR, NEMO, moCluster, iClusterBayes, LRACluster, MCCA, and IntNMF have the most significant p-values in 0, 3, 8, 4, 2, 0, 1, and 2 datasets, respectively.

**Figure 2** shows the distributions of the Cox p-values in the -log<sub>10</sub> scale. Overall, the median -log<sub>10</sub> p-values of SMRT is close to 2 (i.e., median p-value of 0.01) whereas the median -log<sub>10</sub> p-value of the second-best method (NEMO) is close to 1 (i.e., median p-value of 0.1). A Wilcoxon test also confirms that the p-values of SMRT are significantly smaller than the p-values obtained from other methods ( $p = 0.0002$  using the one-tailed Wilcoxon test).

The running time of each method is shown in **Table 2**. The top 39 row shows the running time of each method in each dataset while the last row shows the average running time. On average, SMRT, SNF, NEMO, and MCCA are fast and able to finish each analysis in less than a minute. The remaining methods are slower, especially iClusterBayes and IntNMF, although their analysis is limited to only 2,000 most varied genes.

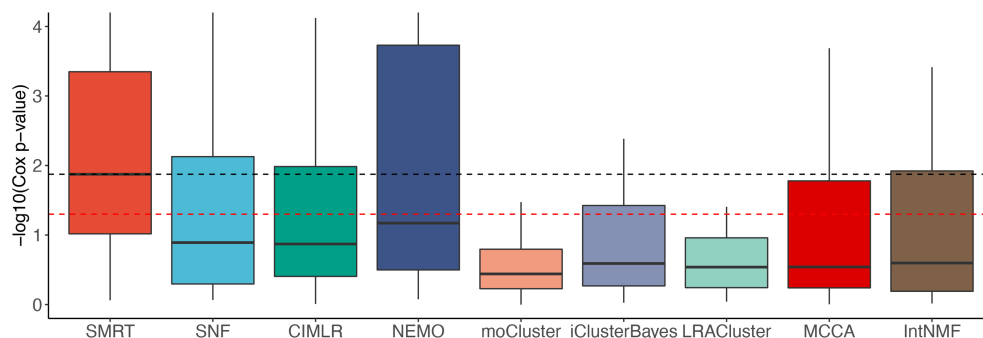
**TABLE 1 |** Cox p-values of subtypes discovered by SNF, CIMLR, NEMO, moCluster, iClusterBayes (iCB), LRAcluster (LRA), MCCA, IntNMF, and SMRT for 37 TCGA datasets and two METABRIC breast cancer datasets (M\_Discovery and M\_Validation).

Dataset	SNF	CIMLR	NEMO	moCluster	iCB	LRA	MCCA	IntNMF	SMRT
1. ACC	4.34e-05	3.96e-01	2.07e-04	2.63e-09	4.26e-03	2.46e-03	1.24e-08	6.11e-03	1.33e-02
2. BLCA	1.09e-01	3.09e-01	6.74e-02	3.13e-01	4.95e-01	7.42e-02	3.57e-01	3.43e-02	1.95e-02
3. BRCA	1.19e-01	4.95e-03	2.93e-02	2.58e-01	3.07e-02	3.90e-01	3.80e-04	2.53e-01	1.96e-03
4. CESC	5.10e-01	1.90e-01	3.33e-01	1.81e-01	1.69e-01	2.90e-01	6.69e-01	8.89e-01	2.95e-02
5. CHOL	5.72e-01	3.35e-01	3.02e-01	5.17e-01	6.51e-01	6.93e-01	4.50e-01	9.63e-01	3.01e-02
6. COAD	1.28e-01	2.52e-01	6.76e-01	3.73e-01	6.47e-01	5.05e-01	6.20e-01	5.35e-01	1.44e-03
7. COADREAD	6.60e-01	1.35e-01	8.11e-01	4.72e-02	2.55e-01	7.47e-01	7.87e-01	4.76e-01	2.89e-03
8. DLBC	7.55e-01	7.44e-01	3.53e-01	9.82e-01	7.42e-01	8.94e-01	8.15e-01	7.28e-01	4.74e-01
9. ESCA	3.92e-01	3.91e-01	3.92e-01	5.01e-01	3.75e-01	1.71e-01	2.25e-01	4.90e-01	3.30e-01
10. GBM	2.08e-02	8.11e-02	1.49e-04	5.12e-01	1.24e-01	5.37e-01	3.69e-01	7.04e-01	8.75e-05
11. GBMLGG	4.75e-14	6.36e-10	2.31e-17	6.46e-16	8.66e-12	8.04e-14	3.83e-07	1.25e-10	7.48e-17
12. HNSC	3.66e-01	6.19e-01	7.41e-05	2.44e-01	1.42e-01	3.27e-01	9.88e-01	1.55e-01	4.56e-02
13. KICH	7.01e-01	4.63e-01	8.14e-14	0.00e+00	4.03e-01	2.10e-01	8.08e-01	6.61e-01	2.77e-02
14. KIPAN	2.11e-07	9.84e-05	4.81e-08	4.04e-13	2.16e-08	4.21e-08	3.82e-03	4.36e-04	1.16e-11
15. KIRC	6.91e-01	9.79e-01	2.46e-01	1.76e-01	6.70e-01	1.76e-01	1.32e-01	7.29e-01	5.98e-05
16. KIRP	5.33e-03	1.85e-02	8.42e-18	1.00e+00	4.60e-02	5.97e-03	2.49e-02	1.93e-01	1.15e-09
17. LAML	1.73e-03	1.24e-02	5.14e-04	7.00e-01	9.38e-01	1.19e-01	1.75e-02	7.78e-02	8.72e-04
18. LGG	1.60e-14	7.14e-15	1.17e-17	3.52e-01	6.08e-03	1.01e-01	1.16e-09	4.04e-02	4.26e-15
19. LIHC	3.34e-01	1.28e-01	1.09e-03	8.25e-01	2.57e-01	2.93e-01	5.04e-01	8.80e-01	7.04e-01
20. LUAD	5.01e-01	3.73e-01	7.51e-03	5.92e-01	2.55e-02	1.49e-01	2.08e-01	8.21e-03	4.66e-01
21. LUSC	8.71e-02	3.91e-02	1.32e-01	7.04e-01	5.11e-01	9.05e-01	2.88e-01	6.75e-01	8.37e-03
22. MESO	4.24e-04	1.72e-02	7.94e-04	7.29e-02	8.66e-05	2.77e-01	5.53e-04	3.85e-04	7.34e-04
23. OV	4.45e-01	5.88e-01	6.95e-01	9.73e-01	4.35e-01	6.47e-01	7.78e-01	9.60e-01	6.81e-01
24. PAAD	7.36e-04	2.03e-03	1.44e-03	2.96e-03	4.19e-03	4.86e-04	3.18e-01	3.45e-02	2.73e-04
25. PCPG	3.32e-01	4.57e-01	2.57e-01	3.11e-01	3.39e-01	1.41e-01	6.63e-01	7.67e-01	8.66e-01
26. PRAD	4.75e-01	6.95e-01	6.61e-01	9.56e-01	3.73e-01	4.97e-01	7.07e-01	3.90e-01	3.49e-01
27. READ	7.62e-01	3.35e-01	6.27e-01	1.00e+00	5.68e-01	2.72e-01	3.53e-01	3.41e-01	2.35e-02
28. SARC	4.37e-02	5.58e-02	7.23e-02	3.37e-02	3.07e-01	6.36e-01	9.54e-02	2.83e-01	3.03e-02
29. SKCM	4.78e-01	7.54e-05	6.37e-04	4.30e-03	4.67e-03	3.92e-02	1.90e-01	1.48e-03	1.05e-01
30. STAD	4.07e-02	5.11e-01	1.02e-01	4.83e-01	6.25e-01	3.08e-01	3.16e-01	5.55e-01	1.86e-04
31. STES	1.57e-01	3.41e-02	1.18e-01	4.97e-01	4.13e-03	5.92e-01	6.35e-02	8.45e-02	1.51e-02
32. TGCT	8.38e-01	8.39e-01	8.38e-01	5.89e-01	2.96e-01	3.74e-01	5.65e-01	5.41e-01	5.31e-01
33. THCA	6.20e-01	3.62e-03	3.87e-02	5.11e-01	7.42e-01	5.51e-01	3.87e-01	1.75e-02	8.82e-02
34. THYM	9.69e-02	1.15e-01	7.11e-02	8.89e-05	7.06e-02	5.96e-01	5.47e-02	1.38e-01	1.33e-02
35. UCEC	1.81e-02	1.70e-01	1.64e-01	6.88e-01	1.65e-01	8.61e-01	1.58e-02	3.02e-03	4.83e-03
36. UCS	8.59e-01	3.59e-01	7.16e-01	1.68e-01	8.76e-01	8.34e-01	5.85e-01	6.27e-01	4.26e-01
37. UVM	1.67e-04	5.80e-04	1.67e-04	5.50e-01	9.19e-02	4.92e-03	2.06e-04	2.20e-05	6.43e-03
38. M_Discovery	2.26e-05	3.15e-12	1.16e-11	2.87e-01	9.16e-01	4.32e-06	4.59e-10	2.01e-07	3.25e-10
39. M_Validation	1.04e-02	4.68e-06	2.75e-07	1.57e-01	1.97e-01	1.28e-01	7.46e-04	9.16e-04	2.66e-05
#Significant	15	15	19	9	11	8	12	14	28

Cells highlighted in yellow have significant Cox p-values at the threshold of 5%. Cells highlighted in green have the most significant Cox p-value in their respective rows. No methods were able to yield subtypes with significantly different survival in 7 data sets (shown with red fonts). SMRT yields subtypes with significantly different survival profiles in 28 out of the 39 datasets. In 12 such datasets, SMRT also p-values more significant than any of those provided by the other eight methods.

To reveal the contribution of each data type, we used SMRT to partition the patients using each of the data types independently. Next, we calculated the Cox p-values obtained from each data type and compared them with those obtained from subtyping the multi-omics data. **Figure 3** shows the distribution of  $-\log_{10}$  p-values of subtypes by each data type for 37 TCGA datasets. The p-values obtained from multi-omics data are substantially more significant than those obtained from individual data types. The median p-value obtained from multi-omics data is close to 0.01 ( $-\log_{10}$  values are close to 2) while the median p-values of each data type are even higher than 0.1 ( $-\log_{10}$  values are close to 1). This demonstrates that SMRT is able to exploit the complementary information available in each data type to determine subtypes with significant survival differences. **Supplementary Section 10** and **Table S15** provide more details on the contribution of individual data types in each dataset.

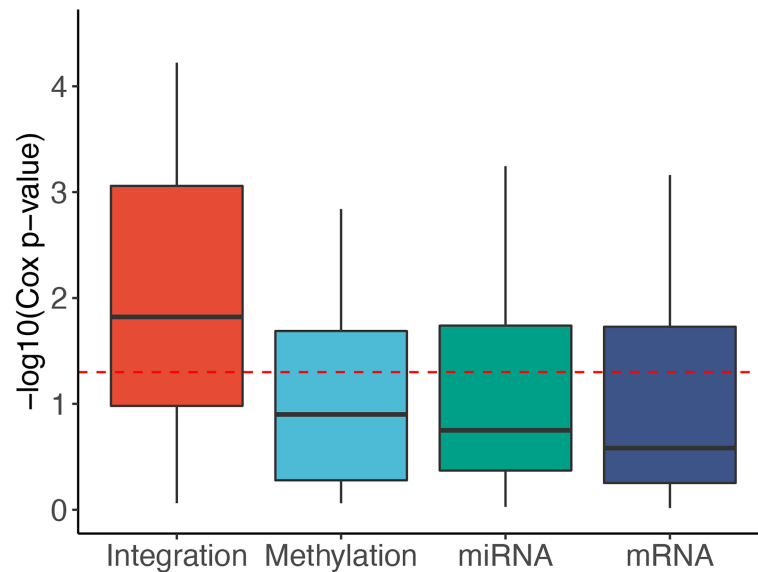
Next, we investigated the association between discovered subtypes and clinical variables. We performed our analysis on gender, age, cancer stage, and tumor grade, which are available for at least 15 datasets. We perform the following analyses: (1) Fisher's exact test to assess the significance of the association between gender (male and female) and the discovered subtypes; (2) ANOVA to assess the age difference between discovered subtypes; and finally (3) calculate the agreement between the discovered subtypes and known cancer stages and tumor grades using Normalized Mutual Information (NMI). The distributions of  $-\log_{10}$  of p-values for gender and age are shown in **Supplementary Figure S8** (see **Supplementary Tables 11-12** for the exact p-values). With the exception of NEMO and iClusterBayes, the clustering methods do not generally yield differences in gender or age in their clustering. For gender, iClusterBayes has significant p-values in 17 out of 31



**FIGURE 2** | Distributions of Cox p-values (in  $-\log_{10}$  scale, higher is better) of the subtypes discovered from 37 TCGA and 2 METABRIC datasets. The red dashed line shows the 5% significance level. Note that all existing methods do not reach this level of significance on average (median). Overall, the Cox p-values obtained from SMRT are substantially more significant than those of other methods ( $p = 0.0002$  using the one-tailed Wilcoxon test).

**TABLE 2** | Running time (in minutes) of SNF, CIMLR, NEMO, moCluster, iClusterBayes (iCB), LRACluster (LRA), MCCA, IntNMF, and SMRT for 37 TCGA and two METABRIC datasets.

Dataset	Size	SNF	CIMLR	NEMO	moCluster	iCB	LRA	MCCA	IntNMF	SMRT
1. ACC	79	0.40	1.14	0.05	0.97	9.09	5.58	0.50	6.64	0.25
2. BLCA	404	0.73	3.71	0.28	7.85	29.57	34.92	0.83	21.94	1.30
3. BRCA	622	1.61	9.44	0.75	24.09	56.39	102.13	1.61	40.07	1.53
4. CESC	304	1.01	3.23	0.28	8.78	30.49	50.41	1.20	20.66	0.90
5. CHOL	36	0.33	0.60	0.02	0.38	5.23	2.02	0.53	4.77	0.10
6. COAD	220	0.93	1.84	0.20	5.28	23.77	30.81	1.07	16.44	0.67
7. COADREAD	294	0.98	4.41	0.30	9.14	29.81	40.10	1.17	21.07	0.96
8. DLBC	47	0.37	0.61	0.03	0.52	6.25	2.66	0.44	4.90	0.16
9. ESCA	183	0.75	2.44	0.14	4.45	16.91	27.54	0.84	12.93	1.20
10. GBM	273	0.05	2.15	0.02	0.46	20.30	1.02	0.19	15.03	0.91
11. GBMLGG	510	0.89	5.33	0.40	11.61	44.30	41.47	0.97	31.08	1.43
12. HNSC	228	0.84	2.24	0.18	5.41	16.32	32.22	1.06	13.51	0.77
13. KICH	65	0.37	1.13	0.03	0.70	5.93	3.47	0.47	4.93	0.33
14. KIPAN	654	1.14	13.77	0.49	14.90	41.54	63.67	1.16	31.39	3.51
15. KIRC	124	0.04	1.14	0.01	0.15	8.53	0.65	0.09	7.76	0.16
16. KIRP	271	0.61	3.93	0.15	3.96	16.85	18.91	0.70	15.96	0.94
17. LAML	164	0.04	1.57	0.01	0.20	10.84	0.68	0.10	8.13	0.13
18. LGG	510	1.29	7.60	0.60	13.95	33.18	83.92	1.37	28.77	1.76
19. LIHC	366	0.80	3.81	0.28	6.54	23.33	34.19	0.94	20.12	0.84
20. LUAD	428	0.81	4.42	0.28	7.95	34.64	39.17	1.02	29.77	1.26
21. LUSC	110	0.04	1.15	0.00	0.11	7.83	0.46	0.09	6.40	0.12
22. MESO	86	0.42	0.85	0.03	0.88	7.67	5.40	0.60	6.98	0.26
23. OV	286	0.36	2.37	0.10	3.14	19.37	16.24	0.53	16.99	0.72
24. PAAD	178	0.46	1.96	0.08	2.23	11.72	12.25	0.67	8.86	0.98
25. PCPG	179	0.55	2.35	0.12	2.52	15.98	14.51	0.64	11.79	0.52
26. PRAD	493	1.51	6.13	0.54	12.52	33.67	79.05	1.29	32.18	1.75
27. READ	74	0.39	0.86	0.03	0.64	6.32	4.24	0.59	5.88	0.22
28. SARC	257	0.54	3.07	0.14	3.29	18.00	17.82	0.63	12.64	1.40
29. SKCM	439	0.83	6.51	0.34	7.71	27.58	35.17	0.78	23.61	1.76
30. STAD	362	0.87	5.07	0.33	5.77	24.99	34.14	0.89	18.61	1.07
31. STES	545	1.55	8.79	0.53	14.11	37.81	88.00	1.22	28.85	1.85
32. TGCT	134	0.85	1.79	0.10	2.01	10.61	18.49	0.93	7.01	0.41
33. THCA	499	1.06	5.90	0.46	8.85	33.01	53.59	0.92	25.35	1.66
34. THYM	119	0.49	0.97	0.07	1.18	8.78	9.76	0.52	7.16	0.28
35. UCEC	234	1.04	2.57	0.19	4.60	19.61	34.42	1.08	14.78	0.88
36. UCS	56	0.47	0.64	0.04	0.49	6.18	3.92	0.62	4.58	0.19
37. UVM	80	0.41	0.73	0.04	0.61	7.91	5.27	0.60	6.25	0.24
38. M_Discovery	997	0.38	17.96	0.21	7.10	60.24	16.17	0.38	49.62	2.42
39. M_Validation	983	0.37	10.14	0.19	6.85	58.11	17.95	0.40	50.87	2.28
Mean	305	0.68	3.96	0.21	5.43	22.53	27.75	0.76	17.80	0.98



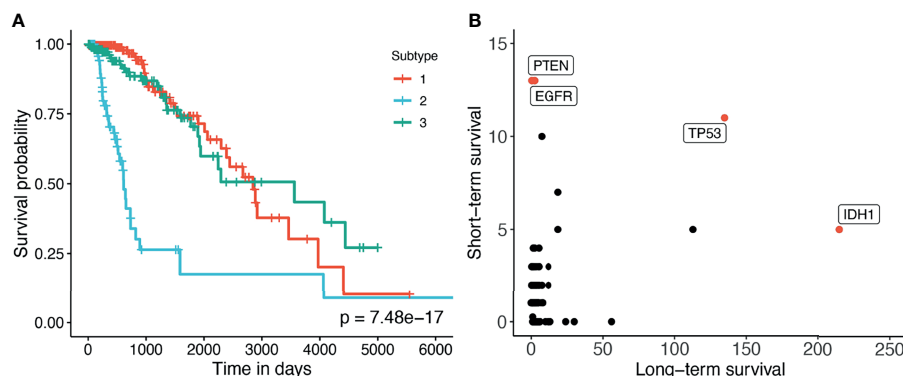
**FIGURE 3** | Distribution of  $-\log_{10}$  Cox  $p$ -values for each data type of the 37 TCGA datasets. The horizontal red line indicates the significant threshold of  $p$ -value = 0.05. The  $p$ -values of subtypes discovered using multi-omics integration are substantially more significant than those obtained from individual data types (mRNA, methylation, miRNA).

datasets. For age, NEMO and iClusterBayes have significant  $p$ -values in 17 and 15 out of 29 datasets, respectively. This result demonstrates that there are meaningful and survival-related molecular signatures inside the data to be discovered, and the methods do not simply separate patients based on some visible clinical variables such as gender or age. **Supplementary Figure S9** and **Supplementary Tables 13, 14** show the NMI values that represent the agreement between the discovered subtypes and known cancer stages and tumor grades. For the cancer stage, the median NMI values of SMRT and NEMO are comparable and are higher than the rest. For tumor grade, SMRT has the highest median NMI. However, for both cancer stage and tumor grade, the NMI values of all methods are low, meaning that there is a

low agreement between the known stages/grades and the discovered subtypes using any of the subtyping methods. In conclusion, the discovered subtypes from SMRT and other subtyping methods have little agreement with clinical variables like gender, age, cancer stage, and tumor grade.

### 3.2 Case Study of the GBMLGG Dataset

Here we perform an in-depth analysis for the GBMLGG (Glioma). **Figure 4A** shows the Kaplan–Meier survival analysis of the discovered subtypes. For this dataset, SMRT discovers three subtypes in which one subtype (group 2) has a very low survival rate where at year 3, the survival probability of patients this group is only at 26% while that number for the patients in



**FIGURE 4** | **(A)** Kaplan–Meier survival analysis of the GBMLGG dataset. The horizontal axis represents the time (days) while the vertical axis represents the estimated survival probability. **(B)** Number of patients in each group for each mutated gene in GBMLGG dataset. The horizontal axis shows the count for other subtypes with high survival rates, and the vertical axis represents the count in the subtype with low survival rates.



the other two subtypes (groups 1 and 3) is 84%. We also perform a variant analysis for the dataset in order to find mutations that highly occur in the short-term-survival patient group (group 2) but not in the long-term-survival patient group (groups 1 and 3) and vice versa. **Figure 4B** shows the mutations of each group in which each point is a gene, and its coordinates represent the number of patients that have that mutation in the corresponding group. In principle, we want to investigate the mutated genes in the top left or bottom right of the figure. In this figure, we can easily identify four marker genes that associate with GBMLGG disease: IDH1, TP53, PTEN, and EGFR. Among those, IDH mutant (bottom-right) is known as a factor driving Low Grade Glioma (LGG) and has been used in the WHO classification system (40) to classify IDH-mutant and IDH-wildtype, which has worse prognoses. On the other hand, EGFR is not a common mutation in LGG but in GBM (Glioblastoma) (41) which has a very low survival rate (42). The amplification of EGFR can cause the mutation of PTEN gene (43) which is a tumor suppressor gene (44). Interestingly, no patient in the long-term-survival group has PTEN mutation. The occurrence of EGFR mutated genes may be another cause that leads to a low survival rate of patients in the short-term-survival group.

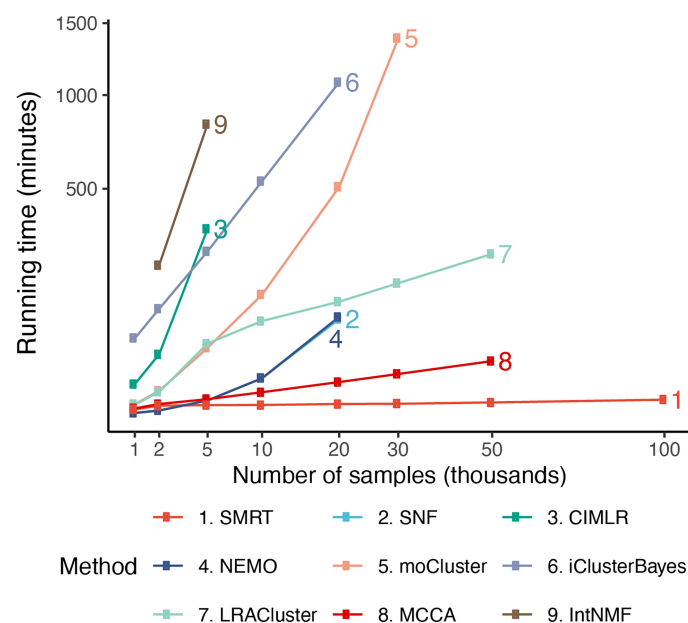
We further conduct pathway analysis using the discovered subtypes on the Consensus Pathway Analysis platform (45) using the FGSEA method (46) and KEGG pathway database. **Supplementary Figure S4** shows the pathways that are significant with a significance threshold of 0.5%. In this connected network, each node is a pathway and there is an edge between two pathways if they have common genes. As shown in the figure, the Glioma pathway is significantly

impacted. Other pathways that have common components with the Glioma pathway, including MAPK signaling pathway, ErbB signaling pathway, Calcium signaling pathway, and Pathway in cancer, are also significantly impacted. This confirms that the subtypes discovered by SMRT have significant differences in the activity of Glioma- and cancer-related pathways. **Supplementary Section 2** and **Figures S1–S4** provide a more detailed analysis of this dataset.

### 3.3 Scalability of the Subtyping Methods

In order to assess the scalability of the nine subtyping methods, we generate a number of simulated datasets with a fixed number of genes/features of 5,000 and varying numbers of samples (from 1,000 to 100,000). In each dataset generated, there are three classes of samples – each with a different set of up-regulated genes. The true class information was used *a posteriori* to assess the accuracy of each clustering method. The memory of our server is limited to 376 GB.

**Figure 5** shows the running time of the methods with varying numbers of samples. The time complexity of SNF, CIMLR, NEMO, and moCluster increases exponentially with respect to sample size. These methods are not able to analyze datasets with more than 30,000 samples (out of memory, produce errors, or take more than 24 hours to analyze a single dataset). MCCA and LRACluster are able to analyze datasets with 50,000 samples but fail to analyze larger datasets. Only SMRT is able to analyze all large datasets, including those with 100,000 samples. SMRT is much faster than other methods and can analyze datasets with 100,000 samples in three minutes. See **Supplemental Section 3**, **Figure S5**, and **Tables 4, 5** for details on simulation and results.



**FIGURE 5** | Running time of the nine subtyping methods with respect to varying numbers of samples and features. SMRT is the only method that can analyze all datasets. Even for large datasets with 100,000 samples, SMRT needs only a couple of minutes to finish the analysis.

## 4 CONCLUSION

In this article, we introduced SMRT, a fast yet accurate method for data integration and subtype discovery. In an extensive analysis using 39 cancer datasets, we showed that SMRT outperformed other state-of-the-art methods in discovering novel subtypes with significantly different survival profiles. We also demonstrated that the method could accurately partition hundreds of thousands of samples in minutes with low memory requirements. At the same time, the provided web application will be extremely useful for life scientists who lack computational background or resources. Although the software was developed for the purpose of cancer subtyping, researchers in other fields can use the web application and R package for unsupervised learning and data integration.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <http://smrt.tinnguyen-lab.com/>.

## REFERENCES

1. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* (2012) 487:330–7. doi: 10.1038/nature11252
2. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. *New Engl J Med* (2016) 375:1109–12. doi: 10.1056/NEJMp1607591
3. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature* (2012) 486:346–52. doi: 10.1038/nature10983
4. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* (2015) 12: e1001779. doi: 10.1371/journal.pmed.1001779
5. 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
6. 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
7. Speicher NK, Pfeifer N. Integrating Different Data Types by Regularized Unsupervised Multiple Kernel Learning With Application to Cancer Subtype Discovery. *Bioinformatics* (2015) 31:268–75. doi: 10.1093/bioinformatics/btv244
8. 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* (2013) 110:4245–50. doi: 10.1073/pnas.1208949110
9. 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* (2018) 19:71–86. doi: 10.1093/biostatistics/kxx017
10. Coretto P, Serra A, Tagliaferri R. Robust Clustering of Noisy High-Dimensional Gene Expression Data for Patients Subtyping. *Bioinformatics* (2018) 34:4064–72. doi: 10.1093/bioinformatics/bty502
11. Ahmad A, Fröhlich H. Towards Clinically More Relevant Dissection of Patient Heterogeneity via Survival-Based Bayesian Clustering. *Bioinformatics* (2017) 33:3558–66. doi: 10.1093/bioinformatics/btx464

## AUTHOR CONTRIBUTIONS

HN and TN conceived of and designed the approach. HN, DT, and BT implemented the method in R, performed the data analysis and computational experiments. MR, AC, and SDa helped with data preparation and some data analysis. HN, DT, SD, and TN wrote the manuscript. All authors reviewed and approved the manuscript.

## FUNDING

This work was partially supported by NIH NIGMS under grant number GM103440, and by NSF under grant numbers 2001385 and 2019609.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2021.725133/full#supplementary-material>

12. Lock EF, Dunson DB. Bayesian Consensus Clustering. *Bioinformatics* (2013) 29:2610–6. doi: 10.1093/bioinformatics/btt425
13. 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
14. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and Individual Variation Explained (Jive) for Integrated Analysis of Multiple Data Types. *Ann Appl Stat* (2013) 7:523. doi: 10.1214/12-AOAS597
15. Meng C, Kuster B, Culhane AC, Gholami AM. A Multivariate Approach to the Integration of Multi-Omics Datasets. *BMC Bioinf* (2014) 15:162. doi: 10.1186/1471-2105-15-162
16. Meng C, Helm D, Frejno M, Kuster B. Mocluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *J Proteome Res* (2016) 15:755–65. doi: 10.1021/acs.jproteome.5b00824
17. Li W, Zhang S, Liu C-C, Zhou XJ. Identifying Multi-Layer Gene Regulatory Modules From Multi-Dimensional Genomic Data. *Bioinformatics* (2012) 28:2458–66. doi: 10.1093/bioinformatics/bts476
18. Yu N, Gao Y-L, Liu J-X, Shang J, Zhu R, Dai L-Y. Co-Differential Gene Selection and Clustering Based on Graph Regularized Multi-View NMF in Cancer Genomic Data. *Genes* (2018) 9:586. doi: 10.3390/genes9120586
19. Liu J, Wang C, Gao J, Han J. Multi-View Clustering via Joint Nonnegative Matrix Factorization. In: *Proceedings of the 2013 SIAM International Conference on Data Mining (SIAM)* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (2013). p. 252–60.
20. 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
21. Yang Z, Michailidis G. A Non-Negative Matrix Factorization Method for Detecting Modules in Heterogeneous Omics Multi-Modal Data. *Bioinformatics* (2016) 32:1–8. doi: 10.1093/bioinformatics/btv544
22. Zhang S, Liu C-C, 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
23. Witten DM, Tibshirani RJ. Extensions of Sparse Canonical Correlation Analysis With Applications to Genomic Data. *Stat Appl Genet Mol Biol* (2009) 8:28. doi: 10.2202/1544-6115.1470
24. 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) 11:333–7. doi: 10.1038/nmeth.2810

25. Yuan Y, Savage RS, Markowitz F. Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. *PloS Comput Biol* (2011) 7:e1002227. doi: 10.1371/journal.pcbi.1002227
26. Shi Q, Zhang C, Peng M, Yu X, Zeng T, Liu J, et al. Pattern Fusion Analysis by Adaptive Alignment of Multiple Heterogeneous Omics Data. *Bioinformatics* (2017) 33:2706–14. doi: 10.1093/bioinformatics/btx176
27. Huo Z, Tseng G. Integrative Sparse K-Means With Overlapping Group Lasso in Genomic Applications for Disease Subtype Discovery. *Ann Appl Stat* (2017) 11:1011. doi: 10.1214/17-AOAS1033
28. Rappoport N, Shamir R. NEMO: Cancer Subtyping by Integration of Partial Multi-Omic Data. *Bioinformatics* (2019) 35:3348–56. doi: 10.1093/bioinformatics/btz058
29. 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
30. 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
31. Tran D, Nguyen H, Le U, Bebis G, Luu HN, Nguyen T. A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis. *Front Oncol* (2020) 10:1052. doi: 10.3389/fonc.2020.01052
32. 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
33. Golub G, Kahan W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *J Soc Ind Appl Mathematics Ser B: Numerical Anal* (1965) 2:205–24. doi: 10.1137/0702016
34. Beygelzimer A, Kakadet S, Langford J, Arya S, Mount D, Li S. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R Package Version 1.1.3. Vienna, Austria: The Comprehensive R Archive Network (2019).
35. Ripley BD. *Modern Applied Statistics With s*. New York, NY, USA: Springer (2002).
36. Kaufman L, Rousseeuw P. Clustering by Means of Medoids. In: Y Dodge, editor. *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Amsterdam: North-Holland (1987). p. 405–16.
37. Langfelder P, Zhang B, Horvath S. Defining Clusters From a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R. *Bioinformatics* (2008) 24:719–20. doi: 10.1093/bioinformatics/btm563
38. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. *Shiny: Web Application Framework for R*. R Package Version 1.4.0.2. Vienna, Austria: The Comprehensive R Archive Network (2020).
39. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY, USA: Springer (2000).
40. Louis DN, Perry A, Reifemberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary. *Acta Neuropathologica* (2016) 131:803–20. doi: 10.1007/s00401-016-1545-1
41. Hao Z, Guo D. Egfr Mutation: Novel Prognostic Factor Associated With Immune Infiltration in Lower-Grade Glioma; an Exploratory Study. *BMC Cancer* (2019) 19:1–13. doi: 10.1186/s12885-019-6384-8
42. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, et al. Effects of Radiotherapy With Concomitant and Adjuvant Temozolomide Versus Radiotherapy Alone on Survival in Glioblastoma in a Randomised Phase III Study: 5-Year Analysis of the EORTC-NCIC Trial. *Lancet Oncol* (2009) 10:459–66. doi: 10.1016/S1470-2045(09)70025-7
43. Ohgaki H, Kleihues P. Genetic Pathways to Primary and Secondary Glioblastoma. *Am J Pathol* (2007) 170:1445–53. doi: 10.2353/ajpath.2007.070011
44. Ali IU, Schriml LM, Dean M. Mutational Spectra of Pten/Mmac1 Gene: A Tumor Suppressor With Lipid Phosphatase Activity. *J Natl Cancer Institute* (1999) 91:1922–32. doi: 10.1093/jnci/91.22.1922
45. Nguyen H, Tran D, Galazka JM, Costes SV, Beheshti A, Draghici S, et al. CPA: A Web-Based Platform for Consensus Pathway Analysis and Interactive Visualization. *Nucleic Acids Res* (2021) 49: gkab421. doi: 10.1093/nar/gkab421
46. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast Gene Set Enrichment Analysis. *BioRxiv* (2021), 060012. doi: 10.1101/060012

**Author Disclaimer:** 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.

**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.

Copyright © 2021 Nguyen, Tran, Tran, Roy, Cassell, Dascalu, Draghici and Nguyen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Gene Co-Expression in Breast Cancer: A Matter of Distance

Alfredo González-Espinoza<sup>1,2</sup>, Jose Zamora-Fuentes<sup>2</sup>, Enrique Hernández-Lemus<sup>2,3</sup>  
and Jesús Espinal-Enríquez<sup>2,3\*</sup>

<sup>1</sup> Department of Biology, University of Pennsylvania, Philadelphia, PA, United States, <sup>2</sup> Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, <sup>3</sup> Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

## OPEN ACCESS

### Edited by:

Maria Rodriguez Martinez,  
IBM Research—Zurich, Switzerland

### Reviewed by:

Sheng Liu,  
Indiana University, United States  
Noemi Eiro,  
Jove Hospital Foundation, Spain  
Kimberly Glass,  
Brigham and Women's Hospital and  
Harvard Medical School, United States

### \*Correspondence:

Jesús Espinal-Enríquez  
jespinal@inmegen.gob.mx

### Specialty section:

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

**Received:** 17 June 2021

**Accepted:** 26 October 2021

**Published:** 17 November 2021

### Citation:

González-Espinoza A,  
Zamora-Fuentes J,  
Hernández-Lemus E and  
Espinal-Enríquez J (2021)  
Gene Co-Expression in Breast  
Cancer: A Matter of Distance.  
Front. Oncol. 11:726493.  
doi: 10.3389/fonc.2021.726493

Gene regulatory and signaling phenomena are known to be relevant players underlying the establishment of cellular phenotypes. It is also known that such regulatory programs are disrupted in cancer, leading to the onset and development of malignant phenotypes. Gene co-expression matrices have allowed us to compare and analyze complex phenotypes such as breast cancer (BrCa) and their control counterparts. Global co-expression patterns have revealed, for instance, that the highest gene-gene co-expression interactions often occur between genes from the same chromosome (*cis*-), meanwhile inter-chromosome (*trans*-) interactions are scarce and have lower correlation values. Furthermore, strength of *cis*- correlations have been shown to decay with the chromosome distance of gene couples. Despite this *loss of long-distance co-expression* has been clearly identified, it has been observed only in a small fraction of the whole co-expression landscape, namely the most significant interactions. For that reason, an approach that takes into account the whole interaction set results appealing. In this work, we developed a hybrid method to analyze whole-chromosome Pearson correlation matrices for the four BrCa subtypes (Luminal A, Luminal B, HER2+ and Basal), as well as adjacent normal breast tissue derived matrices. We implemented a systematic method for clustering gene couples, by using eigenvalue spectral decomposition and the *k*-medoids algorithm, allowing us to determine a number of clusters without removing any interaction. With this method we compared, for each chromosome in the five phenotypes: a) Whether or not the gene-gene co-expression decays with the distance in the breast cancer subtypes b) the chromosome location of *cis*- clusters of gene couples, and c) whether or not the *loss of long-distance co-expression* is observed in the whole range of interactions. We found that in the correlation matrix for the control phenotype, positive and negative Pearson correlations deviate from a random null model independently of the distance between couples. Conversely, for all BrCa subtypes, in all chromosomes, positive correlations decay with distance, and negative correlations do not differ from the null model. We also found that BrCa clusters are distance-dependent, meanwhile for the control phenotype, chromosome location does not determine the clustering. To our knowledge, this is the first time that a dependence on distance is reported for gene

clusters in breast cancer. Since this method uses the whole *cis*- interaction geneset, combination with other -omics approaches may provide further evidence to understand in a more integrative fashion, the mechanisms that disrupt gene regulation in cancer.

**Keywords:** eigenvalue decomposition, gene co-expression clustering, loss of long-distance co-expression, co-expression matrices, breast cancer molecular subtypes

## 1 INTRODUCTION

### 1.1 Breast Cancer: A Complex Disease

Breast cancer is the first cancer-related cause of death in women worldwide. It is also, according to the most recent data (1), the most diagnosed neoplasm in the world. Breast cancer is also the malignant neoplasm with the highest incidence (1). Its diagnosis, response to treatment, relapse, and outcome are strongly determined by the molecular profile underlying the disease (2–4). The PAM50 classifier is among the most relevant methods of classification for breast cancer molecular subtypes (5). This molecular classification is based on the expression signature of 50 genes relevant to the oncogenic phenotype (5–7).

Publicly available massive cohorts of genomic and clinical data in the study of cancer, have allowed the analysis of an immeasurable amount of information. The latter has contributed to a better understanding of the oncogenic process (8). Based on gene expression of hundreds-to-thousands of samples, now it is possible to study such vast experimental information to infer and analyze the whole-genome co-expression landscape, aiming to highlight similarities and differences between cancer and non-cancer samples. Among these efforts, The Cancer Genome Atlas (TCGA) has contributed in an outstanding way (9).

### 1.2 Gene Co-Expression Networks

The study of Cancer within the framework of complex networks has become increasingly relevant in the last years (10–20). Given its size and complexity, genome-wide regulation may include a large number of features (all the genes), potentially inducing a fully connected network, with contributions of very different relevance and certainty. For this reason, several approaches to reduce its dimensionality have been implemented, including the use of threshold methods, to look for the most significant co-expression relationships (18, 21). In particular, in the case of breast cancer molecular subtype networks, the most significant co-expressed pairs have been used as connected nodes in biologically relevant modules (22–25).

Further approaches to determine the optimal network size may analyze a wide range of network scales (13, 26, 27) or backbone-related threshold networks (28), and even use gene co-expression subsets of clinical/biological relevance (29).

In the attempt of reducing the dimensionality of a fully-connected network, identification of groups of genes that behave in a similar way –indicating that their expression profiles are correlated– is a relevant problem and is still an open challenge in network biology (29, 30). The latter point is closely related to the

so-called graph sparsification problem in graph theory. The choice of a significance threshold then becomes relevant.

For instance, in a recent study by Kimura et al. (31), an approach was developed to select parameters in genetic networks by computational methods (mainly Machine Learning and Artificial Intelligence). Other approaches have used the complete set of interactions in order to construct a network backbone (28). There, the authors used the complete matrix of interactions to obtain the most important relationships, preserving those edges with statistically significant deviations with respect to a null model for the local edge's weight assignment.

### 1.3 Gene Co-Expression Is Distance Dependent

In cancer, gene co-expression networks have been used to uncover genes and relationships that may represent crucial elements to determine differences between phenotypes (32). In particular, in breast cancer and breast cancer molecular subtypes (4), gene co-expression networks have been useful to identify the phenomenon of *loss of long-range co-expression* (10, 12, 14, 33): this is, a property observed in cancer networks in which the most significant gene co-expression relationships occur between genes that belong to the same chromosome, i.e., *cis*- interactions. Conversely, inter-chromosome (*trans*-) interactions are often weak in cancer.

Furthermore, the loss of long-range co-expression is not only observed at the level of genes located on different chromosomes. Regarding *cis*- (intra-chromosome) gene interactions, there is an exponential decay of strength of correlations (14) as genes become more distant. This situation could be related to a diminishing of the *accessibility* that a certain region of the genome may have of its environment during the carcinogenic process. Importantly, this lack of accessibility can be attributed to several factors, among which we can mention aberrant expression of transcription factors, copy number alterations, incorrect binding to CTCF, or changes in Topologically Associating Domains (TADs). All of these factors have the potential to alter, both, the structure of DNA and gene expression.

Despite this phenomenon has been discovered not only in breast cancer, but also in clear cell renal carcinoma (13), lung adenocarcinoma and squamous cell lung carcinoma (12), loss of long-range co-expression has been determined for the top highest interactions: a small subset of the most co-expressed gene-gene interactions (tens-to-hundreds of thousands) of the



whole co-expression landscape is observed to be biased to *cis*- interactions.

Since the strength of intra-chromosome interactions have been observed to be the highest ones, it becomes important to evaluate the behavior of the whole intra-chromosome landscape of cancer networks. In these terms, network clustering may provide us with information related to, for example, sets of genes constrained by physical restrictions in certain regions of the genome, genes that act in tandem, events related with the transcriptional process, etc.

To address the questions above, we performed a data-driven clustering analysis using a hybrid algorithm that involves eigenvalue decomposition and *k*-medoids from correlation matrices of each chromosome. These matrices were inferred from RNA-Seq-based gene expression. We evaluated whether or not the loss of long-range co-expression is preserved, by studying all chromosomes for the four breast cancer subtypes as compared with normal tumor-adjacent tissue as control.

With this approach, we constructed co-expression matrices for all chromosomes in adjacent normal breast tissue network, as well as in all four breast cancer subtypes. We analyzed the statistics for their clustering nearest neighbor distributions within each chromosome, comparing each breast cancer molecular subtype as well as the adjacent normal tissue. Additionally, for all phenotypes, we constructed a null model to provide statistical robustness to our analyses. With this, we present a systematic method for intra-chromosome gene clustering, which allows to compare the whole co-expression landscape between a cancerous phenotype with its control counterpart.

## 2 MATERIALS AND METHODS

### 2.1 Data Acquisition

Gene expression data of breast invasive carcinoma was collected from The Cancer Genome Atlas (TCGA) (34). 735 tumor and 113 non-cancerous (adjacent normal), samples were considered, see **Table 1**. Illumina HiSeq RNASeq samples were filtered (biotype, expression mean >10), pre-processed, and  $\log_2$  normalized gene expression values as described in (10). We performed data corrections for transcript length, GC content and RNA composition. Tumor expression values were classified using PAM50 algorithm into the respective intrinsic breast cancer sub-types (Luminal A, Luminal B, Basal, and HER2-Enriched) using the Permutation-Based Confidence for Molecular Classification (35) as implemented in the pbcmc R package (36).

Tumor samples with a non-reliable breast cancer sub-type call were removed from the analysis. To avoid overlapping patterns

among subtype expression values, multidimensional noise reduction was performed using ARSyN R implementation (37), and a multidimensional Principal Component Analysis (PCA) was implemented to confirm noise reduction (14).

Since a crucial part of this work lies in having a highly-confident set of matrices, it is necessary to obtain as many well-characterized samples as possible, for each molecular subtype. Due to this fact, we decided to include all the available samples with a molecular subtype classification i.e., those samples with a molecular subtype label from the original source. Further investigations must be conducted with even more stringent inclusion and exclusion criteria, such as histologically confirmed diagnosis, histopathologically-assessed axillary lymph nodes, metastatic disease at presentation, adjuvant treatment, etc.

In order to provide all the information to reproduce our results, the clinical information about histological data by subtype-samples is now included in the **Supplementary Material S1**. There, for each breast cancer subtype sample we describe: 1) availability of historical adjuvant treatment, 2) lymph node assessment existence, 3) histological type of tumor and 4) axillary lymph-node-stage method type.

To show that those samples with the same molecular subtype are indeed properly classified in their molecular profiles to be included in our correlation matrices, we performed a Principal Component Analysis (PCA) for each subtype (**Supplementary Figure S1**). The PCA groups samples based on the main eigenvalues of the expression profiles. In this case, we present the two main principal components (X and Y axes of the **Supplementary Figure S1**) -though the calculations were made with the full eigenvalue spectra of the matrices. Hence, the PCA could indicate those samples that are not similar to the rest of their class (if any) or if there is any “confounded” or misclassified sample.

As it can be noticed in the **Supplementary Figure S1**, all subtype samples are clearly separated based on the molecular classification. All samples are grouped by its subtype (color). Hence, constructing correlation matrices by using these subtype-separated samples, certainly improves the statistical significance without adding a clear source of noise.

### 2.2 Correlation Matrices

We built intra-chromosomal cross-correlation matrices by estimating the Pearson correlation coefficient between the expression of two genes *i* and *j*, defined as follows:

$$C_{ij} = \frac{\text{Cov}(g_i, g_j)}{\sigma_{g_i} \sigma_{g_j}} = \frac{1}{N_s} \sum_{s=1}^{N_s} \frac{(g_{is} - \mu_{g_i})(g_{js} - \mu_{g_j})}{\sigma_{g_i} \sigma_{g_j}}, \quad (1)$$

where  $g_i$  is the set of  $N_s$  expression samples for gene *i*. By definition, a correlation matrix is symmetric ( $C_{ij} = C_{ji}$ ),

**TABLE 1** | Samples for each subtype.

Control	Basal	Her2	LumA	LumB
113	221	105	217	192

the elements in the diagonal are 1 ( $C_{ii} = 1, \forall i$ ), and its values are bounded to  $-1 \leq C_{ij} \leq 1$ , where  $C_{ij} = 1$  corresponds to perfect correlations,  $C_{ij} = -1$  corresponds to perfect anticorrelations, and  $C_{ij} = 0$  corresponds to uncorrelated gene pairs.

We calculated Pearson correlation between all genes for each chromosome for the five phenotypes. The code for calculation of Pearson correlations can be found in (38).

## 2.3 Spectral Decomposition

Pearson correlation matrices for each chromosome were calculated in order to analyze their spectral properties. Previous works on correlation matrices have shown that their spectral properties carry information about the structure and dynamics of the system (39–48).

For example, in stock market data, the first eigenvectors correspond to clusters of related industries (49, 50). In Electroencephalography measurements, these eigenvectors correspond to different functional regions in the brain (51). However, not all of the eigenvalues carry relevant information about the system. It has been shown that the smallest ones are the most sensitive to noise and some of them correspond to weak interrelations between small components from different clusters (47, 48). To distinguish how many eigenvalues contain useful information to identify clusters, we compared the spectral properties of the empirical correlation matrix to a null model represented by an ensemble of random matrices.

This ensemble of random matrices, is obtained by doing non-biased shuffling over the gene expression values for each sample (in this way, the original distribution of the data is preserved while its correlations will be destroyed) and computing the correlation matrix of each randomized data as in equation 1, we generated an ensemble of  $n_m = 100$  random matrices for each chromosome and phenotype.

The  $k$  deviating eigenvalues of the empirical matrix from the randomized data  $\max(\lambda_R) < \{\lambda_1, \dots, \lambda_k\}$  are the ones containing correlations that cannot be attributed to either the noise in the system or data randomization. It is worth noticing that instead of using the eigenvectors from the spectral decomposition, which can be difficult to separate into independent clusters (52) (see **Supplementary Figure S2**), we used the number of  $k$  deviating eigenvalues as the number of independent clusters for a different clustering method.

## 2.4 Clustering Analysis

We implemented a clustering analysis based on the *k-medoids* algorithm. In a similar fashion to *k-means*, *k-medoids* clustering attempts to minimize the distance between the elements inside a cluster but one element is designated as the center of the cluster. The *k-medoids* algorithm works not exclusively with Euclidean distances, but with general pairwise interactions, this means we can use the correlation values we have estimated for each intra-chromosome matrix. Since correlation values are signed and their magnitude goes from  $-1$  to  $1$ , we define the pairwise interactions between genes  $i$  and  $j$  as:

$$D_{ij} \equiv 1 - |C_{ij}|, \quad (2)$$

with  $0 \leq D \leq 1$ , high correlation or anti-correlation values mean close distance between points, while small correlation values will give higher distances. Finally, for the parameter  $k$  in the clustering algorithm, we considered the number of deviating eigenvalues as obtained from the spectral decomposition.

Given the stochastic nature of the *k-medoids* algorithm, we did  $n_r = 100$  realizations for each clustering computation to ensure statistical significance ( $p < 0.01$ ), choosing the output configuration as the one with the minimum mean distance between the centroids and the elements in each cluster.

In order to compare the clustering results between the control phenotype and any other cancer subtype in a given chromosome, we constructed the intra-cluster Nearest Neighbor Distance (NND) distribution for each subtype. The NND of a given gene  $i$  in a cluster  $k$  is defined as:

$$D_{nn}^i \equiv \min(|j - i|) \quad \forall j \in C_k, \quad (3)$$

where  $C_k$  refers to the cluster  $k$ . To quantify the difference between the clustering in adjacent normal and cancer subtypes we compute Shannon's entropy  $H(x) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$  for the NND distributions, which in this case can be interpreted as how localized or how spread are the genes within each cluster in the chromosome. We also computed the Kolmogorov-Smirnov distance between the adjacent normal case and each of the Cancer subtypes. Given two cumulative distribution functions (CDF) the Kolmogorov-Smirnov distance is defined as:

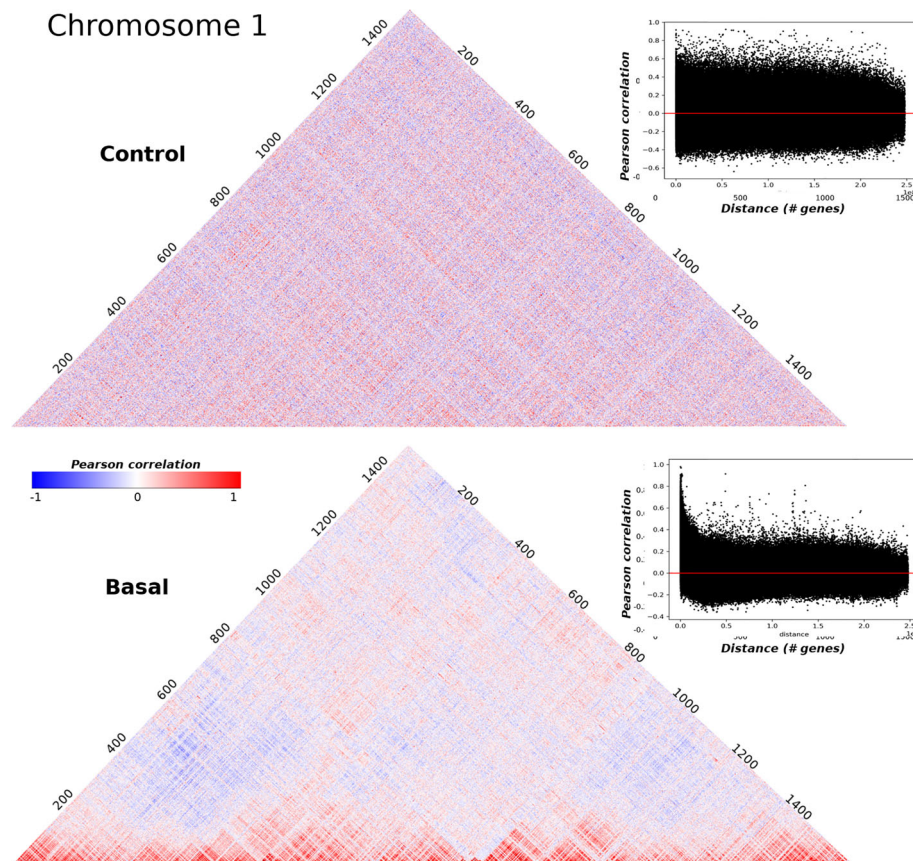
$$D_{KS}(F_n, F_m) = \sup_x |F_n(x) - F_m(x)|, \quad (4)$$

where the functions  $F_n$  and  $F_m$  are the CDFs for two samples  $n$  and  $m$ .

## 3 RESULTS

A correlation matrix of the sort just described, can be visualized as a heatmap as shown in **Figure 1** where correlation matrices for adjacent normal and basal subtype samples in the chromosome 1 are displayed. The axis represent the genes ordered by their physical location in the chromosome. The clearest difference between both matrices seems to be the lowest value of absolute correlation for genes that are physically distant in the basal subtype case. The heatmaps for each chromosome in the five phenotypes can be observed in **Supplementary Materials S2–S6**.

The effect of loss in long range co-expression is consistent with previous works of regulatory networks in breast cancer (10, 12–14, 33, 53). The block-type structure of the basal subtype matrix suggests the utility of clustering analysis to compare the structural properties of the correlation matrices. In what follows, we will present results for these clustering analyses. Through the manuscript, the presented figures will show different chromosomes for the five phenotypes. This has been done, in



**FIGURE 1** | Pearson correlation square matrices for chromosome 1 in control samples (up) and basal breast cancer subtype (down). Genes are placed according to their physical location on the chromosome. Colors represent the correlation value: red corresponds to positive values, meanwhile negative correlations are depicted in blue. The inserts in the right part of both matrices correspond to the scatterplot of Pearson correlations versus distance. The horizontal red line corresponds to Pearson correlation = 0.

order to illustrate the universal nature of the gene clustering in breast cancer molecular subtypes, compared with the adjacent normal tissue.

### 3.1 Co-Expression Decays in All Chromosomes in All Subtypes

We observed a common pattern of distance dependency in all chromosomes in all breast cancer phenotypes. The decay in gene co-expression corresponds exclusively to positive correlations. In the case of negative correlations, such effect is not observed. Conversely, in adjacent normal chromosomes, there is no dependency of distance neither in negative nor positive interactions. Interestingly, this effect is observed in all chromosomes in the four breast cancer molecular subtypes and not observed in adjacent normal breast tissue-derived correlations (**Figure 2** and **Supplementary Materials S7–S11**).

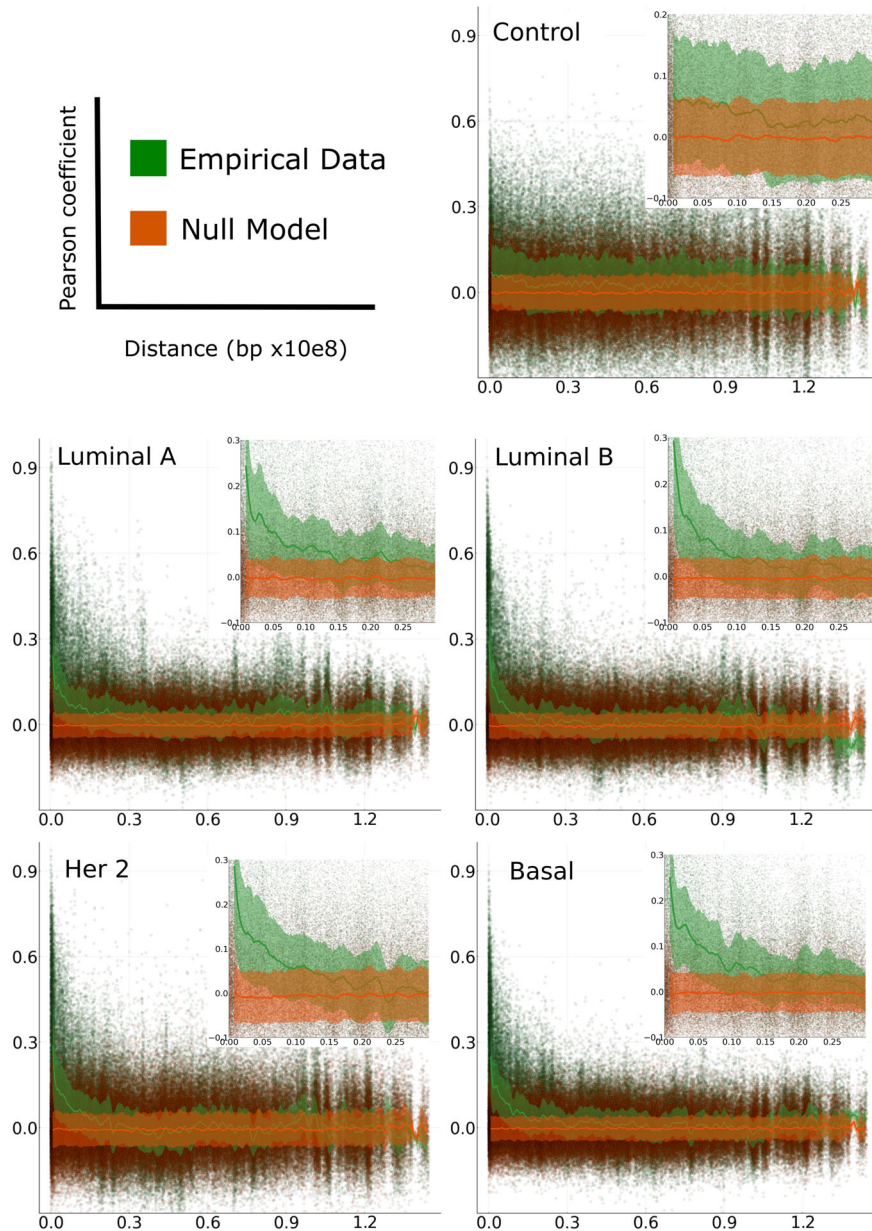
In order to evaluate the differences between the empirical data and the null model, we performed a non-parametric hypothesis test (Kolmogorov-Smirnov) for the correlation values distributions (in all tumor subtypes and adjacent normal

tissue) versus phenotype-specific null models. Additionally we implemented their corresponding significance tests (obtained *via* bootstrap/permutation analysis). The results of the KS test can be observed in **Figure 3**. The results for the rest of chromosomes, as well as their significance p-values, are presented in **Supplementary Materials S12, S13**).

Notice that at short distances, the cancer phenotypes have larger values than the adjacent normal correlations. However, at larger distances, KS for adjacent normal network are larger than those for cancerous phenotypes. The p-values shown in the upper right part of the figure, represent the average of all set of distances.

Based on a null model that lacks the linear correlations from the original data (see Methods), we observed that in adjacent normal chromosomes, positive and negative correlation values seem to be independent of the distance between genes, having significantly higher absolute values when compared with the null model at any distance. In the case of cancer subtypes, negative correlations are non-significant, but a few small regions in specific chromosomes (See **Supplementary Figure S3**).





**FIGURE 2** | Pearson correlation of gene-gene expression versus distance. Plots for adjacent normal and cancer subtypes of chromosome 8 (green) and their respective null model (orange). The solid lines represent the median of a moving average in the distribution of correlation values over each window and the shaded area is the range from its first and third quartiles.

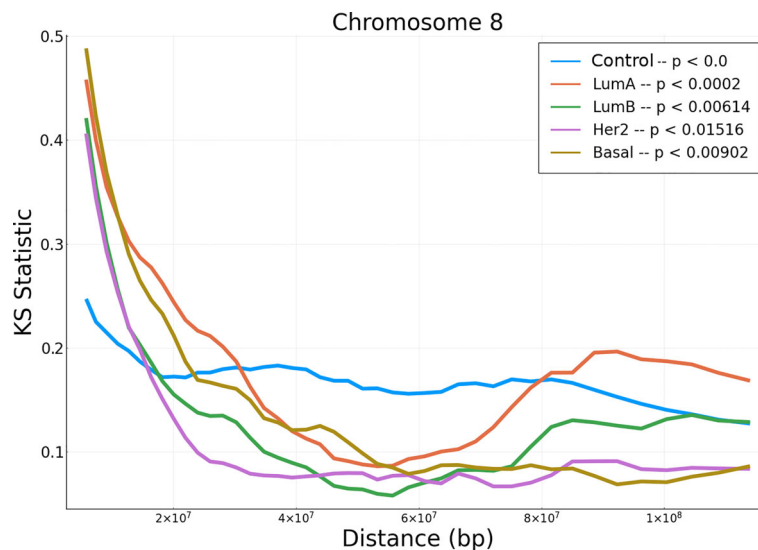
### 3.2 Eigenvalue Decomposition Defines the Number of Clusters in All Phenotypes

We generated an ensemble of ( $N = 100$ ) random matrices and compare the eigenvalue distributions from both, original and random matrices (see *Materials and Methods*). The left panel of **Figure 4** shows the eigenvalue distribution for the ensemble of random matrices, where its shape is the well-known Marchenko-Pastur distribution from random matrix theory (54). Overlapped eigenvalue distributions for the original matrix of chromosome

17 and the ensemble of surrogates are shown in the right panel of **Figure 4**. A set of significant eigenvalues was determined by random matrix permutations ( $p < 0.01$ ) (see box in **Figure 4**).

### 3.3 Gene Clustering Is Distance Dependent in Breast Cancer

With the method referred in Section 3.2 we obtained the full set of clusters for each chromosome in all phenotypes. **Figure 5**



**FIGURE 3** | KS hypothesis test between empirical data from chromosome 8 and null model for the five phenotypes. This plot represents the KS statistic *versus* distance for all phenotypes in chromosome 8. The  $p$ -value for the control phenotype is smaller than  $10^{-5}$ .

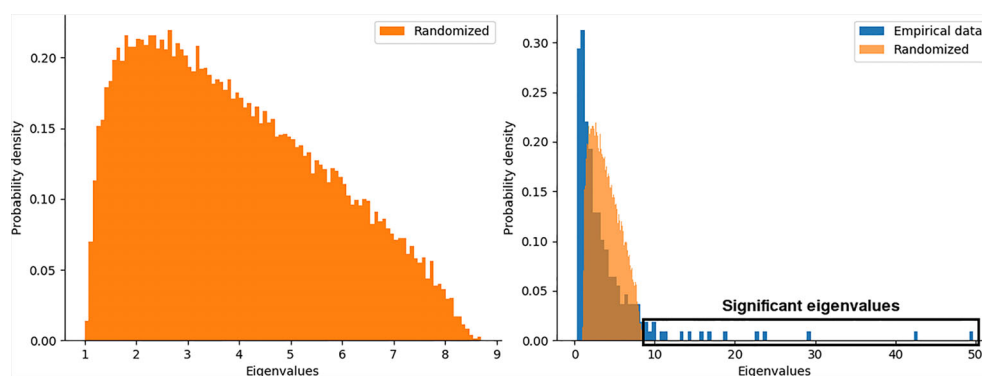
shows Chromosome 19 clusters with genes sorted by gene start base pair position. In the adjacent normal chr19 figure (upper part) we cannot discern a pattern in cluster colors. The distribution of clusters does not seem to depend on the distance between genes. Meanwhile, in basal breast cancer, we can observe cluster panels of colors, clearly detached. In the same figure, in the right panels, we plot the cumulative distribution of genes for each cluster. The larger the slope, the more often contiguous genes belong to the same cluster. All clusters for the five phenotypes in all chromosomes can be found in **Supplementary Material S14**. Cumulative distribution for all clusters can be found in **Supplementary Material S15**.

Cumulative distributions for the Nearest Neighbor Distance (NND) of two different chromosomes are shown in **Figure 6**, which can be interpreted as the probability distribution of the minimum distance between two genes in the same cluster. The

behavior seen in the previous **Figure 5** holds: genes from the same cluster are more likely to be close to each other.

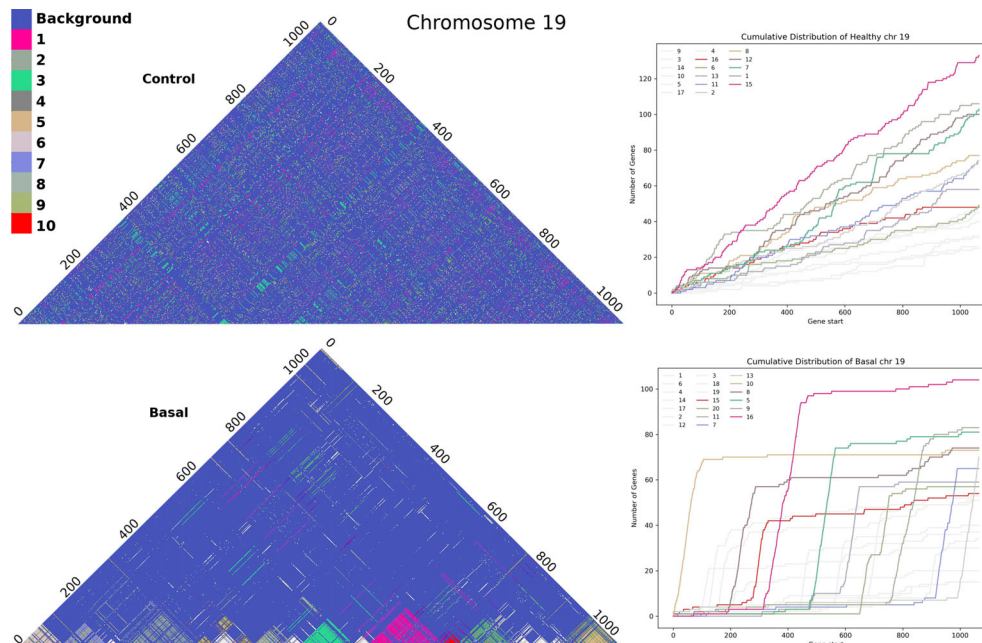
Results for the entropies for the NND distributions are shown on the left panel of **Figure 7**, where a clear trend with the value  $H(x)$  can be identified: Luminal A, Luminal B, HER2+, Basal. It is worth noticing that the aforementioned order coincides with survival rates and metastatic behavior (14, 55, 56). The subtypes with the lowest survival rates and more metastatic behavior also present lower entropy values.

The latter is in agreement with a previously observed trend for the top 0.1% gene co-expression interactions for the four phenotypes: The most aggressive phenotype (basal) has the lowest number of inter-chromosome interactions, meanwhile the Luminal A subtype, which is considered the one with the best prognosis, contains a much larger fraction of interactions between genes from different chromosomes (14).



**FIGURE 4** | Probability distributions of eigenvalues for a) the ensemble of random matrices, b) random and empirical data for the chromosome 17 in the Basal sub-type.





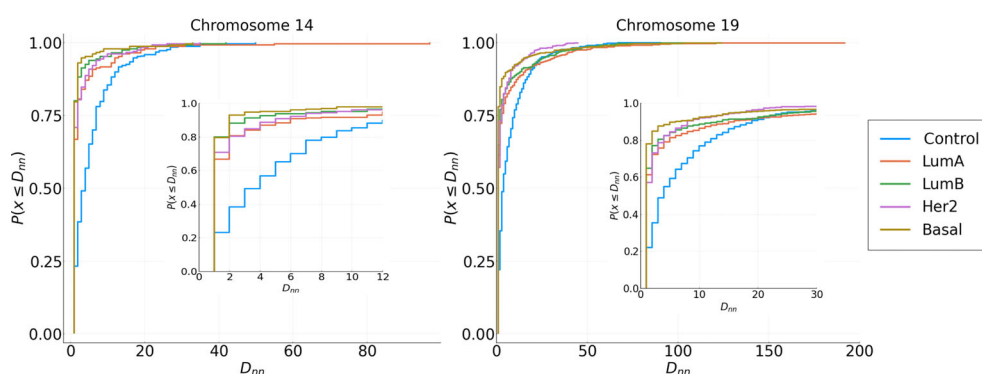
**FIGURE 5** | Cluster assembly of Chromosome 19 for adjacent normal-tissue and Basal breast cancer matrices. Upper part: Heatmap for all clusters in the adjacent normal matrix. Lower part: analog heatmap for Basal breast cancer network. The right panels correspond to the cumulative distribution for each cluster in chromosome 19. Colors represent the top-10 largest clusters.

The decay in the entropy for the NND distribution presents further evidence that in the cancer subtypes, genes co-express in tighter patterns, in contrast with genes in the control phenotype that co-express at broadly scattered distances over the chromosome. A similar trend holds for the KS distance between the control phenotype and each subtype in the right panel of **Figure 7**, where higher values indicate a larger difference in the spatial organization of the clusters. The difference in spatial organization within the clusters in the chromosome is evident with both measures and it is correlated with the survival rate and metastasis of the subtype (14).

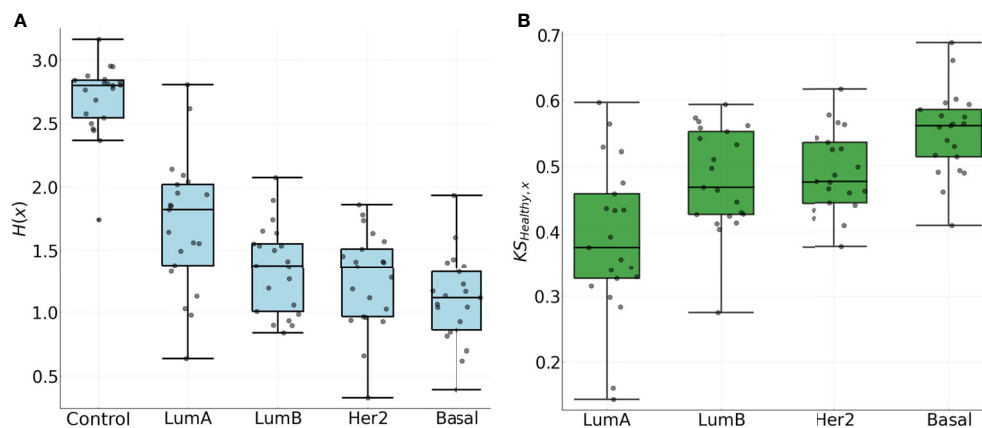
## 4 DISCUSSION

Cancer research increasingly requires comprehensive computational analysis tools. In the search for relevant biological information, it is essential to be able to find selective patterns of individual or collective gene expression. In this sense, clustering methods are becoming a pivotal computational tool.

In this work, we studied the co-expression of genes in breast cancer molecular subtypes. We implemented a method to find the optimal clustering between genes that are co-expressed. We observed a grouping pattern in the case of cancer phenotypes



**FIGURE 6** | Cumulative NND distributions for each subtype in Chromosomes 14 and 19. Distributions of Cancer subtypes have different behavior in short and long distances compared with the adjacent normal tissue.



**FIGURE 7 | (A)** Entropy for the NND distribution in all chromosomes. **(B)** Kolmogorov-Smirnov distance between the CDF of the NND in the adjacent normal phenotype and each of the Cancer subtypes for each chromosome.

with respect to the adjacent normal group. These patterns in the genome indicate that in cancer, physically close genes are co-expressed (*cis*- interactions), while for distant genes (*trans*-interactions) the clustered co-expression is, to a large extent, lost.

The piece-wise Kolmogorov-Smirnov tests for all tumor subtypes and adjacent normal tissue versus phenotype-specific null models and their corresponding significance tests (obtained *via* bootstrap/Permutation analysis) (included in the **Supplementary Materials S12, S13**), show that correlation values at short distances are much more significant for all chromosomes in any cancer phenotype than the adjacent normal network.

It is worth noticing that the significance of the KS tests also decays with the distance for all chromosomes in any cancer phenotype. Conversely, for the adjacent normal network, distance does not exert a considerable influence in the significance of the KS test. Finally, the KS test also show that the significance of differences between the correlation of our empirical data with the null model is unique for each chromosome and for each phenotype.

The fact that genes are highly co-expressed in groups with close positions, may be due to a favored number of nearby transcription sites or to the strong presence of transcription factors. It has been observed for instance, that in Luminal A breast cancer gene co-expression networks, co-factors, CTCF binding sites (57, 58) or copy number alterations (59, 60), may remodel chromatin making it more or less accessible, thus allowing gene transcription of local neighborhoods, resulting in the concomitant high co-expression between those neighboring genes (53). On the other hand, TFs influence more often inter-chromosome edges, meanwhile intra-chromosome interactions are less affected by them (53).

Physical interactions such as CTCF binding sites have captured attention in recent years (61, 62), and more importantly, in breast cancer (63).

For instance, in (53), we constructed an intra-chromosome gene-co-expression network for Luminal A breast cancer samples. There, a community detection method was performed to determine whether CTCF binding sites appeared in the borders of those communities. We observed that there is no link between CTCF binding sites and the border of intra-chromosome communities. In that sense, we argued that, at least for Luminal A breast cancer gene co-expression networks, CTCF binding sites are not determinant for network structure.

Transcription factor (TF) regulation is, of course, one of the central mechanisms for gene regulation. With respect to the role that TFs may exert on gene clustering, we have previously shown that TFs influence genes in a *trans*- fashion, i.e. TFs from a given chromosome regulate genes from different chromosomes. We have shown that in terms of Master Regulators in breast cancer (64, 65), but also in Luminal A breast cancer networks (53). Conversely, for intra-chromosome genes, TFs influence is much less evident.

Finally, Copy Number Variations (CNVs) have been considered as a crucial factor in the rise and development of breast cancer (59). In fact, a correlation between CNVs, protein levels and mRNA gene expression has also been reported previously (66). Hence, high correlations between clusters of physically closed genes appear to be related to copy number alterations.

We have used TCGA-derived CNV data and compared the amplification/deletion peaks with LumA network communities. Interestingly, the community with more overexpressed genes, composed of genes such as FOXM1, HJURP, or CENPA, presented large regions of deletions. The apparently contradictory result suggested that the copy number alterations do not influence the structure of that community. On the other hand, a gene community formed by HLA family genes, presented a common pattern of amplification, but those genes were not differentially expressed (53).

With the aforementioned in mind, we argue that CNVs are not as relevant as one could expect in terms of the gene clustering shown here. Moreover, CNVs influence may be at the expression level, but said effect is more limited regarding co-expression. However, further investigation is necessary to clarify these ideas.

The structure of clustered genes in physically close neighborhoods resembles the images obtained by Hi-C methods (67–69).

In recent times, there has been an increased interest as to how chromosome conformation capture experiments such as Hi-C may lead to relevant clues towards our understanding of further effects in connection with transcriptional regulation. Indeed, we are currently conducting research along these lines in our group. Work is ongoing, however, we can advance that there seems to be important correlations between loss/gain of statistically significant chromosomal contacts and co-expression relationships between genes in the associated genomic regions. It remains to be determined however whether said correlations are significant *via* proper assessment of null models and, more importantly, to determine what may be the biological consequences of these associations.

Preliminary findings from our Hi-C analysis in breast cancer indicate that more relevant contacts are mostly (but not exclusively) on close genomic regions. This is not unlike what we have observed with MI-based gene co-expression networks in which there is a preponderance of co-expression interactions in shorter distances for tumors. Future work undoubtedly will focus on the comparison between the network clusters constructed by this method and those from Hi-C. In particular, the zones/genes between gene groups. The assessment and comparison of both structures will provide us more information regarding the structural alterations during the carcinogenic process.

In brief, after revising the evidence about other mechanisms of gene regulation, we may hypothesize that the ultimate cause of the distance-dependent gene clustering is not a single mechanism, but instead, it could be a non-linear combination of different phenomena. In particular, regarding gene clustering, we have evaluated for the first time the whole set of gene interactions, and the loss of long-distance co-expression remains, which is more evident in the most aggressive subtypes.

Homogeneity/redundancy promotes higher entropy. Systems with redundancy are less likely to fail to catastrophic events. In other words, it seems there are mechanisms that give robustness to gene regulation in a control phenotype. It is still uncertain whether the loss of long range (or gain in short range) gene co-expression is a consequence of cancer, forcing the system to work in a less entropic configuration, but it seems that this preference for a less entropic configuration is common in all cancer subtypes and is consistently progressive with subtype aggressiveness.

As a summary of findings, we may establish the following:

- We used tools previously implemented in time series analysis in the stock market and neuroscience settings (49–51) to develop a systematic, data-driven method for intra-chromosomal gene expression clustering. Using spectral decomposition and a null model, we were able to determine

the number of co-expressed group of genes to perform *k*-medoids algorithm calculations and determine the most accurate clustering configuration. This method allowed us to have significant results, avoiding to set an *a priori* threshold for co-expression values.

- In the adjacent normal phenotype matrices, negative and positive correlations are significant throughout the entire chromosome. On the other hand, in breast cancer, negative correlations are observed in the same rank than those from a null model (see **Figure 2**); furthermore, the positive ones are only out of the null model cloud over short distances.
- In cancer, clustering mostly occurs between nearby genes, unlike what happens in the adjacent normal phenotype matrices. This is a representation of high co-expression over short distances. This fact coincides and corroborate previous results on mutual information-based co-expression networks in these and other types of cancer (10, 12–14).
- The intra-cluster Nearest Neighbor Distance (NND) clearly decays from the adjacent normal network to those cancerous ones. Additionally, the NND for breast cancer networks also decays according to the aggressiveness of the subtype: Luminal A, Luminal B, HER2+ and Basal.
- Analogously to the last point, Kolmogorov-Smirnov (KS) distance between the Cumulative Distribution Function of the NND in the adjacent normal and each breast cancer subtype network, increases with the aggressiveness of the subtype, thus indicating that the larger value of the KS distance, the larger difference between adjacent normal and breast cancer phenotypes' networks.

Clustered genes may be subject to further analyses to reveal, for instance, statistical enrichment of functional categories revealing certain biological functions, additional patterns of coordinated activity, etc. This in turn may lead to the generation of hypotheses to be tested *via* more narrowly targeted assays and interventions.

A closer look at matrices' patterns generated by other type of sorting methods may shed some light on possible mechanisms behind the regulatory changes in co-expression and perhaps even in the establishment of the tumor phenotypes. This is, indeed, still ongoing work.

Further steps towards the understanding of co-expression patterns and the differences in clustering among adjacent normal and cancerous phenotypes may be also based on the usage of multi-layer approaches (11, 70).

There are remaining questions prompted by this study. For example, while it is evident that there is a decay in the strength of correlations depending on the distance in all chromosomes, it is not fully clear what is the origin of the differences in the slope of the aforementioned decays. Also, the negative correlations in adjacent normal network are significant, independently of the position in the chromosome. Is the anti-correlation between genes a possible mechanism of negative feedback? Is that mechanism disrupted in breast cancer? Another important question regarding the clustering in cancer network is the size of the clusters. Is there an "optimal" cluster

size for cancer networks? If so, what is the rationale behind such number?

Finally, the fact that other types of cancer, which have been analyzed in terms of gene co-expression interactions, such as clear cell renal carcinoma (13), lung adenocarcinoma, or lung squamous cell carcinoma (12) have been reported to have the same bias in short-distance interactions, a remaining question is whether the clustering behavior observed in breast cancer subtype networks is a conserved phenomenon along other cancer types.

The above mentioned questions, together with the acquired knowledge on cancer networks, will be eventually answered and that will bring us with complementary information to have a broader point of view on gene regulation in cancer.

## 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 below: <https://github.com/josemaz/gene-matrices>.

## AUTHOR CONTRIBUTIONS

AG-E and JZ-F performed computational analyses, developed methods and implemented programming code, performed pre-processing and low-level data analysis, participated in the writing of the manuscript. EH-L contributed to the design of the study, co-supervised the project, contributed and supervised the writing of the manuscript. JE-E conceived and designed the project, performed calculations, supervised the project, drafted the manuscript. All authors contributed to the article and approved the submitted version. AG-E and JZ-F contributed equally as first author.

## FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México.

## ACKNOWLEDGMENTS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2021.726493/full#supplementary-material>

**Supplementary Figure 1** | Principal component analyses of RNA-seq data clustered by breast cancer subtypes.

**Supplementary Figure 2** | Comparison of the squared components of the four largest eigenvectors from the correlation matrix **b** in **Figure 1**. It can be seen that there is an overlap between the eigenvectors that does not allow to separate the components into clusters.

**Supplementary Figure 3** | Outliers of positive correlations (red dots) in some chromosomes of the breast cancer subtype networks. Plots for chromosomes 1, 9, 19 and 17 for the four subtypes. Arrows indicate sets of positive correlations considered outliers. Notice that the outliers in each plot form almost vertical lines, indicating that those interactions present approximately the same distance. Additionally we can see a null model overlapping (blue).

**Supplementary Material S1** | Excel file containing cross tables between subtype-samples and histological variables.

**Supplementary Material S2** | Heatmaps of Pearson correlation for each chromosome in the adjacent normal phenotype. The color code is the same than in **Figure 1**.

**Supplementary Material S3** | Heatmaps of Pearson correlation for each chromosome in the Luminal A phenotype.

**Supplementary Material S4** | Heatmaps of Pearson correlation for each chromosome in the Luminal B phenotype.

**Supplementary Material S5** | Heatmaps of Pearson correlation for each chromosome in the HER2+ phenotype.

**Supplementary Material S6** | Heatmaps of Pearson correlation for each chromosome in the Basal phenotype.

**Supplementary Material S7 to S11** | Pearson distribution scatter plots for normal adjacent tissue, Basal HER2+, Luminal A and Luminal B, respectively. These plots show correlations sorted by gene start position for the four cancer phenotypes and the adjacent normal network per chromosome.

**Supplementary Material S12** | Kolmogorov-Smirnov significance tests for all chromosomes in the five phenotypes. As in **Figure 3**, the figures represent the KS statistics of 50 equal-area (same number of data-points) intervals for each chromosome. In the figures, the phenotypes are represented by different colors. The associated p-value for the complete set of correlations are depicted in the upper right part of the figures.

**Supplementary Material S13** | Piece-wise permutation p-values of the KS statistics, calculated for all bins obtained in **Supplementary Material S8**, in every chromosomal region for each phenotype.

**Supplementary Material S14** | Clusters by chromosome for the five phenotypes, obtained by eigenvalue decomposition and k-medoids method. The figures are depicted as in the manuscript. Additionally, this material contains files for clusters including the name of the gene, the cluster that the gene belong to, the assignment cost function value, the chromosome location of the gene, and the gene start position of said gene.

**Supplementary Material S15** | Cumulative distribution for the four cancer phenotypes and the adjacent normal network per chromosome. Color code coincides with clusters in **Supplementary Material S7**. For clarity, only the top-ten clusters were colored.



## REFERENCES

- Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2020. *CA: A Cancer J Clin* (2020) 70:7–30. doi: 10.3322/caac.21590
- Kittaneh M, Montero AJ, Glück S. Molecular Profiling for Breast Cancer: A Comprehensive Review. *Biomarkers Cancer* (2013) 5:BIC–S9455. doi: 10.4137/BIC.S9455
- Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, et al. The Prognostic Role of a Gene Signature From Tumorigenic Breast-Cancer Cells. *N Engl J Med* (2007) 356:217–26. doi: 10.1056/NEJMoa063994
- de Anda-Jáuregui G, Velázquez-Caldelas TE, Espinal-Enriquez J, Hernández-Lemus E. Transcriptional Network Architecture of Breast Cancer Molecular Subtypes. *Front Physiol* (2016) 7:568. doi: 10.3389/fphys.2016.00568
- Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular Portraits of Human Breast Tumours. *Nature* (2000) 406:747–52. doi: 10.1038/35021093
- Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses With Clinical Implications. *Proc Natl Acad Sci* (2001) 98:10869–74. doi: 10.1073/pnas.191367098
- Guedj M, Marisa L, De Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A Refined Molecular Taxonomy of Breast Cancer. *Oncogene* (2012) 31:1196–206. doi: 10.1038/onc.2011.301
- Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst* (2019) 9:24–34.e10. doi: 10.1016/j.cels.2019.06.006
- Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-Of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors From 33 Types of Cancer. *Cell* (2018) 173:291–304.e6. doi: 10.1016/j.cell.2018.03.022
- Espinal-Enriquez J, Fresno C, Anda-Jáuregui G, Hernández-Lemus E. RNA-Seq Based Genome-Wide Analysis Reveals Loss of Inter-Chromosomal Regulation in Breast Cancer. *Sci Rep* (2017) 7:1760. doi: 10.1038/s41598-017-01314-1
- Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enriquez J. Multilayer Approach Reveals Organizational Principles Disrupted in Breast Cancer Co-Expression Networks. *Appl Network Sci* (2020) 5:47. doi: 10.1007/s41109-020-00291-1
- Andonegui-Elguera SD, Zamora-Fuentes JM, Espinal-Enriquez J, Hernández-Lemus E. Loss of Long Distance Co-Expression in Lung Cancer. *Front Genet* (2021) 12. doi: 10.3389/fgene.2021.625741
- Zamora-Fuentes JM, Hernández-Lemus E, Espinal-Enriquez J. Gene Expression and Co-Expression Networks Are Strongly Altered Through Stages in Clear Cell Renal Carcinoma. *Front Genet* (2020) 11:578679. doi: 10.3389/fgene.2020.578679
- García-Cortés D, de Anda-Jáuregui G, Fresno C, Hernández-Lemus E, Espinal-Enriquez J. Gene Co-Expression Is Distance-Dependent in Breast Cancer. *Front Oncol* (2020) 10:1232. doi: 10.3389/fonc.2020.01232
- Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene Co-Expression Network Analysis Reveals Common System-Level Properties of Prognostic Genes Across Cancer Types. *Nat Commun* (2014) 5:1–9. doi: 10.1038/ncomms4231
- Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A, et al. Loss of Connectivity in Cancer Co-Expression Networks. *PLoS One* (2014) 9:e87075. doi: 10.1371/journal.pone.0087075
- Deng SP, Zhu L, Huang DS. Predicting Hub Genes Associated With Cervical Cancer Through Gene Co-Expression Networks. *IEEE/ACM Trans Comput Biol Bioinf* (2015) 13:27–35. doi: 10.1109/TCBB.2015.2476790
- Tang J, Kong D, Cui Q, Wang K, Zhang D, Gong Y, et al. Prognostic Genes of Breast Cancer Identified by Gene Co-Expression Network Analysis. *Front Oncol* (2018) 8:374. doi: 10.3389/fonc.2018.00374
- Liao Y, Wang Y, Cheng M, Huang C, Fan X. Weighted Gene Coexpression Network Analysis of Features That Control Cancer Stem Cells Reveals Prognostic Biomarkers in Lung Adenocarcinoma. *Front Genet* (2020) 11:311. doi: 10.3389/fgene.2020.00311
- Yu X, Cao S, Zhou Y, Yu Z, Xu Y. Co-Expression Based Cancer Staging and Application. *Sci Rep* (2020) 10:1–10. doi: 10.1038/s41598-020-67476-7
- Altay G, Emmert-Streib F. Inferring the Conservative Causal Core of Gene Regulatory Networks. *BMC Syst Biol* (2010) 4:1–13. doi: 10.1186/1752-0509-4-132
- Alcalá-Corona SA, Velázquez-Caldelas TE, Espinal-Enriquez J, Hernández-Lemus E. Community Structure Reveals Biologically Functional Modules in Mef2c Transcriptional Regulatory Network. *Front Physiol* (2016) 7:184. doi: 10.3389/fphys.2016.00184
- Alcalá-Corona SA, de Anda-Jáuregui G, Espinal-Enriquez J, Hernández-Lemus E. Network Modularity in Breast Cancer Molecular Subtypes. *Front Physiol* (2017) 8:915. doi: 10.3389/fphys.2017.00915
- de Anda-Jáuregui G, Espinal-Enriquez J, Drago-García D, Hernández-Lemus E. Nonredundant, Highly Connected MicroRNAs Control Functionality in Breast Cancer Networks. *Int J Genomics* (2018) 2018. doi: 10.1155/2018/9585383
- Velázquez-Caldelas TE, Alcalá-Corona SA, Espinal-Enriquez J, Hernández-Lemus E. Unveiling the Link Between Inflammation and Adaptive Immunity in Breast Cancer. *Front Immunol* (2019) 10:56. doi: 10.3389/fimmu.2019.00056
- Liesecke F, De Craene JO, Besseau S, Courdavault V, Clastre M, Vergès V, et al. Improved Gene Co-Expression Network Quality Through Expression Dataset Down-Sampling and Network Aggregation. *Sci Rep* (2019) 9:1–16. doi: 10.1038/s41598-019-50885-8
- Alcalá-Corona SA, Espinal-Enriquez J, De Anda Jáuregui G, Hernandez-Lemus E. The Hierarchical Modular Structure of Her2+ Breast Cancer Network. *Front Physiol* (2018) 9:1423. doi: 10.3389/fphys.2018.01423
- Serrano MA, Boguna M, Vespignani A. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proc Natl Acad Sci* (2009) 106:6483–8. doi: 10.1073/pnas.0808904106
- Perkins AD, Langston MA. Threshold Selection in Gene Co-Expression Networks Using Spectral Graph Theory Techniques. *BMC Bioinf* (2009) 10. doi: 10.1186/1471-2105-10-s11-s4
- Tieri P, Farina L, Petti M, Astolfi L, Paci P, Castiglione F. Network Inference and Reconstruction in Bioinformatics. *Encyclopedia Bioinf Comput Biol (Elsevier)* (2019) 2:805–13. doi: 10.1016/b978-0-12-809633-8.20290-2
- Kimura S, Fukutomi R, Tokuhisa M, Okada M. Inference of Genetic Networks From Time-Series and Static Gene Expression Data: Combining a Random-Forest-Based Inference Method With Feature Selection Methods. *Front Genet* (2020) 11:595912. doi: 10.3389/fgene.2020.595912
- de Anda-Jáuregui G, Alcalá-Corona SA, Espinal-Enriquez J, Hernández-Lemus E. Functional and Transcriptional Connectivity of Communities in Breast Cancer Co-Expression Networks. *Appl Network Sci* (2019) 4:22. doi: 10.1007/s41109-019-0129-0
- Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enriquez J. K-Core Genes Underpin Structural Features of Breast Cancer. *Sci Rep* (2021) 11:16284. doi: 10.1038/s41598-021-95313-y
- Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp Oncol (Poznan Poland)* (2015) 19:A68–77. doi: 10.5114/wo.2014.47136
- Fresno C, González GA, Merino GA, Flesia AG, Podhajcer OL, Llera AS, et al. A Novel Non-Parametric Method for Uncertainty Evaluation of Correlation-Based Molecular Signatures: Its Application on PAM50 Algorithm. *Bioinf (Oxford England)* (2017) 33:693–700. doi: 10.1093/bioinformatics/btw704
- Fresno C, González GA, Llera AS, Fernández EA. Pbcmc: Permutation-Based Confidence for Molecular Classification. *R Package version* (2016) 1.2. doi: 10.18129/B9.bioc.pbcmc
- Nueda MJ, Ferrer A, Conesa A. ARSYN: A Method for the Identification and Removal of Systematic Noise in Multifactorial Time Course Microarray Experiments. *Biostatistics (Oxford England)* (2012) 13:553–66. doi: 10.1093/biostatistics/bxr042
- (2021). Available at: <https://github.com/josemaz/gene-matrices/blob/master/Notebooks/CorrelationVsDistance.ipynb>
- Vinayak, Prosen T, Buča B, Seligman TH. Spectral Analysis of Finite-Time Correlation Matrices Near Equilibrium Phase Transitions. *Epl* (2014) 108:20006. doi: 10.1209/0295-5075/108/20006
- Vinayak V, Seligman TH. Time Series, Correlation Matrices and Random Matrix Models. *AIP Conf Proc* (2014) 1575:196–217. doi: 10.1063/1.4861704
- Gopikrishnan P, Rosenow B, Plerou V, Stanley HE. Quantifying and Interpreting Collective Behavior in Financial Markets. *Phys Rev E - Stat Physics Plasmas Fluids Related Interdiscip Topics* (2001) 64:4. doi: 10.1103/PhysRevE.64.035106
- Luo F, Zhong J, Yang Y, Zhou J. Application of Random Matrix Theory to Microarray Data for Discovering Functional Gene Modules. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2006) 73:1–5. doi: 10.1103/PhysRevE.73.031924

43. Fossion R. A Time-Series Approach to Dynamical Systems From Classical and Quantum Worlds. *AIP Conf Proc* (2014) 1575:89–110. doi: 10.1063/1.4861700
44. Fossion R, Vargas GT, Vieyra JC. Random-Matrix Spectra as a Time Series. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2013) 88:1–4. doi: 10.1103/PhysRevE.88.060902
45. Laloux L, Cizeau P, Bouchaud JP, Potters M. Noise Dressing of Financial Correlation Matrices. *Phys Rev Lett* (1999) 83:1467–70. doi: 10.1103/PhysRevLett.83.1467
46. Zhong J, Gao H, Thompson DK, Luo F, Zhou J, Khan L, et al. Constructing Gene Co-Expression Networks and Predicting Functions of Unknown Genes by Random Matrix Theory. *BMC Bioinf* (2007) 8:299. doi: 10.1186/1471-2105-8-299
47. Rummel C, Müller M, Baier G, Amor F, Schindler K. Analyzing Spatio-Temporal Patterns of Genuine Cross-Correlations. *J Neurosci Methods* (2010) 191:94–100. doi: 10.1016/j.jneumeth.2010.05.022
48. Müller M, Jiménez YL, Rummel C, Baier G, Galka A, Stephani U, et al. Localized Short-Range Correlations in the Spectrum of the Equal-Time Correlation Matrix. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2006) 74:041119. doi: 10.1103/PhysRevE.74.041119
49. Utsugi A, Ino K, Oshikawa M. Random Matrix Theory Analysis of Cross Correlations in Financial Markets. *Phys Rev E - Stat Physics Plasmas Fluids Related Interdiscip Topics* (2004) 70:11. doi: 10.1103/PhysRevE.70.026110
50. Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Guhr T, Stanley HE. Random Matrix Approach to Cross Correlations in Financial Data. *Phys Rev E - Stat Physics Plasmas Fluids Related Interdiscip Topics* (2002) 65:1–18. doi: 10.1103/PhysRevE.65.066126
51. Rummel C. Quantification of Intra- and Inter-Cluster Relations in Nonstationary and Noisy Data. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2008) 77:016708. doi: 10.1103/PhysRevE.77.016708
52. Rummel C, Müller M, Schindler K. Data-Driven Estimates of the Number of Clusters in Multivariate Time Series. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2008) 78:1–12. doi: 10.1103/PhysRevE.78.066703
53. García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. Luminal a Breast Cancer Co-Expression Network: Structural and Functional Alterations. *Front Genet* (2021) 12:629475. doi: 10.3389/fgene.2021.629475
54. Marchenko VA, PL A. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mat Sb. (N.S.)* (1967) 1:457–83. doi: 10.1070/SM1967v001n04ABEH001994
55. Kennecke H, Yerushalmi R, Woods R, Cheang MCU, Voduc D, Speers CH, et al. Metastatic Behavior of Breast Cancer Subtypes. *J Clin Oncol* (2010) 28:3271–7. doi: 10.1200/JCO.2009.25.9820
56. Fallahpour S, Navaneelan T, De P, Borgo A. Breast Cancer Survival by Molecular Subtype: A Population-Based Analysis of Cancer Registry Data. *CMAJ Open* (2017) 5:E734. doi: 10.9778/cmajo.20170030
57. Achinger-Kawecka J, Valdes-Mora F, Luu PL, Giles KA, Caldon CE, Qu W, et al. Epigenetic Reprogramming at Estrogen-Receptor Binding Sites Alters 3d Chromatin Landscape in Endocrine-Resistant Breast Cancer. *Nat Commun* (2020) 11:1–17. doi: 10.1038/s41467-019-14098-x
58. Corces MR, Corces VG. The Three-Dimensional Cancer Genome. *Curr Opin Genet Dev* (2016) 36:1–7. doi: 10.1016/j.gde.2016.01.002
59. Inaki K, Menghi F, Woo XY, Wagner JP, Jacques PÉ, Lee YF, et al. Systems Consequences of Amplicon Formation in Human Breast Cancer. *Genome Res* (2014) 24:1559–71. doi: 10.1101/gr.164871.113
60. Myhre S, Lingjærde OC, Hennessy BT, Aure MR, Carey MS, Alsner J, et al. Influence of DNA Copy Number and mRNA Levels on the Expression of Breast Cancer Related Proteins. *Mol Oncol* (2013) 7:704–18. doi: 10.1016/j.molonc.2013.02.018
61. Achinger-Kawecka J, Clark SJ. Disruption of the 3D Cancer Genome Blueprint. *Epigenomics* (2017) 9:47–55. doi: 10.2217/epi-2016-0111
62. Pugacheva EM, Kubo N, Loukinov D, Tajmul M, Kang S, Kovalchuk AL, et al. Ctf Mediates Chromatin Looping via N-Terminal Domain-Dependent Cohesin Retention. *Proc Natl Acad Sci* (2020) 117:2030–31. doi: 10.1073/pnas.1911708117
63. Fiorito E, Sharma Y, Gilfillan S, Wang S, Singh SK, Satheesh SV, et al. Ctf Modulates Estrogen Receptor Function Through Specific Chromatin and Nuclear Matrix Interactions. *Nucleic Acids Res* (2016) 44:10588–602. doi: 10.1093/nar/gkw785
64. Tovar H, García-Herrera R, Espinal-Enríquez J, Hernández-Lemus E. Transcriptional Master Regulator Analysis in Breast Cancer Genetic Networks. *Comput Biol Chem* (2015) 59:67–77. doi: 10.1016/j.compbiolchem.2015.08.007
65. Tapia-Carrillo D, Tovar H, Velazquez-Caldelas TE, Hernandez-Lemus E. Master Regulators of Signaling Pathways: An Application to the Analysis of Gene Regulation in Breast Cancer. *Front Genet* (2019) 10:1180. doi: 10.3389/fgene.2019.01180
66. Lachmann A. PhD Thesis, Columbia University, New York *Confounding Effects in Gene Expression and Their Impact on Downstream Analysis* (Columbia University). [PhD Thesis]. New York: Columbia University (2016).
67. Soler-Oliva ME, Guerrero-Martínez JA, Bachetti V, Reyes JC. Analysis of the Relationship Between Coexpression Domains and Chromatin 3d Organization. *PLoS Comput Biol* (2017) 13:e1005708. doi: 10.1371/journal.pcbi.1005708
68. Varrone M, Nanni L, Ciriello G, Ceri S. Exploring Chromatin Conformation and Gene Co-Expression Through Graph Embedding. *Bioinformatics* (2020) 36:i700–8. doi: 10.1093/bioinformatics/btaa803
69. Beesley J, Sivakumaran H, Marjaneh MM, Lima LG, Hillman KM, Kaufmann S, et al. Chromatin Interactome Mapping at 139 Independent Breast Cancer Risk Signals. *Genome Biol* (2020) 21:1–19. doi: 10.1186/s13059-019-1877-y
70. Ochoa S, de Anda-Jáuregui G, Hernández-Lemus E. An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation. *Front Genet* (2021) 12:232. doi: 10.3389/fgene.2021.617512

**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.

Copyright © 2021 González-Espinoza, Zamora-Fuentes, Hernández-Lemus and Espinal-Enríquez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Data-Driven Discovery of Mathematical and Physical Relations in Oncology Data Using Human-Understandable Machine Learning

Daria Kurz<sup>1†</sup>, Carlos Salort Sánchez<sup>2†</sup> and Cristian Axenie<sup>3\*†</sup>

<sup>1</sup>Interdisziplinäres Brustzentrum, Helios Klinikum München West, Akademisches Lehrkrankenhaus der Ludwig-Maximilians Universität München, Munich, Germany, <sup>2</sup>TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany, <sup>3</sup>Audi Konfuzius-Institut Ingolstadt Laboratory, Technische Hochschule Ingolstadt, Ingolstadt, Germany

## OPEN ACCESS

### Edited by:

George Bebis,  
University of Nevada, Reno,  
United States

### Reviewed by:

Chang Chen,  
University of Chicago, United States  
Tin Nguyen,  
University of Nevada, Reno,  
United States  
Sokratis Makrogiannis,  
Delaware State University,  
United States

### \*Correspondence:

Cristian Axenie  
cristian.axenie@audi-konfuzius-  
institut-ingolstadt.de

<sup>†</sup>These authors have contributed  
equally to the work

### Specialty section:

This article was submitted to  
Medicine and Public Health,  
a section of the journal  
Frontiers in Artificial Intelligence

**Received:** 23 May 2021

**Accepted:** 08 October 2021

**Published:** 25 November 2021

### Citation:

Kurz D, Sánchez CS and Axenie C  
(2021) Data-Driven Discovery of  
Mathematical and Physical Relations in  
Oncology Data Using Human-  
Understandable Machine Learning.  
Front. Artif. Intell. 4:713690.  
doi: 10.3389/frai.2021.713690

For decades, researchers have used the concepts of rate of change and differential equations to model and forecast neoplastic processes. This expressive mathematical apparatus brought significant insights in oncology by describing the unregulated proliferation and host interactions of cancer cells, as well as their response to treatments. Now, these theories have been given a new life and found new applications. With the advent of routine cancer genome sequencing and the resulting abundance of data, oncology now builds an “arsenal” of new modeling and analysis tools. Models describing the governing physical laws of tumor–host–drug interactions can be now challenged with biological data to make predictions about cancer progression. Our study joins the efforts of the mathematical and computational oncology community by introducing a novel machine learning system for data-driven discovery of mathematical and physical relations in oncology. The system utilizes computational mechanisms such as competition, cooperation, and adaptation in neural networks to simultaneously learn the statistics and the governing relations between multiple clinical data covariates. Targeting an easy adoption in clinical oncology, the solutions of our system reveal human-understandable properties and features hidden in the data. As our experiments demonstrate, our system can describe nonlinear conservation laws in cancer kinetics and growth curves, symmetries in tumor’s phenotypic staging transitions, the preoperative spatial tumor distribution, and up to the nonlinear intracellular and extracellular pharmacokinetics of neoadjuvant therapies. The primary goal of our work is to enhance or improve the mechanistic understanding of cancer dynamics by exploiting heterogeneous clinical data. We demonstrate through multiple instantiations that our system is extracting an accurate human-understandable representation of the underlying dynamics of physical interactions central to typical oncology problems. Our results and evaluation demonstrate that, using simple—yet powerful—computational mechanisms, such a machine learning system can support clinical decision-making. To this end, our system is a representative tool of the field of mathematical and computational oncology

and offers a bridge between the data, the modeler, the data scientist, and the practicing clinician.

**Keywords:** mathematical oncology, machine learning, mechanistic modeling, data-driven predictions, clinical data, decision support system

## 1 INTRODUCTION

The dynamics governing cancer initiation, development, and response to treatment are informed by quantitative measurements. These measurements carry details about the physics of the underlying processes, such as tumor growth, tumor–host cell encounters, and drug transport. Be it through mathematical modeling and patient-specific treatment trajectories—as in the excellent work of Werner et al. (2016)—through tumor’s mechanopathology—systematically described by Nia et al. (2020)—or through hybrid modeling frameworks of tumor development and treatment—identified by Chamseddine and Rejniak (2020)—capturing such processes from data can substantially improve predictions about cancer progression.

Machine learning algorithms are now leveraging automatic discovery of physics principles and governing mathematical relations for such improved predictions. Proof stands the proliferating body of such research—for representative results, see the works of Raissi (2018), Schaeffer (2017), Long et al. (2018), and Champion et al. (2019). However, the naive application of such algorithms is insufficient to infer physical laws underlying cancer progression. Simply positing a physical law or mathematical relation from data is useless without simultaneously proposing an accompanying ground truth to account for the inevitable mismatch between model and observations, as demonstrated in the work of de Silva et al. (2020).

Such a problem is even more important in clinical oncology where, in order to understand the links between the physics of cancer and signaling pathways in cancer biology, we need to describe the fundamental physical principles shared by most, if not all, tumors, as proposed by Nia et al. (2020). Here, mathematical models of the physical mechanisms and corresponding tumor physical hallmarks complement the heterogeneity of the experimental observations. Such a constellation is typically validated through *in vivo* and *in vitro* model systems where the simultaneous identification of both the structure and parameters of the dynamical system describing tumor–host interactions is performed (White et al., 2019).

Given the multidimensional nature of this system identification process, some concepts involved are nonintuitive and require deep and broad understanding of both the physical and biological aspects of cancer. To circumvent this, combining mechanistic modeling and machine learning is a promising approach with high potential for clinical translation. For instance, in a bottom-up approach, fusing cell-line tumor growth curve learning from heterogeneous data (i.e., caliper, imaging, microscopy) and unsupervised extraction of cytostatic pharmacokinetics, the study by Axenie and Kurz (2020a) introduced a novel pipeline for patient-tailored neoadjuvant

therapy planning. In another relevant study, Benzekry (2020) used machine learning to extract model parameters from high-dimensional baseline data (demographic, clinical, pathological molecular) and used mixed-effects theory to combine it with mechanistic models based on longitudinal data (e.g., tumor size measurements, pharmacokinetics, seric biomarkers, and circulating DNA) for treatment individualization.

Yet, despite the recent advances in mathematical and computational oncology, there are only a few systems trying to offer a human-understandable solution, or the steps to reach it—the most relevant are the studies by Jansen et al. (2020) and Lamy et al. (2019). But, such systems lack a rigorous and accessible description of the physical cancer traits assisting their clinical predictions. Our study advocates the improvement of mechanistic modeling with the help of machine learning. Our thesis goes beyond measurements-informed biophysical processes models, as described by Cristini et al. (2017), and toward human-understandable personalized disease evolution and therapy profiles learned from data, as foreseen by Kondylakis et al. (2020).

### 1.1 Study Focus

The purpose of this study is to introduce a system (and a framework) capable of learning human-understandable mathematical and physical relations from heterogeneous oncology data for patient-centered clinical decision support. To demonstrate the versatility of the system, we introduce multiple of its instantiations, in an end-to-end fashion (i.e., from cancer initiation to treatment outcome) for predictions based on available clinical datasets<sup>1</sup>:

- **learning initiation patterns of preinvasive breast cancer** (i.e., ductal carcinoma *in situ* [DCIS]) from histopathology and morphology data available from the studies by Rodallec et al. (2019), Volk et al. (2011), Tan et al. (2015), and Mastri et al. (2019);
- **learning unperturbed tumor growth curves within and between cancer types** (i.e., breast, lung, leukemia) from imaging, microscopy, and caliper data available from the studies by Benzekry et al. (2019) and Simpson-Herren and Lloyd (1970);
- **extracting tumor phenotypic stage transitions** from three cell lines of breast cancer using imaging, immunohistochemistry, and histopathology data available from the studies by Rodallec et al. (2019), Volk et al. (2011), Tan et al. (2015), and Edgerton et al. (2011);

<sup>1</sup>A copy of the used datasets along with the study generating them is included in the codebase associated with the manuscript.



- **simultaneously extracting the drug-perturbed tumor growth and drug pharmacokinetics** for neoadjuvant/adjuvant therapy sequencing using data available from the studies by Kuh et al. (2000), Volk et al. (2011), and Chen et al. (2014);
- **predicting tumor growth/recession** (i.e., estimating tumor volume after each chemotherapy cycle **under various chemotherapy regimens** administered to breast cancer patients, using real-world patient data available from the study by Yee et al. (2020) as well as cell lines studies from Rodallec et al. (2019), Volk et al. (2011), Tan et al. (2015), and Mastro et al. (2019).

In each of the instantiations, we use the same computational substrate (i.e., no specific task parametrization) and compare the performance of our system against state-of-the-art systems capable of extracting governing equations from heterogeneous oncology data from Cook et al. (2010), Mandal and Cichocki (2013), Weber and Wermter (2007), and Champion et al. (2019), respectively. The analysis focuses on (1) the accuracy of the systems in the learned mathematical and physical relations among various covariates, (2) the ability to embed more data and mechanistic models, and (3) the ability to provide a human-understandable solution and the processing steps to obtain that solution.

## 1.2 Study Motivation

In clinical practice, patient tumors are typically described across multiple dimensions from (1) high-dimensional heterogeneous data (e.g., demographic, clinical, pathological, molecular), and (2) longitudinal data (e.g., tumor size measurements, pharmacokinetics, immune screening, biomarkers), to (3) time-to-event data (e.g., progression-free or overall survival analysis), and, in the last years, (4) genetic sequencing that determine the genetic mutations driving their cancer. With this information, the clinical oncologist may tailor treatment to the patient's specific cancer.

But, despite the variety of such rich patient data available, tumor growth data, describing the dynamics of cancer development, from initiation to metastasis has some peculiarities. These features motivated the study and the approach proposed by our system. To summarize, tumor growth data:

- is typically **small**, with only a few data points measured, typically, at days-level resolution (Roland et al., 2009);
- is **unevenly sampled**, with irregular spacing among tumor size/volume observations (Volk et al., 2011);
- has **high variability** between and within tumor types (Benzekry et al., 2014) and type of treatment (Gaddy et al., 2017).
- is **heterogeneous** and sometimes **expensive or difficult to obtain** (e.g., biomarkers, functional magnetic resonance imaging (Abler et al., 2019), fluorescence imaging (Rodallec et al., 2019), flow cytometry, or calipers (Benzekry et al., 2019).

- **determines cancer treatment planning**, for instance, adjuvant versus neoadjuvant chemotherapy (Sarapata and de Pillis, 2014).

Using unsupervised learning, our system seeks to overcome these limitations and provide a human-understandable representation of the mathematical and physical relations describing tumor growth, its phenotype, and, finally, its interaction with chemotherapeutic drugs. The system exploits the temporal evolution of the processes describing growth data along with their distribution in order to reach superior accuracy and versatility on various clinical *in vitro* tumor datasets.

## 2 MATERIALS AND METHODS

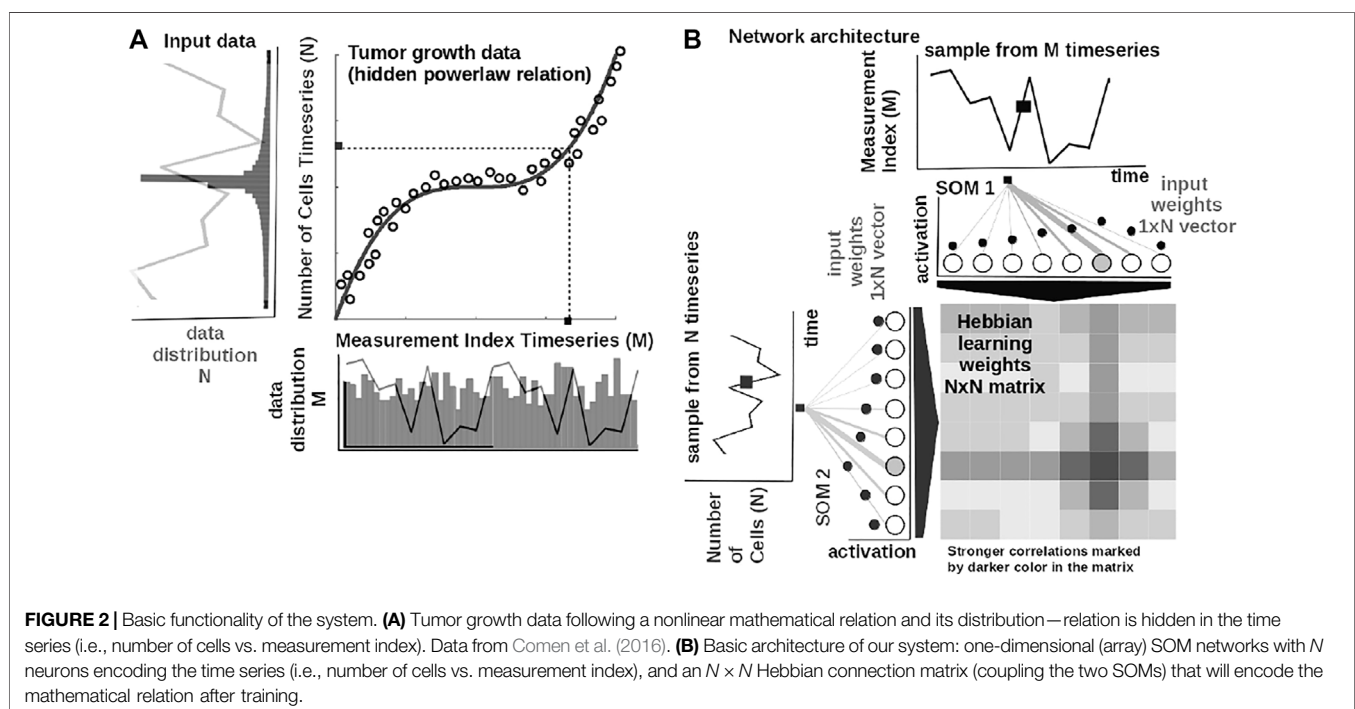
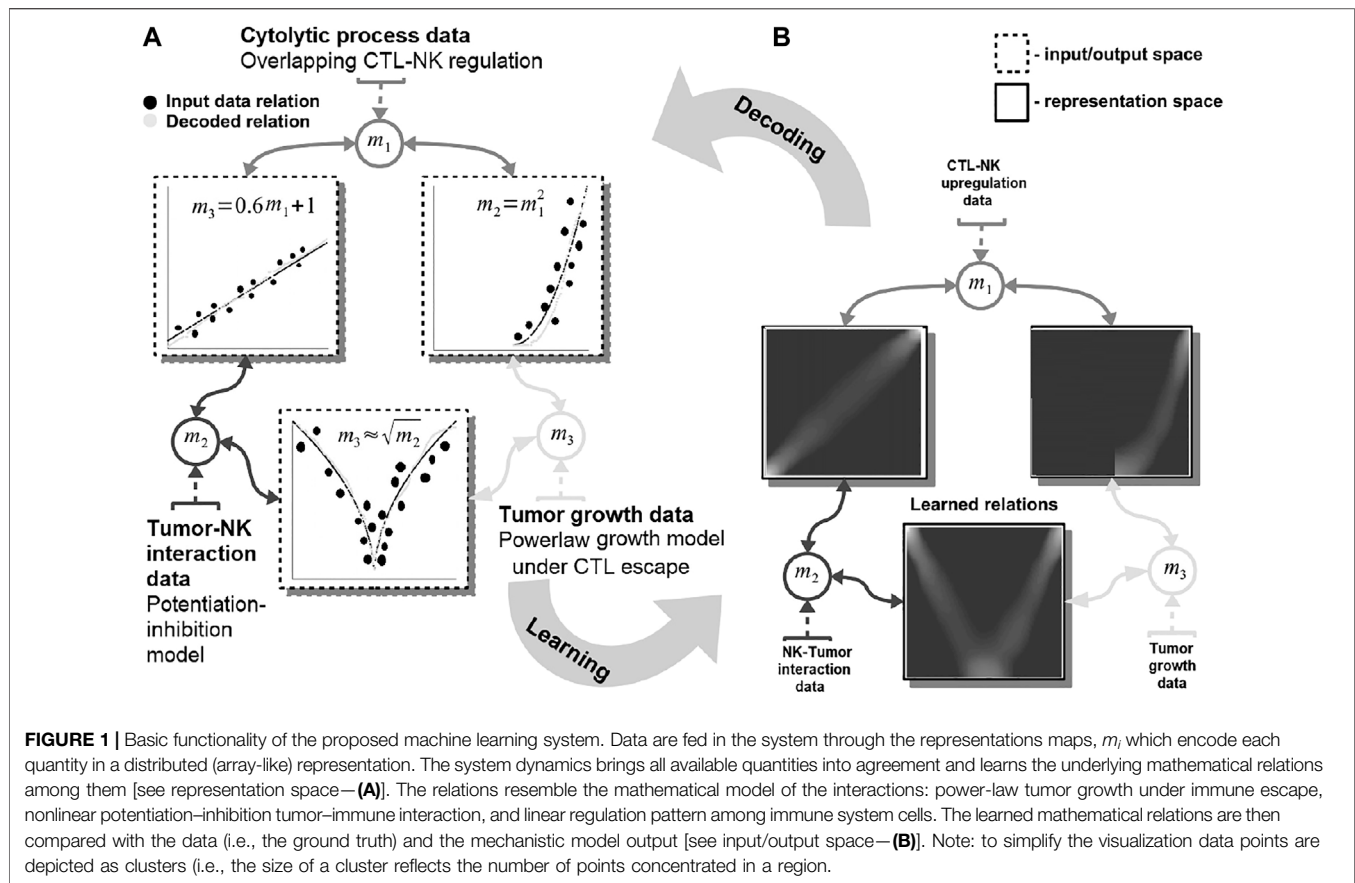
In the current section, we introduce our system through the lens of practical examples of discovering mathematical and physical relations describing tumor–host–drug dynamics. We begin by introducing the basic computational framework as well as the various configurations in which the system can be used. The second part is dedicated to introducing relevant state-of-the-art approaches used in our comparative experimental evaluation.

### 2.1 System Basics

Physical interactions of cancer cells with their environment (e.g., local tissue, immune cells, drugs) determine the physical characteristics of tumors through distinct and interconnected mechanisms. For instance, cellular proliferation and its inherent abnormal growth patterns lead to increased solid stress (Nia et al., 2016). Subsequently, cell contraction and cellular matrix deposition modify the architecture of the surrounding tissue, which can additionally react to drugs (Griffon-Etienne et al., 1999) modulating the stiffness (Rouvière et al., 2017) and interstitial fluid pressure (Nathanson and Nelson, 1994). But such physical characteristics also interact among each other initiating complex dynamics, as demonstrated in Nia et al. (2020).

Our system can capture such complex dynamics through a network-based paradigm for modeling, computation, and prediction. It can extract the mathematical description of the interactions exhibited by multiple entities (e.g., tumor, host cells, cytostatic drugs) for producing informed predictions. For guiding the reader, we present a simple, biologically grounded example in **Figure 1**.

In this example, our system learns simultaneously the power-law tumor growth under immune escape (Benzekry et al., 2014) and the nonlinear potentiation–inhibition model of natural killer (NK) cells–tumor interactions (Ben-Shmuel et al., 2020), while exhibiting the known overlapping cytotoxic T lymphocytes (CTLs)–NK cell mutual linear regulation pattern (Uzhachenko and Shanker, 2019). As shown in **Figure 1**, our system offers the means to learn the mathematical relations governing the physical tumor–immune interactions, without supervision, from available clinical data (**Figure 1**—input data relations and learned and decoded relations). Furthermore, the system can infer unavailable (i.e., expensive to measure) physical quantities (i.e., after learning/training) in order to make predictions on the effects of modifying



the pattern of interactions among the tumor and the immune system. For instance, by not feeding the system with the innate immune response (i.e., the NK cells dynamics), the system infers, based on the CTL–NK cell interaction pattern and the tumor growth pattern, a plausible tumor–NK cell mathematical relation in agreement with observations (**Figure 1B**, squared root nonlinearity).

Basically, our system acts as constraint satisfaction network converging to a global consensus given local (i.e., the impact of the measured data) and global dynamics of the physics governing the interactions (see the clear patterns depicting the mathematical models of interaction in **Figure 1**). The networked structure allows the system to easily interconnect multiple data quantities measuring different biological components (Markowetz and Troyanskaya, 2007) or a different granularity of representation of the underlying interaction physics (Cornish and Markowetz, 2014).

## 2.2 Computational Substrate

The core element of our study is an unsupervised machine learning system based on Self-Organizing Maps (SOMs) Kohonen (1982) and Hebbian learning (HL) Chen et al. (2008). The two components are used in concert to represent and extract the underlying relations among correlated data. In order to introduce the computational steps followed by our system, we provide a simple example in **Figure 2**. Here, we feed the system with data from a cubic growth law (third power-law) describing the effect of drug dose density over 150 weeks of adjuvant chemotherapy in breast cancer (data from Comen et al., 2016). The two data sources (i.e., the cancer cell number and the irregular measurement index over the weeks) follow a cubic dependency (cmp. **Figure 2A**). Before being presented the data, our system has no prior information about the data distribution and its generating process (or model). The system learns the underlying (i.e., hidden) mathematical relation directly from the pairs of input data without supervision.

The input SOMs (i.e., one-dimensional [1-D] lattice networks with  $N$  neurons) extract the probability distribution of the incoming data, depicted in **Figure 2A**, and encode samples in a distributed activity pattern, as shown in **Figure 2B**. This activity pattern is generated such that the closest preferred value of a SOM neuron to the input will be strongly activated and will decay, proportional with distance, for neighboring units. This process is fundamentally benefiting from the quantization capability of SOM. The tasks we solve in this work have low dimensionality, basically allowing a 1-D SOM to provide well-behaved distributed representations. 1-D SOMs are proven mathematically to converge and handling boundary effects. For higher-dimensional data, our system can be coupled with a reduction technique (i.e., principal component analysis, t-Distributed Stochastic Neighbor Embedding) to reduce data to 1-D time series, without a large penalty in complexity. In addition, this process is extended with a dimension corresponding to the latent representation of network resource allocation (i.e., number of neurons allocated to represent the input data space). After

learning, the SOMs specialize to represent a certain (preferred) value in the input data space and learn its probability distribution, by updating its tuning curves shape.

Practically, given an input value  $s^p(k)$  from one time series at time step  $k$ , the network follows the processing stages in **Figure 3**. For each  $i$ th neuron in the  $p$ th input SOM, with preferred value  $w_{in,i}^p$  and tuning curve size  $\xi_i^p(k)$ , the generated neural activation is given by

$$a_i^p(k) = \frac{1}{\sqrt{2\pi}\xi_i^p(k)} e^{\frac{-(s^p(k)-w_{in,i}^p(k))^2}{2\xi_i^p(k)^2}}. \quad (1)$$

The most active (i.e., competition winning) neuron of the  $p$ th population,  $b^p(k)$ , is the one that has the highest activation given the time series data point at time  $k$

$$b^p(k) = \underset{i}{\operatorname{argmax}} a_i^p(k). \quad (2)$$

The competition for highest activation (in representing the input) in the SOM is followed by a cooperation process that captures the input space distribution. More precisely, given the winning neuron,  $b^p(k)$ , the cooperation kernel,

$$h_{b,i}^p(k) = e^{\frac{-\|r_i - r_b\|^2}{2\sigma(k)^2}}, \quad (3)$$

allows neighboring neurons in the network (i.e., found at position  $r_i$  in the network) to precisely represent the input data point given their location in the neighborhood  $\sigma(k)$  of the winning neuron. The topological neighborhood width  $\sigma(k)$  decays in time, to avoid artifacts (e.g., twists) in the SOM. The kernel in **Eq. 3** is chosen such that adjacent neurons in the network specialize on adjacent areas in the input space, by “pulling” the input weights (i.e., preferred values) of the neurons closer to the input data point,

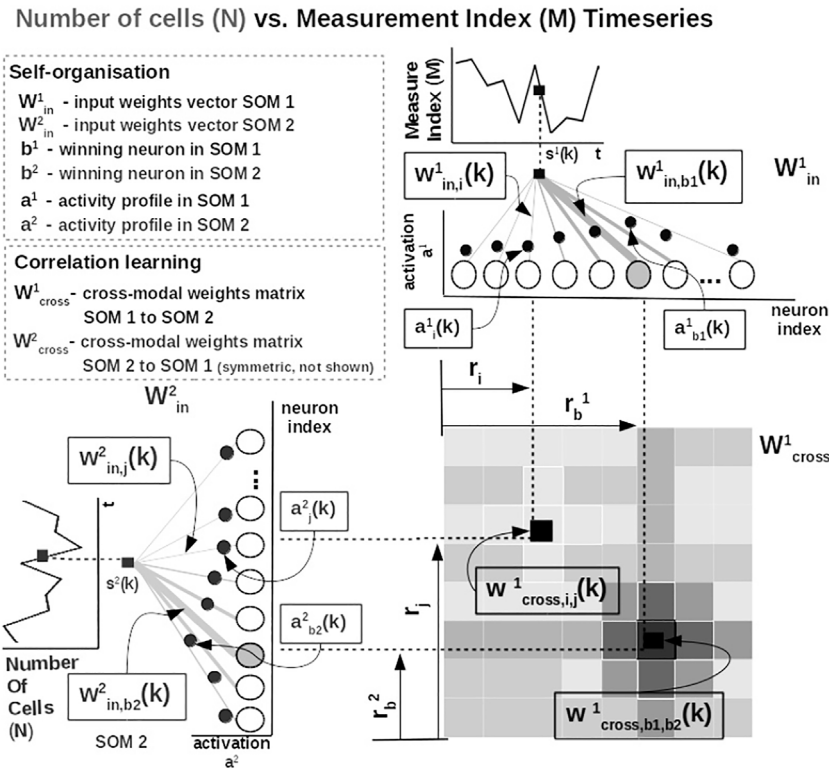
$$\Delta w_{in,i}^p(k) = \alpha(k) h_{b,i}^p(k) (s^p(k) - w_{in,i}^p(k)). \quad (4)$$

This process updates the tuning curves width  $\xi_i^p$  given the spatial location of the neuron in the network, the distance to the input data point, the cooperation kernel size, and a decaying learning rate  $\alpha(k)$ ,

$$\Delta \xi_i^p(k) = \alpha(k) h_{b,i}^p(k) ((s^p(k) - w_{in,i}^p(k))^2 - \xi_i^p(k)^2). \quad (5)$$

To illustrate these mechanisms, we consider the learned tuning curves shapes for five neurons in the input SOMs (i.e., neurons 1, 6, 13, 40, 45) encoding the breast cancer cubic tumor growth law, depicted in **Figure 4**. We observe that higher input probability distributions are represented by dense and sharp tuning curves (e.g., neuron 1, 6, 13 in SOM1), whereas lower or uniform probability distributions are represented by more sparse and wide-tuning curves (e.g., neuron 40, 45 in SOM1).

This way, the system optimally allocates neurons such that a higher amount of neurons represent areas in the input space, which need a finer resolution, and a lower amount for more coarsely represented input space areas. Neurons in the two SOMs are then linked by a fully (all-to-all) connected matrix of synaptic



**FIGURE 3** | Detailed computational steps of our system, instantiated for tumor growth learning given the observed number of cells and the measurement index data from Comen et al. (2016).

connections, where the weights are computed using HL. The connections between uncorrelated (or weakly correlated) neurons in each SOM (i.e.,  $w_{cross}$ ) are suppressed (i.e., darker color), whereas correlated neuron connections are enhanced (i.e., brighter color), as depicted in **Figure 3**. Each connection weight  $w_{cross,i,j}^p$  between neurons  $i, j$  in the input SOMs is updated with an HL rule as follows:

$$\Delta w_{cross,i,j}^p(k) = \eta(k) (a_i^p(k) - \bar{a}_i^p(k)) (a_j^q(k) - \bar{a}_j^q(k)), \quad (6)$$

where

$$\bar{a}_i^p(k) = (1 - \beta(k)) \bar{a}_i^p(k-1) + \beta(k) a_i^p(k), \quad (7)$$

is an exponential decay (i.e., momentum), and  $\eta(k)$ ,  $\beta(k)$  are monotonic (inverse-time) decaying functions. HL ensures a weight increase for correlated activation patterns and a weight decrease for anticorrelated activation patterns. The Hebbian weight matrix encodes the coactivation patterns between the input SOMs, as shown in **Figure 2B**, and, eventually, the learned mathematical relation given the data, as shown in **Figure 4**. Such a representation, as shown in **Figure 4**, demonstrates the human-understandable output of our system that employs powerful, yet simple and transparent, processing principles, as depicted in **Figure 3**.

Input SOM self-organization and Hebbian correlation learning operate at the same time in order to refine both the input data representation and the extracted

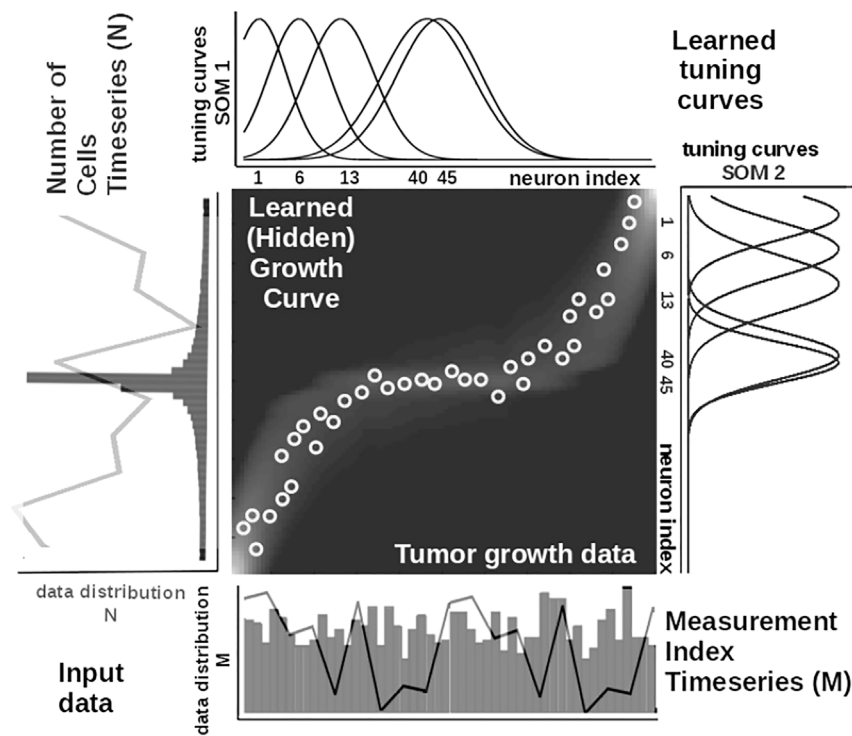
mathematical relation. This is visible in the encoding and the decoding functions where the input activations  $a$  are projected through the input weights  $w_{in}$  (Eq. 1) to the Hebbian matrix and then decoded through the  $w_{cross}$  correlation weights (Eq. 8).

In order to recover the real-world value from the network, we use a decoding mechanism based on (self-learned) bounds of the input data space. The input data space bounds are obtained as minimum and maximum of a cost function of the distance between the current preferred value of the winning neuron (i.e., the value in the input which is closest [in Euclidian space] to the weight vector of the neuron) and the input data point in the SOM (i.e., using Brent's optimization Brent, 2013). Depending on the position of the winning neuron in the SOM, the decoded/recovered value  $y(t)$  from the SOM neurons weights is computed as follows:

$$y(t) = \begin{cases} w_{in,i}^p + d_i^p & \text{if } i \geq \frac{N}{2} \\ w_{in,i}^p - d_i^p & \text{if } i < \frac{N}{2} \end{cases}$$

where  $d_i^p = \sqrt{2\xi_i^p(k)^2 \log(\sqrt{2\pi} a_i^p(k) \xi_i^p(k)^2)}$  for the winning neuron with index  $i$  in the SOM, a preferred value  $w_{in,i}^p$  and  $\xi_i^p(k)$  tuning curve size and  $\Delta a_i^p(k) = w_{cross,i,j}^p(k) a_j^q(k)$ . The activation  $a_i^p(k)$  is computed by projecting one data point through SOM  $q$  and subsequently through the Hebbian matrix





**FIGURE 4 |** Extracted mathematical relation describing the growth law and data statistics for the experimental observations in **Figure 2A** depicting a cubic breast cancer tumor growth law among number of cells and irregular measurement over 150 weeks from Comen et al. (2016). Raw data time series is overlaid on the data distribution and corresponding model encoding tuning curves shapes.

to compute the paired activity (i.e., at the other SOM  $p$ , Eq. (8)) describing the other data quantity.

$$\Delta a_i^p(k) = w_{cross,i,j}^p(k) a_j^q(k) \quad (8)$$

where  $w_{cross,i,j}^p(k) = \text{rot90}(w_{cross,i,j}^q(k))$  and  $\text{rot90}$  is a clockwise rotation. The processes described in the previous equations denote the actual inference process following the training phase (i.e., classifying new data). Basically, after applying the input time series and finding the winner in the input SOM population, the decoding decision is based on the position of the winner. Two bounds (i.e., left and right) are defined with respect to the winner's position such that the recovered value is obtained by running Brent's algorithm between the preferred values of the neurons with indices given by the bounds. The method is guaranteed to converge to global minima (of the cost function), and it is immune to boundary effects, if winners are placed at the extremes of the SOM population. A thorough analysis of the learned relations in the Hebbian matrix demonstrated that because of the asymmetric neighborhood function in the input SOMs, the activity saturated at the edges of the latent representation space. Interestingly, this was also visible in the coactivation pattern, such that the higher activity values characterize the bounds of the Hebbian representation toward the edges. When decoding the activity pattern from the Hebbian matrix, we were able to recover a relatively good probability distribution shape. This, interesting and useful,

behavior emphasizes the joint effect that the SOM distributed representation boundary effects and the Hebbian temporal coactivation have upon the data. The resulting distributions have a convex profile, concentrating a large number of samples toward the edges of the histogram with a large variance, whereas precisely decoded areas follow a relatively uniform distribution. We noticed that the decoder treated equally (i.e., accuracy of decoding) linear relations with strong boundary conditions and symmetric nonlinear relations without boundary conditions. The decoding step is a fundamental aspect contributing to the human-understandable output of our system. This demonstrates that simple operations, such as competition and cooperation in neural networks, can exploit the statistics of clinical data and provide a human-understandable representation of the governing mathematical relations behind tumor growth processes.

## 2.3 Comparable Systems

In this section, we briefly introduce four state-of-the-art approaches that we comparatively evaluated against our system. Ranging from statistical methods, to machine learning, and up to deep learning (DL), the selected systems were designed to extract governing equations from the data.

*Cook et al.* The system of Cook et al. (2010) uses a combination of simple computational mechanisms, like winner-take-all (WTA) circuits, HL, and homeostatic activity regulation, to extract mathematical relations among different data sources.

Real-world values presented to the network are encoded in population code representations. This approach is similar to our approach in terms of the sparse representation used to encode data. The difference resides in the fact that in our model the input population (i.e., SOM network) connectivity is learned. Using this capability, our model is capable of learning the input data bounds and distribution directly from the input data, without any prior information or fixed connectivity. Furthermore, in this system, the dynamics between each population encoded input is performed through plastic Hebbian connections. Starting from a random connectivity pattern, the matrix finally encoded the functional relation between the variables that it connects. The Hebbian linkage used between populations is the correlation detection mechanism used also in our model, although in our formulation we adjusted the learning rule to accommodate both the increase and decrease of the connection weights.

*Weber and Wermter.* Using a different neurally inspired substrate, the system of Weber and Wermter (2007) combines competition and cooperation in a self-organizing network of processing units to extract coordinate transformations. More precisely, the model uses simple, biologically motivated operations, in which coactivated units from population-coded representations self-organize after learning a topological map. This basically assumes solving the reference frame transformation between the inputs (mapping function). Similar to our model, the proposed approach extends the SOM network by using sigma-pi units (i.e., weighted sum of products). The connection weight between this type of processing units implements a logical AND relation. The algorithm produces invariant representations and a topographic map representation.

*Mandal and Cichocki.* Going away from biological inspiration, the system of Mandal and Cichocki (2013) used a type of nonlinear canonical correlation analysis (CCA), namely, alpha-beta divergence correlation analysis (ABCA). The ABCA system extracts relations between sets of multidimensional random variables. The core idea of the system is to first determine linear combinations of two random variables (called canonical variables/variants) such that the correlation between the canonical variables is the highest among all such linear combinations. As traditional CCA is only able to extract linear relations between two sets of multidimensional random variable, the proposed model comes as an extension to extract nonlinear relations, with the requirement that relations are expressed as smooth functions and can have a moderate amount of additive random noise on the mapping. The model employs a probabilistic method based on nonlinear correlation analysis using a more flexible metric (i.e., divergence/distance) than typical CCA.

*Champion et al.* As DL is becoming a routine tool for data discovery, as shown in the recent work of Champion et al. (2019), Raissi (2018), Schaeffer (2017), and de Silva et al. (2020), we also consider a DL system (inspired from Champion et al., 2019) and evaluate it along the other methods. To apply this prediction method to tumor growth, we need to formulate the setup as a time series prediction problem. At any given point, we have the dates and values of previous observations. Using these two features, we

can implement DL architectures that predict the size of the tumor at a future step. Recurrent neural networks (RNNs) are the archetypal DL architectures for time series prediction. The principal characteristic of RNN, compared with simpler DL architectures, is that they iterate over the values that have been observed, obtaining valuable information from it, like the rate at which the objective variable grows, and use that information to improve prediction accuracy. The main drawback of using DL in the medical field is the need of DL models to be presented with large amounts of data. We address this problem by augmenting the data. We use support vector machines (SVMs) for augmenting data, to obtain expected tumor development with normal noise generates realistic measurements. This approach presents the expected average development of a tumor.

### 3 EXPERIMENTAL SETUP AND RESULTS

In order to evaluate our data-driven approach to learn mathematical and physical relations from heterogeneous oncology data, we introduce the five instantiations and their data briefly introduced in the Study Focus section.

#### 3.1 Datasets

In our experiments, we used publicly available tumor growth, pharmacokinetics, and chemotherapy regimens datasets (Table 1), with *in vitro* or *in vivo* clinical tumor volume measurements, for breast cancer (datasets 1, 2, 5, 6, 7) and other cancers (e.g., lung, leukemia—datasets 3 and 4, respectively). This choice is to probe and demonstrate transfer capabilities of the system to tumor growth patterns induced by different cancer types. The choice of the dataset for each of the experiments was determined by the actual task we wanted to demonstrate. For instance, for demonstrating the capability to predict preinvasive cancer volume, we used the DCIS dataset. For the between-cancer predictions, we used four (i.e., two breast and two nonbreast) out of the whole seven datasets, whereas for the within-cancer-type analysis, we only looked at the breast cancer growth prediction (i.e., four datasets). For the *in vivo* experiment, we only considered the I-SPY2 trial data.

It is important to note that tumor cancer types are staged based on the size and spread of tumors, basically their volume. However, because leukemia occurs in the developing blood cells in the bone marrow, its staging is different from solid tumors. In order to emphasize the versatility of the evaluated systems, for the leukemia datasets, we used experiments that monitored human leukemic cell engraftment over time by monitoring tumor volume in scaffolds (Antonelli et al., 2016). For the pharmacokinetics experiments (i.e., mainly focused on taxanes family for experiments on MCF-7 breast cancer cell line from Tan et al. (2015)), we used the data from Kuh et al. (2000) describing intracellular and extracellular concentrations of Paclitaxel during uptake. The datasets and the code for all the systems used in our evaluation are available on GitLab<sup>2</sup>.

<sup>2</sup>Experimental codebase: <https://gitlab.com/akii-microlab/math-comp-ml>.

**TABLE 1** | Description of the datasets used in the experiments.

Experimental dataset setup				
Dataset	Cancer type	Data type	Data points	Data freq.
1	Breast <sup>1</sup> (MDA-MB-231 cell line)	Fluorescence imaging	7	2x/week
2	Breast <sup>2</sup> (MDA-MB-435 cell line)	Digital caliper	14	2x/week
3	Lung <sup>3</sup>	Caliper	10	7x/week
4	Leukemia <sup>4</sup>	Microscopy	23	7x/week
5	Breast <sup>5</sup> (MCF-7 cell line)	Microscopic imaging	8	1x/week
6	Breast <sup>6</sup> (LM2-4LUC + cell line)	Digital caliper	10	3x/week
7	Breast <sup>7</sup> (stage 2/3 cancers)	Functional magnetic resonance imaging	5	1x/week
8	Breast <sup>8</sup> (ductal carcinoma <i>in situ</i> )	Histopathology	5	1x/week

<sup>1</sup>Dataset from the study by Rodallec et al. (2019)

<sup>2</sup>Dataset from the study by Volk et al. (2011)

<sup>3</sup>Dataset from the study by Benzekry et al. (2019)

<sup>4</sup>Dataset from the study by Simpson-Herren and Lloyd (1970)

<sup>5</sup>Dataset from the study by Tan et al. (2015)

<sup>6</sup>Dataset from the study by Mastri et al. (2019)

<sup>7</sup>Dataset from the study by Yee et al. (2020)

<sup>8</sup>Dataset from the study by Edgerton et al. (2011)

### 3.2 Procedures

In order to train the different approaches we considered in our study, basically the datasets were preprocessed to represent two-dimensional dynamics, namely, tumor growth or drug concentration evolution and irregular time evolution, respectively. Each of the two time series was directly encoded in neural distributed neural populations for the work by Cook et al., Weber et al., and our approach, whereas the approaches of Mandal et al. and Champion et al. fused the time series in a single input vector. For the training part, the work by Mandal et al. used alternating conditional expectation algorithm to calculate optimal transformations by fast boxcar averaging the rank-ordered data, whereas the Champion et al. approach used backpropagation. The neurally inspired approaches in Cook et al., Weber et al., and our system used HL, Sigma-Pi (Sum-Product) learning, and a combination of competition and cooperation for correlation learning, respectively. Finally, for inference, we used the systems resulting from the training phase (without modification) for one pass (forward pass) of unseen data through the system (i.e., basically accounting to a series of linear algebra operations).

Our system in all of our experiments, data depicting tumor growth, pharmacokinetics, and chemotherapy regimens are fed to our system, which encodes each time series in the SOMs and learns the underlying relations in the Hebbian matrix. The SOMs are responsible for bringing the time series in the same latent representation space where they can interact (i.e., through their internal correlation). Throughout the experiments, each of the SOM has  $N = 100$  neurons, the Hebbian connection matrix has size  $N \times N$ , and parametrization is done as follows:  $\alpha = [0.01, 0.1]$  decaying,  $\eta = 0.9$ ,  $\sigma = \frac{N}{2}$  decaying following an inverse time law. The training procedure of our system follows the next steps:

- normalize the input dataset;
- set up condition to reach relaxed state (i.e., no more fluctuations  $[\Delta\epsilon]$  in the SOM neural activation and Hebbian matrix);

- for each new data item, go through the pairs of neural populations (i.e., SOMs) and compute activation;
- for cross-connection among SOMs compute the Hebbian matrix entries;
- after convergence (i.e., reached  $\Delta\epsilon$ ), the system comprises the learned relation encoded in the matrix;

The testing procedure of our system follows the next steps:

- decode the encoded relation from the Hebbian matrix;
- denormalize data to match the original input space;
- compare with ground truth.

An important aspect is that for our system, after convergence (i.e., reaching an  $\Delta\epsilon$  of changes in weights), the content of the Hebbian matrix is decoded. This amounts to a process in which the (now) static layout of values in the matrix actually depicts the underlying function  $y = f(x)$ . Our system is basically updating the weights and shapes of the tuning curves (i.e., preferred values) of the SOMs and the cross-SOM Hebbian weights in the training process. After training, for inference and testing, the decoded function (i.e., using Brent's derivative-free optimization method) accounts for a typical regression neural network for which cross-validation is applied. More precisely, we ran a fourfold cross-validation for each dataset.

*Cook et al.* For the neural network system proposed by Cook et al. (2010), in all our experiments, we used neural populations with 200 neurons each, a 0.001 WTA settling threshold, 0.005 scaling factor in homeostatic activity regulation, 0.4 amplitude target for homeostatic activity regulation, and 250 training epochs. More details and the reference codebase are available on GitHub.

*Weber et al.* For the neural network system proposed by Weber and Wermter (2007), in all our experiments, we used a network with 15 neurons, 0.001 learning rate, 200,000 training epochs, and unit normalization factor. The fully parametrized codebase is available, along the other systems reference implementations, on GitHub.

**Mandal et al.** For the CCA-based system proposed by Mandal and Cichocki (2013), in all our comparative experiments, we used a sample size of 100, replication factor 10, 0.5 divergence factor, 1,000 variable permutations, and 1.06 bandwidth for Gaussian kernel density estimate. The full codebase is provided, along the other systems reference implementations, on GitLab.

**Champion et al.** For our DL implementation, we used the system of Champion et al. (2019) as a reference. We then modified the structure to accommodate the peculiarities of the clinical data. In all the experiments, the DL system contained hidden layers of size 128 neurons, trained for 100 epochs, with a mini-batch size of 1, and 50% augmentation percentage. The full codebase is provided, along the other systems reference implementations, on GitLab. Another important implementation aspect is that we use a combination of SVM and DL approaches. While SVM can work with a limited amount of data, DL models tend to perform worse when big data are not available. Therefore, we test multiple approaches to artificially augment the training data:

- **DL with no augmentation, DL.** We train the model directly from the data without further transformations.
- **DL with SVM augmentation, DL + SVM.** We used the SVM model trained beforehand to enhance the data. We set a number of observations that we want to enhance and generate random timestamps we use for prediction using SVM. Then we add those artificial values as new observations for training.
- **DL with SVM augmentation and random noise, DL + SVM + noise.** We follow the same process as in SVM augmentation, but before adding the predictions to the training pool, we add normal noise.

For the SVM we use one input feature, the days passed, and one output feature, the size of the tumor. For DL, we use the gated recurrent units (GRUs; Chung et al. (2014)) as building blocks to design a structure inspired by the work of Champion et al. (2019). The architecture consists on one GRU layer, one ReLU activation, a fully connected layer, and another ReLU activation. We designed a simple architecture to better suit the model to the scarce availability of data inspired by the study by Berg and Nyström (2019). As the DL model is a recurrent model, our input data consist of all data available from a certain patient up to a point. Both models normalize the data (both days and tumor size) by dividing by the maximum value observed. For consistency across methods, we run a fourfold cross-validation for each dataset (except dataset 0, which has only two samples; therefore, we run a twofold cross-validation). We present the average results over the cross-validation. The complete parametrization and implementation are available on GitLab.

### 3.3 Results

As previously mentioned, we evaluate the systems on a series of instantiations depicting various decision support tasks relevant for clinical use. All of the five models were evaluated through multiple metrics (Table 2) on each of the four cell line datasets. In order to evaluate the distribution of the measurement error as a

**TABLE 2 |** Evaluation metrics for data-driven relation learning systems. We consider  $N$ —number of measurements,  $\sigma$ —standard deviation of data,  $p$ —number of parameters of the model.

Metric	Equation
SSE	$\sum_{i=1}^N (\frac{y^i - y_m^i}{\sigma})^2$
RMSE	$\sqrt{\frac{SSE}{N-p}}$
sMAPE	$\frac{1}{N} \sum_{i=1}^N (2 \frac{ y^i - y_m^i }{( y^i  +  y_m^i )})$
AIC	$N \ln(\frac{SSE}{N}) + 2p$
BIC	$N \ln(\frac{SSE}{N}) + \ln(N)p$

function of the measured volumes of the tumors, the work of Benzekry et al. (2014) recommended the following model for the standard deviation of the error  $\sigma_i$  at each measurement time point  $i$ ,

$$\sigma_i = \begin{cases} \sigma (y_m^i)^\alpha, & \text{if } y_m^i \geq y^i \\ \sigma (y^i)^\alpha, & \text{if } y_m^i < y^i \end{cases}$$

This model shows that when overestimating ( $y_m \geq y$ ), the measurement error  $\alpha$  is subproportional, and when underestimating ( $y_m < y$ ), the obtained error is the same as the measured data points. In our experiments, we consider  $\alpha = 0.84$  and  $\sigma = 0.21$  as a good trade-off of error penalty and enhancement. We use this measurement error formulation to calculate the typical performance indices (i.e., sum of squared errors [SSE], root mean squared error [RMSE], symmetric mean absolute percentage error [sMAPE]) and goodness-of-fit and parsimony (i.e., Akaike information criterion [AIC] and Bayesian information criterion [BIC]), as shown in Table 2.

#### 3.3.1 Learning Growth Patterns of Preinvasive Breast Cancer

Analyzing tumor infiltration patterns, clinicians can evaluate the evolution of neoplastic processes, for instance, from DCIS to breast cancer. Such an analysis can provide very important benefits, in early detection, in order to (1) increase patient survival, (2) decrease the likelihood for multiple surgeries, and (3) determine the choice of adjuvant versus neoadjuvant chemotherapy. For a full analysis and in-depth discussion of our system's capabilities for such a task, refer to Axenie and Kurz (2020b). For this task, we assessed the capability of the evaluated systems to learn the dependency between histopathologic and morphological data. We fed the systems with DCIS data from Edgerton et al. (2011), namely, time series of nutrient diffusion penetration length within the breast tissue ( $L$ ), ratio of cell apoptosis to proliferation rates ( $A$ ), and radius of the breast tumor ( $R$ ). The study by Edgerton et al. (2011) postulated that the value of  $R$  depends on  $A$  and  $L$  following a "master equation" Eq. 9

$$A = 3 \frac{L}{R} \left( \frac{1}{\tanh(\frac{R}{L})} - \frac{L}{R} \right) \quad (9)$$

whose predictions are consistent with nearly 80% of *in situ* tumors identified by mammographic screenings. For this



**TABLE 3 |** Evaluation of the data-driven relation learning systems.

Dataset/system	Evaluation metrics		
	SSE	RMSE	sMAPE
Breast (DCIS), Edgerton et al. (2011)			
Cook et al.	56.321	0.4867	0.5901
Weber et al.	59.879	0.5099	0.6512
Mandal et al.	62.346	0.5617	0.6800
Champion et al.	58.645	0.4721	0.6054
Our system	54.216	0.4656	0.5734

initial evaluation of the data-driven mathematical relations learning systems, we consider three typical performance metrics (i.e., SSE, RMSE, and sMAPE, respectively) against the experimental data (i.e., ground truth and Eq. 9):

As one can see in Table 3, our system overcomes the other approaches on predicting the nonlinear dependency between radius of the breast tumor ( $R$ ) given the nutrient diffusion penetration length within the breast tissue ( $L$ ) and ratio of cell apoptosis to proliferation rates ( $A$ ) from real *in vivo* histopathologic and morphological data.

### 3.3.2 Learning Unperturbed Tumor Growth Curves Within and Between Cancer Types

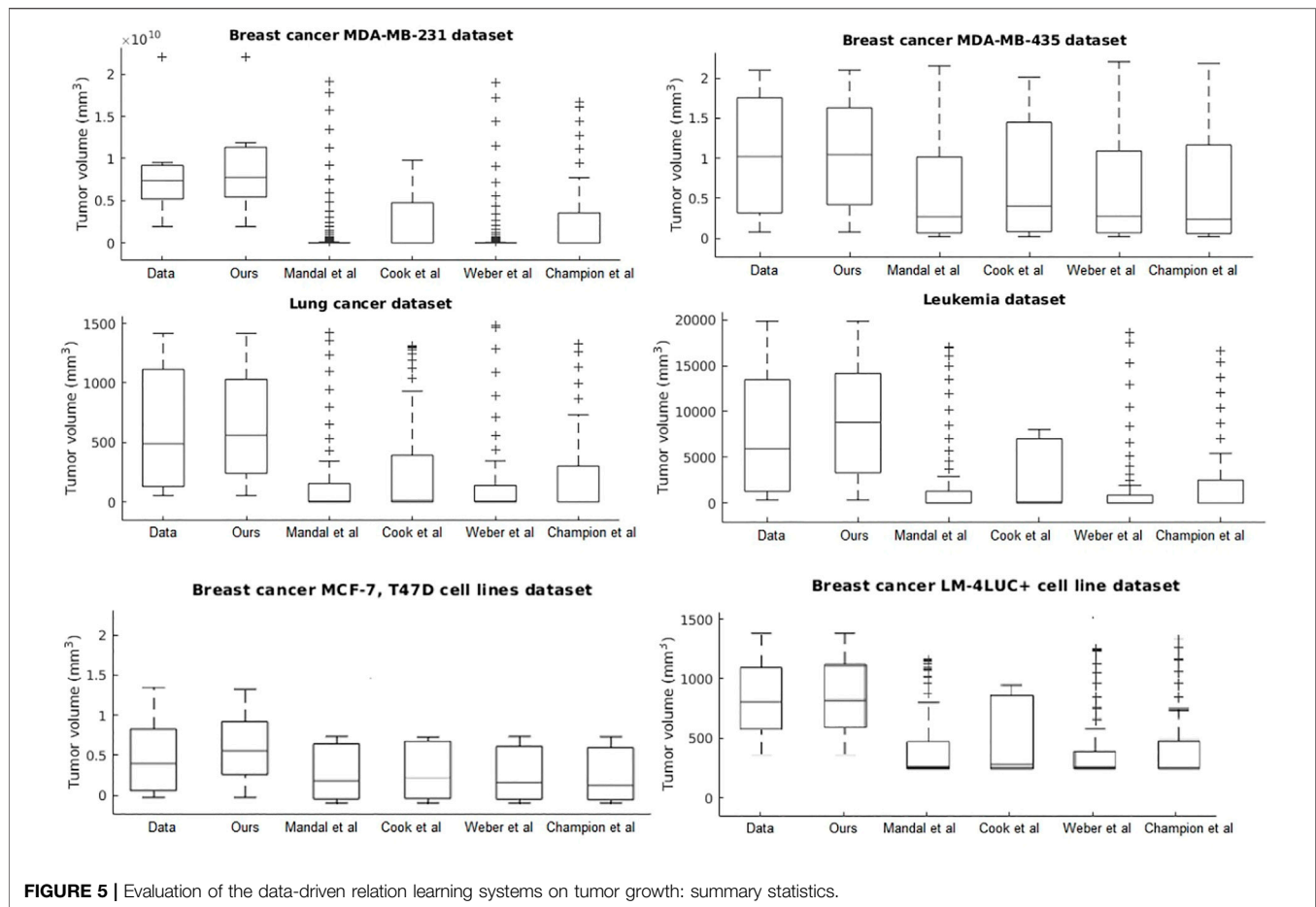
In the second task, we evaluated the systems on learning unperturbed (i.e., growth without treatment) tumor growth curves. The choice of different cancer types (i.e., two breast cell lines, lung, and leukemia) is to probe and demonstrate between- and within-tumor-type prediction versatility.

Our system provides overall better accuracy between- and within-tumor-type growth curve prediction, as shown in Table 4

**TABLE 4 |** Evaluation of the data-driven relation learning systems on tumor growth curve extraction.

Dataset/system	Evaluation metrics (smaller value is better)					
	SSE	RMSE	sMAPE	AIC	BIC	Rank3
Breast4 cancer						
Cook et al.	7,009.6	37.4423	1.7088	52.3639	52.2557	2
Weber et al.	8,004.9	44.7350	1.7088	55.2933	55.1310	5
Mandal et al.	7,971.8	39.9294	1.7088	53.2643	53.1561	4
Champion et al.	6,639.1	40.7403	1.4855	53.9837	53.8215	3
Our system	119.3	4.1285	0.0768	19.8508	19.8508	1
Breast5 cancer						
Cook et al.	0.2936	0.1713	0.1437	-40.5269	-39.5571	4
Weber et al.	0.2315	0.1604	0.1437	-41.3780	-39.9233	2
Mandal et al.	0.3175	0.1782	0.1437	-39.5853	-38.6155	5
Champion et al.	0.2699	0.1732	0.1512	-39.5351	-38.0804	3
Our system	0.0977	0.0902	0.0763	-57.7261	-57.7261	1
Breast6 cancer						
Cook et al.	3.0007	0.7071	1.0606	50.1322	51.2887	2
Weber et al.	3.2942	0.8116	1.6626	56.4133	55.1915	5
Mandal et al.	3.1908	0.7292	1.3506	53.2643	52.5421	4
Champion et al.	3.4772	0.8339	1.1288	53.9837	53.7775	3
Our system	0.7668	0.3096	0.2615	19.3208	19.1298	1
Breast7 cancer						
Cook et al.	45.6031	2.3875	1.2216	-40.0084	-39.9975	4
Weber et al.	56.0738	2.8302	1.8346	-41.2345	-39.1234	2
Mandal et al.	53.2428	2.5797	1.4816	-39.5853	-37.1260	5
Champion et al.	54.7189	2.7958	1.5086	-39.1234	-38.0664	3
Our system	0.2008	0.1417	0.0364	-57.1221	-57.6112	1
Lung cancer						
Cook et al.	44.5261	2.2243	1.5684	19.3800	20.1758	2
Weber et al.	54.1147	2.6008	1.5684	23.5253	24.7190	5
Mandal et al.	53.2475	2.4324	1.5684	21.3476	22.1434	4
Champion et al.	50.6671	2.5166	1.5361	22.8012	23.9949	3
Our system	3.6903	0.5792	0.2121	-12.0140	-12.0140	1
Leukemia						
Cook et al.	223.7271	3.2640	1.6368	56.3235	58.5944	2
Weber et al.	273.6770	3.6992	1.6368	62.9585	66.3649	5
Mandal et al.	259.9277	3.5182	1.6368	59.7729	62.0439	4
Champion et al.	248.5784	3.5255	1.6001	60.7461	64.1526	3
Our system	35.2541	1.2381	0.3232	9.8230	9.8230	1

Notes: 3—Calculated as best in 3/5 metrics; 4—MDA-MB-231 cell line; 5—MDA-MB-435 cell line; 6—MCF-7, T47D cell line; 7—LM2-4LUC + cell line.



**FIGURE 5 |** Evaluation of the data-driven relation learning systems on tumor growth: summary statistics.

and the summary statistics (depicted in **Figure 5**). The superior performance is given by the fact that our system can overcome the other approaches when facing incomplete biological descriptions, the diversity of tumor types, and the small size of the data. Interested readers can refer to Axenie and Kurz (2021) for a deeper performance analysis of our system.

### 3.3.3 Extracting Tumor Phenotypic Stage Transitions

The next evaluation task looks at learning the mathematical relations describing the phenotypic transitions of tumors in breast cancer. For this experiment, we considered the study of 17 breast cancer patients in the study by Edgerton et al. (2011). Typically, in the breast cancer phenotypic state space, quiescent cancer cells (Q) can become proliferative (P) or apoptotic (A). In addition, nonnecrotic cells become hypoxic if the oxygen supply drops below a threshold value. But, hypoxic cells can recover to their previous state or become necrotic, as shown by Macklin et al. (2012).

In this instantiation, we focus on a simplified three-state phenotypic model (i.e., containing P, Q, A states). The transitions among tumor states are stochastic events generated by Poisson processes. Each of the data-driven relation learning systems is fed with time series of raw immunohistochemistry and morphometric data for each of the 17 tumor cases (Edgerton et al., 2011; **Supplementary Tables S1, S2**) as follows: cell cycle

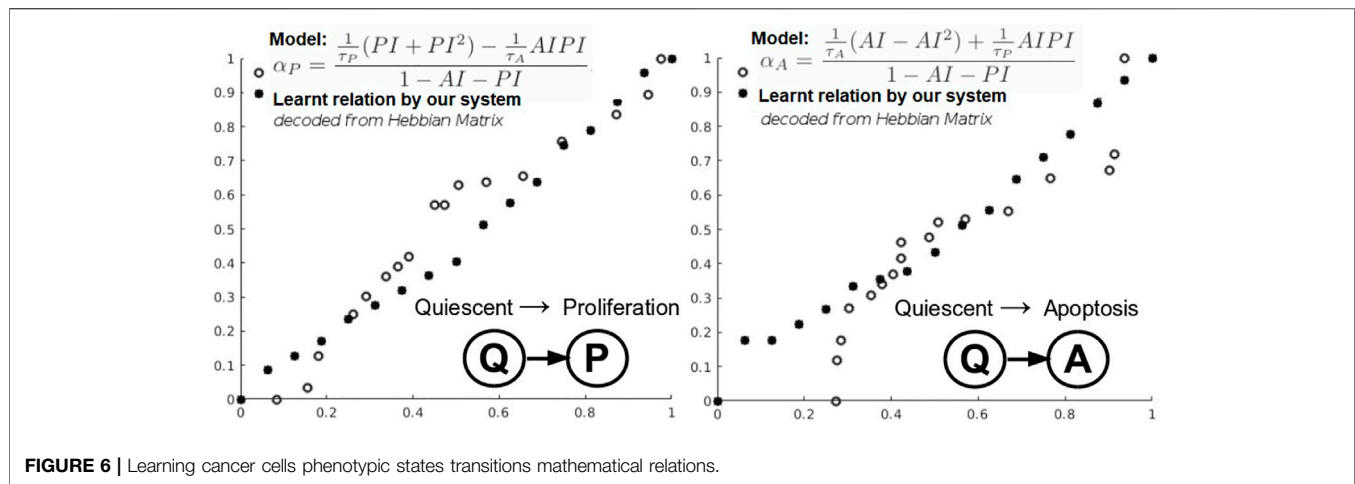
time  $\tau_P$ , cell apoptosis time  $\tau_A$ , proliferation index  $PI$ , and apoptosis index  $AI$ . Given this time series input, each system needs to infer the mathematical relations for  $\alpha_P$ , the mean quiescent-to-proliferation (Q-P) transition rate, and  $\alpha_A$ , the quiescent-to-apoptosis (Q-A) transition rate, respectively (**Figure 6**). Their analytical form state transition is given by:

$$\alpha_P = \frac{\frac{1}{\tau_P} (PI + PI^2) - \frac{1}{\tau_A} AIPI}{1 - AI - PI}, \alpha_A = \frac{\frac{1}{\tau_A} (AI - AI^2) + \frac{1}{\tau_P} AIPI}{1 - AI - PI} \quad (10)$$

Q-A and Q-P state transitions of cancer cells are depicted in **Figure 6**, where we also present the relation that our system learned. Both in **Figure 6** and **Table 5**, we can see that our system is able to recover the correct underlying mathematical function with respect to ground truth (clinically extracted and modeled **Eq. 10** from the study by Macklin et al. (2012)).

### 3.3.4 Simultaneously Extracting Drug-Perturbed Tumor Growth and Drug Pharmacokinetics

Chemotherapy use in the neoadjuvant and adjuvant settings generally provides the same long-term outcome (de Wiel et al., 2017). But what is the best choice for a particular patient? This question points at those quantifiable patient-specific factors (e.g.,



**FIGURE 6 |** Learning cancer cells phenotypic states transitions mathematical relations.

**TABLE 5 |** Evaluation of the data-driven relation learning systems for extracting phenotypic transitions relations.

State transition/system	Evaluation metrics		
	SSE	RMSE	sMAPE
Quiescent(Q) to proliferation(P) transition relation			
Cook et al.	0.820	0.240	0.190
Weber et al.	0.865	0.294	0.196
Mandal et al.	0.904	0.320	0.214
Champion et al.	0.845	0.274	0.189
Our system	0.750	0.210	0.172
Quiescent(Q) to apoptosis(A) transition relation			
Cook et al.	0.421	0.162	0.140
Weber et al.	0.484	0.178	0.154
Mandal et al.	0.490	0.182	0.151
Champion et al.	0.441	0.166	0.147
Our system	0.398	0.153	0.131

Note that none of the evaluated system had prior knowledge of the data distribution or biological assumptions. To have a more detailed overview on the capabilities of our system to capture phenotypic dynamics, refer to Axenie and Kurz (2020c).

tumor growth curve under chemotherapy, drug pharmacokinetics) that influence the sequencing of chemotherapy and surgery in a therapy plan. A large variety of breast cancer tumor growth patterns used in cancer treatments planning were identified experimentally and clinically and modeled over the years (Gerlee, 2013). In addition, progress in pharmacokinetic modeling allowed clinicians to investigate the effect of covariates in drug administration, as shown in the work by Zaheed et al. (2019). Considering breast cancer, paclitaxel is a typical drug choice with broad use in monotherapy as well as immune-combined therapies (Stage et al., 2018).

In the current section, we present the experimental results of all the evaluated systems and consider (1) accuracy in learning the chemotherapy-perturbed tumor growth model and (2) accuracy in learning the pharmacokinetics of the chemotoxic drug (i.e., paclitaxel) dose. For the tumor growth curve extraction, we considered four cell lines of breast cancer (i.e., MDA-MB-231, MDA-MB-435, MCF-7, LM2-LUC + cell lines; Table 1). The evaluation results of the systems in the perturbed tumor

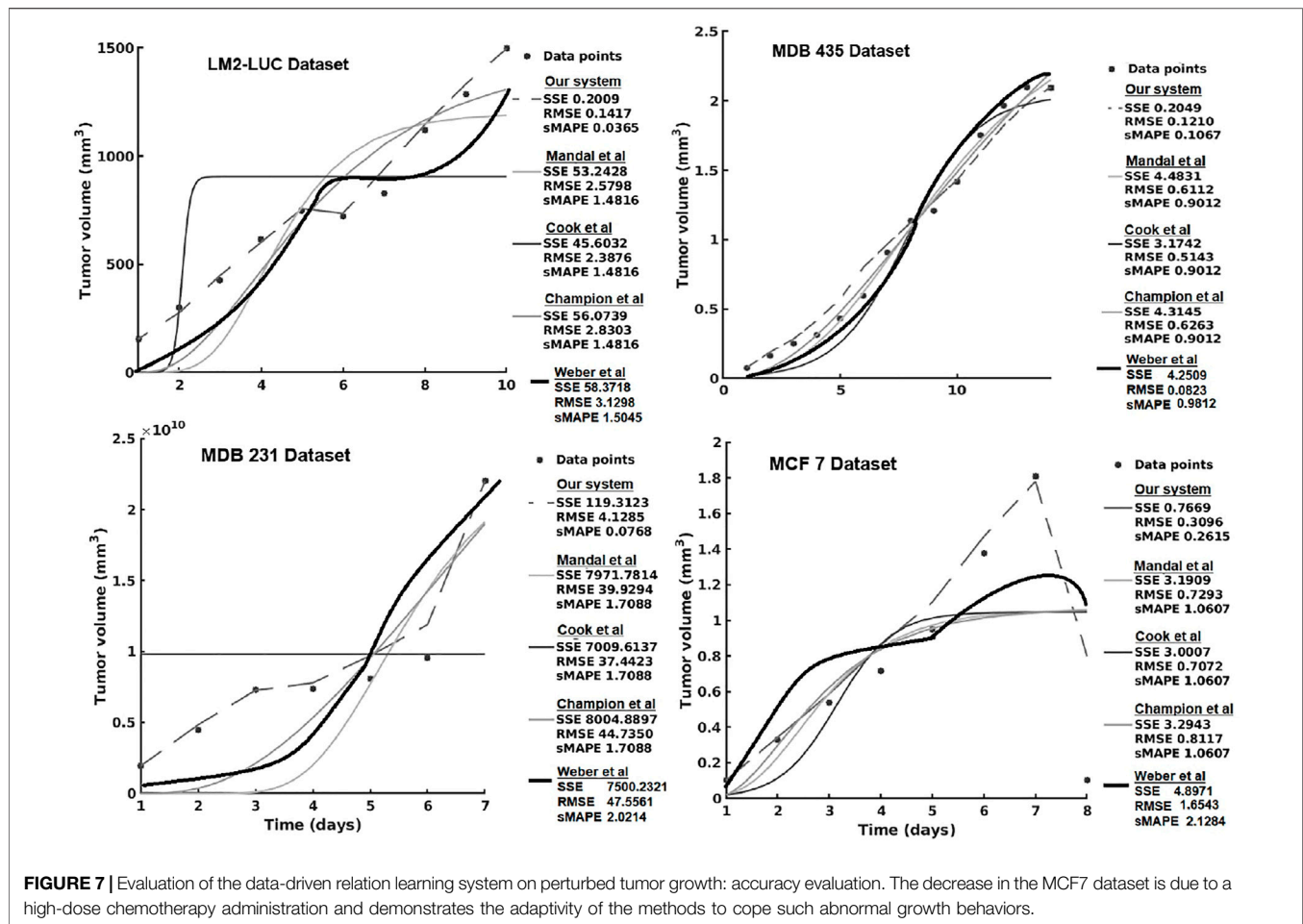
growth scenario are provided in Figure 7. Note that our system learns the temporal relationships among the quantities fed to the two sides of the system (Figure 3), which can, subsequently, be used to infer one (unavailable) quantity based on the one available. For instance, if the system had learned the change in volume at irregular time points, given a next time point, the system will recover the most plausible volume value—basically accounting for a one-step-ahead prediction. For a longer prediction horizon, one can recurrently apply this process for new predictions and so on.

Table 6 presents the results using SVM and the different versions of DL. We can see that usually vanilla DL outperforms SVM. DL is a more complex model, as well as uses more input data, so this result is expected. Once we add the augmentation from SVM, the model has a comparable performance to SVM. Our theory is that this is caused by DL learning to imitate SVM instead of real data. Once we add noise to the augmentation, the data become more realistic and usually yield improvements in performance.

For the pharmacokinetics learning experiments, we used the data from the computational model of intracellular pharmacokinetics of paclitaxel of Kuh et al. (2000) describing the kinetics of paclitaxel uptake, binding, and efflux from cancer cells in both intracellular and extracellular contexts.

As one can see in Figure 8A, the intracellular concentration kinetics of paclitaxel is highly nonlinear. Our system is able to extract the underlying function describing the data without any assumption about the data and other prior information, opposite to the model from (Kuh et al., 2000). Interestingly, our system captured a relevant effect consistent with multiple paclitaxel studies (Stage et al., 2018), namely, that the intracellular concentration increased with time and approached plateau levels, with the longest time to reach plateau levels at the lowest extracellular concentration—as shown in Figure 8.

Analyzing the extracellular concentration in Figure 8B, we can see that our system extracted the trend and the individual variation of drug concentration after the administration of the drug (i.e., in the first 6 h) and learned an accurate fit without any prior or other biological assumptions. Interestingly, our system captured the fact that the intracellular drug concentration



**TABLE 6 |** Description of the DL approach inspired by Champion et al. (2019).

Evaluation of the deep learning approach					
Dataset	Cancer (cell line)	RMSE <sup>SVM</sup>	RMSE <sup>DL</sup>	RMSE <sup>DL+SVM</sup>	RMSE <sup>DL+SVM+noise</sup>
1	Breast (MDA-MB-231)	1.8424	1.5382	1.7544	1.7088
2	Breast (MDA-MB-435)	1.0977	1.5990	0.9584	0.9012
3	Breast (MCF-7)	1.4112	1.7295	1.3632	1.0607
4	Breast (LM2-4LUC+)	1.8945	1.8345	1.7620	1.4816

Note that for all of the evaluation datasets, the best performing DL approach [i.e., inspired by Champion et al. (2019)] is the combined DL—SVM—noise configuration.

increased linearly with extracellular concentration decrease, as shown in **Figure 8**.

The overall evaluation of pharmacokinetics learning is given in **Table 7**.

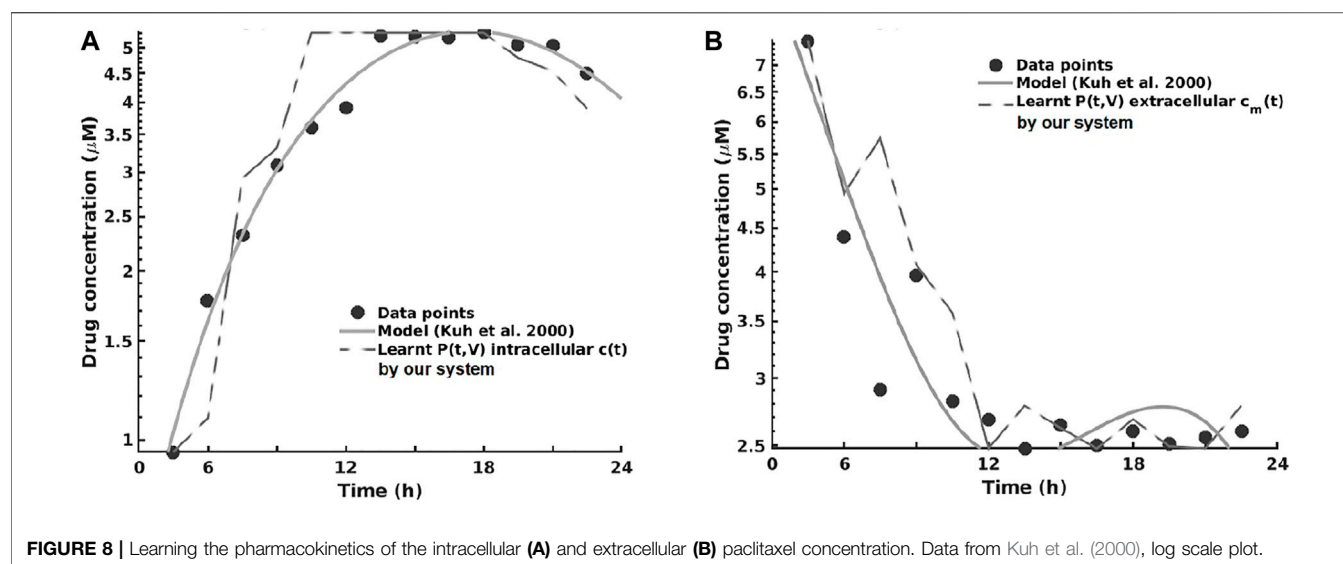
In this series of experiments, all of the systems learned that changes in cell number were represented by changes in volume, which (1) increased with time at low initial total extracellular drug concentrations due to continued cell proliferation and (2) decreased with time at high initial total extracellular drug concentrations due to the antiproliferative and/or cytotoxic drug effects, as reported by Kuh et al. (2000). In order to assess the impact the predictions have on therapy sequencing (i.e., neoadjuvant vs. adjuvant chemotherapy), refer to Axenie and Kurz (2020a).

### 3.3.5 Predicting Tumor Growth/Recession Under Chemotherapy

In the last series of experiments, we used real patient data from the I-SPY 1 TRIAL: ACRIN 6657 (Yee et al., 2020). Data for the 136 patients treated for breast cancer in the I-SPY-1 clinical trial were obtained from the cancer imaging archive<sup>3</sup> and the Breast Imaging Research Program at UCSF. The time series data contained only the largest tumor volume from magnetic resonance imaging measured before therapy, 1 to 3 days after therapy, between therapy cycles, and before surgery,

<sup>3</sup><https://wiki.cancerimagingarchive.net/display/Public/ISPY1>.





**TABLE 7 |** Evaluation of the data-driven relation learning systems for pharmacokinetics extraction.

Pharmacokinetics data/system	Evaluation metrics		
	SSE	RMSE	sMAPE
Intracellular paclitaxel			
Cook et al.	0.6234	0.2812	0.1487
Weber et al.	0.7212	0.3689	0.1794
Mandal et al.	0.6743	0.3046	0.1602
Champion et al.	0.6539	0.2607	0.1500
Our system	0.5960	0.2141	0.1403
Extracellular paclitaxel			
Cook et al.	0.5676	0.2341	0.1213
Weber et al.	0.6674	0.2891	0.1289
Mandal et al.	0.6128	0.2974	0.1366
Champion et al.	0.5790	0.2633	0.1156
Our system	0.5484	0.2054	0.1068

respectively. To summarize, the properties of the dataset are depicted in **Figure 9**.

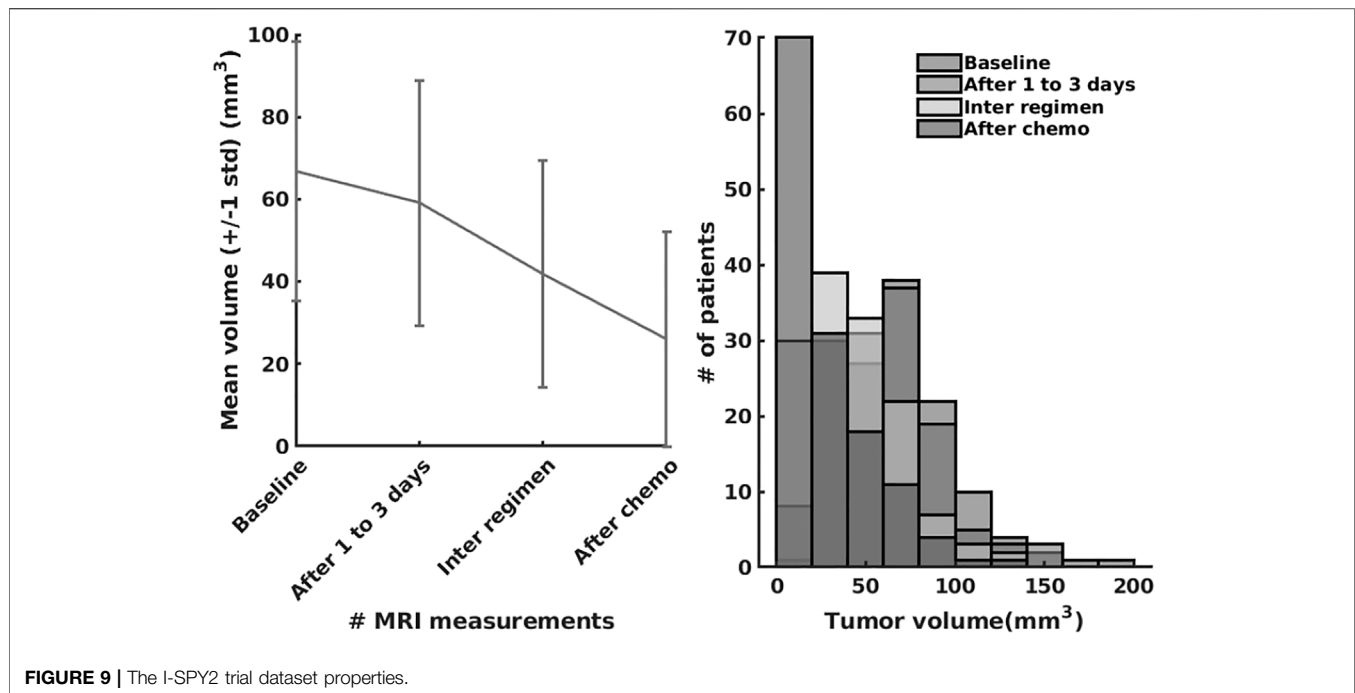
As we can observe in **Table 8**, our system learns a superior fit to the tumor growth data, with respect to the other systems, despite the limited number of samples (i.e., 7 data points for MDA-MD-231 cell line dataset and up to 14 data points for MDA-MD-435 cell line dataset). It is important to note that when analyzing tumor growth functions and response under chemotherapy, we faced the high variability among patients given by the typical constellation of hormone receptor indicators (i.e., HR and HER2neu, which covered the full spectrum of positive and negative values) for positive and negative prognoses. All data-driven learning systems capture such aspects to some extent. Our system learns a superior fit overall the three metrics, capturing the intrinsic impact chemotherapy has upon the tumor growth function, despite the limited number of samples (i.e., 4 data points of the dataset overall evaluation dataset of 20% of patients). An extended evaluation of our system on a broader set of datasets for therapy outcome prediction is given by Kurz and Axenie (2020).

## 4 DISCUSSION

We complement the quantitative evaluation in the previous section with an analysis of the most important features of all the systems capable to extract mathematical relations in the aforementioned clinical oncology tasks. As the performance evaluation was done in the previous section, we will now focus on other specific comparison terms relevant for the adoption of such systems in clinical practice.

One initial aspect is the design and functionality. Using either distributed representations (Cook et al., 2010; Weber and Wermter, 2007; Champion et al., 2019) or compact mathematical forms Mandal and Cichocki (2013), all methods encoded the input variables in a new representation to facilitate computation. At this level, using neural network dynamics (Cook et al., 2010; Weber and Wermter, 2007) or pure mathematical multivariate optimization (Mandal and Cichocki, 2013; Champion et al., 2019), the solution was obtained through iterative processes that converged to consistent representations of the data. Our system employs a lightweight learning mechanism, offering a transparent processing scheme and human-understandable representation of the learned relations as shown in **Figure 4**. Besides the capability to extract the correlation among the two features, the system can simultaneously extract the shape of the distribution of the feature spaces. This is an important feature when working with rather limited medical data samples.

A second aspect refers to the amount of prior information embedded by the designer in the system. It is typical that, depending on the instantiation, a new set of parameters is needed, making the models less flexible. Although less intuitive, the pure mathematical approaches (Mandal and Cichocki, 2013) (i.e., using CCA) need less tuning effort due to the fact that their parameters are the result of an optimization procedure. On the other side, the neural network approaches (Cook et al., 2010; Weber and Wermter, 2007; Champion et al., 2019) need a more judicious parameter tuning, as their dynamics



**FIGURE 9 |** The I-SPY2 trial dataset properties.

**TABLE 8 |** Evaluation of the data-driven relation learning systems on real patient breast cancer data.

Dataset/model	Evaluation metrics		
	SSE	RMSE	sMAPE
I-SPY2 trial Yee et al. (2020)			
Cook et al.	1.3735	1.1439	0.1133
Weber et al.	1.7543	1.2005	0.2539
Mandal et al.	2.963	1.0963	0.7834
Champion et al.	2.0747	1.04100	0.1073
Our system	0.8650	0.4650	0.0389

are more sensitive and can reach either instability (e.g., recurrent networks) or local minima. Except parametrization, prior information about inputs is generally needed when instantiating the system for a certain scenario. Sensory value bounds and probability distributions must be explicitly encoded in the models through explicit distribution of the input space across neurons in the studies by Cook et al. (2010) and Weber and Wermter (2007), linear coefficients in vector combinations (Mandal and Cichocki, 2013), or standardization routines of input variables (Champion et al., 2019). Our system exploits only the available data to simultaneously extract the data distribution and the underlying mathematical relation governing tumor growth processes. Capable of embedding priors (i.e., mechanistic models) in its structure, our system can speed up its computation, through a data-driven model refinement similar in nature with the unsupervised learning process. Basically, in order to combine the learning process with a mechanistic model, the only update will be done in the factorization of the weight update in Eq. 6.

A third aspect relevant to the analysis is the stability and robustness of the obtained representation. The representation of the hidden relation (1) can be encoded in a weight matrix Cook et al. (2010) and Weber and Wermter (2007) such that, after learning, given new input, the representation is continuously refined to accommodate new inputs; (2) can be fixed in vector directions of random variables requiring a new iterative algorithm run from initial conditions to accommodate new input (Mandal and Cichocki, 2013); or (3) can be obtained as an optimization process given the new available input signals (Champion et al., 2019). Given initial conditions, prior knowledge and an optimization criteria (Mandal and Cichocki, 2013) or a recurrent relaxation process toward a point attractor (Cook et al., 2010; Weber and Wermter, 2007; Champion et al., 2019) are required to reach a desired tolerance. Our system exploits the temporal regularities among tumor growth data covariates, to learn the governing relations using a robust distributed representation of each data quantity. The choice of a distributed representation to encode and process the input data gives out the system an advantage in terms of explainability for clinical adoption. As shown in Figure 3, each scalar quantity can be projected in a high dimension where the shape of the distribution can be inferred. Such insights can support the decisions of the system by explaining its predictions.

The capability to handle noisy data is an important aspect concerning the applicability in real-world scenarios. Using either computational mechanisms for denoising (Cook et al., 2010; Weber and Wermter, 2007), iterative updates to minimize a distance metric Mandal and Cichocki (2013), or optimization Champion et al. (2019), each method is capable to cope with moderate amounts of noise. Despite this, some methods have intrinsic methods to cope with noisy data intrinsically, through their dynamics, by recurrently

propagating correct estimates and balancing new samples (Cook et al., 2010). The distributed representation used in our system ensures that the system is robust to noise, and the local learning rules ensure fast convergence on real-world data—as our experiments demonstrated.

Another relevant feature is the capability to infer (i.e., predict/anticipate) missing quantities once the mathematical relation is learned. The capability to use the learned relations to determine missing quantities is not available in all presented systems, such as the system of Mandal and Cichocki (2013). This is due to the fact that the divergence and correlation coefficient expressions might be noninvertible functions that support a simple pass-through of available values to extract missing ones. On the other side, using either the learned co-activation weight matrix (Cook et al., 2010; Weber and Wermter, 2007) or the known standard deviations of the canonical variants (Champion et al., 2019), some systems are able to predict missing quantities. Our system stores learned mathematical relations in the Hebbian matrix, which can be used bidirectionally to recover missing quantities on one side of the input given the other available quantity. This feature is crucial for the predictive aspects of our system. Basically, in its typical operation, the system learns from sets of observations the underlying relations among quantities describing the tumor's state (e.g., growth curve, phenotypic stage, extracellular drug concentration). For prediction purposes, the system is fed with only one quantity (e.g., time index) and, given the learned relation, will recover the most plausible value for the correlated quantity that was trained with (e.g., growth curve) for the next step.

Finally, because of the fact that all methods reencode the real-world values in new representation, it is important to study the capability to decode the learned representation and subsequently measure the precision of the learned representation. Although not explicitly treated in the presented systems, decoding the extracted representations is not trivial. Using a tiled mapping of the input values along the neural network representations, the system of Cook et al. (2010) decoded the encoded value in activity patterns by simply computing the distribution of the input space over the neural population units, whereas Weber and Wermter (2007) used a simple WTA readout, given that the representation was constrained to have a uniquely defined mapping. Given that the model learns the relations in data space through optimization processes, as in the system of Champion et al. (2019), one can use learned curves to simply project available sensory values through the learned function to get the second value, as the scale is preserved. Albeit its capability to precisely extract nonlinear relations from high-dimensional random datasets, the system of Mandal and Cichocki (2013) cannot provide any readout mechanisms to support a proper decoded representation of the extracted relations. This is due to the fact that the method cannot recover the sign and scale of the relations. The human-understandable relation learned by our system is efficiently decoded from the Hebbian matrix back to real-world values. As our experiments demonstrate, the approach introduced through our system

excels in capturing the peculiarities that clinical data carry. Contributing to the explainability features of our system, the read-out mechanism is able to turn the human-understandable visual representation of the learned relation (**Figure 4**) into a function providing the most plausible values of the queried quantities.

## 5 CONCLUSION

Data-driven approaches to improve decision-making in clinical oncology are now going beyond diagnosis. From early detection of infiltrating tumors to unperturbed tumor growth phenotypic staging, and from pharmacokinetics-dictated therapy planning to treatment outcome, data-driven tools capable of learning hidden correlations in the data are now taking the foreground in mathematical and computational oncology. Our study introduces a novel framework and versatile system capable of learning physical and mathematical relations in heterogeneous oncology data. Together with a lightweight and transparent computational substrate, our system provides human-understandable solutions. This is achieved by capturing the distribution of the data in order to achieve superior fit and prediction capabilities between and within cancer types. Supported by an exhaustive evaluation on *in vitro* and *in vivo* data, against state-of-the-art machine learning and DL systems, the proposed system stands out as a promising candidate for clinical adoption. Mathematical and computational oncology is an emerging field where efficient, transparent, and understandable data-driven systems hold the promise of paving the way to individualized therapy. But this can only be achieved by capturing the peculiarities of a patient's tumor across scales and data types.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://gitlab.com/akii-microlab/chimera/-/tree/master/datasets>.

## AUTHOR CONTRIBUTIONS

DK designed the research, collected clinical datasets, and performed data analysis of clinical studies used in the experiments. CS developed the source code for the experiments and the analysis. CA designed the research and developed the source code for the experiments.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2021.713690/full#supplementary-material>

## REFERENCES

- Abler, D., Büchler, P., and Rockne, R. C. (2019). "Towards Model-Based Characterization of Biomechanical Tumor Growth Phenotypes," in *Mathematical and Computational Oncology*. Editors G. Bebis, T. Benos, K. Chen, K. Jahn, and E. Lima (Cham: Springer International Publishing), 75–86. doi:10.1007/978-3-030-35210-3\_6
- Antonelli, A., Noort, W. A., Jaques, J., de Boer, B., de Jong-Korlaar, R., Brouwers-Vos, A. Z., et al. (2016). Establishing Human Leukemia Xenograft Mouse Models by Implanting Human Bone Marrow-like Scaffold-Based Niches. *Blood J. Am. Soc. Hematol.* 128, 2949–2959. doi:10.1182/blood-2016-05-719021
- Axenie, C., and Kurz, D. (2020a). "Chimera: Combining Mechanistic Models and Machine Learning for Personalized Chemotherapy and Surgery Sequencing in Breast Cancer," in *International Symposium on Mathematical and Computational Oncology* (Springer), 13–24. doi:10.1007/978-3-030-64511-3\_2
- Axenie, C., and Kurz, D. (2021). "Glueck: Growth Pattern Learning for Unsupervised Extraction of Cancer Kinetics," in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V* (Springer International Publishing), 171–186. doi:10.1007/978-3-030-67670-4\_11
- Axenie, C., and Kurz, D. (2020b). "Princess: Prediction of Individual Breast Cancer Evolution to Surgical Size," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) (IEEE)*, 457–462. doi:10.1109/cbms49503.2020.00093
- Axenie, C., and Kurz, D. (2020c). "Tumor Characterization Using Unsupervised Learning of Mathematical Relations within Breast Cancer Data," in *International Conference on Artificial Neural Networks* (Springer), 838–849. doi:10.1007/978-3-030-61616-8\_67
- Ben-Shmuel, A., Biber, G., and Barda-Saad, M. (2020). Unleashing Natural Killer Cells in the Tumor Microenvironment-The Next Generation of Immunotherapy? *Front. Immunol.* 11, 275. doi:10.3389/fimmu.2020.00275
- Benzekry, S., Lamont, C., Weremowicz, J., Beheshti, A., Hlatky, L., and Hahnfeldt, P. (2019). *Tumor Growth Kinetics of Subcutaneously Implanted Lewis Lung Carcinoma Cells*. Zenodo, 3572401. doi:10.5281/zenodo.3572401
- Benzekry, S. (2020). Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology. *Clin. Pharmacol. Ther.* 108, 471–486. doi:10.1002/cpt.1951
- Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J. M. L., Hlatky, L., et al. (2014). Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth. *Plos Comput. Biol.* 10 (8), e1003800. doi:10.1371/journal.pcbi.1003800
- Berg, J., and Nyström, K. (2019). Data-driven Discovery of Pdes in Complex Datasets. *J. Comput. Phys.* 384, 239–252. doi:10.1016/j.jcp.2019.01.036
- Brent, R. P. (2013). *Algorithms for Minimization without Derivatives*. Chelmsford, United States: Courier Corporation.
- Champion, K., Lusch, B., Kutz, J. N., and Brunton, S. L. (2019). Data-driven Discovery of Coordinates and Governing Equations. *Proc. Natl. Acad. Sci. USA* 116, 22445–22451. doi:10.1073/pnas.1906995116
- 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
- Chen, N., Li, Y., Ye, Y., Palmisano, M., Chopra, R., and Zhou, S. (2014). Pharmacokinetics and Pharmacodynamics of Nab -paclitaxel in Patients with Solid Tumors: Disposition Kinetics and Pharmacology Distinct from Solvent-based Paclitaxel. *J. Clin. Pharmacol.* 54, 1097–1107. doi:10.1002/jcph.304
- Chen, Z., Haykin, S., Eggermont, J. J., and Becker, S. (2008). *Correlative Learning: A Basis for Brain and Adaptive Systems*, Vol. 49. John Wiley & Sons.
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. arXiv preprint arXiv:1412.3555.
- Comen, E., Gilewski, T. A., and Norton, L. (2016). *Tumor Growth Kinetics*. Hoboken, United States: Holland-Frei Cancer Medicine, 1–11.
- Cook, M., Jug, F., Krautz, C., and Steger, A. (2010). "Unsupervised Learning of Relations," in *International Conference on Artificial Neural Networks* (Springer), 164–173. doi:10.1007/978-3-642-15819-3\_21
- Cornish, A. J., and Markowetz, F. (2014). Santa: Quantifying the Functional Content of Molecular Networks. *Plos Comput. Biol.* 10, e1003808. doi:10.1371/journal.pcbi.1003808
- Cristini, V., Koay, E., and Wang, Z. (2017). *An Introduction to Physical Oncology: How Mechanistic Mathematical Modeling Can Improve Cancer Therapy Outcomes*. Boca Raton: CRC Press.
- de Silva, B. M., Higdon, D. M., Brunton, S. L., and Kutz, J. N. (2020). Discovery of Physics from Data: Universal Laws and Discrepancies. *Front. Artif. Intell.* 3, 25. doi:10.3389/frai.2020.00025
- Edgerton, M. E., Chuang, Y.-L., Macklin, P., Yang, W., Bearer, E. L., and Cristini, V. (2011). A Novel, Patient-specific Mathematical Pathology Approach for Assessment of Surgical Volume: Application to Ductal Carcinomain Situof the Breast. *Anal. Cell. Pathol.* 34, 247–263. doi:10.1155/2011/803816
- Gaddy, T. D., Wu, Q., Arnheim, A. D., and Finley, S. D. (2017). Mechanistic Modeling Quantifies the Influence of Tumor Growth Kinetics on the Response to Anti-angiogenic Treatment. *Plos Comput. Biol.* 13, e1005874. doi:10.1371/journal.pcbi.1005874
- 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
- Griffon-Etienne, G., Boucher, Y., Brekken, C., Suit, H. D., and Jain, R. K. (1999). Taxane-induced Apoptosis Decompresses Blood Vessels and Lowers Interstitial Fluid Pressure in Solid Tumors: Clinical Implications. *Cancer Res.* 59, 3776–3782. doi:10.1158/0008-5472
- Jansen, T., Geleijnse, G., Van Maaren, M., Hendriks, M. P., Ten Teije, A., and Moncada-Torres, A. (2020). "Machine Learning Explainability in Breast Cancer Survival," in *30th Medical Informatics Europe Conference, MIE 2020 (Amsterdam, Netherlands: IOS Press)*, 307–311.
- Kohonen, T. (1982). Self-organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* 43, 59–69. doi:10.1007/bf00337288
- Kondylakis, H., Axenie, C., Bastola, D., Katehakis, D. G., Kouroubali, A., Kurz, D., et al. (2020). Status and Recommendations of Technological and Data-Driven Innovations in Cancer Care: Focus Group Study. *J. Med. Internet Res.* 22, e22034. doi:10.2196/22034
- Kuh, H. J., Jang, S. H., Wientjes, M. G., and Au, J. L. (2000). Computational Model of Intracellular Pharmacokinetics of Paclitaxel. *J. Pharmacol. Exp. Ther.* 293, 761–770.
- Kurz, D., and Axenie, C. (2020). "Perfecto: Prediction of Extended Response and Growth Functions for Estimating Chemotherapy Outcomes in Breast Cancer," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE)*, 609–614. doi:10.1109/bibm49941.2020.9313551
- Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., and Séroussi, B. (2019). Explainable Artificial Intelligence for Breast Cancer: A Visual Case-Based Reasoning Approach. *Artif. intelligence Med.* 94, 42–53. doi:10.1016/j.artmed.2019.01.001
- Long, Z., Lu, Y., Ma, X., and Dong, B. (2018). "Pde-net: Learning Pdes from Data," in *International Conference on Machine Learning (Brookline: Microtome Publishing)*, 3208–3216.
- Macklin, P., Edgerton, M. E., Thompson, A. M., and Cristini, V. (2012). Patient-calibrated Agent-Based Modelling of Ductal Carcinoma *In Situ* (Dcis): from Microscopic Measurements to Macroscopic Predictions of Clinical Progression. *J. Theor. Biol.* 301, 122–140. doi:10.1016/j.jtbi.2012.02.002
- Mandal, A., and Cichocki, A. (2013). Non-linear Canonical Correlation Analysis Using Alpha-Beta Divergence. *Entropy* 15, 2788–2804. doi:10.3390/e15072788
- Markowetz, F., and Troyanskaya, O. G. (2007). Computational Identification of Cellular Networks and Pathways. *Mol. Biosyst.* 3, 478–482. doi:10.1039/b617014p
- Mastri, M., Tracz, A., and Ebos, J. M. (2019). 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
- Nathanson, S. D., and Nelson, L. (1994). Interstitial Fluid Pressure in Breast Cancer, Benign Breast Conditions, and Breast Parenchyma. *Ann. Surg. Oncol.* 1, 333–338. doi:10.1007/bf03187139
- Nia, H. T., Liu, H., Seano, G., Datta, M., Jones, D., Rahbari, N., et al. (2016). Solid Stress and Elastic Energy as Measures of Tumour Mechanopathology. *Nat. Biomed. Eng.* 1, 1–11. doi:10.1038/s41551-016-0004
- Nia, H. T., Munn, L. L., and Jain, R. K. (2020). Physical Traits of Cancer. *Science* 370, eaaz0868. doi:10.1126/science.aaz0868
- Raissi, M. (2018). Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations. *J. Machine Learn. Res.* 19, 932–955.



- Rodallec, A., Giacometti, S., Ciccolini, J., and Fanciullino, R. (2019). *Tumor Growth Kinetics of Human MDA-MB-231 Cells Transfected with dTomato Lentivirus*. Springer. doi:10.5281/zenodo.3593919
- Roland, C. L., Dineen, S. P., Lynn, K. D., Sullivan, L. A., Dellinger, M. T., Sadegh, L., et al. (2009). Inhibition of Vascular Endothelial Growth Factor Reduces Angiogenesis and Modulates Immune Cell Infiltration of Orthotopic Breast Cancer Xenografts. *Mol. Cancer Ther.* 8, 1761–1771. doi:10.1158/1535-7163.MCT-09-0280
- Rouvière, O., Melodelima, C., Hoang Dinh, A., Bratan, F., Pagnoux, G., Sanzalone, T., et al. (2017). Stiffness of Benign and Malignant Prostate Tissue Measured by Shear-Wave Elastography: a Preliminary Study. *Eur. Radiol.* 27, 1858–1866. doi:10.1007/s00330-016-4534-9
- Sarapata, E. A., and de Pillis, L. G. (2014). A Comparison and Catalog of Intrinsic Tumor Growth Models. *Bull. Math. Biol.* 76, 2010–2024. doi:10.1007/s11538-014-9986-y
- Schaeffer, H. (2017). Learning Partial Differential Equations via Data Discovery and Sparse Optimization. *Proc. R. Soc. A.* 473, 20160446. doi:10.1098/rspa.2016.0446
- Simpson-Herren, L., and Lloyd, H. H. (1970). Kinetic Parameters and Growth Curves for Experimental Tumor Systems. *Cancer Chemother. Rep.* 54, 143–174.
- Stage, T. B., Bergmann, T. K., and Kroetz, D. L. (2018). Clinical Pharmacokinetics of Paclitaxel Monotherapy: an Updated Literature Review. *Clin. Pharmacokinet.* 57, 7–19. doi:10.1007/s40262-017-0563-z
- Tan, G., Kasuya, H., Sahin, T. T., Yamamura, K., Wu, Z., Koide, Y., et al. (2015). Combination Therapy of Oncolytic Herpes Simplex Virus Hf10 and Bevacizumab against Experimental Model of Human Breast Carcinoma Xenograft. *Int. J. Cancer* 136, 1718–1730. doi:10.1002/ijc.29163
- Uzhachenko, R. V., and Shanker, A. (2019). Cd8+ T Lymphocyte and Nk Cell Network: Circuitry in the Cytotoxic Domain of Immunity. *Front. Immunol.* 10, 1906. doi:10.3389/fimmu.2019.01906
- Van de Wiel, M., Dockx, Y., Van den Wyngaert, T., Stroobants, S., Tjalma, W. A. A., and Huizing, M. T. (2017). Neoadjuvant Systemic Therapy in Breast Cancer: Challenges and Uncertainties. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 210, 144–156. doi:10.1016/j.ejogrb.2016.12.014
- Volk, L. D., Flister, M. J., Chihade, D., Desai, N., Trieu, V., and Ran, S. (2011). Synergy of Nab-Paclitaxel and Bevacizumab in Eradicating Large Orthotopic Breast Tumors and Preexisting Metastases. *Neoplasia* 13, 327–IN14. doi:10.1593/neo.101490
- Weber, C., and Wermter, S. (2007). A Self-Organizing Map of Sigma-Pi Units. *Neurocomputing* 70, 2552–2560. doi:10.1016/j.neucom.2006.05.014
- Werner, B., Scott, J. G., Sottoriva, A., Anderson, A. R. A., Traulsen, A., and Altrock, P. M. (2016). The Cancer Stem Cell Fraction in Hierarchically Organized Tumors Can Be Estimated Using Mathematical Modeling and Patient-specific Treatment Trajectories. *Cancer Res.* 76, 1705–1713. doi:10.1158/0008-5472.can-15-2069
- White, F. M., Gatenby, R. A., and Fischbach, C. (2019). The Physics of Cancer. *Cancer Res.* 79, 2107–2110. doi:10.1158/0008-5472.can-18-3937
- Yee, D., DeMichele, A. M., Yau, C., Isaacs, C., Symmans, W. F., Albain, K. S., et al. (2020). Association of Event-free and Distant Recurrence-free Survival with Individual-Level Pathologic Complete Response in Neoadjuvant Treatment of Stages 2 and 3 Breast Cancer: Three-Year Follow-Up Analysis for the I-Spy2 Adaptively Randomized Clinical Trial. Chicago: JAMA.
- Zaheed, M., Wilcken, N., Willson, M. L., O'Connell, D. L., and Goodwin, A. (2019). *Sequencing of Anthracyclines and Taxanes in Neoadjuvant and Adjuvant Therapy for Early Breast Cancer*. Hoboken: John Wiley & Sons Ltd.

**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.

Copyright © 2021 Kurz, Sánchez and Axenie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Discriminative Localized Sparse Approximations for Mass Characterization in Mammograms

Sokratis Makrogiannis\*, Keni Zheng and Chelsea Harris

Math Imaging and Visual Computing Lab, Division of Physics, Engineering, Mathematics and Computer Science, Delaware State University, Dover, DE, United States

## OPEN ACCESS

### Edited by:

George Bebis,  
University of Nevada, Reno,  
United States

### Reviewed by:

Parag Kulkarni,  
Tokyo International University, Japan  
Harini Veeraraghavan,  
Memorial Sloan Kettering Cancer  
Center, United States

### \*Correspondence:

Sokratis Makrogiannis  
smakrogiannis@desu.edu

### Specialty section:

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

**Received:** 15 June 2021

**Accepted:** 06 December 2021

**Published:** 30 December 2021

### Citation:

Makrogiannis S, Zheng K and Harris C  
(2021) Discriminative Localized Sparse  
Approximations for Mass  
Characterization in Mammograms.  
Front. Oncol. 11:725320.  
doi: 10.3389/fonc.2021.725320

The most common form of cancer among women in both developed and developing countries is breast cancer. The early detection and diagnosis of this disease is significant because it may reduce the number of deaths caused by breast cancer and improve the quality of life of those effected. Computer-aided detection (CADe) and computer-aided diagnosis (CADx) methods have shown promise in recent years for aiding in the human expert reading analysis and improving the accuracy and reproducibility of pathology results. One significant application of CADe and CADx is for breast cancer screening using mammograms. In image processing and machine learning research, relevant results have been produced by sparse analysis methods to represent and recognize imaging patterns. However, application of sparse analysis techniques to the biomedical field is challenging, as the objects of interest may be obscured because of contrast limitations or background tissues, and their appearance may change because of anatomical variability. We introduce methods for label-specific and label-consistent dictionary learning to improve the separation of benign breast masses from malignant breast masses in mammograms. We integrated these approaches into our Spatially Localized Ensemble Sparse Analysis (SLESA) methodology. We performed 10- and 30-fold cross validation (CV) experiments on multiple mammography datasets to measure the classification performance of our methodology and compared it to deep learning models and conventional sparse representation. Results from these experiments show the potential of this methodology for separation of malignant from benign masses as a part of a breast cancer screening workflow.

**Keywords:** computer-aided diagnosis (CADx), sparse approximation, breast cancer screening, mass classification, mammographic imaging

## 1 INTRODUCTION

The topic of this work is automated classification of breast masses into benign or malignant using mammograms. The diagnosis of breast cancer is an impactful domain of research (1), therefore, automated methods of detection and diagnosis of breast cancer have gained popularity in the past few decades (2–6). Early diagnosis of breast cancer has been shown to reduce mortality related to this disease and significantly improve the quality of life of those affected. To achieve early diagnosis,

mammograms are used to aid in detecting breast cancer. Proper detection and diagnosis of breast abnormalities requires the experience and high levels of expertise of trained radiologists. Computer-aided diagnosis would improve the reproducibility of diagnosis states and reduce the time spent to thoroughly diagnosis breast cancer.

The X-ray mammographic test is a commonly used method for early prediction and diagnosis of breast cancer (7). Therefore, the development of CADe and CADx techniques for breast cancer using mammograms has attracted significant interest. Among these techniques, conventional classification models use specific procedures to craft features for representing and classifying imaging pattern. Such conventional approaches are introduced in (8–13). Features such as shape, texture, and intensity were extracted in (9). Among the extracted features, the genetic algorithm (GA) selected the most relevant features. Additionally, feature extraction through Zernike moments have been used because of their useful ability to well describe shape characteristics (14). In recent years, feature extraction and selection has been achieved through state-of-the-art techniques that use neural networks (NN) (15). A popular group of NN techniques use Convolutional Neural Nets (CNNs) for classification. Key advances in both the design and application of CNNs (16, 17) led to significant improvement in the state-of-the-art object recognition on the Imagenet dataset. A common training method used for CNNs is transfer learning; this technique has been applied to medical imaging for classification tasks (15, 18, 19). In (20), for example, pretrained VGG16, ResNet50, and Inception v3 networks were customized and applied to several mammographic datasets.

The concentration of this research is the diagnosis (CADx) of breast cancer masses into benign or malignant states using sparse representation and dictionary learning techniques. Sparse representation has been applied in the areas of computer vision, signal/image processing, and pattern recognition. The objective of sparse representation methods is to use sparse linear approximations of patterns, or atoms, from a dictionary of signals to represent a specific signal. These sparse approximations can then be used for applications such as compression and denoising of signals/images, classification, object recognition, and other areas. A common area of interest in such techniques is dictionary learning. Dictionary learning focuses on the methods for learning dictionaries in order to obtain optimal representations according to the application objective. Dictionary learning techniques have produced impressive results in a variety of signal and image processing applications (21–30). In more recent years, a widely studied area has been convolutional sparse coding, and its relationship with deep learning techniques (27, 30, 31).

Although there is substantial interest in the aforementioned techniques, their application to the biomedical field remains within limits to the straightforward utilization of sparse representation classification (SRC), or learning of multiple separate dictionaries. Hence motivation remains for the design of methods that leverage the capabilities of dictionary learning and sparse coding using joint discriminative-generative approaches.

Here we propose the integration of discriminative dictionary learning methods into our spatially localized ensemble sparse analysis classification (SLESA) model. Our dictionary learning techniques incorporate class label separation and label consistency and we denote these variations as LS-SLESA and LC-SLESA respectively. We train multiple dictionaries on the same set of ROIs and fuse the residuals of multiple approximations to obtain more robust class estimates than those obtained by single dictionary learning as also supported by (32). Our premise is that optimized spatially localized dictionaries trained using label separation or label consistency constraints, will improve the classification accuracy of our spatially localized sparse analysis. We employ this system for diagnosis of breast cancer in mammograms. We evaluate the performance of our framework and compare it to straightforward sparse representation classification (SRC), and the well-known CNN architectures of Alexnet (16), Googlenet (17), Resnet50 (33), and InceptionV3 (34), after applying transfer learning and data augmentation techniques.

## 1.1 Sparse Analysis

In recent years, the research area of sparse representation of signals has attracted considerable interest. The central focus of sparse analysis is to optimize an objective function. The objective function is comprised of a reconstruction error term and a sparsity term. The reconstruction error term or the residual, produces the measurement of the difference between the signal reconstruction and the test signal. The sparsity term measures the sparsity of the computed solution. The residual term may be set to measure the test signal exactly or within a defined bound of constraint.

In image classification tasks, the sparse representation of a test image is used to assign that image to a class. Sparse representation-based classification has two phases: coding and classification. In the coding phase, an image or signal is collaboratively coded with a dictionary of atoms given a sparsity constraint. The classification of the image is performed based on the coding coefficients and the dictionary. One of the advantages of sparse representation in image classification tasks is its ability to represent a high-dimensional image by few representative samples.

The dictionary  $D$  consists of columns of signals, also called atoms. The design of the dictionary could be simply predefined. For example, a dictionary that consists of all training samples from all classes is considered predefined. However, dictionaries of this form may fail to represent test samples well, if the atoms are inter-correlated, or they do not span the range of the image content. Moreover, very large dictionaries increase the coding complexity.

Sparse analysis solves the following optimization problem: given signals in an  $\mathbb{R}^d$  space, a dictionary  $D \in \mathbb{R}^{d \times n}$  of signals partitioned by class, and a test signal  $y \in \mathbb{R}^d$ , sparse coding seeks to find a coding vector  $\hat{x} \in \mathbb{R}^n$ . The test signal  $y$  is represented as a linear combination of the dictionary atoms and a sparse code. This mathematical optimization problem is expressed by

$$\hat{x} = \arg \min_x \|\hat{x}\|_0 \text{ subject to } y = Dx. \quad (1)$$

Sparsity is represented by the  $\ell_0$  norm, but may also be approximated by the  $\ell_1$  norm, or  $\ell_p$  norms where  $p \in (0,1)$ . Assuming that the signal contains noise, we can introduce  $\epsilon$  as a tolerance parameter and solve the following problem,

$$\hat{x} = \arg \min_x \|\hat{x}\|_0 \text{ subject to } \|y - Dx\| < \epsilon \quad (2)$$

Pursuit algorithms such as basis pursuit (BP) and orthogonal matching pursuit (OMP) are often used to solve the sparse coding problems defined in Equations 1 and 2. Basis pursuit is a linear programming technique that seeks to find the sparsest  $L_1$  solution to the mathematical optimization problem defined in Equation 1. The orthogonal matching pursuit is considered a greedy pursuit algorithm in that it updates the sparse solution vector coefficients using previously updated solution vector atoms. OMP is a more complex and computationally expensive extension of the matching pursuit algorithm (MP), however, can often lead to better sparse solutions.

Early sparse representation techniques such as SRC (35), optimize an objective function of two terms, and design the dictionary  $D$  with the original training images as dictionary columns or atoms. In more recent works, we see an emphasis on the design of the dictionary and task-specific optimization, of which we discuss in the next section.

## 1.2 Dictionary Learning

As discussed before, the dictionary is a key component of the optimization problem. Learning a dictionary from training data has been an area of interest in recent years (25, 36). The goal of such techniques is to construct dictionaries optimized for class representation and separation. Previous works have shown that dictionary learning may improve the performance of image processing and recognition tasks (25). Dictionary learning techniques can be divided into the following groups (23): (i) probabilistic learning methods, (ii) clustering-based learning methods, and (iii) construction methods.

The type, design, and dimensions of the dictionary have a significant effect on the solutions of the sparse optimization problem. The atoms are expected to be able to approximate the variations of the specific image domain and have low correlation with each other. Considering the dictionary dimensions, a dictionary is considered overcomplete when the number of signals within the dictionary ( $n$ ) exceeds the dimension of the signal to be represented ( $d$ ), that is if  $d < n$ . Overcomplete dictionaries are required to produce sparse representations of signals (37).

## 2 METHODOLOGY

In this work, we introduce class label separation and class label consistency into the localized dictionaries within our spatially localized sparse analysis (SLESA) framework. We denote the respective methods by LS-SLESA and LC-SLESA. Our SLESA approach applies localized block decomposition that reduces the

length of the feature vector and helps to build overcomplete dictionaries. In the classification stage, we solve the sparse representation problem for each block using orthogonal matching pursuit (OMP), and fuse the individual block-wise responses to determine the lesion category. LS-SLESA and LC-SLESA aim to further improve the performance of our previous work, SLESA, by finding task-specific dictionaries that utilize the class labels of the training data. We consider two approaches: one calculates separate dictionaries for benign and malignant breast masses, and the other incorporates linear classification errors into the optimization problem. **Figure 1** outlines the main stages of our methodology.

### 2.1 Spatially Localized Block Decomposition

We divide each training image  $I$  into  $m \times n$  px blocks that are spatially ordered. Therefore,  $I = [B^1, B^2, \dots, B^{NB}]$ , where  $B^j$  denotes a block of each training image and  $NB$  is the total number blocks of an image. We construct dictionaries  $D^j$ , where  $j = 1, 2, \dots, NB$ , from the same position of the block  $B^j$  for all  $s$  images of the training set:

$$D^j = [B^j_1, B^j_2, \dots, B^j_s]. \quad (3)$$

Therefore, a number of  $NB$  block dictionaries are constructed, each unique in the spatial information that they provide to classify spatially localized image blocks.

### 2.2 Label Specific Spatially Localized Ensemble Sparse Analysis

We introduce dictionary learning techniques to improve the sparse approximation accuracy and generalizability. We learn a separate dictionary for each type of mass and we then merge the dictionaries to perform sparse coding and classification.

We employ the KSVD algorithm by (21) to learn the dictionary. KSVD updates the atoms of the dictionary by iteratively solving sparse coding problems that alternate between residual and sparsity constraints. The optimized atom in each iteration is computed by Singular Value Decomposition (SVD). This method has been shown to converge to effective solutions and has been widely applied for sparse representation.

After the block decomposition step, we learn  $NB$  discriminative dictionaries using block-based label-separated KSVD. We denote this approach by LS-SLESA.

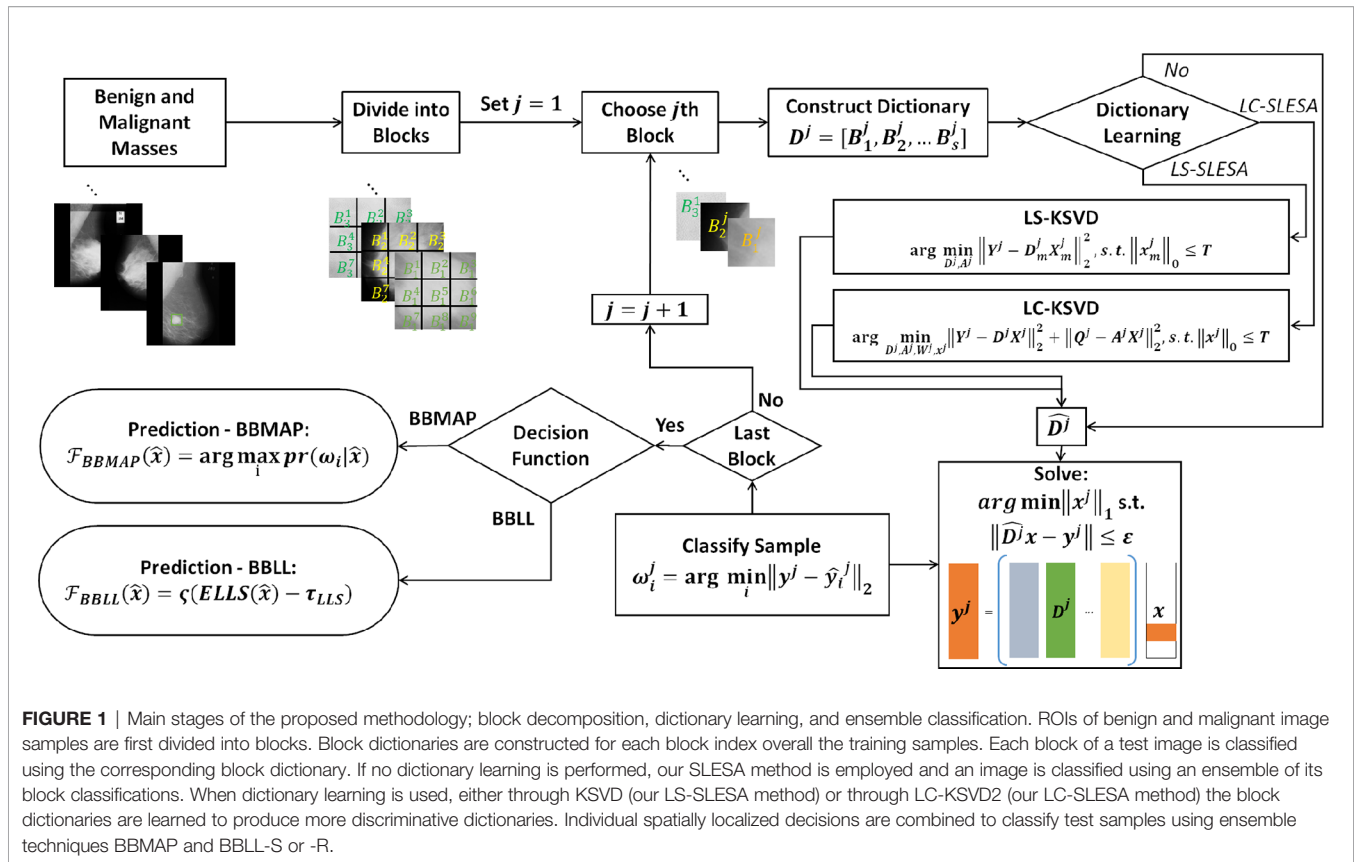
$$\arg \min_{D^j, A^j} \|Y^j - D^j_m X^j_m\|_2^2 \text{ s.t. } \|x^j_m\|_0 \leq T, \quad (4)$$

where  $Y^j$  denotes the training block samples. We solve the above problem for each class index  $m$ , and then concatenate the class-specific dictionaries  $D^j_m$  to form the complete dictionary  $D^j$  for the  $j$ -th block.

### 2.3 Label Consistent Spatially Localized Ensemble Sparse Analysis

Another approach is to learn  $NB$  discriminative dictionaries using the label consistent KSVD algorithm (denoted by LC-KSVD). Sparse coding and sparse classification errors are added





to the optimization problem in order to compute a single discriminative dictionary. We employ LC-KSVD to learn the dictionaries  $D^j$ . The authors in (24) proposed two variants named LC-KSVD1 and LC-KSVD2. In their work, classification performance was consistently greater when the LC-KSVD2 variant is used versus the LC-KSVD1 dictionary learning approach. Thus, we employed the objective function of LC-KSVD2 in our LC-SLESA approach. Thus, omitting the need for ablation experiments on the effectiveness of the loss terms in the LC-KSVD methods.

LC-KSVD2 adds a label consistency regularization term and a joint classification error term to the objective function. The optimization problem is:

$$\arg \min_{D^j, A^j, W^j, x^j} ||Y^j - D^j X^j||_2^2 + ||Q^j - A^j X^j||_2^2 + ||H^j - W^j X^j||_2^2 \quad (5)$$

$$s.t. ||x^j||_0 \leq T.$$

$Q^j$  denotes the class-specific sparse codes for  $Y^j$ , and  $A^j$  is a linear transformation matrix.  $W^j$  symbolizes the parameters of the linear classifier, and  $H^j$  contains the class labels of the training data  $Y^j$ .  $T$  is the sparsity threshold. The term  $||Q^j - A^j X^j||_2^2$  is the discriminative sparse code error that forces patterns from the same class to have similar sparse codes.  $Q^j$  is defined as  $Q^j = [q_1^j, \dots, q_N^j]$  for  $N$  many training samples where the discriminative sparse codes for a sample,  $q_i^j$  contains zero indices where the training sample  $y_i^j \in Y^j$  and its corresponding dictionary do not

share the same class label. The term  $||H^j - W^j X^j||_2^2$  expresses the classification error.

## 2.4 Ensemble Classification

In this stage of our method, we combine the individual spatially localized decisions to classify the test samples. We find the solution  $x^j$  of the regularized noisy  $\ell_1$ -minimization problem, for each test sample  $y^j$  corresponding to the  $j$ th block:

$$\hat{x}^j = \arg \min ||x^j||_1 \text{ subject to } ||D^j x - y^j||_2 \leq \epsilon \quad (6)$$

We propose ensemble learning techniques in a Bayesian probabilistic setting to fuse classifier predictions. We propose a decision function that applies majority voting to individual hypotheses (BBMAP), and an ensemble of log-likelihood scores (BBLL) computed from either the sparsity of the solution (BBLL-S), or approximation residual (BBLL-R).

### 2.4.1 Maximum a Posteriori Decision Function (BBMAP)

The class label of a test sample is determined by the MAP estimate produced by  $NB$  block-based classifiers. The predicted class label  $\hat{\omega}$  is

$$\hat{\omega}_{BBMAP} = \mathcal{F}_{BBMAP}(\hat{x}) \doteq \arg \max_i pr(\omega_i|\hat{x}), \quad (7)$$

where  $pr(\omega_i|\hat{x})$  is the posterior probability for class  $\omega_i$  given  $\hat{x}$ .

### 2.4.2 Log Likelihood Sparsity-Based Decision Function (BLL-S)

This decision function first computes a log-likelihood score based on the relative sparsity scores  $\|\delta_m(\hat{x}^j)\|_1$ ,  $\|\delta_n(\hat{x}^j)\|_1$ , obtained from the sparse representation stage of each classifier

$$LLS(\hat{x}^j) = -\log \frac{\|\delta_m(\hat{x}^j)\|_1}{\|\delta_n(\hat{x}^j)\|_1} \begin{cases} \geq 0, \hat{x}^j \in m\text{th class} \\ < 0, \hat{x}^j \in n\text{th class} \end{cases} \quad (8)$$

We estimate the expectation of  $LLS(\hat{x})$  that we denote by  $ELLS$  over the individual classification scores obtained by (8)

$$\begin{aligned} ELLS(\hat{x}) &\doteq E\{LLS(\hat{x}^j)\} = \frac{1}{NB} \sum_j^{NB} LLS(\hat{x}^j) \\ &= -\frac{1}{NB} \left[ \sum_j^{NB} \log \|\delta_m(\hat{x}^j)\|_1 - \sum_j^{NB} \log \|\delta_n(\hat{x}^j)\|_1 \right]. \end{aligned} \quad (9)$$

We apply a sigmoid function  $\varsigma(\cdot)$  to produce classification scores in the range of  $[-1, 1]$ . We employ a shift parameter  $\tau_{LLS}$  to account for classification bias,

$$\mathcal{F}_{LLS}(\hat{x}) \doteq \varsigma(ELLS(\hat{x}) - \tau_{LLS}). \quad (10)$$

The final decision is given by the sign of  $F_{LLS}(\hat{x})$ :

$$\hat{\omega}_{LLS}(\hat{x}) = \text{Sgn}\{\mathcal{F}_{LLS}(\hat{x})\}. \quad (11)$$

### 2.4.3 Log Likelihood Residual-Based Decision Function (BLL-R)

This function computes a log-likelihood score based on the relative residual scores  $\|\delta_m(\hat{x}^j)\|_1$ ,  $\|\delta_n(\hat{x}^j)\|_1$ , obtained from the sparse representation stage,

$$LLR(\hat{x}^j) = -\log \frac{\|D^j \delta_m(\hat{x}^j) - y^j\|_2}{\|D^j \delta_n(\hat{x}^j) - y^j\|_2} \begin{cases} \geq 0, \hat{x}^j \in m\text{th class} \\ < 0, \hat{x}^j \in n\text{th class} \end{cases} \quad (12)$$

We estimate the expectation of  $LLR(\hat{x})$ , denoted by  $ELLR$ , over all the individual classification scores obtained by (12),

$$ELLR(\hat{x}) \doteq E\{LLR(\hat{x}^j)\} = \frac{1}{NB} \sum_j^{NB} LLR(\hat{x}^j) \quad (13)$$

We apply a sigmoid function  $\varsigma(\cdot)$  with a shift parameter  $\tau_{LLR}$  and a sign function, to determine the state of  $\hat{x}$ , symbolized by  $\hat{\omega}_{LLR}(\hat{x})$ , as in (10, 11).

## 3 EXPERIMENTS AND DISCUSSION

We evaluated our method for classification of breast masses into malignant or benign states on two digital mammographic databases. Next, we describe our experiments and report results produced by our approach. For comparison, we report the results of variants to our proposed method including straightforward sparse representation and multiple strategies for dictionary learning in SLESA, LS-SLESA and LC-SLESA.

These may serve as ablation experiments to evaluate the effect of ensemble classification and the effect of dictionary learning on the performance of our method. We have also validated the performance of widely used convolutional neural networks (16, 17, 33, 34), after applying transfer learning, random resampling, and extensive optimization.

### 3.1 Datasets

The training and testing data used in our experimentation were obtained from the Mammographic Image Analysis Society (MIAS) (2) and the Digital Database for Screening Mammography (DDSM). The Mammographic Image Analysis Society (MIAS) database is one of oldest and the most widely used mammography databases. The resolution of the mammograms is 200-micron pixel edge that is approximately equivalent to  $264.58 \mu\text{m}$  pixel size. The image size after clipping or padding is  $1024 \times 1024$  px. The MIAS dataset consists of 322 digitized mediolateral oblique (MLO) images (68 benign, 51 malignant, 203 normal). We selected mammograms containing 51 malignant and 66 benign masses in total, to evaluate classification performance. The Digital Database for Screening Mammography (DDSM) is a large public database including a total of 10,480 images. CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is a carefully selected and updated subset DDSM (Digital Database in for Screening Mammography). It contains 753 calcification subjects and 891 mass subjects. In our experiments we used the CC view (craniocaudal view) of benign and malignant lesions of CBIS-DDSM (Curated Breast Imaging Subset of DDSM). Thus, the number of malignant cases used in our experiments was narrowed down to 296 malignant and 311 benign cases.

To prepare the data for the first stage of our method, block decomposition, we first selected regions of interest (ROIs) containing the masses. Our method reads-in two key values from radiological readings, that is, the centroid and radius of each mass. It determines a minimum bounding square ROI and select the masses that satisfy a size criterion. In the first approach, we ensured that the majority of the blocks cover the complete mass area. The mass ROI sizes are required to be greater than, or equal to a fixed ROI size. The qualifying masses are center-cropped to generate the ROI data. In the second approach, we selected the complete ROIs including background tissue using the mass centroid and radius. Then we resampled all ROIs to a fixed size, instead of applying a minimum size criterion. In MIAS data we followed both approaches for ROI selection. In the CBIS-DDSM data we followed the second approach. We performed 10- and 30-fold cross-validation on the ROIs to examine the effect of the cross-validation fold size on performance.

### 3.2 Convolutional Neural Networks With Transfer Learning

For comparison purposes, we implemented CNN classifiers using the Alexnet (16), Googlenet (17), Resnet50 (33), and InceptionV3 (34) architectures with transfer learning. All networks were pre-trained on the Imagenet database that contains 1.2 million natural images.

Transfer learning was applied to each network in various ways. To modify Alexnet to our data, we replaced the pre-trained fully connected layers with three new fully connected layers. The learning rates of the pre-trained layers were set to 0 in order to keep the network weights fixed. We only trained the new fully connected layers. For Googlenet, the learning rates of the bottom 10 layers were set to 0, and the top fully connected layer was replaced with a new fully connected layer. We also assigned a greater learning rate factor for the new layer than the pre-trained layers. In Resnet50, we replaced the pre-trained fully connected layers with three new fully connected layers. We set the learning rates of the pre-trained layers to 0, in order to train only the new fully connected layers. In InceptionV3, we replaced the top classification layers with three new fully connected layers. We set the learning rates of the pre-trained layers of InceptionV3 to 0, as we did in Alexnet and Resnet50.

To provide the networks with additional training examples, we applied data resampling using randomly-centered patches inside each ROI. Additionally, we applied data augmentation by rotation, scaling, and horizontal and vertical flipping. Finally, we used Bayesian optimization (38, 39) to tune the learning rate, mini-batch size, and number of epochs.

Due to the ability of deep networks to learn information from the edges of masses and not just the texture, we decided to test our method on 256×256 px ROIs of all masses including the background tissue in the MIAS database (66 benign and 51 malignant). **Table 1** summarizes the results of our cross-validation experiments. Googlenet yields the top ACC of 67.65% and the top AUC of 63.04% for 30-fold cross-validation.

When using DDSM data, we applied the same ROI selection strategy with that of MIAS. The Alexnet architecture yields the

top ACC of 69.59% and the top AUC of 73.04% using 30-fold cross-validation (**Table 2**). We note the increase in classification performance when using DDSM for training and testing. This is expected, because CNNs require a large number of diverse training samples to achieve good performance. DDSM is a larger database than MIAS, therefore CNNs are able to learn more relevant features for classification. Of note is that simpler networks such as Alexnet and Googlenet, with smaller numbers of trainable weights, produce more accurate classifications than deeper networks such as InceptionV3. This is expected because of the limited number of training samples in both datasets.

### 3.3 LS-SLESA and LC-SLESA

Next, we evaluated the performance of our block-based ensemble classification method by 10- and 30-fold cross-validation. In the MIAS section of our experiments, we present results using minimum ROI size of 64×64 pixels, resulting in a dataset of 36 benign and 37 malignant masses. In **Table 3**, we report the classification rates produced for multiple block sizes. When the block size is equal to the ROI size, conventional SRC is performed (35); these results are reported in the first row of **Table 3**. We observe that ACC and AUC generally increase when the number of folds increases, for the same ROI size. The top ACC using 10-fold cross-validation is 72.86% for 8×8 block size by SLESA, and for 64×64 block size by LS-SLESA with BBLL-S decision function. The top AUC for 10-fold CV is 75.35% for 8×8 block size, produced by LS-SLESA. The best overall performance is obtained for 30-fold cross validation. The top accuracy is 90% for 16×16 and 8×8 block sizes by SLESA, and the largest area under the curve is 93.10% for 8×8 block size by SLESA with BBLL-S decision function. In 30-fold cross-validation, 2 or 3

**TABLE 1** | Breast mass classification performance on MIAS data using convolutional neural network classifiers (ROI size: 256 × 256).

Method	k-Fold CV	ROI Size	TPR (%)	TNR (%)	ACC (%)	AUC (%)
Alexnet	10	256 × 256	56.86	72.55	<b>64.71</b>	<b>62.19</b>
	30	256 × 256	58.82	64.71	61.77	60.29
Googlenet	10	256 × 256	64.71	58.82	61.77	57.86
	30	256 × 256	66.67	68.63	<b>67.65</b>	<b>63.04</b>
Resnet50	10	256 × 256	60.78	62.75	61.76	57.32
	30	256 × 256	44.12	55.88	53.6	56.8
InceptionV3	10	256 × 256	58.82	60.78	59.80	58.59
	30	256 × 256	58.82	60.78	59.80	57.44

The top performances of 10- and 30-fold cross-validation are shown in bold.

**TABLE 2** | Breast mass classification performance on DDSM data using convolutional neural network classifiers (ROI size: 256 × 256).

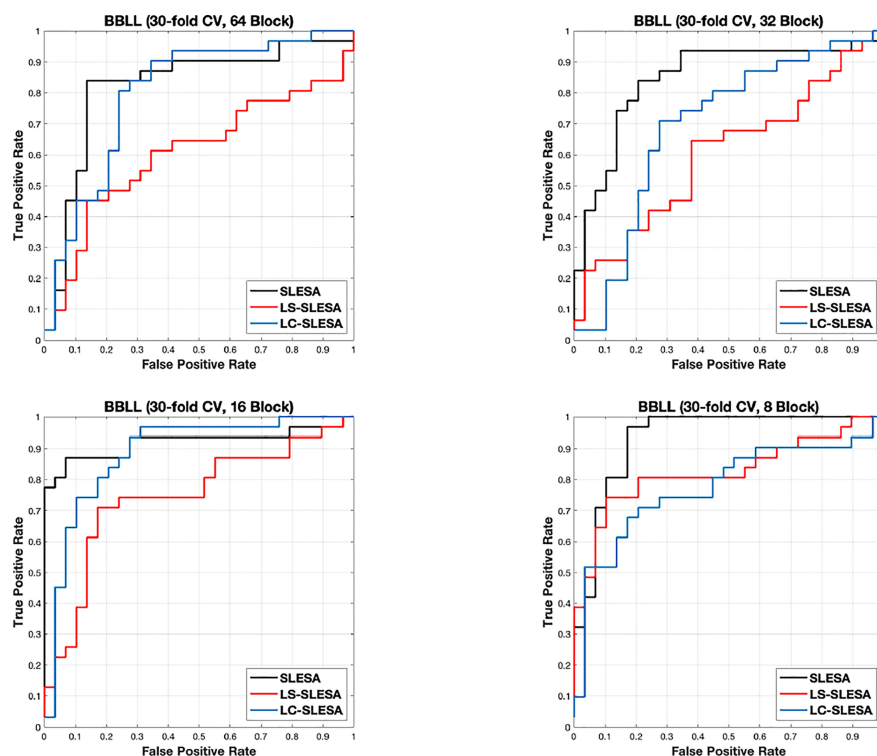
Method	k-Fold CV	ROI Size	TPR (%)	TNR (%)	ACC (%)	AUC (%)
Alexnet	10	256 × 256	67.57	65.88	<b>66.72</b>	<b>69.70</b>
	30	256 × 256	72.64	66.55	<b>69.59</b>	<b>73.04</b>
Googlenet	10	256 × 256	72.64	59.46	66.05	69.55
	30	256 × 256	66.89	64.19	65.5	69.43
Resnet50	10	256 × 256	56.42	75.68	66.05	70.35
	30	256 × 256	60.81	73.31	67.06	71.34
InceptionV3	10	256 × 256	61.82	67.57	64.70	64.70
	30	256 × 256	65.20	64.19	64.70	66.94

The top performances of 10- and 30-fold cross-validation are shown in bold.

**TABLE 3** | Breast mass classification performance on MIAS data using ensembles of block-based sparse classifiers with dictionary learning (ROI size: **64×64**).

Method	k-Fold	Block Size	SLESA	SLESA	SLESA	SLESA	LS-SLESA	LS-SLESA	LS-SLESA	LS-SLESA	LC-SLESA	LC-SLESA	LC-SLESA	LC-SLESA
			TPR (%)	TNR (%)	ACC (%)	AUC (%)	TPR (%)	TNR (%)	ACC (%)	AUC (%)	TPR (%)	TNR (%)	ACC (%)	AUC (%)
BMAP-S	10	64×64	45.95	84.85	64.29	63.55	64.86	81.82	72.86	70.11	75.68	36.36	57.14	53.81
		32×32	51.35	87.88	68.57	69.53	62.16	81.82	71.43	70.84	78.38	36.36	58.57	52.33
		16×16	40.54	90.91	64.29	65.52	59.46	81.82	70.00	69.70	56.76	72.73	64.29	61.26
		8×8	56.76	81.82	68.57	67.90	48.65	81.82	64.29	63.23	62.16	63.64	62.86	60.77
		Mean	48.65	86.37	66.43	66.63	58.78	81.82	69.64	68.47	68.25	52.27	60.72	57.04
		Std Dev	6.98	3.91	2.47	2.63	7.11	0.00	3.76	3.53	10.44	18.75	3.40	4.63
BBLL-S	10	64×64	64.86	72.73	68.57	70.35	72.97	72.73	<b>72.86</b>	71.33	64.86	66.67	65.71	66.42
		32×32	70.27	63.64	67.14	70.02	62.16	81.82	71.43	69.70	70.27	60.61	65.71	68.80
		16×16	59.46	84.85	71.43	<b>74.37</b>	59.46	81.82	70.00	69.94	64.86	75.76	<b>70.00</b>	<b>71.42</b>
		8×8	72.97	72.73	<b>72.86</b>	71.58	59.46	81.82	70.00	<b>75.35</b>	51.35	81.82	65.71	64.78
		Mean	66.89	73.49	70.00	71.58	63.51	79.55	71.07	71.58	62.84	71.21	66.79	67.85
		Std Dev	5.99	8.70	2.61	1.97	6.43	4.55	1.37	2.61	8.07	9.42	2.14	2.89
BMAP-S	30	64×64	22.58	93.10	56.67	52.28	64.52	55.17	60.00	56.62	70.97	62.07	66.67	63.52
		32×32	9.88	100.00	53.33	48.50	48.39	75.86	61.67	59.40	100.00	63.33	63.33	57.17
		16×16	61.29	65.52	63.33	59.96	45.16	82.76	63.33	60.73	75.86	71.67	71.67	69.30
		8×8	38.71	96.55	66.67	61.96	54.84	86.21	70.00	66.07	74.19	55.17	65.00	60.85
		Mean	33.07	88.79	60.00	55.68	53.23	75.00	63.75	60.71	80.26	63.06	66.67	62.71
		Std Dev	22.25	15.77	6.09	6.35	8.54	13.90	4.38	3.97	13.32	6.77	3.60	5.11
BBLL-S	30	64×64	83.87	86.21	85.00	82.09	45.16	86.21	65.00	60.62	90.32	65.52	78.33	79.98
		32×32	83.87	75.86	80.00	84.43	64.52	62.07	63.33	60.78	61.29	75.86	68.33	69.30
		16×16	87.10	93.10	<b>90.00</b>	92.00	70.97	82.76	76.67	74.53	96.77	68.97	<b>83.33</b>	<b>88.43</b>
		8×8	96.77	82.76	<b>90.00</b>	<b>93.10</b>	74.19	89.66	<b>81.67</b>	<b>82.43</b>	67.74	82.76	75.00	77.42
		Mean	87.90	84.48	86.25	87.91	63.71	80.18	71.67	69.59	79.03	73.28	76.25	78.78
		Std Dev	6.10	7.18	4.79	5.47	13.00	12.39	8.93	10.76	17.20	7.65	6.29	7.88

The top performances of 10- and 30-fold cross-validation are shown in bold.

**FIGURE 2** | ROC plots for **64 × 64**, **32 × 32**, **16 × 16**, and **8 × 8** block sizes using the proposed block-based ensemble method on the MIAS dataset with BBLL decision functions and 30-fold CV.



**TABLE 4** | Breast mass classification performance on DDSM data using ensembles of block-based sparse classifiers with dictionary learning (ROI size: **128×128**).

Method	k-Fold CV	Block Size	SLESA TPR (%)	SLESA TNR (%)	SLESA ACC (%)	SLESA AUC (%)	LS-SLESA TPR (%)	LS-SLESA TNR (%)	LS-SLESA ACC (%)	LS-SLESA AUC (%)	LC-SLESA TPR (%)	LC-SLESA TNR (%)	LC-SLESA ACC (%)	LC-SLESA AUC (%)
BBMAP-R	10	128×128	55.97	49.83	52.83	53.12	69.62	43.65	56.33	56.93	57.00	54.07	55.50	55.82
		64×64	40.61	63.52	52.33	51.90	49.15	64.17	56.83	56.70	48.81	66.45	57.83	57.62
		32×32	54.61	55.70	55.17	55.25	59.04	62.22	60.67	61.00	60.07	54.07	57.00	57.21
		16×16	62.12	50.81	56.33	56.71	75.43	36.81	55.67	55.92	57.68	57.00	57.33	57.49
		8×8	60.07	50.81	55.33	55.84	62.12	56.68	59.33	60.05	51.19	66.45	59.00	58.97
		Mean	54.68	54.13	54.40	54.56	63.07	52.71	57.77	58.12	54.95	59.61	57.33	57.42
		Std Dev	8.42	5.73	1.73	1.99	10.08	11.96	2.13	2.25	4.74	6.36	1.27	1.12
BBLL-R	10	128×128	44.30	65.87	54.83	53.35	69.97	43.65	56.50	57.17	34.13	77.85	56.50	56.82
		64×64	73.72	36.16	54.50	54.11	55.97	62.54	59.33	60.93	50.17	69.71	60.17	61.37
		32×32	48.46	68.08	58.50	58.26	45.05	73.94	59.83	62.13	44.37	74.92	60.00	62.31
		16×16	61.43	57.33	59.33	60.37	64.85	58.63	61.67	62.04	63.83	56.35	60.00	61.09
		8×8	47.44	75.24	<b>61.67</b>	<b>62.04</b>	54.61	71.34	<b>63.17</b>	<b>65.34</b>	68.26	57.33	<b>62.67</b>	<b>63.75</b>
		Mean	55.07	60.54	57.77	57.62	58.09	62.02	60.10	61.52	52.15	67.23	59.87	61.07
		Std Dev	12.31	15.05	3.06	3.81	9.66	12.02	2.52	2.94	14.01	9.93	2.20	2.59
BBMAP-R	30	128×128	55.63	48.86	52.17	52.48	60.41	50.49	55.33	55.49	31.40	78.18	55.33	54.87
		64×64	38.91	65.15	52.30	52.02	42.66	65.15	54.17	53.82	48.46	69.38	59.17	58.97
		32×32	52.22	49.84	51.00	51.28	52.56	55.70	54.17	54.31	62.46	56.68	59.50	59.85
		16×16	36.18	80.78	59.00	58.19	69.97	47.23	58.33	59.09	50.85	70.68	61.00	61.30
		8×8	35.84	77.85	57.33	57.21	77.47	43.97	60.33	60.51	49.83	67.10	58.67	58.45
		Mean	43.76	64.50	54.36	54.24	60.61	52.51	56.47	56.64	49.83	67.10	58.73	58.69
		Std Dev	9.44	15.03	3.56	3.21	13.77	8.29	2.75	2.99	11.12	7.76	2.09	2.39
BBLL-R	30	128×128	62.80	43.97	53.17	52.18	18.43	92.83	56.50	54.57	32.08	80.78	57.00	56.75
		64×64	27.65	80.78	54.80	52.30	59.73	57.33	58.50	58.30	47.44	71.34	59.67	61.73
		32×32	83.96	22.48	52.50	51.69	37.88	82.08	60.50	62.64	66.55	57.34	<b>61.83</b>	<b>62.32</b>
		16×16	56.31	64.17	<b>60.33</b>	61.43	48.12	74.27	61.50	61.93	51.53	66.78	59.33	61.40
		8×8	39.25	79.48	59.83	<b>61.82</b>	62.45	60.91	<b>61.67</b>	<b>65.24</b>	63.14	57.65	60.83	62.00
		Mean	53.99	58.18	56.13	55.88	45.32	73.48	59.73	60.54	52.15	66.78	59.73	60.84
		Std Dev	21.75	24.88	3.71	5.25	17.94	14.73	2.20	4.16	13.73	9.86	1.82	2.31

The top performances of 10- and 30-fold cross-validation are shown in bold.

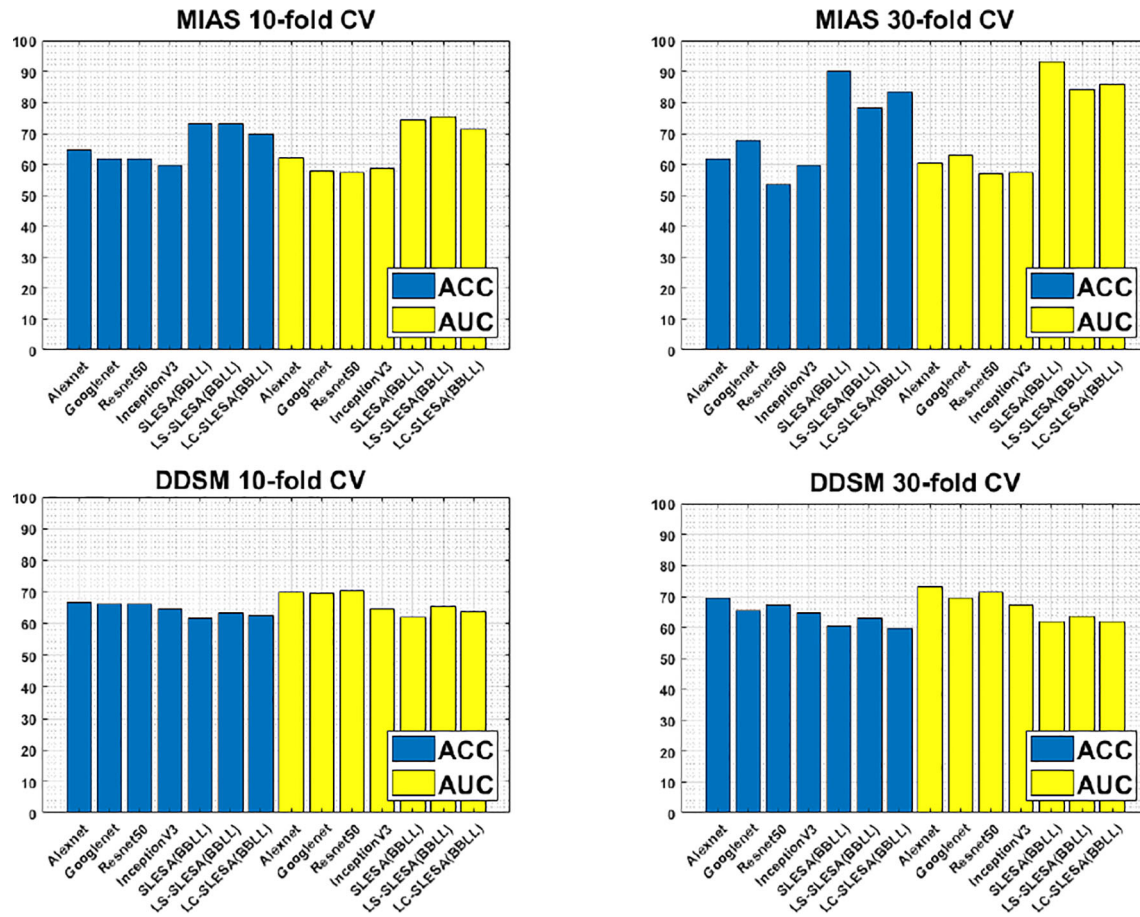
images are tested in each fold. Additionally, in **Table 3** we report true positive rates (TPR) and true negative rates (TNR) for each experiment. Generally, we observe higher true negative rates on average than true positive rates, which is an indication that the positive class, malignant, is more difficult to classify. **Figure 2** displays the receiver operating curves (ROC) by SLESA, LS-SLESA and LC-SLESA for 64,32,16 and 8px block lengths using 30-fold CV. The ROC graphs are consistent with the results in **Table 3**. We compare BBLL-S ROC curves in **Figure 2** among the SLESA methods by applying DeLong's statistical test for 30-fold cross-validation on the MIAS dataset. These tests produced statistically significant differences in AUCs at the level  $\alpha = 0.05$  between SLESA and LS-SLESA for 64,32, and 16px block lengths. These tests determined as significant, AUC differences between SLESA and LC-SLESA for 8px block length, and between LS-SLESA and LC-SLESA for 64px block length. The results indicate that SLESA produced better AUC values in 30-fold CV.

In the DDSM section of our experiments, we selected the complete ROIs including background tissue using the centroid and radius data. Then we resampled all ROIs to the fixed size of 128×128px. **Table 4** contains a summary of the results. LS-SLESA using 8×8 blocks and BBLL-R decision in 10-fold cross-validation, produces the highest AUC and ACC at 65.34% and 63.17% respectively. Overall, label-specific and label-consistent dictionary learning improves the ACC and AUC.

Another general comparison can be made with the cases of equal ROI and block sizes, for example when we use 64×64 block size in MIAS experiments. These cases are equivalent to conventional SRC, proposed by (35) and do not perform ensemble classification. Hence, these are ablation tests for the ensemble stage of our framework. The results indicate that our SLESA techniques outperform conventional SRC in both datasets. This is because block decomposition reduces the dimensionality of the images and enables the creation of multiple overcomplete dictionaries. An additional benefit is that we train multiple dictionaries on the same set of ROIs and fuse the residuals of multiple approximations to improve the classification accuracy.

Furthermore, **Figure 3** compares the ACC and AUC values of Alexnet, Googlenet, Resnet50 and InceptionV3 with SLESA, LS-SLESA and LC-SLESA. We observe that sparse approximations yield clearly better results on MIAS data, while CNNs with transfer learning are a bit more accurate on DDSM data.

We highlight the top AUC performances of CNNs and sparse methods per CV fold and dataset in **Table 5**. Our observations here are consistent with those we made in **Figure 3**. Our SLESA methods significantly outperform the best CNN performance on the MIAS dataset. On the DDSM dataset, the top CNN performances are slightly better than the SLESA counterparts in 10-fold CV, and the difference increases a bit in 30-fold CV. The size of the dataset may play a role in this difference, as neural



**FIGURE 3** | ACC performance comparisons on MIAS (top row) and DDSM (bottom row) datasets using 10- and 30-fold cross-validation.

**TABLE 5** | Top AUC performances of sparse analysis and deep learning methods on MIAS and DDSM datasets.

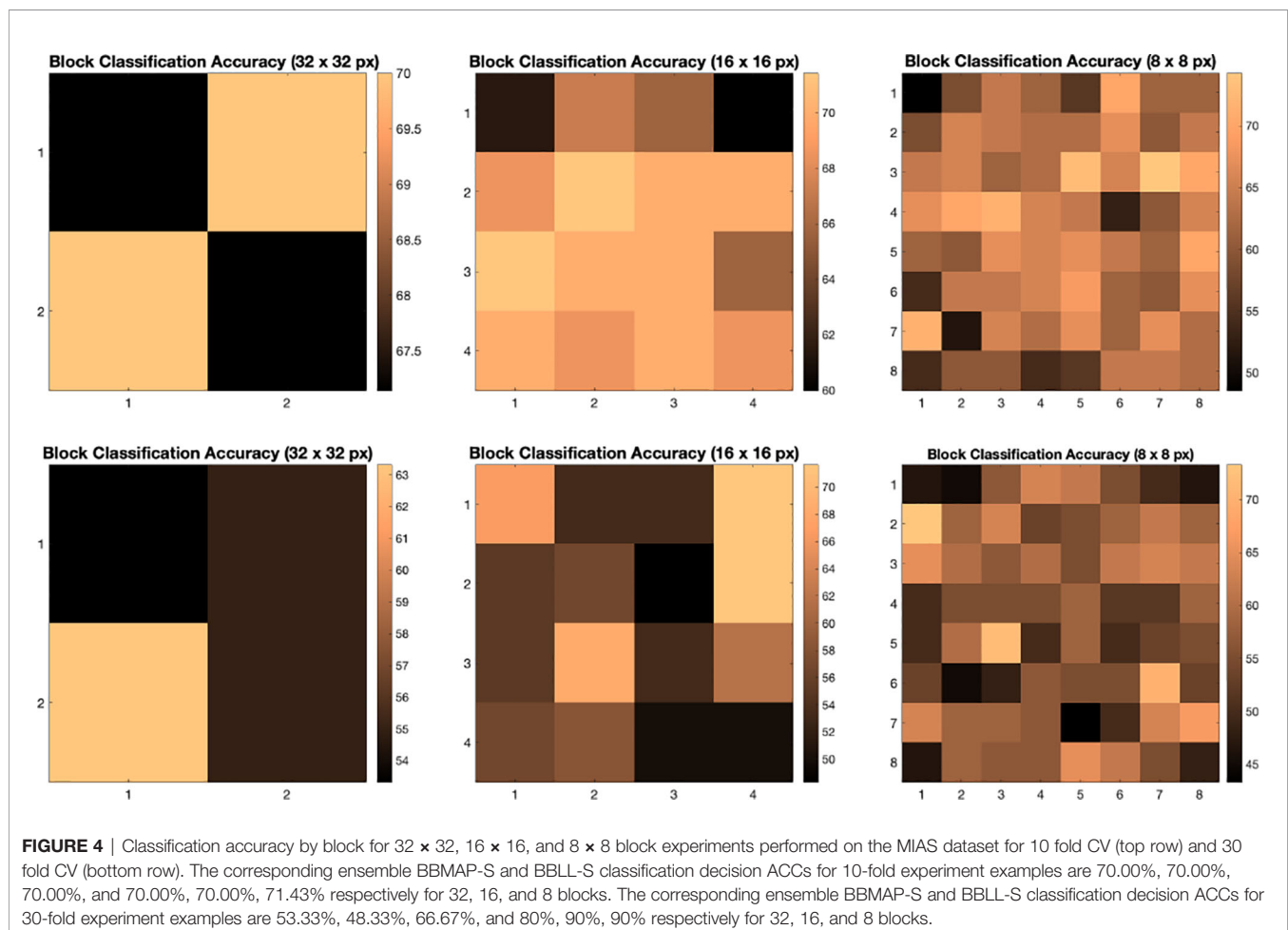
Dataset	k-Fold CV	Method	Block Size	TPR (%)	TNR (%)	ACC (%)	AUC (%)
MIAS	10	Alexnet	N/A	56.86	72.55	64.71	62.19
		SLESA	16 × 16	59.46	84.85	71.43	74.37
		LS-SLESA (BBLL-S)	8 × 8	59.46	81.82	70.00	75.35
		LC-SLESA (BBLL-S)	16 × 16	64.86	75.76	70.00	71.42
		Googlenet	N/A	66.67	68.63	67.65	63.04
MIAS	30	SLESA (BBLL-S)	8 × 8	96.77	82.76	90.00	93.10
		LS-SLESA (BBLL-S)	8 × 8	74.19	89.66	81.67	82.43
		LC-SLESA (BBLL-S)	16 × 16	96.77	68.97	83.33	88.43
		Resnet50	N/A	56.42	75.31	66.05	70.35
		SLESA (BBLL-R)	8 × 8	47.44	75.24	61.67	62.04
DDSM	10	LS-SLESA (BBLL-R)	8 × 8	54.61	71.34	63.17	65.34
		LC-SLESA (BBLL-R)	8 × 8	68.26	57.33	62.67	63.75
		Alexnet	N/A	72.64	66.55	69.59	73.04
		SLESA (BBLL-R)	8 × 8	39.25	79.48	59.83	61.82
		LS-SLESA (BBLL-R)	8 × 8	48.12	74.27	61.67	65.24
DDSM	30	LC-SLESA (BBLL-R)	32 × 32	66.55	57.34	61.83	62.32

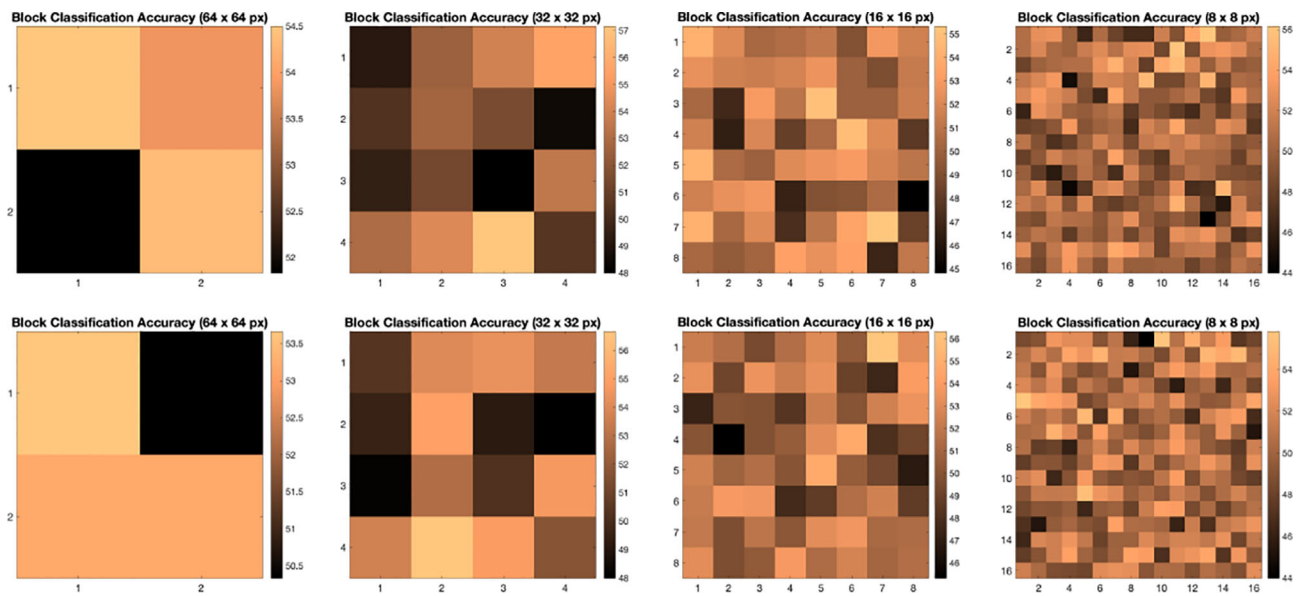
networks learn best with large amounts of data. Additionally, the complexity of finding sparse solution in our sparse analysis methods increases as a larger amount of training samples are learned. Overall, the results indicate that sparse approximations produce good results on both datasets. In addition, they require fewer training data than CNNs, hence can produce better results than CNNs for smaller datasets.

We illustrate the effect of block localized learning on classification by performing block experiments on both datasets and comparing the classification rates per block. We include example block ACC experiment results in **Figures 4** and **5**. In MIAS block ACC experimentation we notice that top block ACC rates increase as the block size decreases, which confirms our expectation. A comparison between the top individual block ACCs and the ensemble BBLL rates reported in both **Figures 4, 5** shows that BBLL is effectively combining block-based predictions to produce equivalent or improved ACC rates. In the block ACC experiments on DDSM (**Figure 5**), we observe consistent patterns of block ACC rates between 10-fold and 30-fold CV for all block sizes except for  $64 \times 64$  px. While ensemble classification has its limitations, such as increased complexity in configuration and training, we see that ensembling reduces the variance and bias of classification.

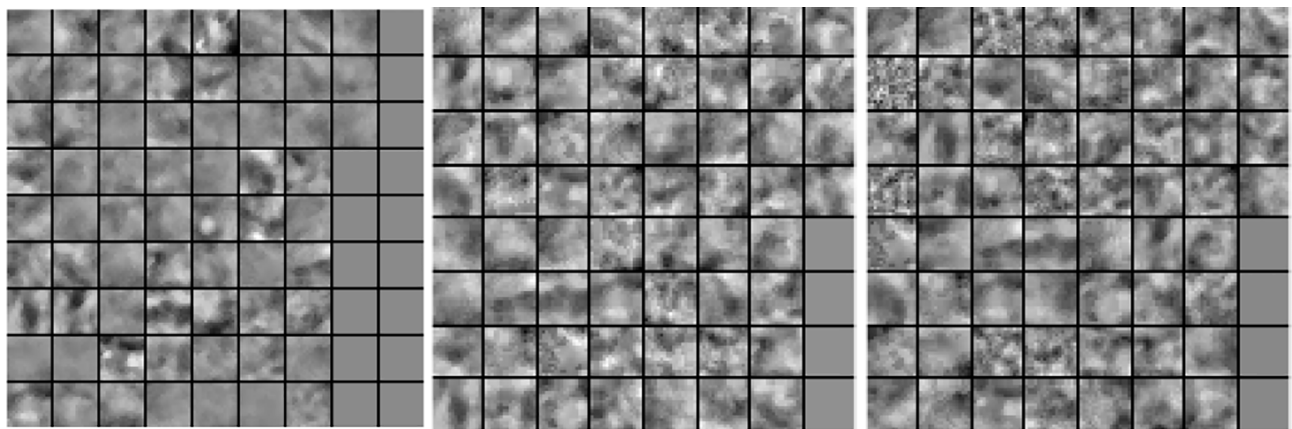
In our next experiment, we explored the dictionaries learned by LS-SLESA and LC-SLESA in terms of visual pattern representation and inter-class separability. **Figure 6** displays examples of dictionaries produced by LS-SLESA and LC-SLESA based on  $16 \times 16$  blocks from  $64 \times 64$  ROIs of the MIAS database. We also display the training set for reference. These blocks correspond to one of the  $D'$  dictionaries defined in (3) and computed by (4) and (5). They were spatially localized –7th in lexicographical order out of a  $4 \times 4$  grid. We see that the dictionary atoms correspond to basic structural patterns of the intensity distribution and texture of the masses.

In **Figure 7** we display the 4-D t-SNE (40) clustering-based embeddings of dictionaries produced under the same conditions as **Figure 6** by LS-SLESA and LC-SLESA. This figure displays pair-wise feature scatterplots and single feature histograms grouped by the mass state. We include a t-SNE clustering plot of the training data without dictionary learning for comparison. We observe greater separation between class dictionaries when dictionary learning is applied to the training data. We also computed the symmetric Kullback Leibler (KL) divergence between the classes of benign and malignant samples in the embedded spaces to measure the level of inter-class separation. The greatest KL divergence of 4.7651 occurs in the third feature





**FIGURE 5** | Classification accuracy by block for  $64 \times 64$ ,  $32 \times 32$ ,  $16 \times 16$ , and  $8 \times 8$  block experiments performed on the DDSM dataset for 10 fold CV (top row) and 30 fold CV (bottom row). The corresponding ensemble BMAP-R and BBLL-R classification decision ACCs for 10-fold experiment examples are 57.33%, 53.67%, 56.17%, 54.67%, and 59.50%, 59.67%, 60.17%, 57.33% respectively for 64, 32, 16, and 8 blocks. The corresponding ensemble BMAP-R and BBLL-R classification decision ACCs for 30-fold experiment examples are 54.67%, 57.00%, 57.67%, 54.50% and 55.00%, 61.33%, 59.50%, 58.83% respectively for 64, 32, 16, 8 blocks.



**FIGURE 6** | Dictionary comparison example for SLES without dictionary learning (left), LS-SLES (middle), and LC-SLES (right).

embedding of the LS-SLES block dictionary and the second highest KL divergence, 4.7252, occurs in the first feature embedding of the LC-SLES block dictionary. The observed separation constitutes the presence of similarities within class specific samples and further illustrates the benefit of dictionary learning on the training samples.

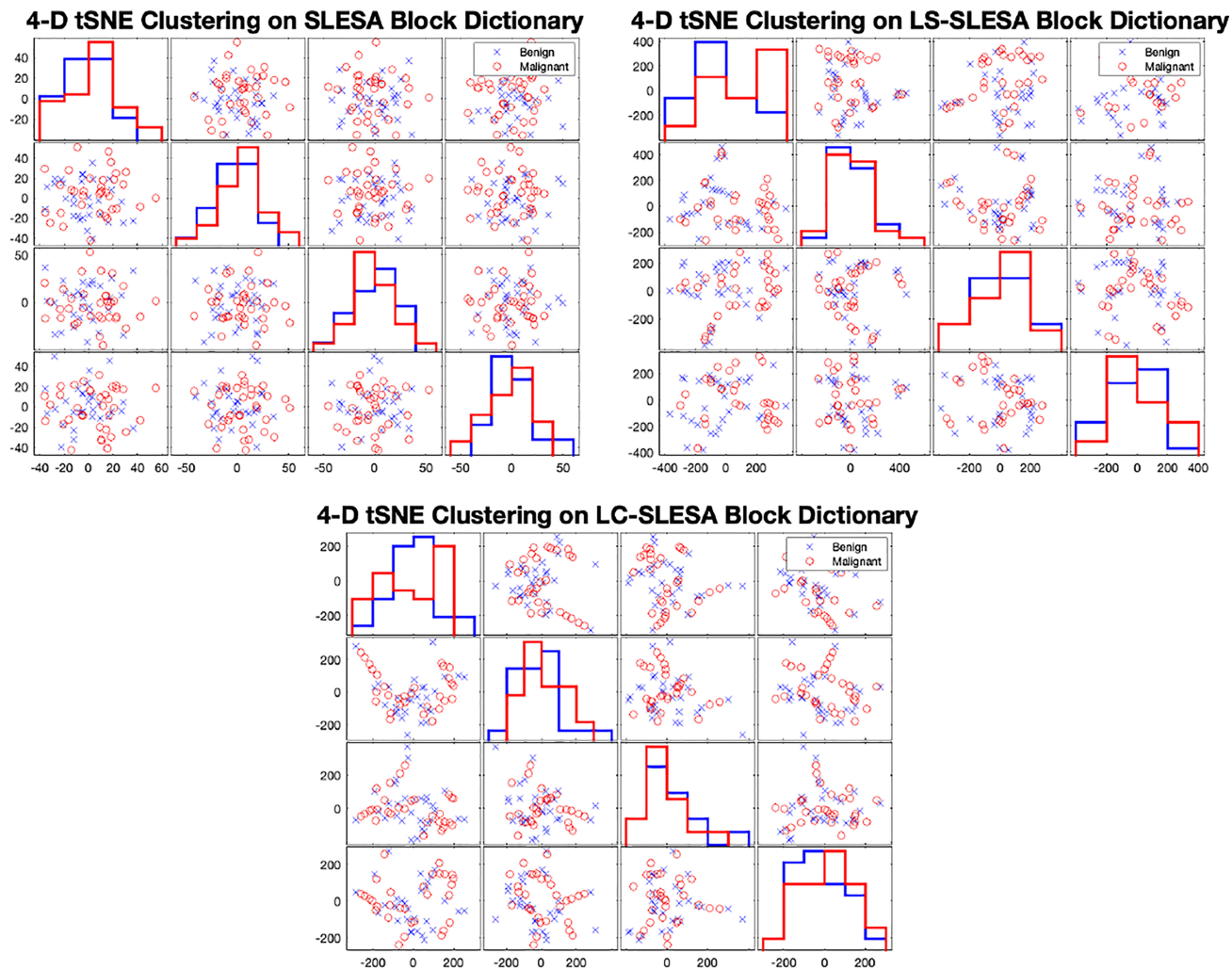
In both the MIAS and DDSM experiments we performed parameter optimization on the sparse techniques using grid search. In SLES we used  $\epsilon$  values of  $\{0.001, 0.01, 0.1, 0.5\}$ . In LS-SLES we added to the search, sparsity levels of

$\{1, 5, 10, 30, 60\}$ , and dictionary sizes of  $\{300, 500\}$  atoms for DDSM. For the MIAS data, we used 60 atoms because of the small sample size. In LC-SLES we added to the search,  $(\sqrt{\alpha}, \sqrt{\beta})$  values of  $\{(4e-4, 2e-4), (4e-3, 2e-3), (0.04, 0.02), (0.4, 0.2)\}$ .

## 4 CONCLUSION

We introduced discriminative localized sparse representations to classify breast masses as benign or malignant using





**FIGURE 7** | t-SNE clustering plots with 4-D embedding of block dictionaries produced by SLESa (top-left), LS-SLESa (top-right), and LC-SLESa (bottom). The greatest KL divergence for SLESa is 3.9353 produced by the first feature. The greatest KL divergence for LS-SLESa is 4.7651 produced by the third feature. The greatest KL divergence for LC-SLESa is 4.7252 produced by the first feature.

mammograms. LS-SLESA and LC-SLESA were designed to incorporate class-based discriminant information into the generative method of sparse representation using dictionary learning. We incorporated these approaches into a spatially localized ensemble learning methodology and extensively evaluated their classification performance. As we observed through our experimentation, these approaches produce sparse approximations that improve the classification accuracy and accomplish 93.1% area under the ROC using 30-fold cross-validation. Our results indicate that this methodology may be applicable for breast mass characterization in a breast cancer screening workflow.

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and

institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

SM designed and implemented the methods, wrote the manuscript, and performed experiments. KZ designed and implemented the methods, and performed experiments. CH implemented methods and performed experiments. All authors contributed to the article and approved the submitted version.

## FUNDING

This research was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) (award no.: SC3GM113754) and by the Army Research Office under grant no. W911NF2010095. We acknowledge the support by Delaware CTR-ACCEL (NIH U54GM104941) and the State of Delaware.

## REFERENCES

1. Ferlay J, Héry C, Autier P, Sankaranarayanan R. *Global Burden of Breast Cancer*. New York, NY: Springer New York (2010). p. 1–19. doi: 10.1007/978-1-4419-0685-41
2. Oliver A, Freixenet J, Marti J, Perez E, Pont J, Denton ER, et al. A Review of Automatic Mass Detection and Segmentation in Mammographic Images. *Med Image Anal* (2010) 14:87–110. doi: 10.1016/j.media.2009.12.005
3. Verma B, McLeod P, Klevansky A. Classification of Benign and Malignant Patterns in Digital Mammograms for the Diagnosis of Breast Cancer. *Expert Syst Appl* (2010) 37:3344–51. doi: 10.1016/j.eswa.2009.10.016
4. Pereira DC, Ramos RP, Do Nascimento MZ. Segmentation and Detection of Breast Cancer in Mammograms Combining Wavelet Analysis and Genetic Algorithm. *Comput Methods Programs Biomed* (2014) 114:88–101. doi: 10.1016/j.cmpb.2014.01.014
5. Huynh BQ, Giger ML, Li H. Digital Mammographic Tumor Classification Using Transfer Learning From Deep Convolutional Neural Networks. *J Med Imaging* (2016) 3(3):034501. doi: 10.1117/1.JMI.3.3.034501
6. Nagarajan R, Upreti M. An Ensemble Predictive Modeling Framework for Breast Cancer Classification Systems Approaches for Identifying Disease Genes and Drug Targets. *Methods* (2017) 131:128–34. doi: 10.1016/j.jymeth.2017.07.011
7. Misra S, Solomon NL, Moffat FL, Koniaris LG. Screening Criteria for Breast Cancer. *Adv Surg* (2010) 44:87–100. doi: 10.1016/j.yasu.2010.05.008
8. Beura S, Majhi B, Dash R. Mammogram Classification Using Two Dimensional Discrete Wavelet Transform and Gray-Level Co-Occurrence Matrix for Detection of Breast Cancer. *Neurocomputing* (2015) 154:1–14. doi: 10.1016/j.neucom.2014.12.032
9. Rouhi R, Jafari M, Kasaei S, Keshavarzian P. Benign and Malignant Breast Tumors Classification Based on Region Growing and CNN Segmentation. *Expert Syst Appl* (2015) 42:990–1002. doi: 10.1016/j.eswa.2014.09.020
10. Rabidas R, Midya A, Chakraborty J. Neighborhood Structural Similarity Mapping for the Classification of Masses in Mammograms. *IEEE J Biomed Health Inf* (2017) 22:826–34. doi: 10.1109/JBHI.2017.2715021
11. Singh SP, Urooj S. An Improved Cad System for Breast Cancer Diagnosis Based on Generalized Pseudo-Zernike Moment and Ada-Dewnn Classifier. *J Med Syst* (2016) 40:105. doi: 10.1007/s10916-016-0454-0
12. Narváez F, Alvarez J, Garcia-Arteaga JD, Tarquino J, Romero E. Characterizing Architectural Distortion in Mammograms by Linear Saliency. *J Med Syst* (2017) 41:26. doi: 10.1007/s10916-016-0672-5
13. George M, Chen Z, Zwiggelaar R. Multiscale Connected Chain Topological Modelling for Microcalcification Classification. *Comput Biol Med* (2019) 114:103422. doi: 10.1016/j.cmpbiomed.2019.103422
14. Sharma MK, Jas M, Karale V, Sadhu A, Mukhopadhyay S. Mammogram Segmentation Using Multi-Atlas Deformable Registration. *Comput Biol Med* (2019) 110:244–53. doi: 10.1016/j.cmpbiomed.2019.06.001
15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. *Med Image Anal* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
16. Krizhevsky A, Sutskever I, Hinton GE. Imagenet Classification With Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst* (2012) 12:1097–105.
17. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper With Convolutions. *Comput Vision Pattern Recog (CVPR)* (2015) 1–9. doi: 10.1109/CVPR.2015.7298594
18. Hepsağ PU, Özel SA, Yazıcı A. Using Deep Learning for Mammography Classification. In: *2017 International Conference on Computer Science and Engineering (UBMK)*. Antalya, Turkey: IEEE (2017). p. 418–23.
19. Zhao W, Xu R, Hirano Y, Tachibana R, Kido S. A Sparse Representation Based Method to Classify Pulmonary Patterns of Diffuse Lung Diseases. *Comput Math Methods Med* (2015) 2015:11. doi: 10.1155/2015/567932
20. Chougrad H, Zouaki H, Alheyane O. Deep Convolutional Neural Networks for Breast Cancer Screening. *Comput Methods Programs Biomed* (2018) 157:19–30. doi: 10.1016/j.cmpb.2018.01.011
21. Aharon M, Elad M, Bruckstein A. K-Svd: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Trans Signal Process* (2006) 54:4311–22. doi: 10.1109/TSP.2006.881199
22. Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S. Sparse Representation for Computer Vision and Pattern Recognition. *Proc IEEE* (2010) 98:1031–44. doi: 10.1109/JPROC.2010.2044470
23. Tosic I, Frossard P. Dictionary Learning. *IEEE Signal Process Mag* (2011) 28:27–38. doi: 10.1109/MSP.2010.939537

24. Jiang Z, Lin Z, Davis LS. Label Consistent K-Svd: Learning a Discriminative Dictionary for Recognition. *IEEE Trans Pattern Anal Mach Intell* (2013) 35:2651–64. doi: 10.1109/TPAMI.2013.88
25. Yang M, Zhang L, Feng X, Zhang D. Sparse Representation Based Fisher Discrimination Dictionary Learning for Image Classification. *Int J Comput Vision* (2014) 109:209–32. doi: 10.1007/s11263-014-0722-8
26. Zhou Y, Chang H, Barner K, Spellman P, Parvin B. Classification of Histology Sections via Multispectral Convolutional Sparse Coding. *Proc IEEE Conf Comput Vision Pattern Recog* (2014), 3081–8. doi: 10.1109/CVPR.2014.394
27. Zhang Z, Xu Y, Yang J, Li X, Zhang D. A Survey of Sparse Representation: Algorithms and Applications. *IEEE Access* (2015) 3:490–530. doi: 10.1109/ACCESS.2015.2430359
28. Plenge E, Klein SS, Niessen WJ, Meijering E. Multiple Sparse Representations Classification. *PloS One* (2015) 10(2015):1–23. doi: 10.1371/journal.pone.0131968
29. Zheng K, Makrogiannis S. Sparse Representation Using Block Decomposition for Characterization of Imaging Patterns. In: G Wu, BC Munsell, Y Zhan, W Bai, G Sanroma, P Coupé, editors. *Patch-Based Techniques in Medical Imaging: Third International Workshop, Patch-MI 2017, Held in Conjunction With MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Proceedings*. Cham: Springer International Publishing (2017). p. 158–66. doi: 10.1007/978-3-319-67434-6\_18
30. Rey-Otero I, Sulam J, Elad M. Variations on the Convolutional Sparse Coding Model. *IEEE Trans Signal Process* (2020) 68:519–28. doi: 10.1109/TSP.2020.2964239
31. Chang H, Han J, Zhong C, Snijders AM, Mao JH. Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications. *IEEE Trans Pattern Anal Mach Intell* (2017) 40:1182–94. doi: 10.1109/TPAMI.2017.2656884
32. Elad M, Yavneh I. A Plurality of Sparse Representations Is Better Than the Sparsest One Alone. *IEEE Trans Inf Theory* (2009) 55:4701–14. doi: 10.1109/TIT.2009.2027565
33. He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks. In: B Leibe, J Matas, N Sebe, M Welling, editors. *Computer Vision – ECCV 2016*. Cham: Springer International Publishing (2016). p. 630–45.
34. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. *IEEE Conf Comput Vision Pattern Recog (CVPR)* (2016). p. 2818–26. doi: 10.1109/CVPR.2016.308
35. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust Face Recognition via Sparse Representation. *IEEE Trans Pattern Anal Mach Intell* (2009) 31:210–27. doi: 10.1109/TPAMI.2008.79
36. Shrivastava A, Patel VM, Pillai JK, Chellappa R. Generalized Dictionaries for Multiple Instance Learning. *Int J Comput Vision* (2015) 114:288–305. doi: 10.1007/s11263-015-0831-z
37. Pappas V, Romano Y, Elad M. Convolutional Neural Networks Analyzed Via Convolutional Sparse Coding. *J Mach Learn Res* (2017) 18:2887–938.
38. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv preprint arXiv* (2012) 25.
39. Mockus J. *Bayesian Approach to Global Optimization: Theory and Applications* Vol. 37. Springer Science & Business Media (2012).
40. Van der Maaten L, Hinton G. Visualizing Data Using T-Sne. *J Mach Learn Res* (2008) 9:2579–605.

**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.

Copyright © 2021 Makrogiannis, Zheng and Harris. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# From Fitting the Average to Fitting the Individual: A Cautionary Tale for Mathematical Modelers

Michael C. Luo<sup>1†</sup>, Elpiniki Nikolopoulou<sup>2†</sup> and Jana L. Gevertz<sup>1\*</sup>

<sup>1</sup> Department of Mathematics and Statistics, The College of New Jersey, Ewing, NJ, United States, <sup>2</sup> School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, United States

## OPEN ACCESS

### Edited by:

George Bebis,  
University of Nevada, Reno,  
United States

### Reviewed by:

Tyler Cassidy,  
Pfizer, United States  
Hsiu-Chuan Wei,  
Feng Chia University, Taiwan

### \*Correspondence:

Jana L. Gevertz  
gevertz@tcnj.edu

### <sup>†</sup>Present address:

Michael C. Luo,  
Department of Mathematical  
Sciences, New Jersey Institute of  
Technology, Newark, NJ,  
United States  
Elpiniki Nikolopoulou,  
Nationwide, Columbus, OH,  
United States

### Specialty section:

This article was submitted to  
Cancer Immunity  
and Immunotherapy,  
a section of the journal  
Frontiers in Oncology

**Received:** 12 October 2021

**Accepted:** 09 March 2022

**Published:** 28 April 2022

### Citation:

Luo MC, Nikolopoulou E and  
Gevertz JL (2022) From Fitting  
the Average to Fitting the  
Individual: A Cautionary Tale  
for Mathematical Modelers.  
Front. Oncol. 12:793908.  
doi: 10.3389/fonc.2022.793908

An outstanding challenge in the clinical care of cancer is moving from a one-size-fits-all approach that relies on population-level statistics towards personalized therapeutic design. Mathematical modeling is a powerful tool in treatment personalization, as it allows for the incorporation of patient-specific data so that treatment can be tailor-designed to the individual. Herein, we work with a mathematical model of murine cancer immunotherapy that has been previously-validated against the average of an experimental dataset. We ask the question: what happens if we try to use this same model to perform personalized fits, and therefore make individualized treatment recommendations? Typically, this would be done by choosing a single fitting methodology, and a single cost function, identifying the individualized best-fit parameters, and extrapolating from there to make personalized treatment recommendations. Our analyses show the potentially problematic nature of this approach, as predicted personalized treatment response proved to be sensitive to the fitting methodology utilized. We also demonstrate how a small amount of the right additional experimental measurements could go a long way to improve consistency in personalized fits. Finally, we show how quantifying the robustness of the average response could also help improve confidence in personalized treatment recommendations.

**Keywords:** cancer, mathematical modeling, personalized therapy, immunotherapy, nonlinear mixed effects modeling

## 1 INTRODUCTION

The conventional approach for developing a cancer treatment protocol relies on measuring average efficacy and toxicity from population-level statistics in randomized clinical trials (1–3). However, it is increasingly recognized that heterogeneity, both between patients and within a patient, is a defining feature of cancer (4, 5). This inevitably results in a portion of cancer patients being over-treated and suffering toxicity consequences from the standard-of-care dose, and another portion being under-treated and not benefiting from the expected efficacy of the treatment (6).

For these reasons, in the last decade there has been much interest in moving away from this ‘one-size-fits-all’ approach to cancer treatment and towards personalized therapeutic design (also called predictive or precision medicine) (1, 2, 7). Collecting patient-specific data has the potential to improve treatment response to chemotherapy (6, 8–11), radiotherapy (12–14), and targeted



molecular therapy (11, 15–17). However, it has been proposed that personalization may hold the most promise when it comes to immunotherapy (18). Immunotherapy is an umbrella term for methods that increase the potency of the immune response against cancer. Unlike other treatment modalities that directly attack the tumor, immunotherapy depends on the interplay between two complex systems (the tumor and the immune system), and therefore may exhibit more variability across individuals (18).

Mathematical modeling has become a valuable tool for understanding tumor-drug interactions. However, just as clinical care is guided by standardized recommendations, most mathematical models are validated based on population-level statistics from preclinical or clinical studies (19). To truly realize the potential of mathematical models in the clinic, these models must be individually parameterized using measurable, patient-specific data. Only then can modeling be harnessed to answer some of the most pressing questions in precision medicine, including selecting the right drug for the right patient, identifying the optimal drug combination for a patient, and prescribing a treatment schedule that maximizes efficacy while minimizing toxicity.

Efforts to personalize mathematical models have been undertaken to understand glioblastoma treatment response (20, 21), to identify optimal chemotherapeutic and granulocyte colony-stimulating factor combined schedules in metastatic breast cancer (22), to identify optimal maintenance therapy chemotherapeutic dosing for childhood acute lymphoblastic leukemia (9), and to identify optimized doses and dosing schedules of the chemotherapeutic everolimus with the targeted agent sorafenib for solid tumors (23). Interesting work has also been done in the realm of radiotherapy, where individualized head and neck cancer evolution has been modeled through a dynamic carrying capacity informed by patient response to their last radiation dose (24).

Beyond these examples, most model personalization efforts have focused on prostate cancer, as prostate-specific antigen is a clinically measurable marker of prostate cancer burden (25) that can be used in the parameterization of personalized mathematical models. The work of Hirata and colleagues has focused on the personalization of intermittent androgen suppression therapy using retrospective clinical trial data (26, 27). Other interesting work using clinical trial data has been done by Agur and colleagues, focusing on individualizing a prostate cancer vaccine using retrospective phase 2 clinical trial data (25, 28), as well as androgen deprivation therapy using data from an advanced stage prostate cancer registry (29). Especially exciting work on personalizing prostate cancer has been undertaken by Gatenby and colleagues, who used a mathematical model to discover patient-specific adaptive protocols for the administration of the chemotherapeutic agent abiraterone acetate (30). Among the 11 patients in a pilot clinical trial treated with the personalized adaptive therapy, they observed the median time to progression increased to at least 27 months as compared to 16.5 months observed with standard dosing, while also using a cumulative drug amount that was 47% less than the standard dosing (17).

Despite these examples, classically mathematical models are not personalized, but are validated against the average of experimental data. In particular, modelers choose a single fitting methodology, a single cost function to minimize, and find the best-fit parameters to the average of the data. Using the best-fit parameters and the mathematical model, treatment optimization can be performed. Recognizing the limitations of this approach in describing variable treatment response across populations, modelers have begun employing virtual population cohorts (31–33). There is much value in this population-level approach to study variability, but it is not equivalent to looking at individualized treatment response.

In this work, we explore the consequences of performing individualized fits using a minimal mathematical model previously-validated against the average of an experimental dataset. In *Materials and Methods*, we describe the preclinical data collected by Huang et al. (34) on a mouse model of melanoma treated with two forms of immunotherapy, and our previously-developed mathematical model that has been validated against population-level data from this trial (35). Individual mouse volumetric time-course data is fit to our dynamical systems model using two different approaches detailed in *Materials and Methods*: the first fits each mouse independent of the other mice in the population, whereas the second constrains the fits to each mouse using population-level statistics. In *Results*, we demonstrate that the treatment response identified for an individual mouse is *sensitive to the fitting methodology utilized*. We explore the causes of these predictive discrepancies and how robustness of the optimal-for-the-average treatment protocol influences these discrepancies. We conclude with actionable suggestions for how to increase our confidence in mathematical predictions made from personalized fits.

## 2 MATERIALS AND METHODS

### 2.1 Data Set

The data in this study considers the impact of two immunotherapeutic protocols on a murine model of melanoma (34). The first protocol uses oncolytic viruses (OVs) that are genetically engineered to lyse and kill cancer cells. In (34) the OVs are immuno-enhanced by inserting transgenes that cause the virus to release 4-1BB ligand (4-1BBL) and interleukin (IL)-12, both of which result in the stimulation of the tumor-targeting T cell population (34). The preclinical work of Huang et al. has shown that oncolytic viruses carrying 4-1BBL and IL-12 (which we will call Ad/4-1BBL/IL-12) can cause tumor debulking *via* virus-induced tumor cell lysis, and immune system stimulation from the local release of the immunostimulants (34).

The second protocol utilized by Huang et al. are dendritic cell (DC) injections. DCs are antigen-presenting cells that, when exposed to tumor antigens *ex vivo* and intratumorally injected, can stimulate a strong adaptive immune response against cancer cells (34). Huang et al. showed that combination of Ad/4-1BBL/IL-12 with DC injections results in a stronger antitumor response than either treatment individually (34). Volumetric trajectories

of individual mice treated with three doses of Ad/4-1BBL/IL-12 on days 0, 2 and 4, and three doses of DCs on days 1, 3, 5, along with the average trajectory, are shown in **Figure 1**.

## 2.2 Mathematical Model

Our model contains the following five ordinary differential equations:

$$\frac{dU}{dt} = rU - \beta \frac{UV}{N} - (\kappa_0 + c_{kill}I) \frac{UT}{N}, U(0) = U_0, \quad (1)$$

$$\frac{dI}{dt} = \beta \frac{UV}{N} - \delta_I I - (\kappa_0 + c_{kill}I) \frac{IT}{N}, I(0) = 0, \quad (2)$$

$$\frac{dV}{dt} = u_V(t) + a\delta_I I - \delta_V V, V(0) = 0 \quad (3)$$

$$\frac{dT}{dt} = c_T I + \chi_D D - \delta_T T, T(0) = 0, \quad (4)$$

$$\frac{dD}{dt} = u_D(t) - \delta_D D, D(0) = 0 \quad (5)$$

where  $U$  is the volume of uninfected tumor cells,  $I$  is the volume of OV-infected tumor cells,  $V$  is the volume of free OVs,  $T$  is the volume of tumor-targeting T cells,  $D$  is the volume of injected dendritic cells, and  $N$  is the total volume of cells (tumor cells and T cells) at the tumor site. When all parameters and time-varying terms are positive, this models captures the effects of tumor growth and response to treatment with Ad/4-1BBL/IL-12 and DCs (35). By allowing various parameters and time-varying terms to be identically zero, other treatment protocols tested in Huang et al. (34) can also be described.

This model was built in a hierarchical fashion, details of which have been described extensively elsewhere (32, 35–37).

Here, we briefly summarize the full model. Uninfected tumor cells grow exponentially at a rate  $r$ , and upon being infected by an OV convert to infected cancer cells at a density-dependent rate  $\beta UV/N$ . These infected cells get lysed by the virus or other mechanisms at a rate of  $\delta_I$ , thus acting as a source term for the virus by releasing an average of  $\alpha$  free virions into the tissue space. Viruses decay at a rate of  $\delta_V$ .

The activation/recruitment of tumor-targeting T cells can happen in two ways: 1) stimulation of cytotoxic T cells due to 4-1BBL or IL-12 (modeled through  $I$ , at a rate of  $c_T$ , as infected cells are the ones to release 4-1BBL and IL-12), and 2) production/recruitment due to the externally-primed dendritic cells at a rate of  $\chi_D$ . These tumor-targeting T cells indiscriminately kill uninfected and infected tumor cells, with the rate of killing that depends on IL-12 and 4-1BBL production (again, modeled through  $I$  in the term  $(\kappa_0 + c_{kill}I)$ ), and they can also experience natural death at a rate of  $\delta_T$ . The time-dependent terms,  $u_V(t)$  and  $u_D(t)$ , represent the source of the drug and are determined by the delivery and dosing schedule of interest.

## 2.3 Fitting Methodologies

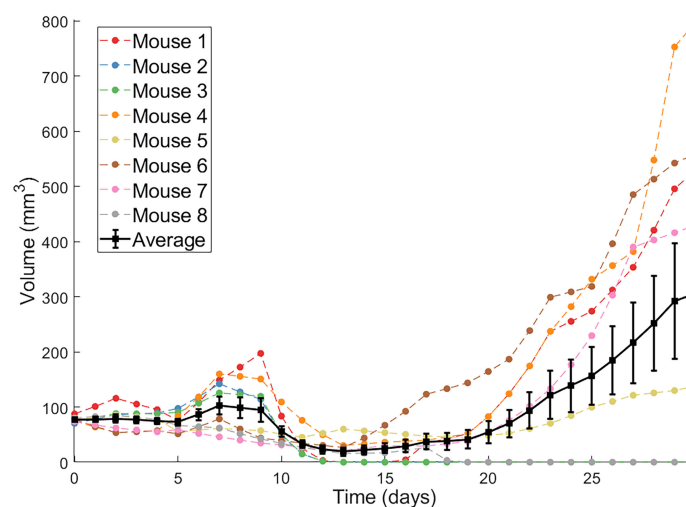
For both fitting methodologies, the full set of model parameters  $\{r, \beta, \alpha, \delta_V, \kappa_0, \delta_T, \chi_D, \delta_I, c_{kill}, c_T, \delta_D, U_0\}$ , which includes the initial uninfected tumor volume, is fit to each individual mouse.

### 2.3.1 Independently Fitting Individuals

Our first attempt at individualized fitting is to find the parameter set that minimizes the square of the  $\ell^2$ -norm between the model and the individual mouse data:

$$\zeta = \sum_{t=0}^n (V_{model}(t) - V_{data}(t))^2 \quad (6)$$

where  $V_{model}(t) = U(t) + I(t)$  is the volumetric output predicted by our model in eqns. (1)–(5),  $V_{data}(t)$  represents the volumetric



**FIGURE 1** | Individual volumetric trajectories are shown for eight mice treated with Ad/4-1BBL/IL-12 (on days 0, 2, 4) + DCs (on days 1, 3, 5). The average, with standard error bars, is also shown in black (34).

data for an individual mouse, and  $n$  is the last time point at which the volume is measured in the experiments.

To independently fit an individual mouse, 12-dimensional space is first quasi-randomly sampled (with each point sampled in the range  $[0,1]$ ) using high-dimensional Sobol' Low Discrepancy Sequences (LDS). LDS are designed to give rise to quasi-random numbers that sample points in space as uniformly as possible, while also (typically) having faster convergence rates than standard Monte Carlo sampling methods (38). Each randomly sampled point is then scaled to be in a biologically plausible range for the corresponding parameter value. For those parameters that were previously-fit to the average of the experimental data  $(r, \beta, \chi_D, c_T, c_{kill})$ , the range was set using the lower and upper-bound of the 95%-credible interval for the parameter, as determined in (35). For parameters not fit to the average in prior work, the minimum and maximum values were set to 50% and 200% of the value the parameter was fixed to for the average, respectively. See **Supplementary Table 1** for details.

After the best-fit parameter set has been selected among the  $10^6$  randomly sampled sets chosen by LDS, the optimal is refined using simulated annealing (39). Having observed that the landscape of the objective function near the optimal parameter set does not contain local minima, we randomly perturb the LDS-chosen parameter set, and accept any realistic parameter changes that decrease the value of the objective function - making the method equivalent to gradient descent. We consider a parameter set realistic at this stage if all parameter values are non-negative. This random perturbation process is repeated until no significant change in  $\zeta$  can be achieved, which we defined as the relative change in  $\zeta$  for the last five accepted parameter sets being less than  $10^{-5}$ . We call this final parameter set the optimal parameter set. More details can be found in **Supplementary Algorithm 1**.

It is important to note that, by approaching fitting in this way, the parameters for Mouse  $i$  depend only the volumetric data for Mouse  $i$ ; that is, the volumetric data for the other mice are not accounted for.

### 2.3.2 Fitting Individuals with Population-Level Constraints

Nonlinear mixed effects (NLME) models incorporate fixed and random effects to generate models to analyze data that are non-independent, multilevel/hierarchical, longitudinal, or correlated (40). Fixed effects refer to parameters that can generalize across an entire population. Random effects refer to parameters that differ between individuals that are randomly sampled from a population. To employ NLME for our mathematical model, for each mouse  $i$  we define the structural model  $T(t_{ij}, \psi_i) = U(t_j) + I(t_j)$ . We assume that each parameter  $\psi_{i,k}$  in the parameter set  $\psi_i$  is lognormally distributed with mean  $\bar{\psi}_{i,k}$  and standard deviation  $\omega_{i,k}$ :

$$\log(\psi_{i,k}) \sim \mathcal{N}(\log(\bar{\psi}_{i,k}), \omega_{i,k}^2) \quad (7)$$

We also assume that the error is a scalar value proportional to our structural model. Our resultant mixed effects model is:

$$y_{ij} = T(t_{ij}, \psi_i) + bT(t_{ij}, \psi_i) \varepsilon_{i,j}, i = 1, \dots, M, j = 0, \dots, n_i - 1, \quad (8)$$

where  $y_{ij}$  is the predicted tumor volume at each day  $j$  for each individual  $i$  (that is, at time  $t_{ij}$ ),  $M = 8$  is the number of mice,  $n_i = 31$  is the number of observations per mouse, and  $\varepsilon_{ij}$  is the random noise term, which we assume to follow a standard normal distribution.

Typically, NLME models attempt to maximize the likelihood of the parameter set given the available data. There does not exist a general closed-form solution to this maximization problem (41), so numerical optimization is often needed to find a maximum likelihood estimate. In this work, we employ Monolix (42), which uses a Markov Chain Monte Carlo method to find values of the model parameters that optimize the likelihood function. To implement NLME in Monolix, we first processed and arranged our experimental data consisting of tumor volume and dosing schedule in a Monolix-specified spreadsheet. The data is then censored to avoid overfitting very small tumor volumes, as detailed in (43). To understand why this overfitting occurs in uncensored data, consider the scenario where the model predicts a volume  $10^{-4} \text{ mm}^3$  at a time point whereas the experimental measurement is  $0 \text{ mm}^3$ . The parameter set corresponding to this prediction is assigned a lower likelihood, despite the fact that  $10^{-4}$  is a perfectly reasonable model prediction of an experimental measurement of 0. To avoid penalizing insignificant prediction errors at very small tumor volumes, the data has been censored so that when the negative log likelihood is computed, instead of calculating the likelihood the model gives exactly the value of 0, it computes the likelihood the model predicts a value between 0 and 1. While this censoring was necessary to prevent NLME from over-fitting data points of volume zero at the expense of the fits to the nonzero data points, such censoring was not required for the independent fitting approach, as there we are just minimizing the sum of the square error. That is, in the independent fitting approach, when the model predicts a very small volume and the experimental measurement is 0, the contribution to the sum of the square error is negligible and thus censoring is not needed.

In order to solve this NLME model in Monolix, initial guesses are needed for the mean and standard deviation of our lognormally-distributed parameters. Based on previous fits to the average of the data in (37), we used the following set of initial guesses for the mean of each parameter:

$$\begin{aligned} & [r, \beta, \alpha, \delta_v, \kappa_0, \delta_T, \chi_D, \delta_I, c_{kill}, c_T, \delta_D, U_0] \\ & = [0.32, 1, 3, 2.3, 2, 0.35, 5.5, 1, 0.51, 1.2, 0.35, 55.6], \end{aligned}$$

and after numerical exploration, we ended up choosing the initial standard deviations as:

$$\begin{aligned} & [\omega_r, \omega_\beta, \omega_\alpha, \omega_{\delta_v}, \omega_{\kappa_0}, \omega_{\delta_T}, \omega_{\chi_D}, \omega_{\delta_I}, \omega_{c_{kill}}, \omega_{c_T}, \omega_{\delta_D}, \omega_{\kappa_0}] \\ & = [0.25, 0.5, 1, 0.1, 1, 0.1, 0.25, 0.1, 0.5, 0.5, 0.1, 5] \end{aligned}$$

## 2.4 Practical Identifiability via the Profile Likelihood Method

It is well-established that estimating a unique parameter set for a mathematical model can be challenging due to the limited

availability of often noisy experimental data (44). A non-identifiable model is one in which multiple parameter sets give “good” fits to the experimental data. Here, we will study the practical identifiability of our system in eqns. (1) - (5) using the profile likelihood approach (45, 46).

A single parameter is profiled by fixing it across a range of values, and subsequently fitting all other model parameters to the data (44). To execute the profile likelihood method, let  $p$  be the vector that contains all parameters of the model,  $\theta$  be one parameter of interest contained in the vector  $p$ . The profile likelihood  $PL$  for the parameter  $\theta$  is defined in (47) as:

$$PL(\theta) = \min_{p \in \{p | p_k = \theta\}} \left( \sum_{t=0}^n \left( \frac{V_{data}(t) - V_{model}(t; p)}{\sigma(t)} \right)^2 \right) \quad (9)$$

where  $V_{model}(t; p) = U(t) + I(t)$  is the volumetric output predicted by our model for parameter set  $p$ , and  $\bar{V}_{data}$  represents the average volume across all mice at that time point with corresponding standard deviation  $\sigma(t)$ . For normally distributed observational noise this corresponds to the maximum likelihood estimate of  $\theta$ :

$$PL(\theta) = \min_{p \in \{p | p_k = \theta\}} (-2LL(p; V_{data}(0), \dots, V_{data}(n))) \quad (10)$$

where  $LL(p; V_{data}(0), \dots, V_{data}(n))$  is the log of the likelihood function. The likelihood function represents the likeliness of the measured data  $V_{data}$  given a model with parameters  $p$  (48). This likelihood is higher for a parameter set that is more likely given the available data, and it is smaller for parameter sets that are less likely given the data. The profile likelihood curve for any parameter of interest  $\theta$  is found using the following process:

1. Determine a range for the parameter values of  $\theta$ .
2. Fix  $\theta = \theta^*$  at a value in the range.
3. Find the value of the non-fixed parameters that minimize the objective function in eqn. (9). The quasi-random Monte Carlo method with gradient descent was used for the fitting, as detailed previously.
4. Evaluate the objective function at those optimum values for the fixed value of  $\theta^*$ .
5. Repeat the process described in steps 2-4 for a discrete set of values in the range of the parameter  $\theta$ . This yields the profile likelihood function for the parameter  $\theta$ .

This process results in a population-level (not individual) profile likelihood curve for each parameter. Once  $PL(\theta)$  is determined, the confidence interval for  $\theta$  at a level of significance  $\alpha$  can be computed using:

$$PL(\theta) - 2LL(p_k^*) \leq \Delta_\alpha \quad (11)$$

where  $\Delta_\alpha$  denotes the  $\alpha$  quantile of the  $\chi^2$  distribution with  $df$  degrees of freedom (which represents the number of fit model parameters when calculating  $PL(\theta)$ ) (44). We use  $\alpha = 0.95$  for a 95% confidence interval. The intersection points between the threshold  $2LL(p_k^*) + \Delta_\alpha$  and  $PL(\theta)$  result in the bounds of the confidence interval. A parameter is said to be practically identifiable if the shape of the profile likelihood plot is close to

quadratic on a finite confidence interval (49). Otherwise, a parameter is said to be practically unidentifiable.

## 3 RESULTS

### 3.1 Personalized Fits

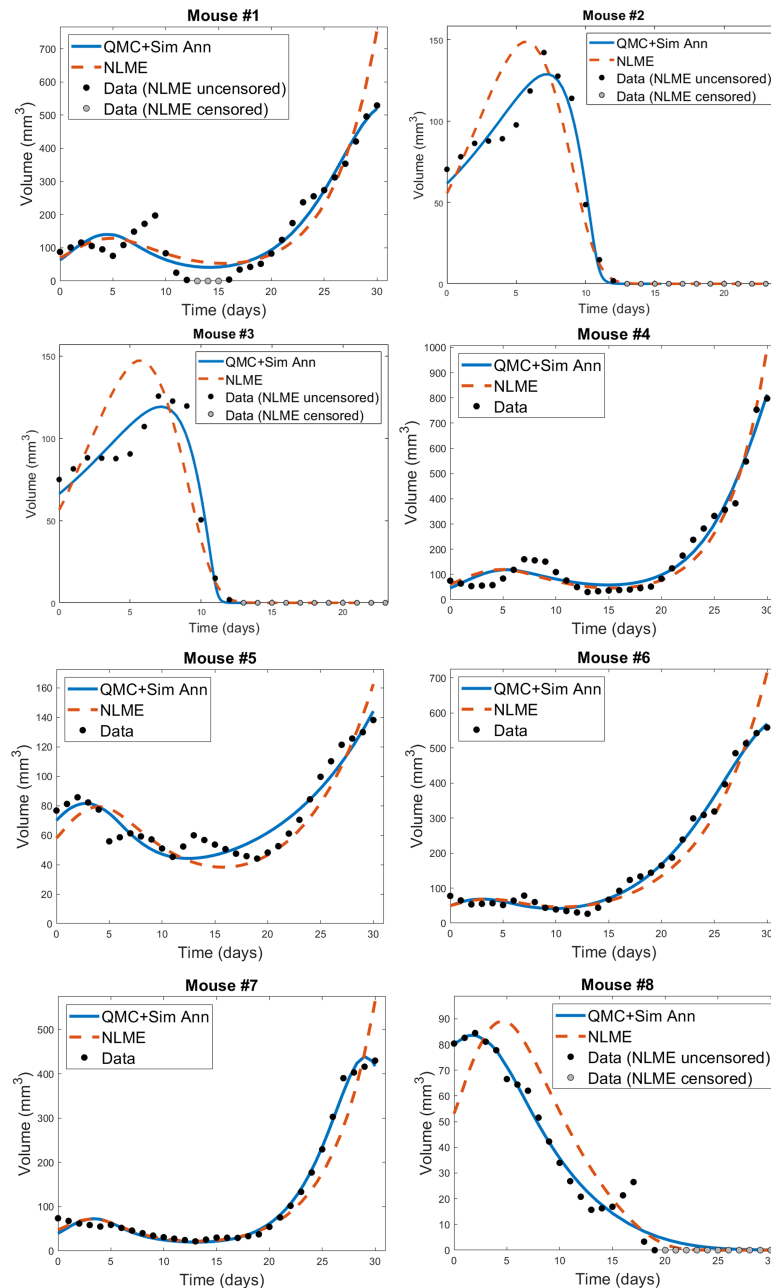
The individual mouse data in response to treatment with Ad/4-1BBL/IL-12 + DCs (34) is fit using the two methodologies discussed previously: 1) quasi-Monte-Carlo method with gradient descent in which each mouse is fit independently (which we will call the “QMC” method for short), and 2) nonlinear mixed effects modeling in which population-level statistics constrain individual fits. In **Figure 2**, we can see the best-fit for each mouse using the two fitting approaches.

We do observe some shortcomings in the fits, particularly at earlier time points. These are most-pronounced in Mouse 1 and 4, where the model cannot capture the early-time decrease in tumor volume. This highlights that a model validated against the average of a dataset may not be fully sufficient at describing individual trajectories. That said, we overall find that the model is able to capture the trends in the volumetric data despite the heterogeneity across individual mice. While a more detailed model could potentially pick up some early-time trends our model did not capture, this would come at the expense of introducing more (likely non-identifiable) parameters.

For each mouse, the QMC algorithm results in a fit that more accurately captures the dynamics in the experimental data. The differences between the two fitting methodologies explain why this is occurring. NLME assumes each parameter is sampled from a lognormal distribution whose mean and variance are determined by the full population of mice. The estimated lognormal distributions for each model parameter are shown in **Figure S1**. On the other hand, the QMC algorithm fits each mouse independently, and despite the initial bounds set on the parameters when sampling parameter space, gradient descent relaxes these constraints and the end result is that non-negativity is the only constraint imposed. This allows the QMC algorithm to explore a larger region of parameter space, resulting in better fits. The potential downside, as we will show, is that the QMC algorithm can select parameter values that deviate more significantly from the average value. This variability may represent the true variability across individual mice, or may be a consequence of doing independent fits.

In **Figure 3** and **Figure S2** we show the best-fit parameter value for each mouse and fitting methodology relative to the best-fit parameter value for the average mouse. For example, the best-fit value of the tumor growth rate  $r$  to the average of the control data has been shown to be  $r = 0.3198$  (35). Since Mouse 1 has a relative value of 1.0916 when fitting is done using QMC, the value of  $r$  predicted for that Mouse is 9.16% larger than the value for the average mouse, meaning QMC predicts  $r = 0.3491$  for Mouse 1. On the other hand, the relative value is 0.7512 when fitting is done using NLME, meaning the predicted value is  $r = 0.2402$ , which is 24.88% less than the value for the average mouse.

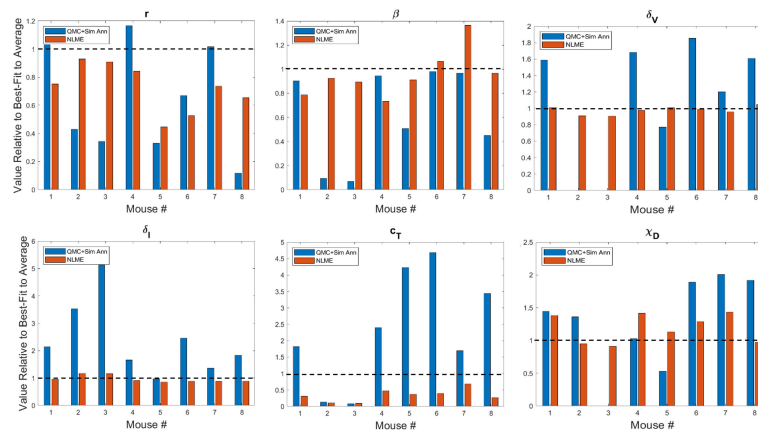




**FIGURE 2** | Best-fit for each mouse treated with Ad/41BBL/IL-12 and DCs in the order VDVVD at a dose of  $2.5 \times 10^9$  OV and  $10^6$  DCs (34). The QMC fits (in which each mouse is treated independently of the others) are shown in blue, and the NLME fits are shown in red. The experimental data (black if uncensored for NLME fitting, grey if censored) is also provided on each plot.

A study of the values a parameter can take on across methodologies reveals that while most values are of the same order of magnitude, differences can exist across methodologies. As expected due to the constraining lognormal distribution, NLME-associated parameters exhibit smaller variations from the best-fit parameter for the average mouse than QMC-associated parameters. Generally speaking, the variation seen

could be explained by a heterogeneous response to the treatment protocol across mice. For instance, a very small value of  $\chi_D$  in Mouse 3 indicates the DCs are not successfully stimulating the production of tumor-targeting T cells. As another example, a very small value of  $\kappa_0$  in Mouse 2 indicates that in the absence of immunostimulation, the T cells are unable to target and destroy cancer cells. There is one scenario that emerges in both Mouse 2



**FIGURE 3** | Best-fit values of tumor growth rate parameter  $r$ , virus infectivity parameter  $\beta$ , viral decay rate  $\delta_v$ , infected cell lysis rate  $\delta_i$ , T cell stimulation term by immunostimulants  $c_T$ , and T cell stimulation term by DCs  $\chi_D$ . The best-fit values are shown for each mouse and are presented relative to the best-fit value of the parameter in the average mouse (35). Therefore, a value of 1 (shown in the dashed black line) means the parameter value is equal to that in the average mouse, less than 1 is a smaller value, and greater than 1 is a larger value. Values for other model parameters are shown in **Figure S2**.

and 3, however, that cannot be explained by a heterogeneous treatment response. In particular, the QMC approach predicts that these mice have  $\delta_v = 0$ , indicative that the virus will not decay over the 30-day experimental time period. As this scenario is highly unlikely, we also refit these mice using the QMC approach, assuming a (somewhat arbitrary) lower bound on the viral decay rate of  $\delta_v = 0.46 \text{ day}^{-1}$ , which assumes the decay rate can never be smaller than a quarter of the average value of  $\delta_v = 2.3 \text{ day}^{-1}$  (37). Refitting both mice with the QMC algorithm and this additional constraint resulted in the best-fit value of  $\delta_v$  being this strict lower bound. All treatment predictions presented in this manuscript were identical whether Mouse 2 or 3 was analyzed using the parameter set with  $\delta_v = 0$  or the parameter set where  $\delta_v$  was a quarter the maximum value.

Looking across methodologies, parameter disparities are the most pronounced in  $c_T$ , the rate of cytotoxic T cell stimulation from 4-1BBL and IL-12. The QMC-predicted parameters cover a much larger range of values relative to the average mouse. According to the QMC fits,  $c_T$  can range anywhere from 92.15% below the value in the average mouse to 4.69 times higher than the value in the average mouse. Compare this to the NLME-predicted values of  $c_T$ , which can range from 90.29% below the value in the average mouse to 31.87% below the value for the average mouse. What is clear from looking at the best-fit parameter values across methodologies is that it is not differences in a single or small set of parameter values that explain the difference in fits. The nonlinearities in the model simply do not allow the effects of one parameter to be easily teased out from the effects of the other parameters.

### 3.2 Personalized Treatment Response at Experimental Dose

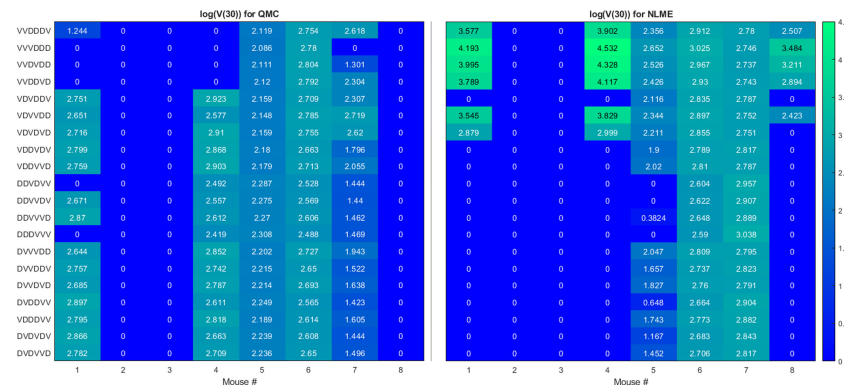
Here we seek to determine if the two sets of best-fit parameters for a single individual yield similar personalized predictions about tumor response to a range of treatment protocols. The

treatment protocols we consider are modeled after the experimental work in (34). Each day consists of only a single treatment, which can be either an injection of Ad/4-1BBL/IL-12 at  $2.5 \times 10^9$  viruses per dose, or a dose of  $10^6$  DCs. Treatment will be given for six consecutive days, with three days of treatment being Ad/4-1BBL/IL-12, and three days being DCs. If only one dose can be given per day, there are exactly 20 treatment protocols to consider. The 20 protocols are shown on the vertical axis in **Figure 4**, where  $V$  represents a dose of Ad/4-1BBL/IL-12, and  $D$  represents a dose of dendritic cells.

To quantify predicted tumor response, we will simulate mouse dynamics using the determined best-fit parameters for each of the 20 6-day protocols. Unless otherwise stated, we will use the predicted tumor volume after 30 days,  $V(30)$ , to quantify treatment response. For each fitting methodology, mouse, and protocol we display the  $\log(V(30))$  in a heatmap (as in **Figure 4**). For all  $V(30) \leq 1 \text{ mm}^3$ , we display the logarithm as 0, as showing negative values would hinder cross-methodology comparison and overemphasize insignificant differences in treatment response. We consider any tumor with  $V(30) < 1 \text{ mm}^3$  to be effectively treated by the associated protocol. Any nonzero values correspond to the value of  $\log(V(30))$  when  $V(30) > 1 \text{ mm}^3$ , and we assume these tumors have not been successfully treated. The resulting heatmap at the experimental dose of  $2.5 \times 10^9$  viruses per dose, and  $10^6$  DCs per dose is shown in **Figure 4**.

Ideally, we would find that treatment response to a protocol for a given mouse is independent of the fitting methodology utilized, at least in the binary sense of treatment success or failure. However, that does not generally appear to be the case for our data, model and fitting methodologies, as we elaborate on here.

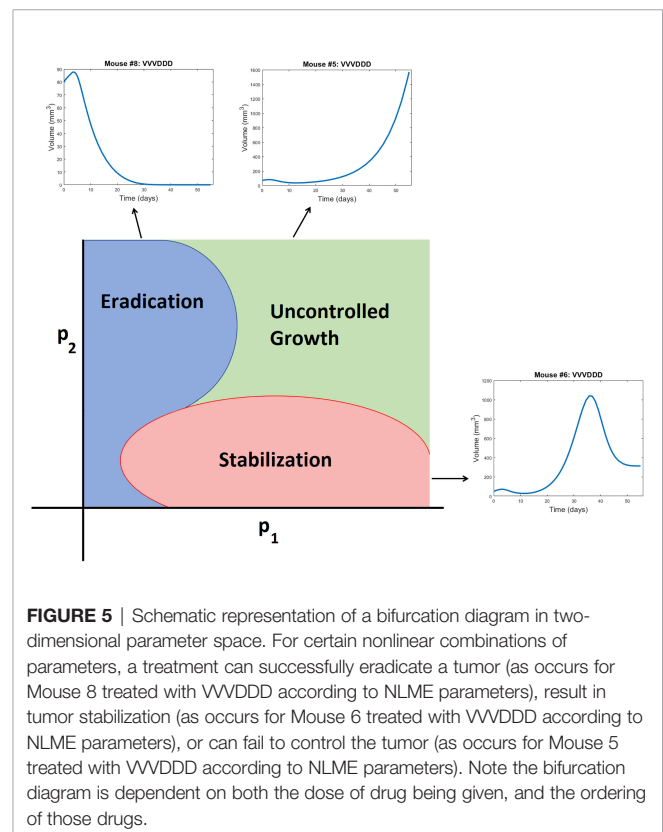
- **Cumulative statistics on consistencies across methodologies.** As shown in **Figure 4**, the two fitting methodologies give the



**FIGURE 4** | Heatmaps showing the log of the tumor volume measured at 30 days, at the OV and DC dose used in (34). If  $\log(V(30)) \leq 1$ , its value is shown as 0 on the heatmap. Left shows predictions when parameters are fit using QMC, and right shows NLME predictions.

same qualitative predictions for 73.75% (118/160) of the treatment protocols. Of the 118 agreements, 57 consistently predict treatment success whereas 61 consistently predict treatment failure. It is of note that these numbers only change slightly if we use  $V(80)$  as our measurement for determining treatment success or failure (81.875% agreement with 78/131 consistently predicting eradication and 53/131 consistently predicting failure - see **Figure S3**). Mouse 2, 3 and 6 have perfect agreement across fitting methodologies, and Mouse 7 has 95% agreement across methodologies. For these mice, treatment response is generally not dependent on dosing order. For instance, Mouse 2 and 3 are successfully treated by all twenty protocols considered, whereas Mouse 6 cannot be successfully treated by any protocol. In fact,  $V(30)$  for Mouse 6 is highly conserved across dosing order, suggesting that the ordering itself is having minimal impact on treatment response. While performing a bifurcation analysis in 11D parameter space is not feasible, what is clear is that for the mice with significant agreement across methodologies, the best-fit parameters must be sufficiently far from the bifurcation surface, as shown in the schematic diagram in **Figure 5**. As a result, predicted treatment response is not sensitive to changes in the parameter values that result from using a different fitting methodology. While not equivalent, they also do not appear to be sensitive to dosing order.

- Cumulative statistics on inconsistencies across methodologies.** The two fitting methodologies give different qualitative predictions for 26.25% (42/160) of the treatment protocols (see **Figure 4**). Mouse 1 and 4 are largely responsible for these predictive discrepancies, with Mouse 1 having inconsistent predictions for 75% of protocols, and Mouse 4 having inconsistent predictions for 90% of protocols. Note that each methodology must agree for the protocol VDVDVD, as this was the experimental protocol that was used for parameter fitting. So, 95% is the maximum disagreement rate we can see across methodologies for a given mouse. We observe that the QMC-associated parameter set is much more likely to predict treatment failure for these mice, whereas the NLME parameter set is more



**FIGURE 5** | Schematic representation of a bifurcation diagram in two-dimensional parameter space. For certain nonlinear combinations of parameters, a treatment can successfully eradicate a tumor (as occurs for Mouse 8 treated with VVDDD according to NLME parameters), result in tumor stabilization (as occurs for Mouse 6 treated with VVDDD according to NLME parameters), or can fail to control the tumor (as occurs for Mouse 5 treated with VVDDD according to NLME parameters). Note the bifurcation diagram is dependent on both the dose of drug being given, and the ordering of those drugs.

likely to predict treatment success. Contrary to the mice for which there is significant cross-methodology agreement, we see a high dependency of treatment response to dosing order for Mouse 1 and 4. From the perspective of the high dimensional bifurcation diagram, these parameters must fall sufficiently close to the bifurcation surface so that parametric changes that result from using different fitting methodologies can lead to wildly different predictions about treatment response (see schematic in **Figure 5**).

In turn, this appears to make these mice significantly more sensitive to dosing order.

Though the results in this paper are presented for one best-fit parameter set per methodology, we have also explored how parametric uncertainty influences treatment predictions. In particular, for the QMC fitting method, for each mouse we identified suboptimal parameter sets by performing Sobol sampling in a 10% range about the optimal parameter set. Any sampled parameter set that gives a goodness-of-fit within 10% of the optimal is considered a suboptimal parameter set (see **Figure S4**). For all such suboptimal parameter sets, treatment response to the 20 protocols was determined. This allows us to study if binary treatment response is insensitive to the precise best-fit parameters used. In **Figure S5** we show the probability a treatment is effective for each mouse across all suboptimal parameter sets. Overwhelmingly, treatment response predicted for an individual mouse and protocol shows excellent agreement across suboptimal parameter sets. Besides treatment response to the protocols VDVVDD and VDVDVD for Mouse 7, predicted treatment response across suboptimal parameter sets agrees over a minimum of 95% of the suboptimal parameter sets. This is seen in **Figure S6** by the probabilities of an effective treatment being either  $>0.95$  or  $<0.05$ . As small parametric perturbations that result in “good” fits to the data do not significantly influence predicted treatment response, we conclude it is reasonable to compare the prediction across methodologies using only the best-fit parameters.

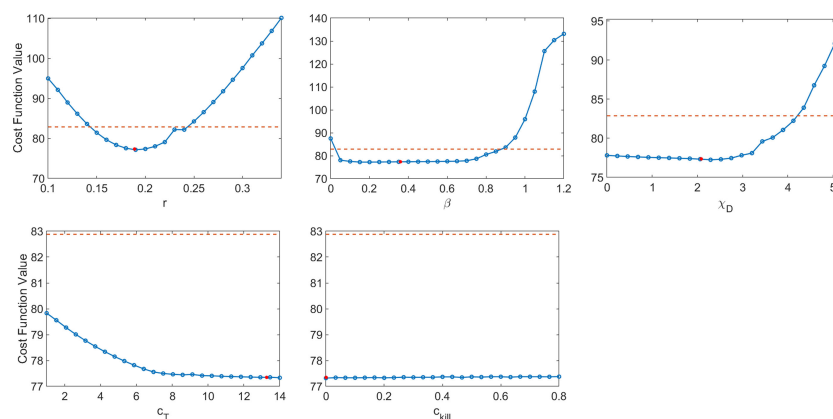
### 3.3 Exploring Predictive Discrepancies Between Fitting Methodologies

The predictive discrepancies across fitting methodologies begs the question of whether the parameters we are fitting are actually practically identifiable given the available experimental data. To explore this question, we generated profile likelihood curves for fitting the *average* tumor growth data, following the

methodology detailed in Section 2. As a first step, we fixed the parameters whose values we could reasonably approximate from experimental data:  $\delta_I = 1$ ,  $\alpha = 3000$ ,  $\delta_V = 2.3$ ,  $\kappa_0 = 2$ ,  $\delta_T = 0.35$ , and  $\delta_D = 0.35$  (37). This means we are using  $df = 5$  in the calculation of the threshold, as the generation of each profile likelihood curve requires fitting four model parameters plus the initial condition  $U(0)$ .

The resulting profile likelihood curves in **Figure 6** show that, even under the assumption that six of the eleven non-initial condition parameters are known, several of the fit model parameters lack practical identifiability. The tumor growth rate  $r$  and the infectivity parameter  $\beta$  are both practically identifiable, ignoring slight numerical noise. The T cell activation parameters  $\chi_D$  and  $c_T$  lack practical identifiability as they have profiles with a shallow and one-sided minimum (44). The profile for  $c_{kill}$  demonstrates that the model can equally well-describe the data over a large range of values for this enhanced cytotoxicity parameter. The flat likelihood profile is indicative of (local) structural unidentifiability, which also results in the parameter being practically unidentifiable (44). It is worth noting that the original work fitting to the average mouse was done in a *hierarchical* fashion (35, 37), and this circumvented the identifiability issues that emerge when doing simultaneous parameter fitting.

As we are unable to exploit the benefits of hierarchical fitting when performing personalized fits, this lack of practical identifiability poses significant issues for treatment personalization. We have already seen the consequences of this when we observed that despite both giving good fits to the data, QMC and NLME make consistent qualitative predictions in only 73.75% of the treatment protocols tested across all individuals. While the lack of practical identifiability helps explain why this can happen, it does not explain the mechanisms that drive predictive differences. To this end, we will now focus on the simulated dynamics of Mouse 4 in more detail, as this was the mouse with the most predictive discrepancies across methodologies.



**FIGURE 6** | Profile likelihood curves. Top row: tumor growth rate  $r$ , infectivity rate  $\beta$ , T cell activation rate by DCs  $\chi_D$ . Bottom row: T cell stimulation rate by immunostimulants  $c_T$ , and rate at which immunostimulants enhance cytotoxicity of T cells  $c_{kill}$ . The threshold (red dashed line) is calculated using  $df = 5$  and a 95% confidence interval.



As shown in **Figure 7**, when we simulate the model ten days beyond the data-collection window, we see that the QMC and NLME parameters fall on different sides of the bifurcation surface. In particular, in the QMC-associated simulation, at around 34 days the tumor exhibits a local maximum in volume and continues to shrink from there (**Figure 7**, left). This is in comparison to the NLME-associated simulation, where the tumor grows exponentially beyond the data-collection window. To uncover the biological mechanism driving these extreme differences, we look at the “hidden” variables in our model – that is, variables for which we have no experimental data. As shown in **Figure 7**, despite the similar fits to the volumetric data, the two parameters sets predict drastically different dynamics for the OV's and T cells. For the NLME-associated parameters, the virus and T cell population die out, eventually resulting in unbounded tumor growth. On the other hand, the virus and T cell population remain endemic throughout the simulation when using the QMC-associated parameters, driving the tumor population towards extinction.

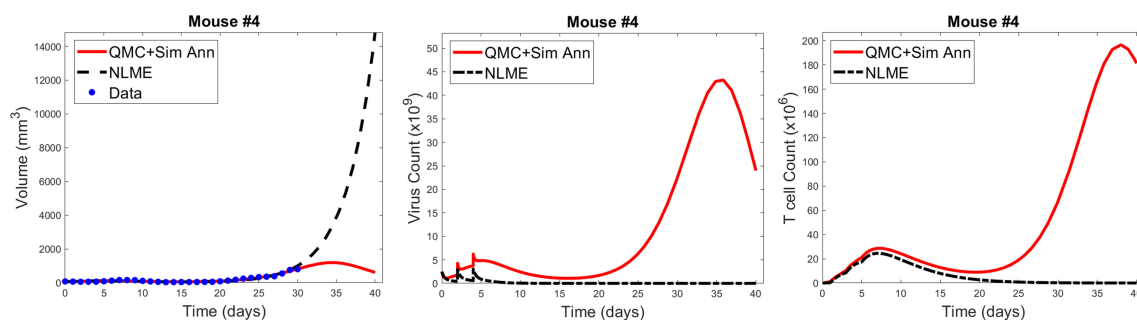
It is common knowledge that more data improves parameter identifiability. Not all data is created equal, however. We could get a lot more time-course data on total tumor volume over the 30-day window, but that would not necessarily improve parameter identifiability. Instead, we have identified that the addition of a single data point, for the right variable, at the right time, could go a long way in reconciling predictive discrepancies across fitting methodologies. To make this concrete, suppose we had data that, for Mouse 4, no tumor-targeting T cells are detected at 30 days. If we introduced a modified cost function that weighed both the contribution of the tumor volume and this T cell measurement, the parameter set identified by QMC would no longer be optimal, as it predicts a T cell burden on the order of  $10^8$  ( $100 \times 10^6$ ). The optimal parameter set should be one for which  $T \rightarrow 0$ , and once this occurs, there is no mechanism to control the tumor in the long-term. As a result, the tumor must regrow, just as predicted for the NLME-associated parameters. While this thought experiment does not suggest all practical identifiability issues would be reconciled by having this one data point, it does indicate why the predictive discrepancies we see for Mouse 4 (and also Mouse 1) would be at least partially resolved by the addition

of a single data point on tumor-targeting T cell volume. This highlights that although one must be quite cautious in using mathematical models to make personalized predictions, models can help us determine precisely what additional data is needed so that we can have more trust in our mathematical predictions.

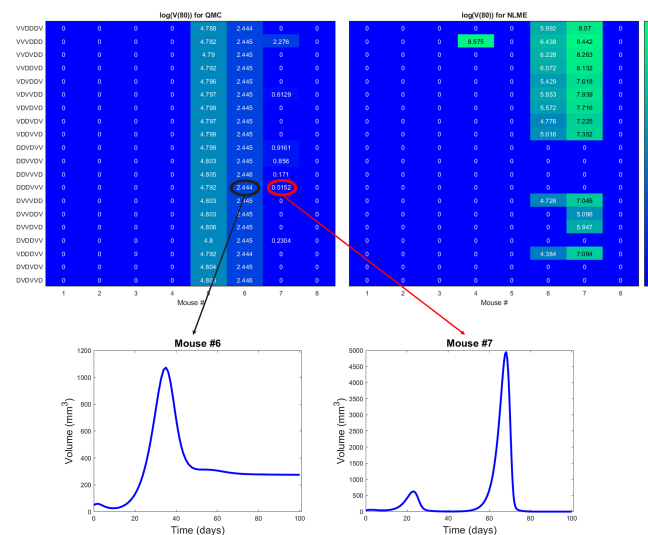
### 3.4 Personalized Treatment Response to the Optimal for Average Protocol

Ideally, when an optimal prediction is made for the average of a population, that optimal treatment protocol would also well-control the tumors of individual patients in the population. However, it is well known and supported by our earlier work with virtual populations that this is not necessarily the case. In (32) we showed that the experimental dose being considered herein is *fragile or non-robust*. We define a dosing regime as robust if virtual populations that deviate somewhat from the average population have the same qualitative response to the optimal-for-the-average protocol. Otherwise, a protocol is called fragile. The ability to classify fragility/robustness relies on the generation of a virtual population cohort that mimics a broad spectrum of individuals with different disease dynamics (31–33). By determining treatment response for each individual in the virtual cohort, we arrive at a statistic describing the likelihood the considered treatment is effective across heterogeneous individuals in the virtual population. We previously classified the optimal-for-the-average protocol of VVVDDD as fragile because this protocol eradicates the average tumor (37), yet only 30% of individuals in our virtual cohort were successfully eradicated by this treatment (32). Importantly, fragility is a probabilistic population-level descriptor, and not an individual descriptor. While it tells us that populations that deviate somewhat from the average are less likely to behave like the average, it tells us nothing about individuals, particularly if the individuals have behavior that deviates significantly from the average (which is often the case, as shown in **Figure 1**). Though, it seems natural to hypothesize that in such a fragile region we may have to be more careful about applying a prediction for the average of a population to any one individual in that population.

We will explore that hypothesis here by looking at statistics on how individual mice respond to VVVDDD, the predicted optimal



**FIGURE 7** | Left: QMC and NLME-associated fits to Mouse 4 treated with VDVDD, with model predictions extended 10 days beyond the data-collection window. Center and right: Predicted virus and T cell counts associated with each fitting methodology, respectively.



**FIGURE 8** | Heatmaps showing the log of the tumor volume measured at 80 days, at the high DC (50% greater than experimental dose), low OV (50% lower than experimental dose) region of dosing space. Left shows predictions if parameters are fit using QMC, and right shows NLME predictions. Inserts show time course of predicted treatment response for Mouse 6 and 7 to the optimal-for-the-average protocol of DDDVVV.

treatment protocol for the average mouse. While this protocol was effectively able to eradicate the average tumor in the population, its success across individual mice varies significantly across fitting methodologies. For the QMC-associated predictions, this protocol eradicates tumors in 75% of the individual mice (second row of the heatmaps in **Figure 4**, left). Compare this to the NLME-associated predictions, in which this protocol eradicates tumors in only 25% of the individual mice (second row of the heatmaps in **Figure 4**, right). As shown in **Figure S3**, this prediction is unchanged if we determine treatment success or failure at day 80 instead of day 30.

We can also compare response to the optimal-for-the-average protocol across methodologies. We see a qualitative agreement across methodologies (eradication or treatment failure) in only 50% of the mice (Mouse 2, 3, 5, 6). Mouse 7 is particularly interesting, as there was 95% agreement across methodologies when using  $V(30)$  to measure treatment success or failure, and the optimal for the average of VVVDDD is the only protocol for which treatment response differed (with QMC predicting tumor eradication, and NLME predicting treatment failure). As a further sign of caution, notice how for Mouse 1 and 4 (the cases with significant predictive discrepancies across methodologies), and Mouse 8 (intermediate case with 25% predictive discrepancies), VVVDDD eradicates the tumor with the QMC-associated parameters yet is the worst protocol that could be given (largest  $\log(V(30))$ ) for the NLME-associated parameters. This is particularly unsettling as it means the population-level optimal treatment recommendation could be the worst protocol for some individuals. This confirms our hypothesis that a population-level prediction should be applied to individuals very cautiously when in a fragile region of dosing space.

This raises the question: what if we were assessing individualized response to a protocol in a robust region of dosing space, wherein

treatment response across individuals in a virtual population is statistically similar to the treatment response in the population average? In (32), we previously classified the optimal-for-the-average protocol of DDDVVV as robust in the high DC (50% greater than experimental dose), low OV (50% lower than experimental dose) region of dosing space. It was classified as robust because this protocol eradicates the average tumor, and it also eradicates 84% of the individuals in our virtual cohort (32). This probabilistic population-level assessment of robustness naturally leads to the hypothesis that in a robust region of dosing space, we may have more success with the optimal-for-the-average treatment in individual mice. We will explore that hypothesis here.

The robust population-level optimal of DDDVVV yields a successful treatment response in all eight mice for the NLME-associated parameters. This holds whether we use  $V(30)$ , our original measure for establishing treatment success (as shown in **Figure S6**), or if we use  $V(80)$  as shown in **Figure 8**. This is consistent with the robust nature of this region of dosing space, as the NLME-associated parameters are less likely to wildly deviate from the average mouse due to population-level distributions constraining the value of these parameters. In comparison, the QMC-associated predictions show that only 62.5% of the individual mice are successfully treated by the optimal for the average in an 80-day window (**Figure 8**, top left). That said, if we look at the data more closely, we can see that Mouse 7 has essentially been eradicated even though 80 days was not quite long enough to drive  $V(80) < 1 \text{ mm}^3$ , our threshold for eradication. **Figure 8** also shows that the tumor volume for Mouse 6 has stabilized. Thus, we see that the QMC-associated predictions actually agree with the optimal-for-the-average response in 75% of cases (or, 87.5% if you consider the stabilization of Mouse 6 to be a “success” rather than a “failure”).

In closing, we have confirmation of our hypothesis that there is a significant benefit to working with a robust optimal-for-the-average protocol, even in the absence of all model parameters being practically identifiable. In the presence of robustness, we predict that one could generally apply the optimal-for-the-average protocol and expect a qualitatively similar response in most individuals. While this does not mean each individual is treated with their personalized optimal protocol, this has important consequences for determining when a population-level prediction will be effective in an individual.

## 4 DISCUSSION

In this work, we demonstrated that computational challenges can arise when using individualized model fits to make treatment recommendations. In particular, we showed that treatment response can be sensitive to the fitting methodology utilized when lacking sufficient patient-specific data. We found that for our model and preclinical dataset, predictive discrepancies can be at least somewhat explained by the lack of practical identifiability of model parameters. This can result in the dangerous scenario where an effective treatment recommendation according to one fitting methodology is predicted to be the worst treatment option according to a different fitting methodology. This raises concerns regarding the utility of mathematical models in personalized oncology when individual data is limited.

While it is well-established that more data improves parameter identifiability, here we highlight how we can identify precisely what data would improve the reliability of model predictions. In particular, we see how having a single additional measurement on the viral load or T cell count at the end of the data collection window would go a long way to reducing the predictive discrepancies across fitting methodologies (**Figure 7**). While the full benefits of this observation are not realized in a retrospective study, they could be realized in a scenario where data collection and modeling are occurring simultaneously. In this scenario, an experimentalist could collect data on a small number of individuals (like the eight mice shown in **Figure 1**). A mathematical model validated against this data can be used to identify any predictive challenges that emerge within this dataset, and what data would be needed to overcome these predictive challenges. This would inform the experimentalist of what data to collect in the next cohort of individuals in order to have more confidence in personalized treatment predictions.

When additional data is not available, an alternative option to personalization is simply treating with the population-level optimal. Here we showed the dangers of applying the optimal-for-the-average for a fragile protocol, and we demonstrated that such a one-size-fits all approach is much safer to employ for a robust optimal protocol. Therefore, even when data is lacking to make personalized predictions, establishing the robustness of treatment response can be a powerful tool in predictive oncology.

It is of note that this study uses just one mathematical model, with one set of assumptions, to reach our cautionary conclusion regarding the fitting methodology utilized and the resulting biological predictions. And this model is quite a simple one,

ignoring many aspects of the immune system, and spatial aspects of immune infiltration (as done in (50), among many other references). The model used herein was chosen because it has been previously validated against the average of the available experimental data. A more complex model would be problematic here, as there is simply not the associated experimental data to validate such a model. While this study certainly does not guarantee that similar issues will arise when working with other models and datasets, it highlights the need for caution when using personalized fits to draw meaningful biological conclusions.

As we enter the era of healthcare where personalized medicine becomes a more common approach to treating cancer patients, harnessing the power of mathematical models will only become more essential. Understanding the identifiability of model parameters, what data is needed to achieve identifiability and/or predictive confidence, and whether treatment response is robust or fragile are all important considerations that can greatly improve the reliability of personalized predictions made from mathematical models.

## DATA AVAILABILITY STATEMENT

Data and code availability requests can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

All authors substantially contributed to the analysis of the data in this work, and the drafting of the manuscript. Each author also provides approval for publication of the content.

## FUNDING

JG and ML acknowledge use of the ELSA high-performance computing cluster at The College of New Jersey for conducting the research reported in this paper. This cluster is funded in part by the National Science Foundation under grant numbers OAC-1826915 and OAC-1828163.

## ACKNOWLEDGMENTS

JG would like to thank Dr. Joanna Wares and Dr. Eduardo Sontag for the many discussions that helped to develop the ideas in this manuscript. The authors are also grateful to Dr. Chae-Ok Yun for the exposure and access she provided to the rich dataset utilized in this work. EN would like to acknowledge Dr. Yang Kuang for his support at the early stages of this project.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.793908/full#supplementary-material>

## REFERENCES

- Deisboeck TS. Personalizing Medicine: A Systems Biology Perspective. *Molec Sys Biol* (2009) 5:249. doi: 10.1038/msb.2009.8
- Agur Z, Elishmereni M, Kheifetz Y. Personalizing Oncology Treatments by Predicting Drug Efficacy, Side-Effects, and Improved Therapy: Mathematics, Statistics, and Their Integration. *WIREs Syst Biol Med* (2014) 6:239–53. doi: 10.1002/wsbm.1263
- Barbolosi D, Ciccolini J, Lacarelle B, Barlési F, André N. Computational Oncology - Mathematical Modelling of Drug Regimens for Precision Medicine. *Nat Rev Clin Oncol* (2016) 13:242–54. doi: 10.1038/nrclinonc.2015.204
- Malaney P, Nicosia S, Davé V. One Mouse, One Patient Paradigm: New Avatars of Personalized Cancer Therapy. *Cancer Lett* (2014) 344:1–12. doi: 10.1016/j.canlet.2013.10.010
- Bryne A, Alferez D, Amant Fea. Interrogating Open Issues in Cancer Precision Medicine With Patient-Derived Xenografts. *Nat Rev Cancer* (2017) 17:254–68. doi: 10.1038/nrc.2016.140
- Engels F, Loos W, van der Bol J, de Bruijn P, Mathijssen R, Verweij J, et al. Therapeutic Drug Monitoring for the Individualization of Docetaxel Dosing: A Randomized Pharmacokinetic Study. *Clin Cancer Res* (2011) 17:353–62. doi: 10.1158/1078-0432.CCR-10-1636
- Lorenzo G, Scott MA, Tew K, Hughes TJR, Zhange YJ, Liu L, et al. Tissue-Scale, Personalized Modeling and Simulation of Prostate Cancer Growth. *Proc Natl Acad Sci* (2016) 113:E7663–71. doi: 10.1073/pnas.1615791113
- Walko C, McLeod H. Pharmacogenomic Progress in Individualized Dosing of Key Drugs for Cancer Patients. *Nat Clin Pract Oncol* (2009) 6:153–62. doi: 10.1038/ncponc1303
- Noble S, Sherer E, Hammemann R, Ramkrishna D, Vik T, Rundell A. Using Adaptive Model Predictive Control to Customize Maintenance Therapy Chemotherapeutic Dosing for Childhood Acute Lymphoblastic Leukemia. *J Theor Biol* (2010) 264:990–1002. doi: 10.1016/j.jtbi.2010.01.031
- Patel J. Personalizing Chemotherapy Dosing Using Pharmacological Methods. *Cancer Chemother Pharmacol* (2015) 76:879–96. doi: 10.1007/s00280-015-2849-x
- Chantal P, Hopkins B, Prandi D, Shaw R, Fedrizzi T, Sboner A, et al. Personalized *In Vitro* and *In Vivo* Cancer Models to Guide Precision Medicine. *Cancer Discov* (2017) 7:462–77. doi: 10.1158/2159-8290.CD-16-1154
- Ree A, Redalen K. Personalized Radiotherapy: Concepts, Biomarkers Andtrial Design. *Br J Radiol* (2015) 88:20150009. doi: 10.1259/bjr.20150009
- Caudell J, Torres-Roca J, Gillies R, Enderling H, Kim S, Rishi A, et al. The Future of Personalised Radiotherapy for Head and Neck Cancer. *Lancet Oncol* (2017) 18:e266–73. doi: 10.1016/S1470-2045(17)30252-8
- Sunasee E, Tan D, Ji N, Brady R, Moros E, Caudell J, et al. Proliferation Saturation Index in an Adaptive Bayesian Approach to Predict Patient-Specific Radiotherapy Responses. *Int J Radiat Biol* (2019) 95:1421–6. doi: 10.1080/09553002.2019.1589013
- Kim E, Herbst R, Wistuba I, Lee J, Blumenschein G, Tsao A, et al. The BATTLE Trial: Personalizing Therapy for Lung Cancer. *Cancer Discov* (2011) 1:44–53. doi: 10.1158/2159-8274.CD-10-0010
- Tsimberidou AM, Iskander N, Hong D, Wheler J, Falchook G, Fu S, et al. Personalized Medicine in a Phase I Clinical Trials Program: The MD Anderson Cancer Center Initiative. *Clin Cancer Res* (2012) 18:6373–83. doi: 10.1158/1078-0432.CCR-12-1627
- Zhang J, Cunningham J, Brown J, Gatenby R. Integrating Evolutionary Dynamics Into Treatment of Metastatic Castrate-Resistant Prostate Cancer. *Nat Commun* (2017) 8:1816. doi: 10.1038/s41467-017-01968-5
- Agur Z, Halevi-Tobias K, Kogan Y, Shlagman O. Employing Dynamical Computational Models for Personalizing Cancer Immunotherapy. *Expert Opin Biol Ther* (2016) 16:1373–85. doi: 10.1080/14712598.2016.1223622
- Agur Z, Vuk-Pavlović S. Mathematical Modeling in Immunotherapy of Cancer: Personalizing Clinical Trials. *Molec Ther* (2012) 20:1–2. doi: 10.1038/mt.2011.272
- Kogan Y, Forýs U, Shukron O, Kronik N, Agur Z. Cellular Immunotherapy for High Grade Gliomas: Mathematical Analysis Deriving Efficacious Infusion Rates Based on Patient Requirements. *SIAM J Appl Math* (2010) 70:1953–76. doi: 10.1137/08073740X
- Hawkins-Daruud A, Johnston S, Swanson K. Quantifying Uncertainty and Robustness in a Biomathematical Model-Based Patient-Specific Response Metric for Glioblastoma. *JCO Clin Cancer Inform* (2019) 3:1–8. doi: 10.1200/CCL18.00066
- Vainas O, Ariad S, Amir O, Mermershtain W, Vainstein V, Kleiman M, et al. Personalising Docetaxel and G-CSf Schedules in Cancer Patients by a Clinically Validated Computational Model. *Br J Cancer* (2012) 107:814–22. doi: 10.1038/bjc.2012.316
- El-Madani M, Hénin E, Lefort T, Tod M, Freyer G, Cassier P, et al. Multiparameter Phase I Trials: A Tool for Model-Based Development of Targeted Agent Combinations—Example of Evesor Trial. *Future Oncol* (2015) 11:1511–8. doi: 10.2217/fon.15.49
- Zahid MU, Mohsin N, Mohamed AS, Caudell JJ, Harrison LB, Fuller CD, et al. Forecasting Individual Patient Response to Radiotherapy in Head and Neck Cancer With a Dynamic Carrying Capacity Model. *Int J Radiat OncologyBiologyPhysics* (2021) 111(3):693–704 doi: 10.1016/j.ijrobp.2021.05.132
- Kronik N, Kogan Y, Elishmereni M, Halevi-Tobias K, Vuk-Pavlović S, Agur Z. Predicting Outcomes of Prostate Cancer Immunotherapy by Personalized Mathematical Models. *PloS One* (2010) 5:e15482. doi: 10.1371/journal.pone.0015482
- Hirata Y, Morino K, Akakura K, Higano CS, Bruchovsky N, Gambol T, et al. Intermittent Androgen Suppression: Estimating Parameters for Individual Patients Based on Initial Psa Data in Response to Androgen Deprivation Therapy. *PloS One* (2015) 10:e0130372. doi: 10.1371/journal.pone.0130372
- Hirata Y, Morino K, Akakura K, Higano CS, Aihara K. Personalizing Androgen Suppression for Prostate Cancer Using Mathematical Modeling. *Sci Rep* (2018) 8:2563. doi: 10.1038/s41598-018-20788-1
- Kogan Y, Halevi-Tobias K, Elishmereni M, Vuk-Pavlović S, Agur Z. Reconsidering the Paradigm of Cancer Immunotherapy by Computationally Aided Real-Time Personalization. *Cancer Res* (2012) 72:2218–27. doi: 10.1158/0008-5472.CAN-11-4166
- Elishmereni M, Kheifetz Y, Shukrun I, Bevan GH, Nandy D, McKenzie KM, et al. Predicting Time to Castration Resistance in Hormone Sensitive Prostate Cancer by a Personalization Algorithm Based on a Mechanistic Model Integrating Patient Data. *Prostate* (2016) 76:48–57. doi: 10.1002/pros.23099
- Gatenby R, Silva A, Gillies R, Frieden B. Adaptive Therapy. *Cancer Res* (2009) 69:4894–903. doi: 10.1158/0008-5472.CAN-08-3658
- Allen R, Rieger T, Musante C. Efficient Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models. *CPT Pharmacometrics Syst Pharmacol* (2016) 5:140–6. doi: 10.1002/psp4.12063
- Barish S, Ochs M, Sontag E, Gevertz J. Evaluating Optimal Therapy Robustness by Virtual Expansion of a Sample Population, With a Case Study in Cancer Immunotherapy. *Proc Natl Acad Sci* (2017) 114:E6277–86. doi: 10.1073/pnas.1703355114
- Cassidy T, Craig M. Determinants of Combination Gm-Csf Immunotherapy and Oncolytic Virotherapy Success Identified Through *In Silico* Treatment Personalization. *PloS Comput Biol* (2019) 15:1–16. doi: 10.1371/journal.pcbi.1007495
- Huang JH, Zhang SN, Choi KJ, Choi IK, Kim JH, Lee M, et al. Therapeutic and Tumor-Specific Immunity Induced by Combination of Dendritic Cells and Oncolytic Adenovirus Expressing IL-12 and 4-1BBL. *Mol Ther* (2010) 18:264–274. doi: 10.1038/mt.2009.205
- Gevertz J, Wares J. Developing a Minimally Structured Model of Cancer Treatment With Oncolytic Viruses and Dendritic Cell Injections. *Comp Math Meth Med* (2018) 2018:8760371. doi: 10.1155/2018/8760371
- Kim P, Crivelli J, Choi IK, Yun CO, Wares J. Quantitative Impact of Immunomodulation Versus Oncolysis With Cytokine-Expressing Virus Therapeutics. *Math Biosci Eng* (2015) 12:841–58. doi: 10.3934/mbe.2015.12.841
- Wares J, Crivelli J, Yun CO, Choi IK, Gevertz J, Kim P. Treatment Strategies for Combining Immunostimulatory Oncolytic Virus Therapeutics With Dendritic Cell Injections. *Math Biosci Eng* (2015) 12:1237–56. doi: 10.3934/mbe.2015.12.1237
- Kucherenko S, Albrecht D, Saltelli A. Exploring Multi-Dimensional Spaces: A Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques. *arXiv* (2015) 1505:02350. doi: 10.48550/arXiv.1505.02350
- Torquato S. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. New York: Springer-Verlag (2002).
- Olofsen E, Dinges D, Van Dongen H. Nonlinear Mixed-Effects Modeling: Individualization and Prediction. *Aviat Space Environ Med* (2004) 75:A134–140.
- Myung JI. Tutorial on Maximum Likelihood Estimation. *J Math Psychol* (2003) 47:90–100. doi: 10.1016/S0022-2496(02)00028-7



42. Lixoft. *Monolix* (2021). Available at: <https://lixoft.com/products/monolix/>.
43. Lixoft. *Monolix* (2021). Available at: <https://monolix.lixoft.com/data-and-models/censoreddata/>.
44. Eisenberg M, Harsh E. A Confidence Building Exercise in Data and Identifiability: Modeling Cancer Chemotherapy as a Case Study. *J Theor Biol* (2017) 431:63–78. doi: 10.1016/j.jtbi.2017.07.018
45. Venzon D, Moolgavkar S. A Method for Computing Profile-Likelihood Based Confidence Intervals. *Appl Stat* (1988) 37:87–94. doi: 10.2307/2347496
46. Murphy S, van der Vaart A. On Profile Likelihood. *J Am Stat Assoc* (2000) 95:449–85. doi: 10.1080/01621459.2000.10474219
47. Raue A, Maiwald K, Bachmann J, Schilling M, Klingmüller U, Timmer J. 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
48. Sivia D, Skilling J. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press (2006).
49. Maiwald T, Hass H, Steiert B, Vanlier J, Engesser R, Raue A, et al. Driving the Model to its Limit: Profile Likelihood Based Model Reduction. *PLoS One* (2016) 11:1–18. doi: 10.1371/journal.pone.0162366
50. Zhang S, Gong C, Ruiz-Martinez A, Wang H, Davis-Marcisak E, Deshpande A, et al. Integrating Single Cell Sequencing With a Spatial Quantitative Systems Pharmacology Model Spqsp for Personalized Prediction of Triple-

Negative Breast Cancer Immunotherapy Response. *ImmunoInformatics* (2021) 1–2: 100002. doi: 10.1016/j.immuno.2021.100002

**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.

Copyright © 2022 Luo, Nikolopoulou and Gevertz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# 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](http://www.frontiersin.org)

Contact us: [frontiersin.org/about/contact](http://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