INTEGRATION OF MULTISOURCE HETEROGENOUS OMICS INFORMATION IN CANCER

EDITED BY: Victor Jin, Junbai Wang and Binhua Tang
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88963-448-4
DOI 10.3389/978-2-88963-448-4

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: researchtopics@frontiersin.org

INTEGRATION OF MULTISOURCE HETEROGENOUS OMICS INFORMATION IN CANCER

Topic Editors:
Victor Jin, The University of Texas Health Science Center at San Antonio, United States
Junbai Wang, Oslo University Hospital, Norway
Binhua Tang, Hohai University, China

Multisource heterogenous omics data can provide unprecedented perspectives and insights into cancer studies, but also pose great analytical problems for researchers due to the vast amount of data produced. This Research Topic aims to provide a forum for sharing ideas, tools and results among researchers from various computational cancer biology fields such as genetic/epigenetic and genome-wide studies.

Citation: Jin, V., Wang, J., Tang, B., eds. (2020). Integration of Multisource Heterogenous Omics Information in Cancer. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88963-448-4

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 December 2018
doi: 10.3389/fgene.2018.00611

Long Noncoding RNA FAM201A Mediates the Radiosensitivity of Esophageal Squamous Cell Cancer by Regulating ATM and mTOR Expression via miR-101

Mingqiu Chen[1,2,3†], Pingping Liu[4†], Yuangui Chen[5†], Zhiwei Chen[6], Minmin Shen[4], Xiaohong Liu[4], Xiqing Li[4], Anchuan Li[5], Yu Lin[7], Rongqiang Yang[8], Wei Ni[8], Xin Zhou[8], Lurong Zhang[7], Ye Tian[2,3], Jiancheng Li[7] and Junqiang Chen[7*]*

[1] Department of Radiation Oncology, Fujian Medical University Union Hospital and Fujian Provincial Platform for Medical Laboratory Research of First Affiliated Hospital, Fujian, China, [2] Department of Radiation Oncology, The Second Affiliated Hospital of Soochow University, Suzhou, China, [3] Institute of Radiotherapy & Oncology, Soochow University, Suzhou, China, [4] Shengli Clinical Medical College, Fujian Medical University, Fuzhou, China, [5] Department of Radiation Oncology, Fujian Medical University Union Hospital, Fuzhou, China, [6] Fuzhou Center for Disease Control and Prevention, Fuzhou, China, [7] Department of Radiation Oncology, Fujian Cancer Hospital & Fujian Medical University Cancer Hospital, Fuzhou, China,, [8] Cancer and Genetics Research Complex, Department Molecular Genetics and Microbiology, College Medicine, University of Florida, Gainesville, FL, United States

OPEN ACCESS

Edited by:
Binhua Tang,
Hohai University, China

Reviewed by:
Shaoli Das,
National Institutes of Health (NIH), United States
Suman Ghosal,
National Institutes of Health (NIH), United States

***Correspondence:**
Jiancheng Li
jianchengli6@126.com
Junqiang Chen
junqiangc@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 13 August 2018
Accepted: 19 November 2018
Published: 05 December 2018

Citation:
Chen M, Liu P, Chen Y, Chen Z, Shen M, Liu X, Li X, Li A, Lin Y, Yang R, Ni W, Zhou X, Zhang L, Tian Y, Li J and Chen J (2018) Long Noncoding RNA FAM201A Mediates the Radiosensitivity of Esophageal Squamous Cell Cancer by Regulating ATM and mTOR Expression via miR-101. Front. Genet. 9:611. doi: 10.3389/fgene.2018.00611

Background: The aim of the present study was to identify the potential long non-coding (lnc.)-RNA and its associated molecular mechanisms involved in the regulation of the radiosensitivity of esophageal squamous cell cancer (ESCC) in order to assess whether it could be a biomarker for the prediction of the response to radiotherapy and prognosis in patients with ESCC.

Methods: Microarrays and bioinformatics analysis were utilized to screen the potential lncRNAs associated with radiosensitivity in radiosensitive (n = 3) and radioresistant (n = 3) ESCC tumor tissues. Reverse transcription-quantitative polymerase chain reaction (RT-qPCR) was performed in 35 ESCC tumor tissues (20 radiosensitive and 15 radioresistant tissues, respectively) to validate the lncRNA that contributed the most to the radiosensitivity of ESCC (named the candidate lncRNA). MTT, flow cytometry, and western blot assays were conducted to assess the effect of the candidate lncRNA on radiosensitivity *in vitro* in ECA109/ECA109R ESCC cells. A mouse xenograft model was established to confirm the function of the candidate lncRNA in the radiosensitivity of ESCC *in vivo*. The putative downstream target genes regulated by the candidate lncRNA were predicted using Starbase 2.0 software and the TargetScan database. The interactions between the candidate lncRNA and the putative downstream target genes were examined by Luciferase reporter assay, and were confirmed by PCR.

Results: A total of 113 aberrantly expressed lncRNAs were identified by microarray analysis, of which family with sequence similarity 201-member A (FAM201A) was identified as the lncRNA that contributed the most to the radiosensitivity of ESCC. FAM201A was upregulated in radioresistant ESCC tumor tissues and had a poorer short-term response to radiotherapy resulting in inferior overall survival. FAM201A

knockdown enhanced the radiosensitivity of ECA109/ECA109R cells by upregulating ataxia telangiectasia mutated (ATM) and mammalian target of rapamycin (mTOR) expression via the negative regulation of miR-101 expression. The mouse xenograft model demonstrated that FAM201A knockdown improved the radiosensitivity of ESCC.

Conclusion: The lncRNA FAM201A, which mediated the radiosensitivity of ESCC by regulating ATM and mTOR expression via miR-101 in the present study, may be a potential biomarker for predicting radiosensitivity and patient prognosis, and may be a therapeutic target for enhancing cancer radiosensitivity in ESCC.

Keywords: ATM, esophageal squamous cell carcinoma, FAM201A, long noncoding RNA, miR-101, mTOR, radiosensitivity

INTRODUCTION

Globally, esophageal cancer (EC) is one of the most common types of cancer, with the 7th highest incidence rate and 6th greatest rate of cancer-associated death (Bray et al., 2018). Surgery still plays an important role in the treatment of EC (Pennathur et al., 2013; Rustgi and El-Serag, 2014). However, due to the patients' physiological conditions, the tumor location or the tumor stage, only ∼25% of newly diagnosed patients are suitable for surgery (Short et al., 2017). For patients with unresectable EC, radiotherapy (RT) combined with chemotherapy is considered to be the optimal treatment (Sasaki and Kato, 2016).

However, predominantly because of local failure (Lloyd and Chang, 2014; Versteijne et al., 2014) which has been associated with intrinsic and/or acquired radioresistance (Chen X. et al., 2017), the survival rate in EC patients following RT is as low as 10–30% after 5 years (Cooper et al., 1999; Gwynne et al., 2011). Therefore, how to predict the radiosensitivity and resensitize patients is imperative in patients with EC treated with RT. Unfortunately, as the molecular mechanism of radioresistance, which is known to involve DNA repair proteins (Zafar et al., 2010), cell signal pathways (Dumont and Bischoff, 2012), angiogenesis (Francescone et al., 2011), cancer stem cells (Moncharmont et al., 2012), and autophagy (Chaachouay et al., 2011), is intricate and has not been elucidated thoroughly, there are currently no accurate biomarkers to predict radioresistance or therapeutic targets to enhance the radiosensitivity of EC.

Long non-coding RNAs (lncRNAs) are a new class of non-protein-coding transcripts that are longer than 200 nucleotides (Qi and Du, 2013). A number of previous studies have demonstrated that lncRNAs are important regulators of gene expression, that control both physiological and pathological processes in development and diseases such as cancer (Kung et al., 2013). Recent studies have reported that lncRNAs also function as regulators of tumor radiosensitivity and may serve as biomarkers for tumor response to RT (Spizzo et al., 2012; Yu et al., 2012). However, radiosensitivity-associated lncRNAs in esophageal squamous cell carcinoma (ESCC) are rarely reported (Tong et al., 2014; Zhang et al., 2015; Li et al., 2016; Zhou et al., 2016).

In the present study, we demonstrated that the lncRNA family with sequence similarity 201-member A (FAM201A) contributed the most to the radioresistance of ESCC. Furthermore, functional and mechanistic analyses revealed that FAM201A contributed to radioresistance by upregulating ataxia telangiectasia mutated (ATM) and mammalian target of rapamycin (mTOR) expression via actions as a miR-101 sponge. This study first established a FAM201A-miR-101-ATM/mTOR regulatory network in ESCC, revealing a promising therapeutic strategy for treating ESCC with radioresistance.

MATERIALS AND METHODS

Patients and Tissue Specimens

The present study was approved by the Fujian Medical University Union Hospital Institutional Review Board (No. 2014KY001). All of the patients signed informed consent prior to treatment, and all of the information was anonymized prior to its analysis. The pretreatment work-up and eligibility criteria, details of radiotherapy and chemotherapy, criteria for toxicity, and short-term response, follow-up and the statistical analysis of survival were presented in our previous study (Chen M. Q. et al., 2017).

Between July 2015 and March 2017, a total of 41 patients with ESCC who received RT were recruited. Tissue specimens obtained during pretreatment with esophagogastroduodenoscopy were histopathologically examined by two independent pathologists and were snap

Abbreviations: ATM, ataxia telangiectasia mutated; AUC, area under the curve; CASC2, cancer susceptibility candidate 2; CR, complete response; CTV, clinical target volume; DLEU2, deleted in lymphocytic leukemia 2; DLX6-AS1, DLX6 antisense RNA 1; DNA, deoxyribonucleic acid; DNA-PKcs, DNA-dependent protein kinase catalytic subunit; ECOG, Eastern Cooperative Oncology Group; ESCC, esophageal squamous cell carcinoma; FAM201A, family with sequence similarity 201-member A; GTV, gross tumor volume; HRR, homologous recombination repair; IC, Induction chemotherapy; lncRNA, long non-coding RNA; MCF2L-AS1, MCF2L antisense RNA 1. mTOR, mammalian target of rapamycin; MTT, 3-(4,5)-dimethylthiahiazo(-z-y1)-3,5-di-phenytetrazoliumromide; miR, microRNA; NHEJ, non-homologous end joining; OS, overall survival; PCR, polymerase chain reaction; PD, progression of disease; PR, partial response; RT-qPCR, reverse transcription-quantitative polymerase chain reaction; RI-DSB, radiation-induced double-strand breaks; RNA, ribonucleic acid; ROC, receiver operating characteristic; RT, radiotherapy; SD, stable disease; sh-RNA, short hairpin RNA; siRNA, small interfering RNA; TP, platinum compound plus taxane.

frozen in liquid nitrogen and then stored at −80°C until RNA extraction.

Tissue specimens were divided into a radiosensitive group ($n = 23$) and a radioresistant group ($n = 18$) based on short-term response to RT. The short-term responses to RT were classified as a clinically complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD) according to the Japanese Classification of Esophageal Cancer guidelines (Japan Esophageal Society, 2017). Of these, the CR and PR were termed radiosensitive group and the SD and PD were termed radioresistant group in the current study.

Microarray Screening and Bioinformatics Analysis

Microarray profiling was performed using three radiosensitive ESCC tumor tissues and three radioresistant ESCC tumor tissues. RNA extraction and sequential microarray hybridization were conducted by Biotechnology Company (Shanghai, China), and the detected human genome transcripts were obtained by the Human lncRNA array V6.0 (4x180 K; Agilent Technologies, Inc., Santa Clara, CA, USA). Bioinformatics analysis was performed using GeneSpring Software to obtain differentially expressed lncRNAs correlated with ESCC radiosensitivity.

Cell Lines and Culture

The ESCC cell line Eca109 was obtained from Chinese Academy of Sciences (Beijing, China). The corresponding radioresistant cells (Eca109R) were established from the parental cell line Eca109 by stepwise X-ray irradiation at 30 Gy in three fractions (10 Gy per fraction) (Da et al., 2017). Cells were cultured in RPMI-1640 medium (HyClone; GE Healthcare Life Sciences, Logan, UT, USA) with 10% (v/v) fetal bovine serum (Thermo Fisher Scientific, Inc., Waltham, MA, USA) and antibiotics (100 U/mL penicillin and 100 μg/mL streptomycin; HyClone) in an atmosphere of 95% air/ 5% CO_2 at 37°C.

RNA Isolation and Reverse Transcription-Quantitative Polymerase Chain Reaction (RT-qPCR)

Total RNAs from either tissue samples or cultured cells were extracted with TRIzol reagent (Thermo Fisher Scientific, Inc.) according to the manufacturer's instructions. The RNA concentration and quality were measured using a NanoDrop ND-2000 spectrophotometer which measured the absorbance at 260 and 280 nm. Samples with an A_{260}:A_{280} ratio ≥2.0 were selected for further analysis.

First strand cDNA for the potential lncRNAs and putative micro (mi)-RNA were synthesized using the PrimeScript™ RT reagent kit with gDNA Eraser (Takara Biotechnology, Co., Ltd., Dalian, China) according to the manufacturer's protocol. Briefly, 1 μg total RNA, 2 μl 5X gDNA Eraser Buffer, 1 μl gDNA Eraser and RNase Free dH_2O, were combined in a total reaction volume of 10 μl and incubated at 42°C for 2 min to eliminate the genomic DNA. A total of 10 μl of the RT reaction mixture (consisting of 4 μl 5X PrimeScript Buffer 2, 1 μl PrimeScript RT Enzyme Mix 1, 1 μl RT Primer Mix, and 4 μl RNase Free dH_2O) was then added, and the mixture was incubated at 37°C for 15 min, followed by 85°C for 5 s to generate the cDNA.

The expression of the potential lncRNAs in the radiosensitive tumor tissues, compared with the radioresistant tumor tissues, was quantified using SYBR® Premix Ex Taq (Takara Biotechnology Co., Ltd.) according to the manufacturer's instructions on the ABI 7500 Real-Time PCR System (Applied Biosystems; Thermo Fisher Scientific, Inc.). Briefly, the 20 μl reaction mixtures were incubated at 95°C for 30 s for the initial denaturation, followed by 40 cycles at 95°C for 5 s and 60°C for 34 s. The expression levels of lncRNAs were calculated using the ΔCt method, where $\Delta Ct = Ct_{target} - Ct_{reference}$, a smaller ΔCt value indicates a greater expression. The relative expression of lncRNAs was analyzed using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001); data was normalized to the endogenous control GAPDH. Each sample was examined in triplicate. The primers and oligonucleotides of the plasmid were synthesized by Invitrogen (Thermo Fisher Scientific, Inc.), the sequences are presented in **Table 1**. The aberrant lncRNA that had the greatest sensitivity and specificity for predicting ESCC radiosensitivity (in radiosensitive and radioresistant tissues), as identified by receiver operating characteristic (ROC) curves, and was associated with survival, was identified as the candidate lncRNA for further study.

Transient Transfection

Small interfering RNA (siRNA) specifically targeting candidate lncRNA (si-candidate-lncRNA) and putative-miRNA, negative control (NC) si-candidate-lncRNA and si-putative-miRNA, candidate-lncRNA mimic, putative-miRNA mimic, and the inhibitor control were constructed by Nanjing Dongji Biotechnology Company (Nanjing, China). Ectopic expression of the candidate lncRNA was achieved by introducing the candidate lncRNA sequence into a pcDNA3.1 vector (Thermo Fisher Scientific, Inc.). Eca109/Eca109R cells were seeded into 6-well plates at a density of 1×10^6 cells/well and cultured overnight prior to transfection. Then, transient transfection with oligonucleotides or plasmids into Eca109/Eca109R cells was performed using Lipofectamine 2000™ (Thermo Fisher Scientific, Inc.). Cells were harvested 48 h post-transfection for subsequent analysis. PCR was used to validate the efficacy of Eca109/Eca109R cell transfection with si-candidate-lncRNA and candidate-lncRNA-mimic.

Western Blot Analysis

Protein samples from tissues or cells were subjected to 10% SDS-PAGE and transferred to PVDF membranes. Following blocking in 5% skim milk for 2 h, the membranes were incubated overnight at 4°C with the primary antibodies against P-glycoprotein (P-gp; 1:1,000), glutathione S-transferase π (GST-π; 1:500), ATM (1:750), mTOR (1:1,000), and β-actin (1:5,000) purchased from Zen Bioscience Biotechnology, Inc. (Chengdu, China), followed by incubation with horseradish peroxidase-conjugated goat anti-rabbit secondary antibodies for 2 h (1:5,000). The antigen-antibody complexes were visualized using chemiluminescence.

TABLE 1 | The primer sequences used in reverse transcription-quantitative polymerase chain reaction.

Primers used for RT-qPCR	Forward (5′-3′)	Reverse (3′-5′)
FAM201A	TCTCTGATGGGAGCCTCTTTA	CAAGCCACAGACGGAGAAA
CASC2	GTCCGCATGGTAAGGAATCA	GACTGCGTTTATCAAGTCCAAAG
DLEU2	TGGCGCAGTCGGTTTAAT	TTCCTTGCAGTACACCTTTCA
DLX6-AS1	TCTCCTCCTACCTAGCATCTTC	CCTTTGAAGCTCCTACTCCTTT
MCF2L-AS1	TTGAGCCTGGGCAATGTAG	CTTCCTGCTGGAATTCTCTCTC
GAPDH	CAGGGCTGCTTTTAACTCTGGTAA	GGGTGGAATCATATTGGAACATGT
FAM201A mimic	GGGGTACCGAGTGCACCTGGCCTGAGAG	GGAAGCCTTTTGTGGTTAGATATTTGAAAT
OLIGONUCLEOTIDES OF PLASMID		
siFAM201A	GATCTTTCGTCCATTTACTtt	
NC-siFAM201A	GCCTTATTTCTATCTTACGtt	
FAM201A-cDNA	GTACCTCGATCTTTCGTCCATTTACTTCAAGAGA GTAAATGGACGAAAGATCTTTTTGGAAA	AGCTTTTCCAAAAAGATCTTTCGTCCATTTACTCT CTTGAAGTAAATGGACGAAAGATCGAG
NC-FAM201A-cDNA	GTACCTCGCCTTATTTCTATCTTACGTCAAGAGC GTAAGATAGAAATAAGGCTTTTTGGAAA	AGCTTTTCCAAAAAGCCTTATTTCTATCTTACGCT CTTGACGTAAGATAGAAATAAGGCGAG
miR-101	AAGUCAAUAGUGUCAUGACAU	
miR-590	GACGUGAAAAUACUUAUUCGAG	
Negative control	UUCUCCGAACGUGUCACGUUU	

Radiosensitivity Assay

Radiosensitivity was assessed by 4-5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) assay. ECA109/ECA109R cells (5,000/well) were incubated for 48 h prior to exposure to various doses of radiation (0 Gy, 2 Gy, 4 Gy, 6 Gy, 8 Gy, and 10 Gy). Subsequently, 10 μl of 5 mg/mL MTT was added to each well for a further 3 h, followed by the addition of 150 μl DMSO to dissolve the generated formazan crystals. The absorbance at a wavelength of 570 nm was detected using a microplate reader.

Flow Cytometry Analysis of Apoptosis

ECA109/ECA109R cells (5,000/well) were incubated for 48 h prior to exposure to various doses of radiation (0 Gy, 2 Gy, 4 Gy, 6 Gy, 8 Gy, and 10 Gy). The ratio of apoptotic cells was detected using an Annexin V-FITC Apoptosis Detection Kit (BD Bioscience, Franklin Lakes, NJ, USA) and analyzed using a BD Calibur flow cytometer with CellQuest software (BD Biosciences).

Candidate lncRNA Downstream Target Genes and Luciferase Reporter Assay

The potential target genes downstream of the candidate lncRNA were predicted using Starbase 2.0 software (http://starbase.sysu.edu.cn/starbase2/index.php) and the TargetScan (www.targetscan.org/vert_71/) database.

The full fragments of the candidate lncRNA or its mutant containing the putative miRNA-binding sites were synthesized and cloned downstream of the firefly luciferase gene in pGL3 plasmids (Promega Corporation, Madison, WI, USA), and were termed the pGL3-candidate lncRNA-wild type (Wt) and pGL3-candidate lncRNA-mutant (Mut). Eca109 and Eca109R cells were maintained in 96-well plates and co-transfected with 400 ng of the constructed luciferase reporter plasmids, 50 ng of *Renilla* luciferase reporter vector and 50 nM of the putative miRNA mimic, miR-con, or putative miRNA-vector using Lipofectamine 3000^{TM} (Thermo Fisher Scientific, Inc.). Cells were harvested at 48 h after transfection, and luciferase activity was determined using a Dual Luciferase Reporter Assay Kit (Promega Corporation). *Renilla* luciferase activities were used as the internal control for the normalization of firefly luciferase activity.

In vivo Experiments

The animal experiments were approved by the Animal Care and Use Committee of Fujian Medical University Union Hospital and were performed in accordance with the Institutional Guide for the Care And Use Of Laboratory Animals. Lentiviral vector [Lenti-short hairpin (sh)-candidate lncRNA] for stable silenced expression of the candidate lncRNA was obtained from Shanghai GenePharma Co., Ltd. (Shanghai, China) and transfected into Eca109/Eca109R cells. The success of transfection was detected by PCR and the survival of the cells was determined by an MTT assay. Then, equal numbers of siRNA-candidate lncRNA-transfected Eca109, NC and control cells were implanted into 8-week old nude mice (n = 5 per group; Model Animal Research Center of Nanjing University) by subcutaneous injection.

At two weeks after the injection (to allow for tumor growth), the tumors were irradiated by X-ray at 10 Gy. Tumor size was measured every 3 days with a caliper, and tumor volume was calculated according to the following formula: Volume = (length x width^2)/2. All mice were sacrificed on day 42 after inoculation. The resected tumor masses were harvested for subsequent weight measurements. A growth curve was constructed to determine tumor radiosensitivity and the effect of the siRNA of the candidate lncRNA on tumorigenicity in nude mice was analyzed.

Statistical Analysis

The overall survival data was analyzed using SPSS software 23.0 (IBM Corp., Armonk, NY, USA). Survival curves were established through the Kaplan-Meier method and compared by a log rank test.

A multivariable analysis of patient demographic and clinical parameters (gender, age, ECOG score, tumor location, clinical T and N stages, the radiotherapy doses for GTV and CTV, and the tumor response to treatment) was performed using the Cox proportional hazards model.

Experimental data are presented as $\bar{x} \pm s$ from independent experiments performed in triplicate. For comparisons, paired or independent Student's *t*-tests, Chi-square tests or ANOVA with *post hoc* tests (Tukey's) were performed. ROC curves were used for selecting an optimal cut-off point for each test and for comparing the accuracy of diagnostic tests. Two-tailed $P < 0.05$ ($^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$) was considered to indicate a statistically significant difference.

RESULTS

Patient Characteristics, Treatment Response, and Survival

Between July 2015 and March 2017, a total of 41 ESCC patients treated with RT combined with chemotherapy were enrolled in the present study. After RT, a total of 4 patients achieved CR, 19 patients reached PR, 9 patients maintained SD and 9 cases had PD. There were no significant differences between radiosensitive (4 CR and 19 PR) and radioresistant (9 SD and 9 PD) patients regarding the distributions of gender, age, ECOG score, tumor location, and clinical stage (**Table 2**).

Differential Expression of lncRNAs Potentially Correlated With Radiosensitivity

A total of 113 aberrantly expressed lncRNAs were identified in the microarray analysis using three radiosensitive ESCC tumor tissues and three radioresistant ESCC tumor tissues, of which 71 lncRNA transcripts were upregulated (fold change >2, $P < 0.05$) and 42 lncRNA transcripts were downregulated (fold change < 0.5, $P < 0.05$) in the radiosensitive ESCC tumor tissues when compared with the radioresistant ESCC tumor tissues. The lncRNAs CASC2, FAM201A, DLEU2, DLX6-AS1, and MCF2L-AS1 were considered to be the potential lncRNAs related to radiosensitivity when analyzed using GeneSpring Software 12.6 (Agilent Technologies, Inc.) (**Figure 1A**, **Supplementary File 1**).

Tumor tissues from the remaining 35 enrolled patients (20 radiosensitive patients and 15 radioresistant patients, respectively) were collected to detect the expression of the lncRNAs CASC2, FAM201A, DLEU2, DLX6-AS1, and MCF2L-AS1 by RT-qPCR. The results revealed that the differential expression of CASC2, FAM201A, and DLX6-AS1 between the radioresistant and radiosensitive groups were significantly different, while the difference in the DLEU2 and MCF2L-AS1 expressions were not significantly different when comparing the groups (**Figure 1B**; **Supplementary File 2**).

FAM201A Is a Novel lncRNA With a Potential Function in the Radiosensitivity and Survival of ESCC

Based on above data, the ROC curve of the lncRNAs CASC2, FAM201A, and DLX6-AS1 was applied to identify the lncRNA that was the most correlated to radiosensitivity and survival using the area under curve (AUC) were 0.783 (95%CI: 0.609–957, $P = 0.005$), 0.817 (95%CI: 0.673–960, $P = 0.002$), and 0.340 (95%CI: 0.150–530, $P = 0.110$); respectively. Compared with the lncRNA DLX6-AS1, FAM201A, and CASC2 yielded a superior AUC with specificity and sensitivity for distinguishing radiosensitive ESCC tumor tissues from radioresistant ESCC tumor tissues (**Figure 2A**).

CASC2 was associated with short-term response to RT but not with survival, while FAM201A was correlated with both the short-term response and survival (**Figures 2B,C**). This indicated that FAM201A, as opposed to CASC2, may be a suitable biomarker of ESCC treated with RT.

To analyze whether FAM201A functions as a biomarker for radiosensitivity and survival in ESCC or not, the maximum Youden index method (Fluss et al., 2005) was performed to establish the cutoff value of FAM201A in the ROC curve. A total of 22 patients were termed as FAM201A-low with an average ΔCt expression value of 6.155, whereas, the remaining 13 patients, named the FAM201A-high expression group, had an average ΔCt expression value of 8.437 (**Supplementary File 3**).

Compared with the FAM201A-low group, the FAM201A-high group exhibited a poorer short-term response to RT and lower survival time. However, neither high or low FAM201A expression was correlated with tumor stage, regardless of whether it was T or N stage (**Table 3**). Furthermore, univariate and multivariate analysis indicated that FAM201A was the only independent risk factor for survival (OR, 0.642; 95% CI, 0.4668–0.885; $P = 0.007$). These data suggested that FAM201A could be a robust molecular marker for predicting RT sensitivity and survival in patients with ESCC.

FAM201A Regulated Radiosensitivity *in vitro*

Based on the above results, the effects of FAM201A regulated radiosensitivity in ESCC cancer cells were further explored by performing an X-ray irradiation experiment using Eca109/Eca109R cells transfected with si-FAM201A and FAM201A-mimic (**Figures 3A,B**; **Supplementary File 4**).

The results revealed that the survival rates of both Eca109 and Eca109R cells decreased with the increasing X-ray irradiation dose, and the percentage of apoptotic cells in each line increased with the increasing X-ray irradiation dose (**Figures 3C,D**; **Supplementary File 4**). The decrease in survival was more pronounced with the increase in X-ray irradiation dose in ECA109 cells when compared with ECA109R cells, demonstrating that the Eca109R cells were more resistant to X-ray irradiation.

In Eca109 cells, when compared with the control cells, FAM201A-mimic exhibited a significant promotion in cell proliferation, while si-FAM201A exhibited a significant

TABLE 2 | Clinicopathological characteristics of the entire cohort of 41 patients with ESCC.

Characters	Radiosensitive	Radioresistant	Total	*p*
Gender				0.706
Male	17	15	32	
Female	6	3	9	
Age (range)	61 (47-70)	63 (47-70)	61 (47-70)	0.406
ECOG score				0.767
0	13	11	24	
1	10	7	17	
Tumor location				0.515
Cervical	4	3	7	
Upper	5	6	11	
Middle	11	8	19	
Lower	3	1	4	
T stage				0.112
2	1	1	2	
3	11	3	14	
4	11	14	25	
N stage				0.164
0	2	0	2	
1	14	8	22	
2	7	10	17	
M stage [a]				0.542
0	20	15	35	
1	3	3	6	
Clinical stage [b]				0.112
II	1	1	2	
III	11	3	14	
IV	11	14	25	
GTV (cGy, range)	6000 (4000–6600)	6000 (5040–6600)	6000 (4000–6600)	0.128
CTV (cGy, range)	5040 (4000–5040)	5040 (4500–5040)	5040 (4000–5040)	0.300
IC				0.574
None	7	8	15	
PF	5	2	7	
TL	1	0	1	
TP	10	8	18	

There were no significant differences between radiosensitive and radioresistant patients regarding the distributions of gender, age, ECOG score, tumor location and clinical stage. ECOG, eastern cooperative oncology group; GTV, gross tumor volume; CTV, clinical target volume; PF, platinum plus fluorouracil; TP, platinum plus taxane; a, M1 means Supraclavicular lymphatic node metastasis; IC, Induction chemotherapy; b, according to the 7th AJCC TNM staging system.

increase in proliferation inhibition, indicating that for Eca109 cells, upregulated FAM201A expression likely resulted in cell radioresistance to X-rays (**Figure 3E**; **Supplementary File 4**).

In Eca109R cells, when compared with the control cells, si-FAM201A exhibited a significant inhibition of cell proliferation, while FAM201A-mimic did not exhibit the increased cell proliferation that was observed in ECA109 cells, indicating that the expression level of FAM201A in Eca109R cells was already at a high level, and thus, further elevation of FAM201A expression was not possible to enhance its radioresistance. These results indicated that, whether in cases of intrinsic or acquired radioresistance, si-FAM201A may enhance ESCC cell radiosensitivity, which may therefore be a novel effective target strategy for sensitizing ESCC to radiotherapy (**Figure 3F**; **Supplementary File 4**).

FAM201A Knockdown Enhanced the Radiosensitivity of ESCC *in vivo*

To confirm the efficacy of si-FAM201A on radiosensitivity *in vivo*, a xenograft tumor mouse model was established. A total of 15 mice with similar weights and dates of birth were selected in the present study (male: female = 8:7). When compared with the control groups, FAM201A knockdown (sh-FAM201A) significantly blocked tumor growth (decreased tumor volume and weight), suggesting that the silenced FAM201A expression enhanced radiosensitivity, thereby confirming that

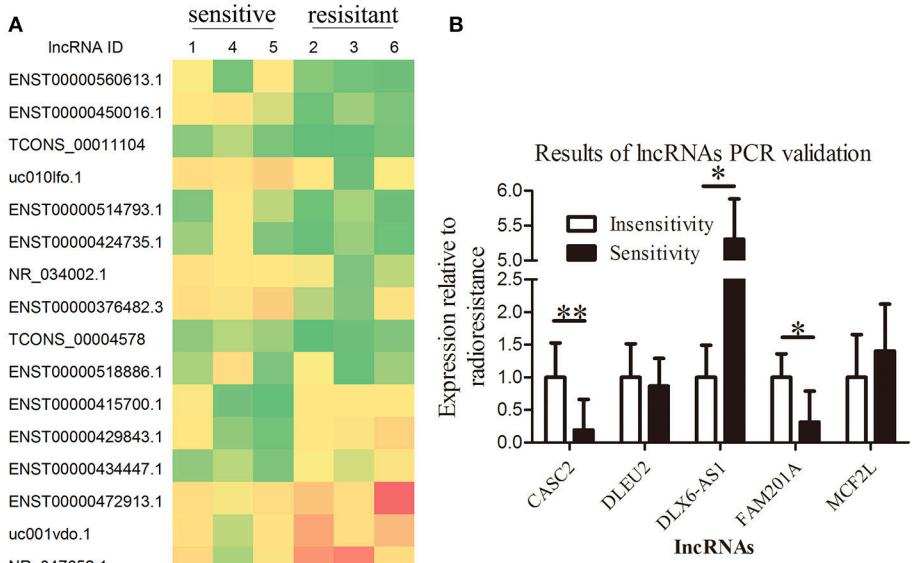

FIGURE 1 | Overexpression of lncRNA FAM201A is highly correlated with the radiosensitivity of ESCC and is associated with poor survival. **(A)** A heatmap presenting the gene expression levels in RNA samples isolated from three radiosensitive and three radioresistant ESCC tumor tissues by microarray assays. **(B)** Differential expression of the potential lncRNAs related to radiosensitivity (CASC2, FAM201A, DLEU2, DLX6-AS1, and MCF2L-AS1) in radiosensitive ($n = 20$) and radioresistant ($n = 15$) ESCC tumor tissues by reverse transcription-quantitative polymerase chain reaction. $^*P < 0.05$, $^{**}P < 0.01$.

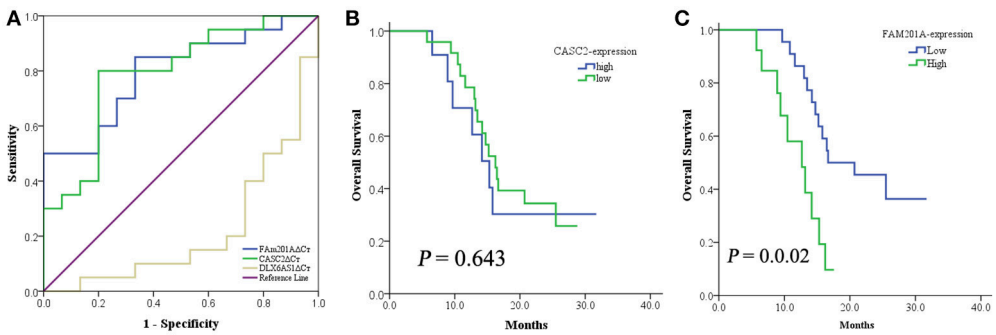

FIGURE 2 | (A) The ROC curve of lncRNA CASC2, FAM201A, and DLX6-AS1. When compared with the lncRNA DLX6-AS1, FAM201A, and CASC2 yielded a superior AUC with specificity and sensitivity for distinguishing radiosensitive ESCC tumor tissues from radioresistant ESCC tumor tissues. **(B)** The 1-year OS rate between patients with low- ($n = 24$) and high-expression ($n = 11$) of CASC2, was not different. **(C)** The 1-year OS rate between patients with low- ($n = 22$) and high-expression ($n = 13$) of FAM201A was significantly different ($P = 0.001$).

FAM201A could induce radiosensitivity *in vivo* (**Figures 3G,H**; **Supplementary File 4**).

FAM201A Negatively Regulated the Expression of miR-101

Using Starbase 2.0, miR-101 and miR-590 were predicted to have complementary base pairings with FAM201A. Accordingly, luciferase reporter vectors containing the Wt or a Mut FAM201A binding site were established and co-transfected with miR-101 into Eca109 cells. The same process was performed for miR-590.

The results demonstrated that the ectopic expression of miR-101 was markedly suppressed by co-transfection with the FAM201A mutant sequence in the Eca109 cell luciferase activity reporter assay. However, neither pGL3-FAM201A-Wt reporter nor pGL3-FAM201A-Mut transfection in Eca109 cells affected miR-590 expression (**Figures 4A,B**; **Supplementary File 5**).

TABLE 3 | Treatment results in the high and low FAM201A expression groups.

Variable	Low FAM201A expression, *n*	High FAM201A expression, *n*	Total *n*	*P*-value
T stage				0.161
2	0	2	2	
3	11	5	16	
4	11	6	17	
N stage				0.998
0	3	2	5	
1	10	6	16	
2	7	4	11	
3	2	1	3	
M stage				0.388
0	20	13	33	
1	2	0	2	
Tumor response, *n* (%)				0.001
CR	1	0	1	
PR	17	2	19	
SD	3	6	9	
PD	1	5	6	
Pattern of failure, *n*				0.177
Locoregional alone	9	4	13	
Locoregional and distant	0	2	2	
Distant alone	4	4	8	
1-year overall survival rate (%)	45.5	9.7		0.002

The level expression of FAM201A was not correlated with the tumor stage, whatever in term of T stage or N stage. Compared with low expression FAM201A, patients with high expression of FAM201A resulted in poorer short-term response to RT. CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

FAM201A Upregulated ATM and mTOR Expression by Acting as a miR-101 Sponge

To further evaluate the regulatory relationship between FAM201A and miR-101, Eca109 cells were transfected with si-FAM201A and FAM201A-mimic sequences and matched controls. The results revealed that miR-101 expression was significantly downregulated in FAM201A-mimic Eca109/Eca109R cells, and was notably upregulated in si-FAM201A-transfected Eca109/Eca109R cells (**Figures 4C,D**). Taken together, these results indicated that FAM201A suppressed the expression of miR-101 (**Supplementary File 6**).

Using TargetScan, ATM and mTOR were predicted to be the downstream targets of miR-101. In Eca109/Eca109R cells, the expression of ATM and mTOR was increased while that of miR-101 was decreased in FAM201A-mimic cells when compared with control cells. When FAM201A expression was decreased, the expression of ATM and mTOR was downregulated while that of miR-101 was increased. Compared with non-irradiated cells, the expression of ATM and mTOR increased after X-ray irradiation. Western blotting confirmed the results of PCR (**Figure 5**; **Supplementary File 6**).

DISCUSSION

The earliest study on lncRNAs associated with radiosensitivity in ESCC was reported by Tong et al. in 2014 (Tong et al., 2014). In this study, they revealed that, when compared with normal para-carcinoma tissue, tumor tissues with a low expression of lncRNA LOC285194 exhibited a larger tumor size, poorer histological grade, had an advanced TNM stage, more lymph node and distant metastases, and was significantly negatively correlated with the pathological response to RT than the LOC285194-high group. Subsequently, researchers have revealed another three lncRNAs related to ESCC radiosensitivity, including BOKAS (Zhang et al., 2015), MALAT1 (Li et al., 2016), and AFAP1-AS1 (Zhou et al., 2016). However, clinical trials for evaluating such lncRNAs related to ESCC radiosensitivity are lacking as the mechanism for how lncRNAs regulate radiosensitivity has yet to be fully elucidated, and so no promising lncRNAs have been applied in the clinic.

In the present study, we identified that the lncRNA FAM201A contributed the most to the radioresistance of ESCC regardless of the tumor stage. The FAM201A gene is a 2.9 Kbp long gene located in genomic 9p13.1 (Humphray et al., 2004) that results in RNA transcripts without ORFs, which means that it has no protein-coding potential. FAM201A in human diseases has been reported crudely in Obsessive-compulsive disorder and Tourette's syndrome by Yu et al. (2015), while it was first mentioned in cancer (colorectal) by Matsumura et al. (2017). Recently, Huang et al. revealed that the biofunction of FAM201A was involved in the development of Osteonecrosis of the femoral head (Huang et al., 2018). However, the molecular mechanism of lncRNA FAM201A function has not been studied. To the best of our knowledge, the present study was the first to report on the correlation of FAM201A with ESCC radiosensitivity and to investigate its potential molecular mechanism, in order to elucidate whether it may be a biomarker for the prognosis and prediction of the patient's response to RT.

The results revealed that patients with FAM201A overexpression had poorer radiosensitivity and inferior survival. Conversely, lower FAM201A expression in ESCC was associated with improved radiosensitivity and a good prognosis, indicating that lnc-FAM201A may serve as a predictor of radiosensitivity in ESCC.

Subsequently, we performed experiments *in vitro* and *in vivo* to confirm the functions of FAM201A. *In vitro*, the overexpression of FAM201A was demonstrated to promote Eca109 cell proliferation; while decreasing FAM201A expression inhibited cell proliferation. The difference in radioresistance following the overexpression of FAM201A in Eca109 and Eca109R cells indicated that FAM201A upregulation likely resulted in cell radioresistance to X-rays irradiation. In addition, the similar levels of radiosensitivity following the reduction in FAM201A expression in Eca109 and Eca109R cells suggested that si-FAM201A may enhance the radiosensitivity of both intrinsically and acquired-radioresistant tumor cells, indicating that siFAM201A may serve as an effective sensitizing molecular strategy for ESCC. *In vivo*, when compared with control groups, FAM201A knockdown significantly blocked xenograft tumor

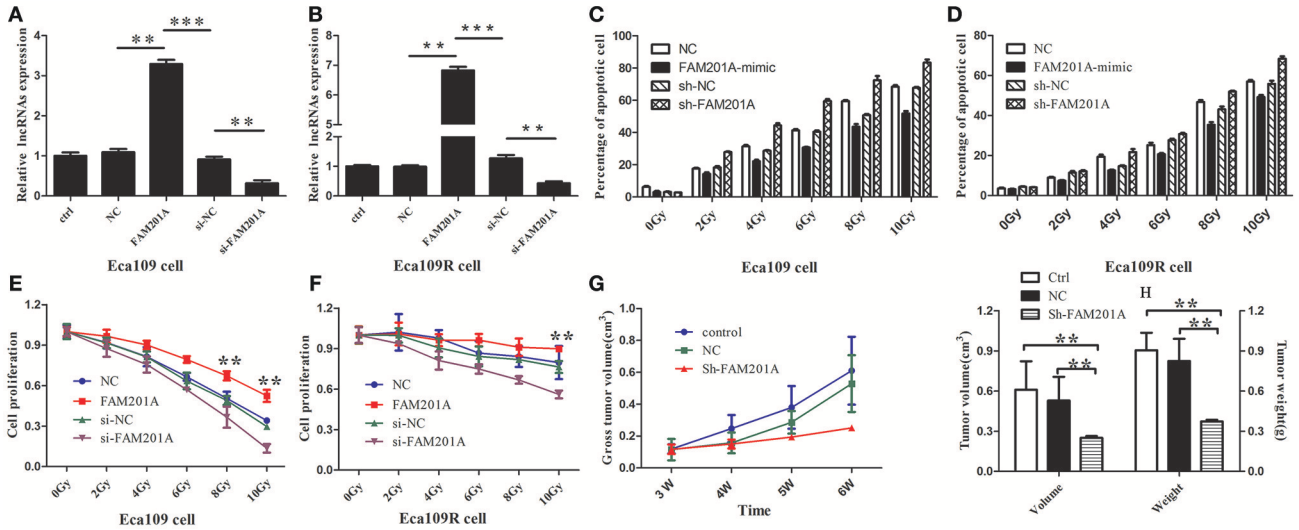

FIGURE 3 | Reverse transcription-quantitative polymerase chain analysis confirmed the efficiency of transfected **(A)** Eca109 or **(B)** Eca109R cells with si-FAM201A and FAM201A-mimic. In both **(C)** Eca109 or **(D)** Eca109R cancer cells transfected with si-FAM201A or FAM201A-mimic, the percentage of apoptotic cells in each line increased with the increasing X-ray irradiation dose. In contrast to the levels of apoptosis, cell proliferation decreased with increasing radiation doses in **(E)** Eca109 or **(F)** Eca109R. The effect of shFAM201A on Xenograft tumor survival was also evaluated: **(G)** tumor survival curve, **(H)** tumor volume and weight. $^{**}P < 0.01$, $^{***}P < 0.001$.

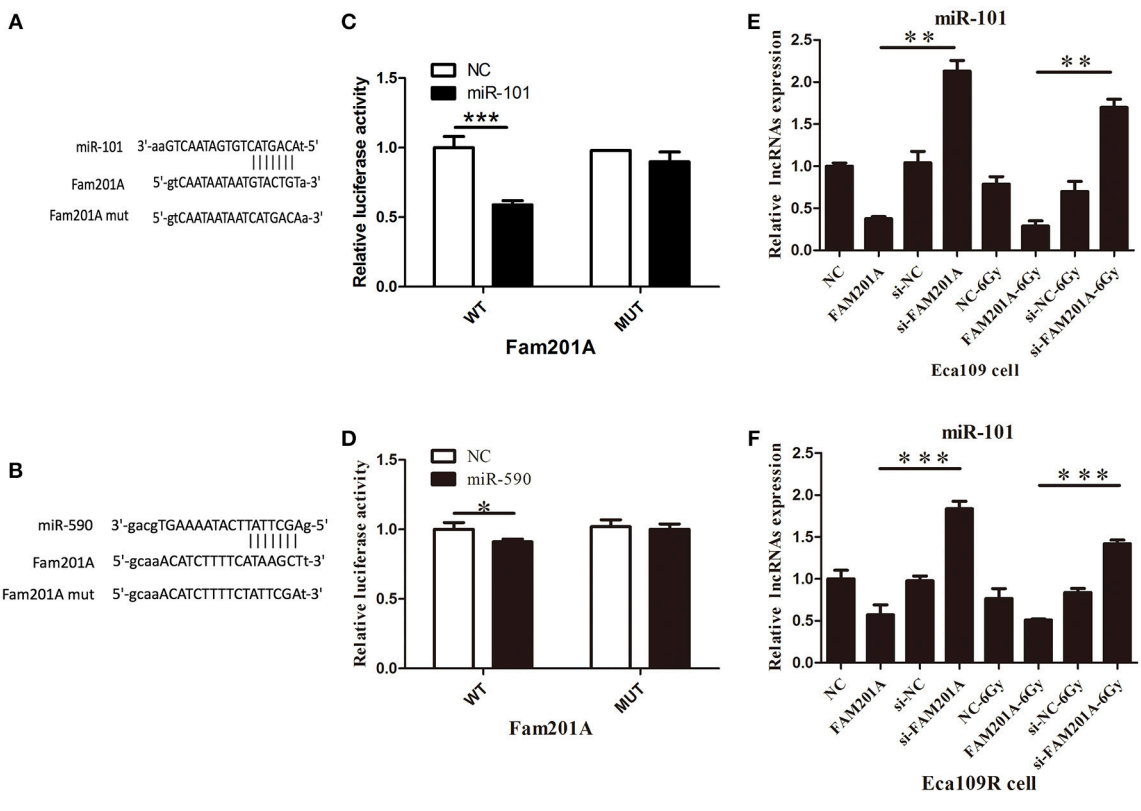

FIGURE 4 | (A) miR-101 and **(B)** miR-509 were predicted to have complementary base pairings with FAM201A. The relative luciferase activity of the wild-type and mutated FAM201A were compared between **(C)** miR-101 and **(D)** miR-590. miR-101 expression was negatively regulated by FAM201A in **(E)** Eca109 and Eca1 **(F)** 109R cells. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

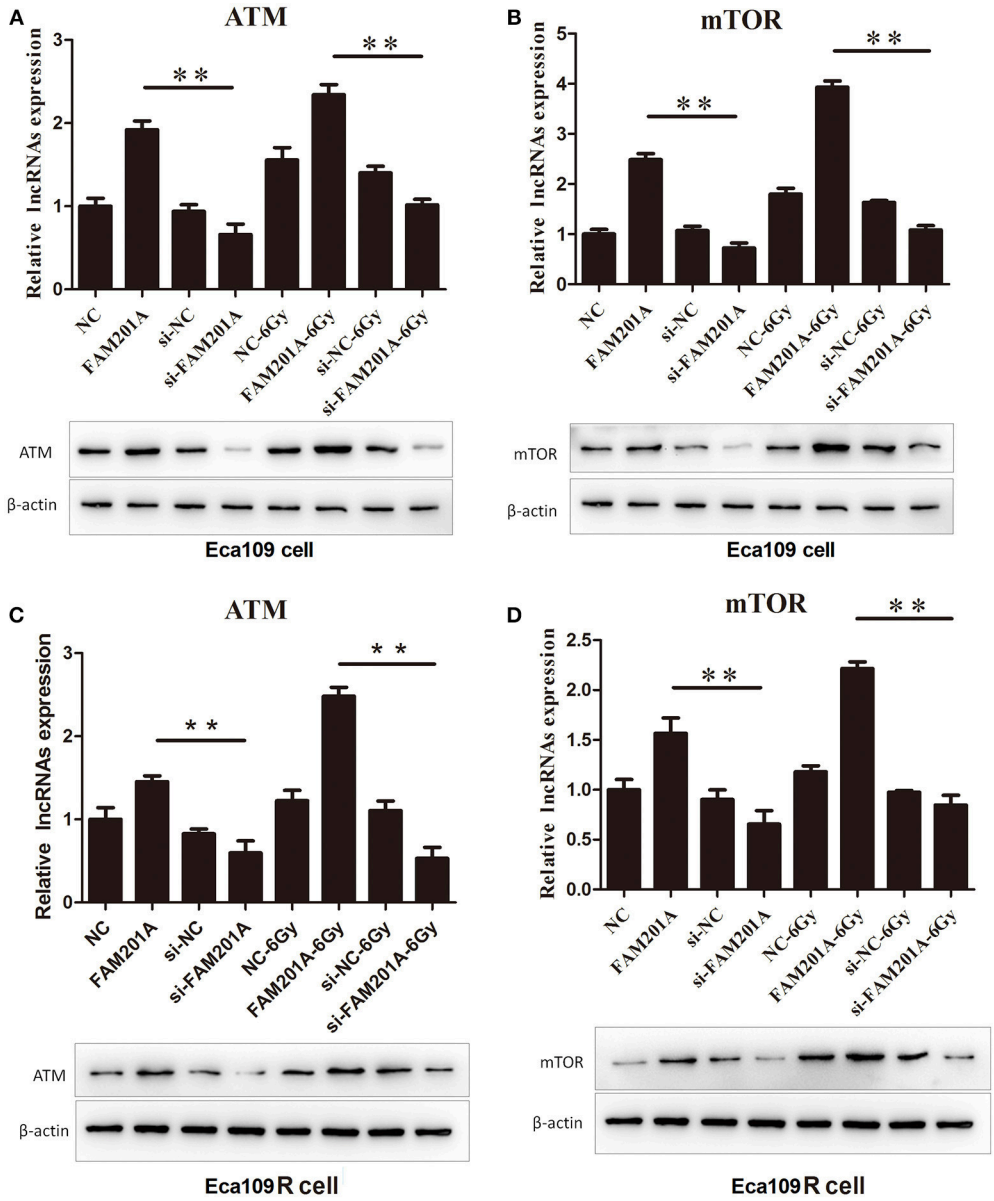

FIGURE 5 | Effects of overexpressed- and si-FAM201A on the expression of miR-101, ATM and mTOR in **(A, B)** Eca109 and **(C, D)** Eca109R cells before and after X-ray irradiation. Western blotting validation of ATM and mTOR in Eca109 and Eca109R cells. $^{**}P < 0.01$.

growth (decreased tumor volume and weight), which confirmed that siFAM201A was able enhance radiosensitivity.

Recently, a competing endogenous RNAs hypothesis proposed that lncRNAs may exert their biological function by acting as a molecular sponge for miRNAs, in turn leading to derepression of miRNA targets (Tay et al., 2014). To explore the molecular mechanism of FAM201A-modulated radiosensitivity in ESCC, we used the online software Starbase 2.0 to predict the downstream target genes, and found that miR-101 and miR-590 had complementary base pairings with FAM201A. Only miR-101, and not miR-590, was observed to directly interact with FAM201A, as determined by the luciferase reporter assay. The qPCR analysis further demonstrated that FAM201A overexpression downregulated miR-101 expression while si-FAM201A transfection upregulated miR-101. These results suggested that FAM201A may modulate target gene expression by serving as a "sponge" for miR-101 (Kung et al., 2013).

Further, the role of miRNAs usually depends on what genes they target. The TargetScan analysis showed that ATM and mTOR were the targets of miR-101. Furthermore, qPCR revealed that overexpression of FAM201A leads to the downregulation of miR-101, the upregulation of ATM and mTOR, and resulted in radioresistance; however, depletion of FAM201A led to the upregulation of miR-101, downregulation of ATM, and mTOR, and resulted in radiosensitivity. Additionally, western blotting confirmed these PCR results.

ATM is the major repair protein involved in the homologous recombination repair (HRR) of ionizing radiation-induced double-strand breaks (RI-DSB). ATM deficiency leads to HRR disorders, increased apoptosis and radiosensitivity (Cliby et al., 1998; Cuddihy and Bristow, 2004; Hammond and Muschel, 2014). Therefore, we hypothesize that FAM201A may regulate ESCC radiosensitivity via a "FAM201A-miRNA101-ATM-HRR" axis.

HRR occurs only in the S and G2 phases of DNA replication, due to the requirement of homologous sister chromatids as a template (Pâques and Haber, 1999). DSBs during the absence of homologous sequence chromosomes requires non-homologous end joining (NHEJ) to achieve DNA repair, which is a repair function performed throughout the cell cycle and was initially considered to be the primary mechanism of RI-DSB repair (Branzei and Foiani, 2008; Beucher et al., 2009). Yan et al. (2010) reported that miR-101 regulates the radiosensitivity of cells by regulating the DNA-dependent protein kinase catalytic subunit, an important member of the NHEJ machinery, via mTOR. Therefore, we hypothesize that lncRNA-FAM201A may also modulate cell ionizing radiosensitivity via a "FAM201A-miR-101-mTOR-NHEJ" axis. In future research, we will focus on the upstream mechanism underlying FAM201A upregulation in regulating ESCC radiosensitivity.

CONCLUSIONS

In conclusion, the present study revealed that lncRNA FAM201A may be a potential biomarker for predicting radiosensitivity and prognosis, as well as a therapeutic target for enhancing cancer radiosensitivity in ESCC. FAM201A contributed to radioresistance through a FAM201A-miR-101-ATM/mTOR regulatory network in ESCC. However, the upstream mechanism for FAM201A upregulation in regulating ESCC radiosensitivity requires further study.

ETHICS STATEMENT

This study was subject to approval by the Fujian Medical University Union Hospital Institutional Review Board (No. 2014KY001). All patients signed an informed consent prior to treatment, and all information was anonymized and deidentified prior to its analysis.

AUTHORS CONTRIBUTIONS

MC, PL, YT, JC, and JL conceived the study, manuscript, and statics analysis. YL, XiaL, MS, XiqL, and AL assistance with collecting clinical data. RY, WN, XZ, YC, and LZ provided assistance with study design and revisions of the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported in part by grants from the Fujian Provincial Health & Family Planning Commission (Project Number: 2016-ZQN-32), the Fujian Provincial Department of Science & Technology (Project Number: 2018J01306), the Fujian Provincial Department of Science & Technology (Project Number: 2017Y9079), the Fujian Provincial Platform for Medical Laboratory Research, and Key Laboratory for Tumor Individualized Active Immunity (Project Number: FYKFKT-2017015).

ACKNOWLEDGMENTS

The authors thank all patients who participated in the present study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00611/full#supplementary-material

File S1 | Results of lncRNA chip microarray.

File S2 | Results of PCR of candidate lncRNAs.

File S3 | FAM201A as a novel lncRNA with a potential function in radiosensitivity.

File S4 | Validation of FAM201A *in vitro* and *in vivo*.

File S5 | The results of the luciferase assay.

File S6 | Overexpression of FAM201A and its effect on miR-101/ATM/mTOR in Eca109 and Eca109R cells.

REFERENCES

Beucher, A., Birraux, J., Tchouandong, L., Barton, O., Shibata, A., Conrad, S., et al. (2009). ATM and artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2. *EMBO J.* 28, 3413–3427. doi: 10.1038/emboj.2009.276

Branzei, D., and Foiani, M. (2008). Regulation of DNA repair throughout the cell cycle. *Nat. Rev. Mol. Cell Biol.* 9, 297–308. doi: 10.1038/nrm2351

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424 doi: 10.3322/caac.21492

Chaachouay, H., Ohneseit, P., Toulany, M., Kehlbach, R., Multhoff, G., Rodemann, H. P. (2011). Autophagy contributes to resistance of tumor cells to ionizing radiation. *Radiother Oncol.* 99, 287–292. doi: 10.1016/j.radonc.2011.06.002

Chen, M. Q., Lin, Q. L., Chen, Y. G., Guo, J. H., Xu, B. H., and Tian, Y. (2017). Neoadjuvant chemotherapy may not benefit esophageal squamous cell carcinoma patients treated with definitive chemoradiotherapy. *J. Chin. Med. Assoc.* 80, 636–643. doi: 10.1016/j.jcma.2017.06.014

Chen, X., Liao, R., Li, D., and Sun, J. (2017). Induced cancer stem cells generated by radiochemotherapy and their therapeutic implications. *Oncotarget* 8, 17301–17312. doi: 10.18632/oncotarget.14230

Cliby, W. A., Roberts, C. J., Cimprich, K. A., Stringer, C. M., Lamb, J. R., Schreiber, S. L., et al. (1998). Overexpression of a kinase-inactive ATR protein causes sensitivity to DNA-damaging agents and defects in cell cycle checkpoints. *EMBO J.* 17, 159–169. doi: 10.1093/emboj/17.1.159

Cooper, J. S., Guo, M. D., Herskovic, A., Macdonald, J. S., Martenson, J. A. Jr, Al-Sarraf, M., et al. (1999). Chemoradiotherapy of locally advanced esophageal cancer: long-term follow-up of a prospective randomized trial (RTOG 85–01). radiation therapy oncology group. *JAMA* 281, 1623–1627. doi: 10.1001/jama.281.17.1623

Cuddihy, A. R., and Bristow, R. G. (2004). The p53 protein family and radiation sensitivity: yes or no? *Cancer Metastasis Rev.* 23, 237–257. doi: 10.1023/B:CANC.0000031764.81141.e4

Da, C., Wu, L., Liu, Y., Wang, R., and Li, R. (2017). Effects of irradiation on radioresistance, HOTAIR and epithelial-mesenchymal transition/cancer stem cell marker expression in esophageal squamous cell carcinoma. *Oncol. Lett.* 13, 2751–2757. doi: 10.3892/ol.2017.5774

Dumont, F. J., and Bischoff, P. (2012). Disrupting the mTOR signaling network as a potential strategy for the enhancement of cancer radiotherapy. *Curr. Cancer Drug Targets*. 12, 899–924. doi: 10.2174/156800912803251243

Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biom. J.* 47, 458–472. doi: 10.1002/bimj.200410135

Francescone, R. A., Scully, S., Faibish, M., Taylor, S. L., Oh, D., Moral, L., et al. (2011). Role of YKL-40 in the angiogenesis, radioresistance, and progression of glioblastoma. *J. Biol. Chem*. 286, 15332–15343. doi: 10.1074/jbc.M110.212514

Gwynne, S., Hurt, C., Evans, M., Holden, C., Vout, L., and Crosby, T. (2011). Definitive chemoradiation for oesophageal cancer–a standard of care in patients with non-metastatic oesophageal cancer. *Clin. Oncol. (R. Coll. Radiol)* 23, 182–188. doi: 10.1016/j.clon.2010.12.001

Hammond, E. M., and Muschel, R. J. (2014). Radiation and ATM inhibition: the heart of the matter. *J. Clin. Invest.* 124, 3289–3291. doi: 10.1172/JCI77195

Huang, G., Zhao, G., Xia, J., Wei, Y., Chen, F., Chen, J., et al. (2018). FGF2 and FAM201A affect the development of osteonecrosis of the femoral head after femoral neck fracture. *Gene* 652, 39–47. doi: 10.1016/j.gene.2018.01.090

Humphray, S. J., Oliver, K., Hunt, A. R., Plumb, R. W., Loveland, J. E., Howe, K. L., et al. (2004). DNA sequence and analysis of human chromosome 9. *Nature* 429, 369–374. doi: 10.1038/nature,02465

Japan Esophageal Society (2017). Japanese classification of esophageal cancer, 11th edition: part II and III. *Esophagus* 14, 37–65. doi: 10.1007/s10388-016-0556-2

Kung, J. T., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics* 193, 651–669. doi: 10.1534/genetics.112.146704

Li, Z., Zhou, Y., Tu, B., Bu, Y., Liu, A., and Kong, J. (2016). Long noncoding RNA MALAT1 affects the efficacy of radiotherapy for esophageal squamous cell carcinoma by regulating Cks1 expression. *J. Oral Pathol. Med*. 46, 583–590. doi: 10.1111/jop.12538

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta delta C(T)) method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lloyd, S., and Chang, B. W. (2014). Current strategies in chemoradiation for esophageal cancer. *J. Gastrointest. Oncol.* 5, 156–165. doi: 10.3978/j.issn.2078-6891.2014.033

Matsumura, K., Kawasaki, Y., Miyamoto, M., Kamoshida, Y., Nakamura, J., Negishi, L., et al. (2017). The novel G-quadruplex-containing long non-coding RNA GSEC antagonizes DHX36 and modulates colon cancer cell migration. *Oncogene* 36, 1191–1199. doi: 10.1038/onc.2016.282

Moncharmont, C., Levy, A., Gilormini, M., Bertrand, G., Chargari, C., Alphonse, G., et al. (2012). Targeting a cornerstone of radiation resistance: cancer stem *Cell* 322, 139–147. doi: 10.1016/j.canlet.2012.03.024

Pâques, F., and Haber, J. E. (1999). Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev*. 63, 349–404.

Pennathur, A., Gibson, M. K., Jobe, B. A., and Luketich, J. D. (2013). Oesophageal carcinoma. *Lancet* 381, 400–412. doi: 10.1016/S0140-6736(12)60643-6

Qi, P., and Du, X. (2013). The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine. *Mod. Pathol.* 26, 155–165. doi: 10.1038/modpathol.2012.160

Rustgi, A. K., and El-Serag, H. B. (2014). Esophageal carcinoma. *N. Engl. J. Med.* 371, 2499–2509. doi: 10.1056/NEJMra1314530

Sasaki, Y., and Kato, K. (2016). Chemoradiotherapy for esophageal squamous cell cancer. *JPN. J. Clin. Oncol.* 46:805. doi: 10.1093/jjco/hyw082

Short, M. W., Burgers, K. G., and Fry, V. T. (2017). Esophageal Cancer. *Am. Fam. Physician* 95, 22–28.

Spizzo, R., Almeida, M. I., Colombatti, A., and Calin, G. A. (2012). Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* 31:4577. doi: 10.1038/onc.2011.621

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352. doi: 10.1038/nature12986

Tong, Y. S., Zhou, X. L., Wang, X. W., Wu, Q. Q., Yang, T. X., Lv, J., et al. (2014). Association of decreased expression of long non-coding RNA LOC285194 with chemoradiotherapy resistance and poor prognosis in esophageal squamous cell carcinoma. *J. Transl. Med.* 12:233. doi: 10.1186/s12967-014-0233-y

Versteijne, E., Van Laarhoven, H. W., van Hooft, J. E., Van Os, R. M., Geijsen, E. D., Van Berge Henegouwen, M. I., et al. (2014). Definitive chemoradiation for patients with inoperable and/or unresectable esophageal cancer: locoregional recurrence pattern. *Dis. Esophaqus*. 28, 453–459. doi: 10.1111/dote.12215

Yan, D., Ng, W. L., Zhang, X., Wang, P., Zhang, Z., Mo, Y. Y., et al. (2010). Targeting DNA-PKcs and ATM with miR-101 sensitizes tumors to radiation. *PLoS ONE* 5:e11397. doi: 10.1371/journal.pone.0011397

Yu, D., Mathews, C. A., Scharf, J. M., Neale, B. M., Davis, L. K., Gamazon, E. R., et al. (2015). Cross-disorder genome-wide analyses suggest a complex genetic relationship between tourette's syndrome and OCD. *Am. J. Psychiatry* 172, 82–93. doi: 10.1176/appi.ajp.2014

Yu, Z. Q., Zhang, C., Lao, X. Y., Wang, H., Gao, X. H., Cao, G. W., et al. (2012). Long non-coding RNA influences radiosensitivity of colorectal carcinoma cell lines by regulating cyclin D1 expression. *Zhonghua Wei Chang Wai Ke Za Zhi* 15, 288–291.

Zafar, F., Seidler, S. B., Kronenberg, A., Schild, D., and Wiese, C. (2010). Homologous recombination contributes to the repair of DNA double-strand breaks induced by high-energy iron ions. *Radiat. Res.* 173, 27–39. doi: 10.1667/RR1910.1

Zhang, H., Luo, H., Hu, Z., Peng, J., Jiang, Z., Song, T., et al. (2015). Targeting WISP1 to sensitize esophageal squamous cell carcinoma to irradiation. *Oncotarget* 6:6218. doi: 10.18632/oncotarget.3358

Zhou, X. L., Wang, W. W., Zhu, W. G., Yu, C. H., Tao, G. Z., Wu, Q. Q., et al. (2016). High expression of long non-coding RNA AFAP1-AS1 predicts chemoradioresistance and poor prognosis in patients with esophageal squamous cell carcinoma treated with definitive chemoradiotherapy. *Mol. Carcinog.* 55, 2095–2105. doi: 10.1002/mc.22454

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

frontiers in Genetics

METHODS
published: 19 February 2019
doi: 10.3389/fgene.2019.00121

CaDrA: A Computational Framework for Performing Candidate Driver Analyses Using Genomic Features

*Vinay K. Kartha[1,2], Paola Sebastiani[1,3], Joseph G. Kern[4], Liye Zhang[5], Xaralabos Varelas[4] and Stefano Monti[1,2,3]**

[1] Bioinformatics Program, Boston University, Boston, MA, United States, [2] Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, United States, [3] Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States, [4] Department of Biochemistry, Boston University School of Medicine, Boston, MA, United States, [5] School of Life Sciences and Technology, ShanghaiTech University, Shanghai, China

The identification of genetic alteration combinations as drivers of a given phenotypic outcome, such as drug sensitivity, gene or protein expression, and pathway activity, is a challenging task that is essential to gaining new biological insights and to discovering therapeutic targets. Existing methods designed to predict complementary drivers of such outcomes lack analytical flexibility, including the support for joint analyses of multiple genomic alteration types, such as somatic mutations and copy number alterations, multiple scoring functions, and rigorous significance and reproducibility testing procedures. To address these limitations, we developed Candidate Driver Analysis or CaDrA, an integrative framework that implements a step-wise heuristic search approach to identify functionally relevant subsets of genomic features that, together, are maximally associated with a specific outcome of interest. We show CaDrA's overall high sensitivity and specificity for typically sized multi-omic datasets using simulated data, and demonstrate CaDrA's ability to identify known mutations linked with sensitivity of cancer cells to drug treatment using data from the Cancer Cell Line Encyclopedia (CCLE). We further apply CaDrA to identify novel regulators of oncogenic activity mediated by Hippo signaling pathway effectors YAP and TAZ in primary breast cancer tumors using data from The Cancer Genome Atlas (TCGA), which we functionally validate *in vitro*. Finally, we use pan-cancer TCGA protein expression data to show the high reproducibility of CaDrA's search procedure. Collectively, this work demonstrates the utility of our framework for supporting the fast querying of large, publicly available multi-omics datasets, including but not limited to TCGA and CCLE, for potential drivers of a given target profile of interest.

Keywords: oncogenic driver analysis, stepwise search, TCGA, CCLE, R package

OPEN ACCESS

Edited by:
Binhua Tang,
Hohai University, China

Reviewed by:
Ao Li,
University of Science and Technology of China, China
Samir B. Amin,
The Jackson Laboratory for Genomic Medicine, United States

***Correspondence:**
Stefano Monti
smonti@bu.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 October 2018*
Accepted: *04 February 2019*
Published: *19 February 2019*

Citation:
Kartha VK, Sebastiani P, Kern JG, Zhang L, Varelas X and Monti S (2019) CaDrA: A Computational Framework for Performing Candidate Driver Analyses Using Genomic Features. Front. Genet. 10:121. doi: 10.3389/fgene.2019.00121

Abbreviations: BRCA, breast carcinomas; CaDrA, candidate driver analysis; CCLE, Cancer Cell Line Encyclopedia; COSMIC, Catalogue of Somatic Mutations in Cancer; FDR, false discovery rate; FPR, false positive rate; KS, Kolmogorov–Smirnov; qRT-PCR, quantitative real-time polymerase chain reaction; RPPA, reverse phase protein array; SCNA, somatic copy number alteration; TCGA, The Cancer Genome Atlas; TN, triple-negative; TPR, true positive rate.

INTRODUCTION

Advances in high-throughput sequencing technology has led to a rapid rise in the availability of large multi-omic datasets through compendia such as the CCLE, TCGA, the Genotype-Tissue Expression (GTEx), and others (Barretina et al., 2012; Chang et al., 2013; Ardlie et al., 2015). These data include genetic alterations, comprising SCNAs and somatic mutations, epigenetic information, such as microRNA expression and DNA methylation, as well as gene expression profiling through microarray or RNA-sequencing (RNASeq) technology, across tens of thousands of samples representing varying biological contexts. Concomitantly, several computational methods have been developed and applied to effectively query and integrate different types of genome-wide datasets in order to make meaningful predictions about the biological processes driving the phenotypes of interest (Drier et al., 2013; Kristensen et al., 2014). An important application of such methods is the identification of recurrent genomic alterations, and their potential effects on downstream pathway activity or phenotypes associated with development and disease states. For example, in many cancers, samples exhibiting elevated activity of a given oncogenic signature may be enriched for, or driven by functionally relevant somatic mutations or SCNAs. Identifying such associations may help elucidate underlying mechanisms contributing to abnormal pathway activity, further enabling disease subtyping and sample classification (Bea et al., 2005; Savage et al., 2003; Monti et al., 2012). Alternatively, linking these genomic features with their close interactors through protein-protein interaction networks, gene function annotations or phenotypic readouts such as drug sensitivity may support the discovery of novel druggable targets and further guide precision medicine regimens (Bild et al., 2006; Heiser et al., 2011; Daemen et al., 2013; Hou and Ma, 2014; Jia and Zhao, 2014).

Recently, computational methods and models have been developed for performing driver gene analyses applied to high-dimensional 'omics' data from cancer cell lines and patients. These are typically motivated either by frequency or exclusivity of alterations across samples (Youn and Simon, 2011; Ciriello et al., 2012; Dees et al., 2012; Vandin et al., 2012; Lawrence et al., 2013; Leiserson et al., 2013; Kim et al., 2016), or their functional interplay based on biological interaction networks and pathway ontology (Ng et al., 2012; Creixell et al., 2015; Leiserson et al., 2015; Cho et al., 2016). Indeed, certain approaches integrate interactome and functional information to further guide driver gene prioritization in cancer (Chen et al., 2014; Xi et al., 2017; Sanchez-Vega et al., 2018). Some of these tools have been proposed to specifically identify subsets or combinations of genomic features that are collectively associated with a given phenotypic response, explaining a larger fraction of the biological context than any individual feature alone (Kim et al., 2016). These methods, while useful, do not offer simultaneous support for: (i) the joint analyses of multi-type features, including SCNAs and somatic mutations, with possible extension to other genomic data, (ii) multiple feature scoring functions and, most importantly, (iii) rigorous assessment of the statistical significance of the discovered associations. Of equal relevance, a user-friendly and flexible programming package supporting the rapid screening for candidate drivers given a set of ranked genomic features is currently lacking, and would prove extremely useful for incorporation in analytical pipelines aimed at the generation of novel biological hypotheses.

Here, we present CaDrA, a methodology that searches for the set of genomic alterations, here denoted as *features* (mutations, SCNAs, translocations, etc.), associated with a user-provided ranking of samples within a dataset. Our method specifically employs a stepwise heuristic search to identify a subset of features whose union is maximally associated with the observed sample ranking, and carries out rigorous statistical significance testing based on sample permutation, thereby allowing for the identification of candidate genetic drivers associated with aberrant pathway activity or drug sensitivity, while still exploiting aspects of feature complementarity and sample heterogeneity. To highlight the method's overall performance, along with its relevance and ability to select sets of genomic features that indeed drive certain oncogenic phenotypes in cancer, we perform extensive evaluation of CaDrA based on simulated data, as well as real genomic data from cancer cell lines and primary human tumors. The results from simulations show that CaDrA has high sensitivity for mid- to large-sized datasets, and high specificity for all sample sizes considered. Using genomic data drawn from CCLE and TCGA, we demonstrate CaDrA's capacity to correctly identify well-characterized driver mutations in cancer cell lines and primary tumors spanning multiple cancer types, along with its ability to discover novel features associated with invasive phenotypes in human breast cancer samples, which we functionally validate *in vitro*. Our framework, which is publicly available as an R package, will allow for rapidly mining numerous multi-omics datasets for candidate drivers of user-specified molecular readouts, such as pathway activity, drug sensitivity, protein expression, or other quantitative measurements of interest, further enabling targeted queries and novel hypothesis generation.

RESULTS

CaDrA Overview

An overview of CaDrA's workflow is summarized in **Figure 1**. CaDrA implements a step-wise heuristic approach that searches through a set of binary features [each represented as a 1/0-valued vector, indicating the presence/absence of a SCNA, somatic mutation, or other (epi)genetic alterations across samples, respectively], and returns a final subset of features whose union (logical OR) defines an alteration 'meta-feature' that is maximally associated with the defined sample ranking provided as input (see section "Methods"). The strength of the association of a meta-feature with a sample ranking is a function of the agreement between the skewness of the alterations' occurrences and the sample ranking. The input sample ranking is usually a function of a sample-specific measurement, e.g., the activity level of a pathway, the response to a targeted treatment, the expression level of a given transcript or protein, etc. Therefore, the meta-feature returned by the search is the set of features maximally

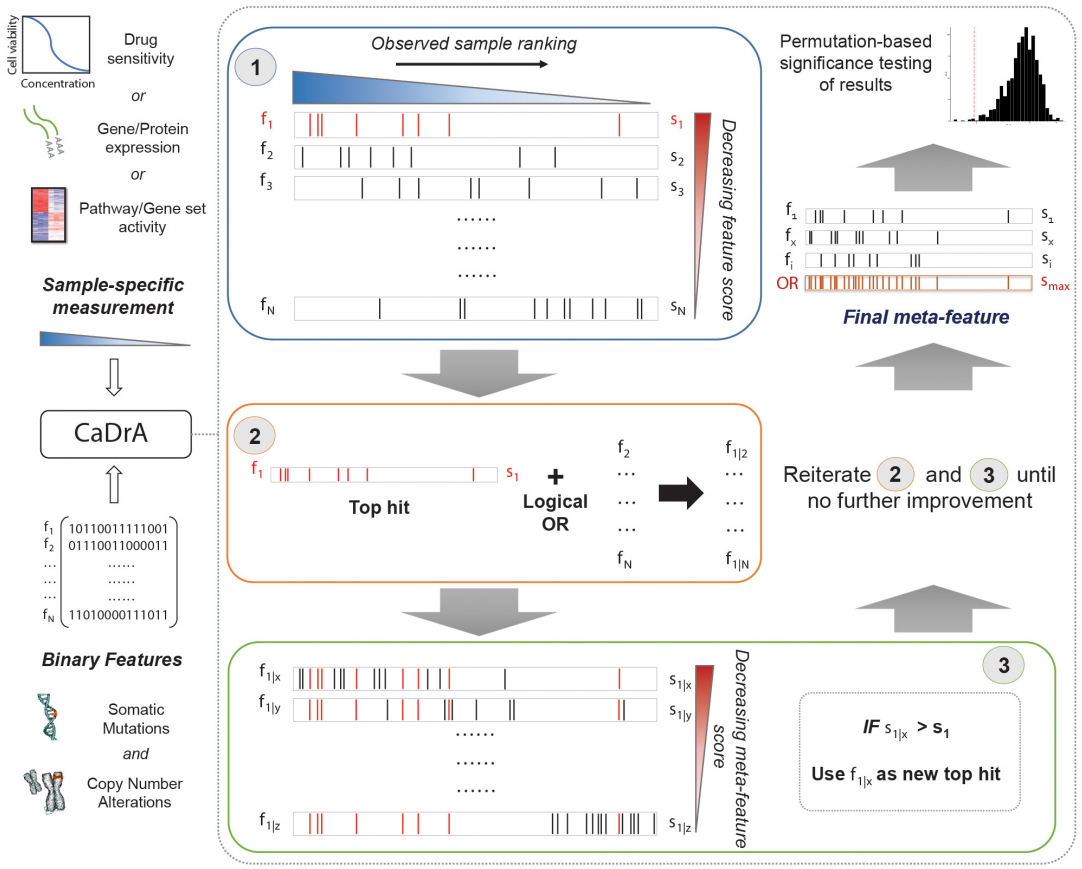

FIGURE 1 | Overview of CaDrA workflow and implementation. CaDrA takes as input a sample-specific measurement to rank samples, and a matrix of binary features of the same samples. In Step 1 (blue box), CaDrA begins by choosing a starting feature, which is either the single feature having the best score based on its left-skewness, or a user-specified start feature. In the next step (Step 2; orange box), the union (logical OR) of this feature with each of the remaining features in the dataset is taken, yielding 'meta-features' with their corresponding scores. If any meta-feature has a better score than the hit from the previous step (Step 3; green box), CaDrA uses this new meta-feature as a reference for the next iteration, repeating Steps 2 and 3 until no further improvement in scores can be obtained. The final output is a set of features (meta-feature) whose union has the (local) maximum score and its permutation-based *p*-value.

predictive of that same sample-specific measurement variable. The logical OR operator used in the iterative search framework specifically takes advantage of heterogeneity seen across samples (i.e., samples harboring similar phenotypes but different drivers of the given outcome), thus enabling the potential identification of complementary drivers of target phenotypes (Kim et al., 2016). CaDrA allows for multiple modes to query ranked binary datasets with user-specified parameters defining search criteria, enables rigorous permutation-based significance testing of results, and reduced computation time by exploiting pre-computed score distributions and parallel computing, when available (see section "Methods").

Analysis of Simulated Data to Evaluate CaDrA Performance

To assess the overall performance of CaDrA to recover (statistically) significantly associated meta-features, we simulated two types of datasets for a range of sample sizes: (i) the *true-positive datasets* consist of both left-skewed (i.e., true positive with skewness concordant with sample ranking) as well as uniformly distributed (i.e., null) features; and (ii) the *null datasets* consist of null features only (see section "Methods" and **Supplementary Figure S1**). This enabled us to estimate the overall sensitivity and specificity of CaDrA using the true positive and null datasets, respectively. By running CaDrA on multiple simulated datasets of different sample sizes (n = 500 true positive and null datasets for each sample size), we first evaluated the resulting meta-features based on the number of true positive features and the total number of features contained within each returned meta-feature (i.e., the meta-feature size; **Figures 2A,B**). The true positive datasets had a maximum of five positive features to be detected, while the maximum number of features CaDrA was allowed to add was set to 7, to evaluate the ability of the search to recover all but no more than the positive features. With progressively higher sample sizes, we observed an increase in the fraction of CaDrA-identified meta-features that include all 5 true positive features (**Figure 2A**). The TPR and FPR of CaDrA on the simulated positive and null data, respectively, for different sample sizes are shown in **Figures 2C,D**, and was calculated as the fraction of searches

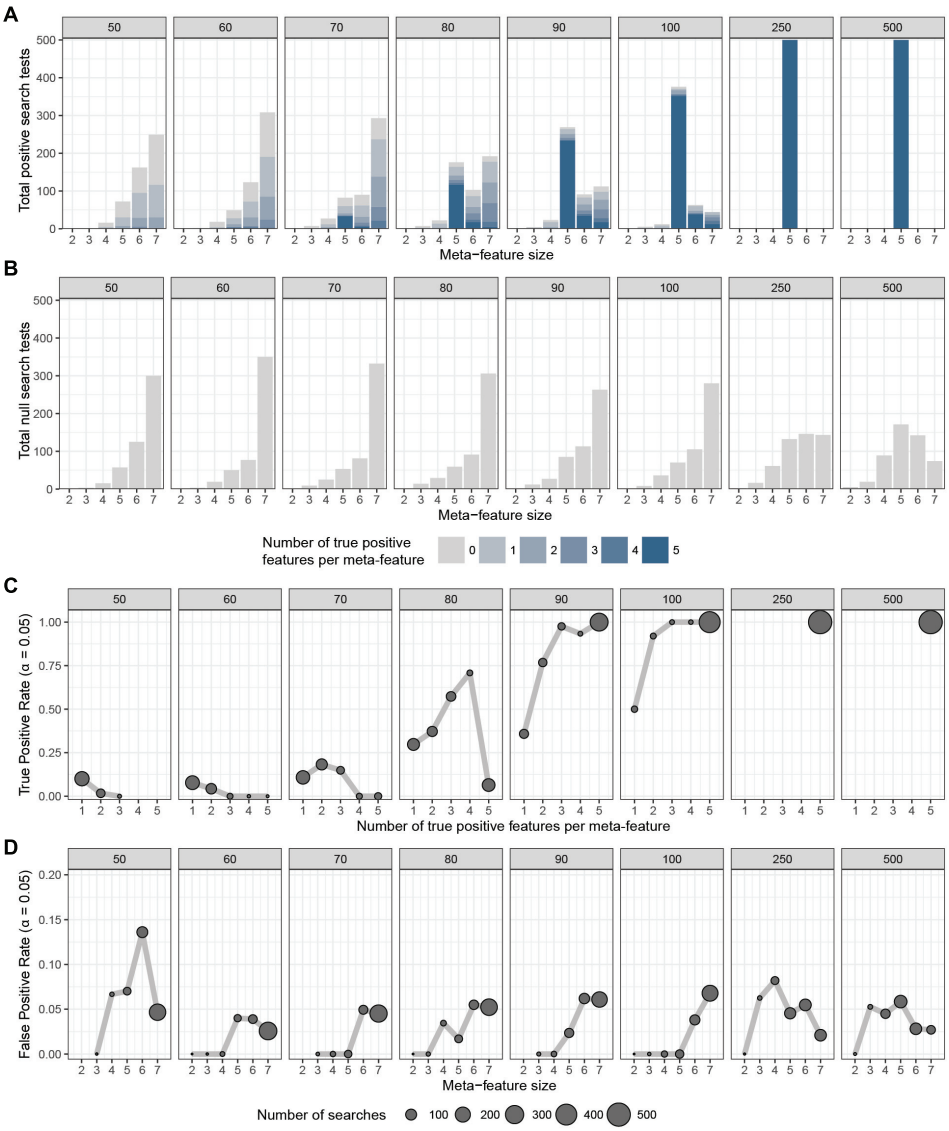

FIGURE 2 | CaDrA performance on simulated data. CaDrA was run on 500 independent simulated datasets containing **(A)** both positive and null, and **(B)** only null features with sample sizes ranging between 50 and 500 samples (number in gray box above each sub-panel). In each case, the distribution of the number of features per meta-feature (i.e., the meta-feature size) returned by CaDrA is shown **(A,B)** as well as the number and fraction of searches that yielded significance for $\alpha = 0.05$ **(C,D)**, corresponding to the true positive rate (TPR) and false positive rate (FPR), respectively.

returning meta-features with permutation *p*-value significant at $\alpha = 0.05$ (**Supplementary Figure S2**). The TPR was estimated for different numbers of recovered true positive features (in the true positive datasets), while the FPR was estimated for different numbers of returned features (by definition, false positives) in the null datasets, and is summarized in **Table 1**. CaDrA returned all of the simulated true positive features with 100% TPR for sample sizes larger than $N = 100$. CaDrA also yielded a very high mean TPR of >95% at $N = 100$, with the sensitivity dropping to 7.7% only at the smallest sample size of $N = 50$ (**Table 1**). Further, when applied to the null datasets (**Figure 2B**), the majority of meta-features returned by CaDrA were correctly deemed as non-significant at $\alpha = 0.05$, with a maximum mean FPR of 7.2% for the lowest sample size analyzed (**Figure 2D** and **Table 1**).

These results suggest that CaDrA requires mid- to large-sized datasets for sufficient sensitivity, while maintaining high specificity at all sample sizes assessed.

CaDrA Identifies Known Regulators of Ras/Raf/Mek/ERK Signaling Sensitivity in Cancer Cell Lines

The mitogen-activated protein kinase (MAPK) kinase (MEKK)/extra-cellular signal-regulated kinase (ERK) pathway is a well-conserved kinase cascade known to play a regulatory

TABLE 1 | Overall true positive rate (TPR) and false positive rate (FPR) of CaDrA based on simulated data.

Sample Size (*N*)	Mean TPR (%)	Mean FPR (%)
50	7.69	7.2
60	5.76	2.8
70	11.53	3.8
80	30.72	4.6
90	87.55	5
100	96.51	4.6
250	100	4.6
500	100	4.2

Weight-averaged TPR and FPRs were computed per sample size for true positive and null simulated datasets, respectively (n = 500 simulated datasets per sample size; see section "Methods").

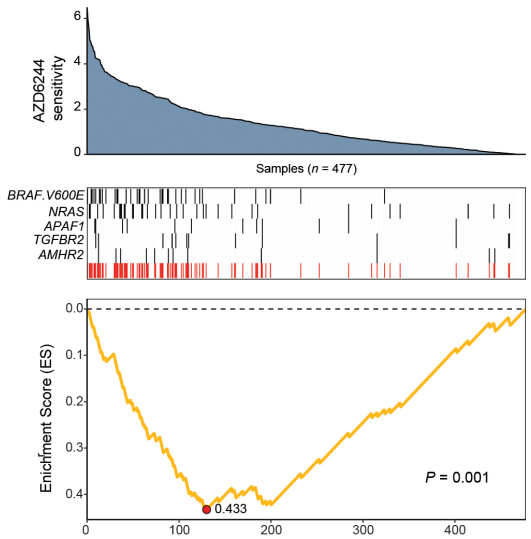

FIGURE 3 | CaDrA identifies mutations in MAPK/ErK signaling genes that contribute to hyper-sensitivity to MEK inhibition *in vitro*. ActArea measurements reflecting sensitivity to MEK inhibitor AZD6244 were used to rank CCLE cell lines (n = 477). CaDrA was then run to identify sets of genomic features that were most-associated with decreasing ActArea (i.e., increasing sensitivity) scores. Through step-wise search iterations, CaDrA identified somatic mutations in known regulators upstream of MEK, including an activating mutation in *BRAF* ($BRAF^{V600E}$) and *NRAS*, as well as those in *APAF1*, *TGFBR2,* and *AMHR2*, before terminating the search process. The resulting meta-feature (red track) and its corresponding enrichment score (ES) is shown.

role in cell proliferation, differentiation, and survival in response to extracellular signaling (Kim and Choi, 2010; Cargnello and Roux, 2011; Burotto et al., 2014). Increased MAP/ERK kinase (MEK) activity is a feature of many cancers, and is often triggered by missense mutations in *BRAF* and *NRAS*, two upstream oncogenes and potent regulators of Ras/Raf/Mek/ERK signaling (Cantwell-Dorris et al., 2011; Burotto et al., 2014). Small molecules targeting these mutated proteins have been shown to be effective in treating these cancers via inactivation of Ras/Raf/Mek/ERK signaling (Roberts and Der, 2007; Chapman et al., 2011; Barretina et al., 2012; Johnson and Puzanov, 2015). To highlight CaDrA's ability to recover independent genomic features that may confer hypersensitivity of cancer cells to targeted small molecule treatment, we utilized drug sensitivity profiles for MEK inhibitor AZD6244 (Yeh et al., 2007), along with matched genomic data from CCLE. Specifically, we used per-sample estimates of 'ActArea' or area under the fitted dose response curve, a metric that has been shown to accurately capture drug response behavior (Jang et al., 2014), to rank cell lines from high to low sensitivity, as well as data comprising somatic mutations and SCNAs as the binary feature matrix (see section "Methods"). CaDrA was then run to look for a subset of features associated with increased sensitivity to treatment with AZD6244 (i.e., increased ActArea scores).

The resulting feature set (i.e., meta-feature) is shown in **Figure 3**. Remarkably, CaDrA selected the $BRAF^{V600E}$ and *NRAS* somatic mutations in the first two iterations, respectively. Subsequent iterations identified mutations in *APAF1*, *TGFBR2,* and *AMHR2*, before terminating the search process ($P \leq 0.001$). APAF1 is a pro-apoptotic factor and known regulator of cell survival and tumor development (Ferraro et al., 2003), the depleted expression of which has been observed in malignant melanoma cell lines and specimens (Soengas et al., 2006). TGFBR2 and AMHR2 are both type II receptors functioning as part of the transforming growth factor (TGF)/bone morphogenetic protein (BMP) superfamily, together serving as mediators of cellular differentiation, proliferation and survival, and play important roles in directing epithelial-mesenchymal transition (EMT) (Rojas et al., 2009; Stone et al., 2016). Notably, MAPK signaling activity can also be regulated by TGF/BMP stimulation (Derynck and Zhang, 2003; Moustakas and Heldin, 2005; Chapnick et al., 2011), suggesting that these mutations are potential independent drivers of increased MEK signaling, and hence, of increased sensitivity to treatment with AZD6244. We next extended our analysis of cancer cell line sensitivity profiles to alternative small molecules targeting MEK (PD-0325901), as well as RAF (PLX4720 and RAF265). The meta-features associated with increased sensitivity to each of the four drug treatments assessed are shown in **Supplementary Figure S3** and summarized in **Table 2**. Importantly, both $BRAF^{V600E}$ and NRAS mutations were identified as candidate drivers of sensitivity to MEK inhibition by AZD6244 and PD-0325901. Furthermore, the $BRAF^{V600E}$ mutation was returned by CaDrA for all four independent queries, highlighting its association with increased sensitivity to inhibitors targeting the same protein (BRAF) as well as its downstream effector (MEK).

Collectively, these results confirm CaDrA's capability to accurately identify upstream drivers of cellular response to treatment that are both components of independently linked pathways, as well as part of the same signaling branch, which in turn suggests their role in driving the disease state of interest.

CaDrA Identifies Hallmark Drivers Associated With Protein Biomarkers in Human Cancers

Protein abundance levels have widely been utilized to histologically classify several human tumor subtypes, with

TABLE 2 | Summary of mutation subsets identified by CaDrA as associated with elevated Mek and Raf inhibition in cancer cell lines.

Target	Treatment	CaDrA hits	*P*-value
MEK	AZD6244	*BRAF.V600E, NRAS, APAF1, TGFBR2, AMHR2*	0.001
MEK	PD-0325901	*BRAF.V600E, NRAS, TRIM33*	0.001
RAF	PLX4720	*BRAF.V600E*	0.001
RAF	RAF265	*TTK, BRAF.V600E, ZMYM2, IL21R, BCL11B, MAP3K5, TAF15*	0.005

Mutation meta-features identified as associated with increased sensitivity to inhibitors targeting Mek (AZD6244, PD-0325901) and Raf (PLX4720) are shown, along with the corresponding permutation p-value of each search result.

relevant diagnostic and therapeutic implications. Epidermal Growth Factor Receptor (EGFR) expression, for instance, together with *EGFR* mutation status can be used to predict response to existing anti-EGFR treatments in patients with lung cancers (Pao et al., 2004; Mascaux et al., 2011). To demonstrate CaDrA's targeted search mode when identifying genomic alterations that track with a pre-defined starting feature, we ran CaDrA using phosphorylated EGFR ($EGFR^{Tyr1068}$) protein expression levels to stratify TCGA lung adenocarcinomas (LUAD), and seeded the search process with EGFR mutations. Subsequent search iterations selected well-known regulators of EGFR activity in lung cancers, including mutations in epithelial-to-mesenchymal transition mediators *SMAD4* and *LAMC2*, as well as *ERBB2* (Liu et al., 2015; Moon et al., 2015), with the meta-feature being statistically significant based on the permuted null background obtained for the same search criterion ($P \leq 0.02$; **Supplementary Figure S4**).

We then wished to more systematically determine whether CaDrA can identify known drivers of target profiles previously associated with oncogenic and tumor-suppressive markers in human cancers. To do so, we queried TCGA expression profiles of proteins encoded by a set of hallmark genes that are defined in the COSMIC database (Forbes et al., 2017), along with genomic data from nine different cancer types in TCGA (Forbes et al., 2017). Briefly, for each cancer type, a CaDrA query was performed with respect to each of the proteins corresponding to the COSMIC-defined oncogenes or tumor suppressor genes (n = 57). In particular, CaDrA was applied to search for sets of genomic features associated with elevated protein expression for each protein under consideration. The features selected by CaDrA were then pooled across all protein queries, and the resulting feature set was tested for enrichment against the reference COSMIC list of frequently mutated oncogenes and tumor suppressor genes (n = 554; see section "Methods"). We observed a significant enrichment of the reference cancer driver mutations among the CaDrA-identified features in all cancer types tested (Hyper-enrichment FDR < 0.05; **Figure 4** and **Supplementary Table S1**). These results validate CaDrA's ability to identify independently cataloged, functionally relevant genomic drivers in primary human malignancies.

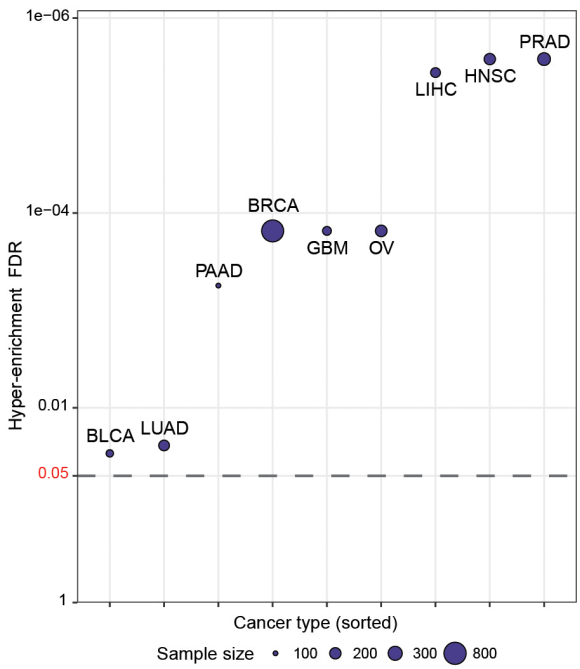

FIGURE 4 | CaDrA systematically identifies known drivers of onco-proteins and tumor suppressor proteins in human cancers. TCGA genomic data for nine different cancer types were queried using the expression of distinct proteins mapping to hallmark genes included in COSMIC (n = 57) for sample ranking. Resulting meta-features identified by CaDrA were then pooled across all protein queries and tested for enrichment against a reference COSMIC-defined gene list (n = 554). FDR-adjusted gene set enrichment p-values are shown, with cancer types sorted in decreasing order of FDR q-value. BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinomas; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma. Points are plotted in $-\log_{10}$ space.

CaDrA Reveals Novel Drivers of Oncogenic YAP/TAZ Activity in Human Breast Cancer

Next, we tested whether our framework can be applied to the discovery of novel drivers of oncogenic pathways in cancer. The Hippo signaling pathway is a highly conserved developmental pathway known to play an essential role in cell proliferation and survival (Varelas, 2014). YAP (Sudol, 1994), and TAZ (Kanai et al., 2000) serve as central downstream transcriptional effectors of the pathway. Aberrant nuclear YAP/TAZ localization and transcriptional activity is associated with a range of cancers, including BRCAs (Hiemer et al., 2015; Moroishi et al., 2015; Zanconato et al., 2015, 2016). To identify alternative genetic events that can potentially explain

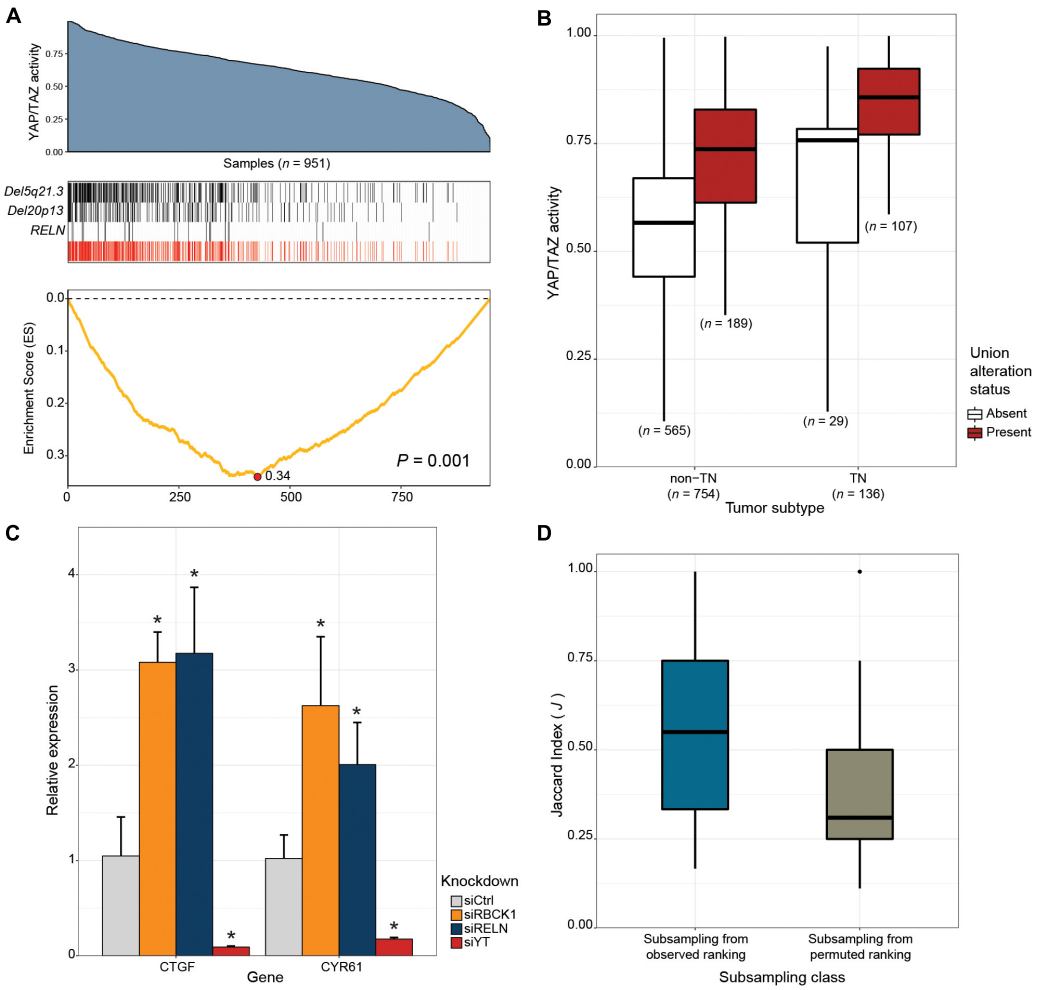

FIGURE 5 | CaDrA identifies novel drivers of oncogenic YAP/TAZ activity in human breast carcinomas. **(A)** TCGA BRCA RNASeq data (n = 951) was projected onto the space of YAP/TAZ-activating genes (blue area plot; see section "Methods"). A CaDrA search for features associated with elevated YAP/TAZ activity identified two chromosomal deletions (*Del5q21.3*, *Del20p13*), and a somatic mutation in *RELN* (black tracks). The union of the three features (red track) and the corresponding running enrichment score (ES) is also shown. **(B)** Box plot of YAP/TAZ activity estimates for triple negative (TN) and non-TN TCGA BRCA samples. Sample groups are further stratified by the presence or absence of the union alteration status of the meta-feature identified by CaDrA (panel a, red track). Only samples with known TN status were considered **(C)** siRNA-mediated knockdown of 20p13-harboring gene *RBCK1*, and *RELN* in HS578T cells resulted in significant increase in the expression levels of canonical YAP/TAZ targets CTGF and CYR61, as indicated by their relative qRT-PCR expression, confirming the identified CaDrA hits as potential regulators of BRCA-associated YAP/TAZ activity. **(D)** Sub-sampling-based reproducibility assessment for candidate drivers of YAP/TAZ activity compared to a CaDrA query for a random profile ranking in TCGA BRCAs. Jaccard (J) indices of the returned meta-features obtained with and without sub-sampling (repeated for n = 100 independent sub-sampling iterations) were computed and compared for the two queries, yielding a significantly higher J index distribution for the original query relative to the permuted ranking query (Wilcox $P < 0.0001$). Ctrl: Scrambled control; YT: YAP/TAZ; * FDR < 0.05; two-tailed Student's t-test.

the elevated YAP/TAZ activity exhibited in some human breast cancers, we applied CaDrA using genomic data from the TCGA BRCA sample cohort, along with corresponding per-sample estimates of YAP/TAZ activity derived using a gene expression signature of YAP/TAZ knockdown in MDA-MB-231 cells (see section "Methods"). Samples with available RNASeq, somatic mutation and SCNA profiles (n = 957) were first ranked in decreasing order of their overall YAP/TAZ activity estimates. The ranked binary matrix of mutation and SCNA features were then used as input to CaDrA. In the first iteration, CaDrA identified the top scoring genomic feature to be a deletion on chromosomal locus chr5q21.3 (**Figure 5A**), harboring tyrosine kinase receptor-encoding gene *EFNA5*. *EFNA5*, a member of the Eph receptor family, has been hypothesized to function as a tumor suppressor, whose expression has been shown to be reduced in human BRCAs relative to normal epithelial tissue (Fu et al., 2010). Advancing to a second iteration, CaDrA then identified an additional deletion of chr20p13 as the next-best feature (**Figure 5A**). The chr20p13 genomic deletion spans multiple genes (**Supplementary Table S2**), including *RBCK1*, whose reduced expression has been shown to be associated with increased tumor cell proliferation and survival, as well as with poor prognosis in breast cancer (Donley et al., 2014). CaDrA then proceeded to identify somatic mutations in the

RELN gene, before terminating the search process ($P \leq 0.001$; **Figure 5A**). Loss of *RELN* expression has indeed been shown to induce cell migration in esophageal carcinoma, and to be associated with poor prognosis in breast cancer (Stein et al., 2010; Yuan et al., 2012). To ensure that the derived meta-feature association is not a spurious consequence of correlation with tumor subtype, we tested for the association of YAP/TAZ activity with the meta-feature while controlling for BRCA TN status using a linear regression model. The results confirmed that the positive association between YAP/TAZ activity and the occurrence of these genomic alterations is independent of BRCA patho-histology (linear regression meta-feature coefficient $P < 0.0001$; **Figure 5B**). Analysis of YAP/TAZ activity based on the same knockdown signature in CCLE BRCA cell lines ($n = 59$; **Supplementary Figure S5A**) shows that *RBCK1* and *RELN* display the highest anti-correlation between their gene expression and YAP/TAZ activity (**Supplementary Figure S5B**). In order to assess whether these identified candidates indeed drive the elevated YAP/TAZ activity phenotype, we performed siRNA-mediated knockdown of *RELN* or *RBCK1* in HS578T breast cancer cells, followed by expression quantification of YAP/TAZ canonical targets, which serves as a read-out of nuclear YAP/TAZ activity (Piccolo et al., 2014). HS578T cells which, similar to MDA-MB-231 cells from which the gene signature was derived, are TN BRCA cells but display lower overall YAP/TAZ activity (rank 7/59) compared to the latter (rank 54/59). Importantly, knockdown of either of these candidate drivers in these cells yielded a significant increase in expression levels of YAP/TAZ targets CTGF and CYR61 (FDR < 0.05; two-tailed Student's *t*-test), validating the association of their loss of function with increased YAP/TAZ transcriptional activity (**Figure 5C**).

Thus, application of CaDrA to the analysis of YAP/TAZ activity in primary BRCA samples identified multiple new candidate drivers, with *in vitro* validation confirming the causal role of the top two candidates, *RBCK1* and *RELN*, in driving this activity. These results highlight our tool's ability to discover novel oncogenic genomic drivers.

Evaluation of CaDrA Reproducibility

Next, we sought to determine CaDrA's reproducibility, and how this may be influenced by the statistical significance of the returned meta-feature (as determined by permutation *p*-value). To do so, we implemented a sub-sampling procedure and applied it to the search for YAP/TAZ activity drivers in TCGA BRCAs. Specifically, the original meta-feature returned by the search on the full dataset, and the meta-feature returned when performing the same search on a random subset (80%) of samples were compared by the Jaccard (*J*) index (see section "Methods"). We performed this sub-sampling search procedure both with respect to the original sample ranking (**Figure 5A**), and with respect to a permuted sample ranking ($n = 100$ iterations each). Comparison of the resulting *J* index distributions yielded a significantly higher reproducibility of results when sub-sampling from the original sample ranking, than from the randomly permuted one (Wilcox $P < 0.0001$; **Figure 5D**). These results support the conclusion that the CaDrA-based significance testing is a strong predictor of a search result reproducibility, and a rigorous criterion to discriminate between true and false positives.

To systematically validate this conclusion, we extended the sub-sampling analysis to CaDrA queries of protein expression profiles across the nine different cancer types previously described. Briefly, for each cancer type we assessed whether the meta-features corresponding to the top five most-significant CaDrA protein queries (CaDrA $P \leq 0.05$) were more reproducible than those corresponding to a randomly selected subset of five non-significant protein queries (CaDrA $P > 0.05$). To this end, the *J* index distribution obtained upon sub-sampling from the significant queries ($n = 100$ iterations each) was compared to the equivalent distribution from the non-significant queries, and a significantly higher reproducibility of the former was observed in all nine cancer types tested (Wilcox FDR < 0.001; **Figure 6**).

Taken together, these results show that CaDrA-based significance testing is a strong predictor of a search result reproducibility. Most importantly, it provides for a statistically rigorous decision rule, which would not be available based on the sub-sampling results alone.

DISCUSSION

Identifying (epi)genetic drivers of molecular readouts is of fundamental importance to determining alternative mechanisms influencing the phenotype in question. Existing methods attempting to extract functionally relevant sets of genomic alterations associated with a given context either do not support the analysis of data beyond somatic mutations, do not incorporate multiple feature scoring functions and search modes, or do not implement rigorous statistical significance testing of the obtained results. Importantly, a computational framework package bundling all of these features does not exist, and can significantly help identify novel drivers of signature activity.

Here, we presented CaDrA as a tool that determines the subset of queried binary features most associated with a phenotypic signature of interest by specifically exploiting a stepwise heuristic search method. CaDrA was applied to identify both known and novel genomic drivers of sample signature activity, comprising drug sensitivity, protein expression and gene set activity estimates, using publicly available multi-omics datasets from cancer cell lines and primary tumors. Querying CCLE data for features associated with increased sensitivity to Mek/Raf inhibitors, CaDra recovered known driver mutations in oncogenes known to be gate-keepers of MEK pathway activity, including *NRAS* and *BRAF*. Importantly, $BRAF^{V600E}$ mutations account for >90% of BRAF mutations and is generally found to be mutually exclusive to *NRAS* mutations (Sensi et al., 2006; Cantwell-Dorris et al., 2011), as also observed in the CCLE, highlighting CaDrA's ability to identify features exhibiting mutual exclusivity. Further, the large-scale investigation of expression profiles of annotated hallmark proteins in tumors from nine different cancer types in TCGA confirmed CaDrA's ability to systematically identify known mutations of oncogenes and

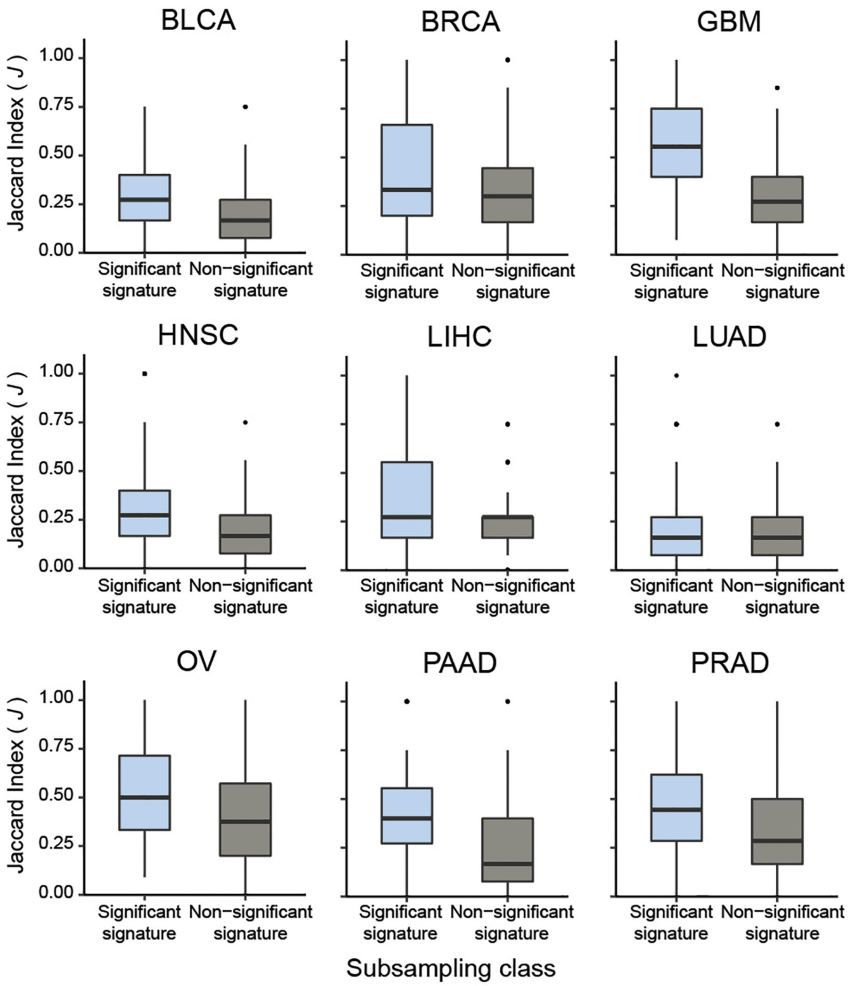

FIGURE 6 | Pan-cancer sub-sampling analysis confirms agreement between CaDrA search significance and reproducibility of identified meta-features. CaDrA was applied to search for genomic alterations associated with elevated protein expression for all proteins profiled using RPPAs, for nine different cancer types in TCGA. Reproducibility by sub-sampling was then assessed for the top 5 significant (CaDrA $P \leq 0.05$), and 5 non-significant (CaDrA $P > 0.05$) protein queries (see text). Consistency of CaDrA results was computed by the Jaccard (*J*) index of the returned meta-feature obtained with and without sub-sampling for each iteration, with the *J* indices pooled for the 5 significant and non-significant results, respectively. Box plots highlight a significantly higher *J* index coefficient among the significant protein queries compared to the non-significant queries across all cancer types investigated (Wilcox FDR < 0.001).

tumor suppressor genes in human cancers, as defined in the COSMIC database.

Through our extensive evaluation on simulated data, we were able to highlight CaDrA's high sensitivity for mid-to-large sized datasets ($N > 90$), and high specificity for all sample sizes considered. Importantly, multi-omics datasets produced by networks such as CCLE and TCGA, also presented in this study, are well above this sample size limit. CaDrA's specificity was further evident when querying genetic drivers of increased sensitivity to treatment with PLX4720, a potent and selective inhibitor designed to preferentially inhibit active B-Raf protein bearing the V600E allele (Tsai et al., 2008). In this scenario, the search process correctly identified the $BRAF^{V600E}$ mutation as the sole feature associated with elevated sensitivity to treatment, in agreement with the known specificity of the small molecule inhibitor, with the feature association being highly statistically significant. It is important to emphasize that the evaluation of CaDrA's sensitivity and specificity crucially relied on the statistical testing procedure we defined, a feature missing in most of the other existing methods.

We were also able to demonstrate the utility of our framework in the discovery of novel drivers in human breast cancers. Specifically, we asked whether there were genomic alterations associated with elevated activity of Hippo pathway co-activators YAP/TAZ, known to control pro-tumorigenic signals in multiple cancer types (Hiemer et al., 2015; Moroishi et al., 2015; Zanconato et al., 2016). The mechanisms contributing to dysregulated YAP/TAZ activity in cancer remain poorly understood. To date, very few genomic alterations have been associated with driving tumorigenic YAP/TAZ activity (Harvey et al., 2013). Our CaDrA search with respect to a sample ranking of decreasing YAP/TAZ activity, as measured by the coordinated

expression of YAP/TAZ-activated genes, yielded a meta-feature consisting of chromosomal deletions of 5q21.3 and 20p13, and mutations in the *RELN*. Subsequent functional validation by knockdown of select targets, namely RELN and RBCK1, in HS578T BRCA cells exhibiting low YAP/TAZ-activity resulted in a significant increase in the expression of canonical YAP/TAZ targets CTGF and CYR61. These results confirmed the selected targets' involvement in the regulation of YAP/TAZ-mediated activity, and the capability of CaDrA to identify new drivers of pathway activity. Importantly, this case study highlights the capability of the method to integrate information, and discover targets pertaining to multiple DNA alteration types.

A sub-sampling-based assessment of CaDrA's results show that the ability to recover reproducible meta-features was higher for the true (significant) YAP/TAZ activity ranking, compared to a randomly permuted sample ranking. This sub-sampling procedure was independently assessed using a systematic pan-cancer comparison of reproducibility results from significant and non-significant protein queries, which revealed a significantly higher concordance of the former compared to the latter in all cases tested. Together, these results confirm the agreement between the estimated permutation *p*-values and the reproducibility of the meta-features identified by CaDrA, and emphasize the importance of our statistical testing procedure in supporting normative decision making.

Previously developed methods have indeed been shown to aid in the selection of functionally relevant genomic features in cancer (Ciriello et al., 2012; Vandin et al., 2012; Leiserson et al., 2013, 2015; Kim et al., 2016). However, CaDrA is to our knowledge the only method performing *rank-based* prediction in this context, which we believe is well-suited to: (i) model the noisy relationship between (epi)genetic alterations and a functional readout, and (ii) privilege the accurate prediction of highly ranked samples over lowly ranked samples, a desirable feature when modeling oncogenic activity. Furthermore, the framework as defined is flexible enough such that non-rank-based scoring functions can be easily incorporated. We emphasize that using rank-based scoring functions, while advantageous for the reasons mentioned, rely on accurate stratification of samples based on the dependent variable to yield concordant associations for a given biological question. Thus, the soundness of predictions is dependent on the quality of signatures used to query the target profile of interest.

The method that most-resembles CaDrA in its approach is REVEALER (Kim et al., 2016), an iterative search algorithm that functions in a similar fashion to CaDrA, while specifically seeking only those features that are mutually exclusive given the sample context. We note that a direct and rigorous comparison between CaDrA and REVEALER was not possible given the lack of a formal procedure to estimate statistical significance of results in the latter. We further emphasize that our tool defines a flexible framework capable of incorporating additional feature scoring functions, including the mutual information criterion implemented in REVEALER. Indeed, the incorporation of such scoring functions would benefit from the statistical significance estimation module built into CaDrA.

Current implementations of CaDrA and other similar methods are limited to the use of summarized input genomic features that are treated as binary events, denoting the presence or absence of a given mutation or SCNA in a sample. As we have demonstrated, this summarization approach is indeed sufficient to identifying genomic feature sets that may drive the target profile of interest. However, since different types of point mutations (missense, truncating, etc.) may impose differing functional impacts in oncogenes versus tumor suppressor genes, we surmise that these methods could be further improved by qualitatively differentiating between the different types of alterations being considered. One possibility would be to separate mutations by predicted gain or loss-of-function, as well as to distinguish between low (1) and high ($\geq$2) DNA copy number gains or losses, although this may lead to excessive sparsity in the input matrix for low-frequency point mutations and SCNAs.

While our evaluations focused on somatic mutations and SCNAs, CaDrA's search functionality can be applied to additional sequencing readouts capturing regulatory features, including and not limited to, DNA methylation and microRNA expression, albeit with proper discretization of these continuous features. A joint analysis of these additional data types might provide insight into epigenetic mechanisms that complement the assessed genetic features in driving phenotypic variation. Furthermore, we envision the adoption of CaDrA for the study of germ-line variation as well, thus contributing to move beyond the "one feature at a time" paradigm typical of GWAS studies, although issues of computational efficiency in that problem space will likely become more challenging.

CONCLUSION

CaDrA enables the efficient identification of subsets of genomic features, including somatic mutations and SCNAs, as candidate drivers of a pre-defined phenotypic variable. Given the rapid rise in the availability of multi-omics datasets, as well as an increased need to interrogate targeted molecular readouts within these contexts, we believe that our methodology will accelerate feature prioritization for further follow-up and consideration, in turn aiding in the discovery of potential drivers of the phenotype of interest. Thus, we propose CaDrA as a tool for both targeted hypotheses testing, and novel hypothesis generation.

METHODS

The CaDrA Algorithm

An overview of CaDrA's workflow is summarized in **Figure 1**. CaDrA takes as input the sample ranking induced by a sample-specific measurement, a matrix of binary features (1/0 indicating the presence/absence of a given feature in a sample), and a scoring method specification to measure the significance of the concordance between the occurrence of alteration events and the defined sample ranking. The pre-defined sample ranking can be based on quantitative estimates of a gene expression, a signature or pathway activity, or other experimentally derived

measurements. Each row in the matrix of binary features denotes the presence or absence of a somatic alteration (mutation, CNA, or other) in each of the samples in the ranked cohort. The score function is a measure of the *left-skewness* of a binary vector with respect to the sample ranking. The more the occurrences of an alteration are skewed toward higher rankings (i.e., the more the 1's in the feature vector are skewed toward the left), the higher the score. The scores currently implemented are the KS test (default), and the Wilcoxon rank-sum test, but additional scoring functions can easily be added.

Given the sample ranking, the matrix of binary features, and the score of choice (KS or Wilcoxon), CaDrA implements a step-wise greedy search: it begins by first selecting the single feature that maximizes the score (Step 1; **Figure 1**). It then generates the union (logical OR) of this starting feature with every other remaining feature in the dataset and computes scores for the obtained 'meta-features' (Step 2; **Figure 1**); it selects a 2nd feature that, added to the first (as a union), maximally increases the score – which will then serve as the new top reference hit (Step 3; **Figure 1**). Repeating this process until no further improvement to the cumulative score can be attained, the search output is a set of features (i.e., a meta-feature) whose union has the (local) maximum skewness score with respect to the input sample ranking. The significance of a CaDrA search and its cumulative score are determined by generating an empirical null distribution of scores based on the exact same data and search parameters, but with randomly permuted sample rankings, providing a permutation *p*-value per search result. Since the CaDrA algorithm specifically returns feature-sets maximally left-skewed given the provided sample ranking variable, it can be applied to identify features that are either positively correlated or anti-correlated with the continuous variable of interest by ranking samples in decreasing or increasing order of that variable, respectively.

CaDrA Features

Search Modes

CaDrA supports multiple search modalities: it allows for the selection of a user-specified feature from which to start the search (rather than selecting the feature with highest score as depicted in Step 1 of **Figure 1**); alternatively, since the greedy search is not guaranteed to find the global maximum, it also allows for a "top-N" search modality, whereby the search is started from each of the first N features (as measured by their individual skewness scores), and the result of the best search can be determined by selecting the set of features with the best cumulative score over the top-N runs.

Visualization of Search Results

For a given search, CaDrA outputs a set of features (meta-feature), which can be visualized as a 'meta-plot'. This includes (panels from top to bottom): an area plot of the sample-specific measurements used to obtain the sample ranks; a color-coded matrix of all features in the meta-feature (in the step-wise order that they were added), one feature per row, with the corresponding union of the meta-feature (red) last; and a corresponding enrichment score (ES) plot below. Additionally, top-N search results can be visualized for overlapping features to evaluate robustness across different search starting points.

Parallelization Support

The generation of the empirical null distribution for significance testing is typically done for $\geq$500 iterations (i.e., permuted sample ranks). In order to speed up this potentially time-consuming task, CaDrA supports exploiting parallel computing with the help of the parallel R package functionality, should multiple compute cores be available to users.

Permutation Caching

Since the generation of the null distribution used for significance testing is a time-consuming step, and since the null distribution of scores depends solely on the feature dataset and the search parameters specified (scoring method, starting feature versus top-N search mode etc.), and not on the input sample ranking, we can implement cacheing of the null distribution corresponding to each dataset and search parameters. When submitting multiple subsequent queries (each with its own sample ranking) that utilize the same dataset and search criteria, CaDrA can then fetch the corresponding cached null distribution to generate permutation *p*-values almost instantaneously, avoiding the need for repetitive computation, thus significantly reducing overall query run time.

Data Availability and Processing

CaDrA is freely available for download and use as a documented R package under the git repository https://github.com/montilab/CaDrA, and will further be deposited and maintained for future use under Bioconductor, including complete code and example use-cases.

DNA copy number (GISTIC2), mutation and RPPA data for TCGA analyses were obtained using Firehose v0.4.3 corresponding to the Jan 28th, 2016 (SCNA and somatic mutations) and Jul 15th, 2016 (RPPA) Firehose release. Somatic mutation data was processed at the gene level by assigning either 1 or 0 based on the presence or absence of any given mutation in that gene, respectively (excluding synonymous mutations). Annotated Level 3 RPPA data was used for all protein-related TCGA data queries. For pan-cancer analyses, these three data sets were obtained for nine cancer types, including bladder urothelial carcinoma (BLCA), breast invasive carcinomas (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), and prostate adenocarcinoma (PRAD). RNASeq version 2 data processed as Level 3 RSEM-normalized gene expression values corresponding to the Feb 4th, 2015 Firehose release was used for the TCGA BRCA analysis. CCLE genomic data were downloaded from https://portals.broadinstitute.org/ccle and processed as previously described (Kim et al., 2016). Somatic mutation binary calls per gene were used as is, and SCNA data was processed using GISTIC2 (Mermel et al., 2011) with all default parameters barring the confidence level, which was set to 99%. ActArea estimates pertaining to drug treatment

sensitivity across CCLE samples was used as previously described (Barretina et al., 2012).

In all cases presented, SCNA and somatic mutation data were jointly analyzed as a single input dataset to CaDrA, thereby including samples for which both data were available. All input data to CaDrA were further pre-filtered so as to exclude alteration frequencies below 3% and above 60% to reduce feature sparsity and redundancy, respectively, across samples (CaDrA's default feature pre-filtering settings).

Simulated Data Generation

To evaluate both the sensitivity and specificity of CaDrA, we generated simulated data to represent cases where there was a mix of left-skewed ("true positive") and randomly distributed ("null") features, as well as cases where there were only null features. The left-skewness of a feature is a measure of its association with the sample ranking, since samples are sorted from left (high rank) to right (low rank). The design and parameter specification of the simulated data matrix is shown in **Supplementary Figure S1**. Each feature/row is a binary (0/1) vector, with 1 (0) in the *i*th position denoting the occurrence (non-occurrence) of the genetic event (e.g., SCNA or mutation) in the *i*th sample. This simulation of binary features relies on the following parameters:

N: Dataset sample size (number of columns in the matrix).
n: Total number of features in the dataset (number of rows in the matrix).
p: Number of true positive features generated per dataset [a positive feature is a feature whose distribution of events (i.e., the number of 1's) is significantly associated with the sample ranking, i.e., left-skewed].
f: Left-skew proportion. The proportion of samples that are *cumulatively* left-skewed in the sample ranking.
λ: The mean (and variance) of the Poisson distribution from which the number of events in the null features is sampled. This is equal to the number of 1's per skewed positive feature. A Poisson distribution is used so that we can partially control (through the mean) the number of 1's in a null feature, which are then uniformly distributed across samples (see description of Null feature generation below).

The resulting simulated binary data matrix will consist of two main types of features:

True Positive (TP) Features: A total of *p* TP features are generated. Events (i.e., 1's) are assigned to the TP features in a mutually exclusive fashion, with each of these features having $(f \times N)/p$ entries set to 1, with their cumulative OR yielding an N-sized vector with the left-most $f \times N$ entries set to 1's. For example, if we generate data for 100 samples and 5 positive features, with the left-skew proportion set to 0.5, each non-overlapping feature will have 10 among the 50 left-most entries (columns) set to 1, such that the union (logical OR) of the 5 features will have 1's in the first 50 entries.

Null Features: Null features are generated for a total of $(n–p)$ features. To generate these features, we sample the number of 1's per null feature based on a Poisson distribution with mean parameter $\lambda = (f \times N)/p$. In this fashion, the number of 1's in the null features will have a distribution centered on the corresponding number for the TP features. For instance, if we generate data for 100 samples and 5 TP features with left-skew proportion $f = 0.5$, then each of the TP features will have ten 1's, and each of the remaining 995 null features will have a number of 1's sampled from Poisson ($\lambda = 10$), uniformly distributed over the *N* samples.

A schematic representation of this data, along with the parameters that define its composition is shown in **Supplementary Figure S1**.

Evaluation of CaDrA Performance on Simulated Data

Evaluation of CaDrA performance was performed considering two main scenarios: (a) True positive datasets: Data containing both true positive and null features (where the sensitivity of CaDrA is tested); and (b) Null datasets: Data containing only null features (where the specificity of CaDrA is tested), with the following parameter specifications for data generation:

$N = \{50, 60, 70, 80, 90, 100, 250, \text{and } 500\}$
$n = 1000$
$p = 5$
$f = 0.5$

CaDrA was run using default input parameters, returning a meta-feature which had the best score, along with a permutation *p*-value based on the empirical null search distribution (**Supplementary Figure S2**). These results were then used to determine performance estimates for different sample sizes, composition (i.e., distribution of TP versus null features per returned meta-feature), size (i.e., the number of features within the returned meta-feature) and statistical significance of the returned meta-features. Mean TPR percentages shown in **Table 1** are a result of weight-averaging TPRs corresponding to different number of true positive features per meta-feature, weighted by the total searches returning such meta-features (gray circles **Figure 2C**). Mean FPR percentages shown in **Table 1** are a result of weight-averaging FPRs corresponding to different meta-feature sizes, weighted by the total searches returning such meta-features (gray circles **Figure 2D**).

COSMIC Enrichment Analyses

For enrichment analyses, RPPA protein data for the nine cancer types (see section "Data Availability and Processing") was first restricted to those proteins representing hallmark oncogene or tumor suppressor genes included in the COSMIC v84 database ($n = 57$)[1] (Forbes et al., 2017). For each cancer type, a CaDrA query was then performed with respect to the protein expression-induced sample ranking, using somatic mutation and copy number alteration data as input features, in order to search

[1]https://cancer.sanger.ac.uk/census

for features associated with elevated protein expression of each of the hallmark proteins queried. The features selected thereof were then pooled across all queries, and the resulting gene list tested for significant enrichment (based on the hyper-geometric distribution) with respect to a set of annotated oncogenes and tumor suppressor genes in COSMIC (n = 554), compared to the pooled list of non-selected features.

Sub-Sampling Analyses

For all sub-sampling analyses presented, CaDrA was run after sub-sampling 80% of the original data, with consistency of CaDrA results computed as the Jaccard (J) index of the returned meta-feature obtained with and without sub-sampling (repeated for n = 100 independent sub-sampling iterations). To assess reproducibility of drivers associated with YAP/TAZ activity, the search was repeated by either preserving the observed ranking (decreasing YAP/TAZ activity), or by taking a permuted ranking. J indices were then compared between the original and permuted ranking cases using a Wilcox rank sum test. For the pan-cancer protein query analysis, all available proteins profiled as part of the RPPA data were used, with J indices similarly computed for the top 5 protein queries that yielded significant meta-features ($P \leq 0.05$), and 5 queries randomly selected from the non-significant list ($P > 0.05$) in each cancer type. J indices were then pooled for the five significant, and non-significant results, respectively, and compared using a Wilcox rank sum test. FDR correction was used for all pan-cancer analyses tests of significance.

YAP/TAZ Signature Projection and Assessment in TCGA BRCAs

A signature comprising YAP/TAZ-activating genes (n = 717) in MDA-MB-231 cells was obtained based on a previous study (Enzo et al., 2015). The TCGA BRCA RNASeq data (n = 1,186 samples) was projected onto the signature genes and per-sample estimates of YAP/TAZ activity were derived using ASSIGN (Shen et al., 2015), which was then used as a continuous ranking variable with CaDrA. The association of YAP/TAZ activity with the CaDrA-derived meta-feature, and with BRCA subtype (i.e., TN status) was determined using a linear regression model.

Cell Culture, siRNA Knockdown and qRT-PCR

HS578T BRCA cells were purchased from ATCC and cultured using media and conditions suggested by ATCC. For RNA interference, cells were transfected using RNAiMAX (Thermo Fisher) with control siRNA (Qiagen, 1027310) or an equal molar mixture of siRNA targeting RELN (Sigma), RBCK1 (Sigma), or TAZ and YAP (Hiemer et al., 2014). 48 h post transfection, RNA was extracted from cells using RNeasy kit (Qiagen) and the synthesis of cDNA was performed as previously described (Hiemer et al., 2014). Quantitative real-time PCR (qRT-PCR) was performed using Taqman Universal master mix II (Thermo Fisher) and measured on ViiA 7 real-time PCR system. Taqman probes used included those recognizing CTGF (Thermo Fisher Hs00170014_m1), CYR61 (Thermo Fisher Hs00155479_m1), RELN (Thermo Fisher Hs01022646_m1), RBCK1 (Thermo Fisher Hs00934608_m1), WWTR1 (Thermo Fisher Hs01086149_m1), and YAP (Thermo Fisher Hs00902712_g1) and GAPDH (Thermo Fisher 4326317E). Expression levels of each gene were calculated using the $\Delta\Delta$Ct method and normalized to GAPDH. Knockdown efficiency of YAP, TAZ, RELN, and RBCK1 was verified for each experiment. Mean transcriptional knockdown of YAP, TAZ, and RBCK in HS578T cells was >80%. Basal RELN levels in HS578T cells were low, and relative knockdown in these cells was 28.3% ($\pm$14.1). Data from qRT-PCR experiments are shown as mean $\pm$ S.D., with each knockdown compared with respect to the scrambled siRNA control (siCtl) using an unpaired, two-tailed Student's t-test.

CaDrA Search Parameters

For evaluation using genomic data, CaDrA was run in the top-N mode using the default of N = 7, choosing the best resulting meta-feature (see section "Methods"; CaDrA features: Search modes). For evaluation of simulated data, only the top-scoring feature was considered as a starting feature per search run (i.e., N = 1). The "ks" method was chosen for evaluating skewness of features at each step in all cases presented. All other default input search parameters were used for all cases presented.

AVAILABILITY OF DATA AND MATERIAL

The datasets generated and/or analyzed during the current study are available in the TCGA repository (https://tcga-data.nci.nih.gov/docs/publications/tcga), and CCLE repository (https://portals.broadinstitute.org/ccle), and are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

VK developed the R package and conducted the analyses. VK and SM wrote the manuscript, with input from PS and XV. JK performed the siRNA and qRT-PCR experiments. LZ assisted in obtaining the gene expression signature for TCGA data projection. PS assisted in the evaluation of CaDrA on simulated data. SM and VK designed the CaDrA framework and features, and interpreted the results. XV designed the experimental validation of novel candidate drivers, and interpreted the results thereof. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Institutes of Health NIDCR fellowship F31 DE025536 (VK), CDMRP grant W81XWH-14-1-0336 (XV), the Dahod breast cancer research program at Boston University School of Medicine (XV and SM), as well as the Clinical and Translational Science Institute (supported by Clinical and Translational Research Award CTSA grant UL1-TR001430) at Boston University School of Medicine (SM).

The funding sources played no role in the design of the study and collection, analysis, and interpretation of data and in the writing of this manuscript.

ACKNOWLEDGMENTS

We would like to thank Joshua Klein for making suggestions toward the implementation of specific package features. We further acknowledge dbGap for granting access to the TCGA data (phs000178.v9.p8).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00121/full#supplementary-material

REFERENCES

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., et al. (2015). The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Bea, S., Zettl, A., Wright, G., Salaverria, I., Jehn, P., Moreno, V., et al. (2005). Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression – based survival prediction. *Hematology* 106, 3183–3190. doi: 10.1182/blood-2005-04-1399

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357. doi: 10.1038/nature04296

Burotto, M., Chiou, V. L., Lee, J. M., and Kohn, E. C. (2014). The MAPK pathway across different malignancies: a new perspective. *Cancer* 120, 3446–3456. doi: 10.1002/cncr.28864

Cantwell-Dorris, E. R., O'Leary, J. J., and Sheils, O. M. (2011). BRAFV600E: implications for carcinogenesis and molecular therapy. *Mol. Cancer Ther.* 10, 385–394. doi: 10.1158/1535-7163.MCT-10-0799

Cargnello, M., and Roux, P. P. (2011). Activation and function of the MAPKs and their substrates, the MAPK-activated protein kinases. *Microbiol. Mol. Biol. Rev.* 75, 50–83. doi: 10.1128/MMBR.00031-10

Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., et al. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* 364, 2507–2516. doi: 10.1056/NEJMoa1103782

Chapnick, D. A., Warner, L., Bernet, J., Rao, T., and Liu, X. (2011). Partners in crime: the TGFβ and MAPK pathways in cancer progression. *Cell Biosci.* 1:42. doi: 10.1186/2045-3701-1-42

Chen, J. C., Alvarez, M. J., Talos, F., Dhruv, H., Rieckhof, G. E., Iyer, A., et al. (2014). Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* 159, 402–414. doi: 10.1016/j.cell.2014.09.021

Cho, A., Shim, J. E., Kim, E., Supek, F., Lehner, B., and Lee, I. (2016). MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* 17:129. doi: 10.1186/s13059-016-0989-x

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., et al. (2015). Pathway and network analysis of cancer genomes. *Nat. Methods* 12, 615–621. doi: 10.1038/nmeth.3440

Daemen, A., Griffith, O. L., Heiser, L. M., Wang, N. J., Enache, O. M., Sanborn, Z., et al. (2013). Modeling precision treatment of breast cancer. *Genome Biol.* 14:R110. doi: 10.1186/gb-2013-14-10-r110

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111.22

Derynck, R., and Zhang, Y. E. (2003). Smad-dependent and Smad-independent pathways in TGF-β family signalling. *Nature* 425, 577–584. doi: 10.1038/nature02006

Donley, C., McClelland, K., McKeen, H. D., Nelson, L., Yakkundi, A., Jithesh, P. V., et al. (2014). Identification of RBCK1 as a novel regulator of FKBPL: implications for tumor growth and response to tamoxifen. *Oncogene* 33, 3441–3450. doi: 10.1038/onc.2013.306

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6388–6393. doi: 10.1073/pnas.1219651110

Enzo, E., Santinon, G., Pocaterra, A., Aragona, M., Bresolin, S., Forcato, M., et al. (2015). Aerobic glycolysis tunes YAP/TAZ transcriptional activity. *EMBO J.* 34, 1349–1370. doi: 10.15252/embj.201490379

Ferraro, E., Corvaro, M., and Cecconi, F. (2003). Physiological and pathological roles of Apaf1 and the apoptosome. *J. Cell. Mol. Med.* 7, 21–34. doi: 10.1111/j.1582-4934.2003.tb00199.x

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783. doi: 10.1093/nar/gkw1121

Fu, D.-Y., Wang, Z.-M., Wang, B.-L., Chen, L., Yang, W.-T., Shen, Z.-Z., et al. (2010). Frequent epigenetic inactivation of the receptor tyrosine kinase EphA5 by promoter methylation in human breast cancer. *Hum. Pathol.* 41, 48–58. doi: 10.1016/j.humpath.2009.06.007

Harvey, K. F., Zhang, X., and Thomas, D. M. (2013). The Hippo pathway and human cancer. *Nat. Rev. Cancer* 13, 246–257. doi: 10.1038/nrc3458

Heiser, L. M., Sadanandam, A., Kuo, W., Benz, S. C., Goldstein, T. C., Ng, S., et al. (2011). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2724–2729. doi: 10.1073/pnas.1018854108

Hiemer, S. E., Szymaniak, A. D., and Varelas, X. (2014). The transcriptional regulators TAZ and YAP direct transforming growth factor B-induced tumorigenic phenotypes in breast cancer cells. *J. Biol. Chem.* 289, 13461–13474. doi: 10.1074/jbc.M113.529115

Hiemer, S. E., Zhang, L., Kartha, V. K., Packer, T. S., Almershed, M., Noonan, V., et al. (2015). A YAP/TAZ-regulated molecular signature is associated with oral squamous cell carcinoma. *Mol. Cancer Res.* 13, 957–968. doi: 10.1158/1541-7786.MCR-14-0580

Hou, J. P., and Ma, J. (2014). DawnRank: discovering personalized driver genes in cancer. *Genome Med.* 6:56. doi: 10.1186/s13073-014-0056-8

Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., and Margolin, A. A. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac. Symp. Biocomput.* 2014, 63–74. doi: 10.1055/s-0029-1237430

Jia, P., and Zhao, Z. (2014). VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput. Biol.* 10:e1003460. doi: 10.1371/journal.pcbi.1003460

Johnson, D. B., and Puzanov, I. (2015). Treatment of NRAS-mutant melanoma. *Curr. Treat. Options Oncol.* 16:15. doi: 10.1007/s11864-015-0330-z

Kanai, F., Marignani, P. A., Sarbassova, D., Yagi, R., Hall, R. A., Donowitz, M., et al. (2000). TAZ: a novel transcriptional co-activator regulated by interactions with 14-3-3 and PDZ domain proteins. *EMBO J.* 19, 6778–6791. doi: 10.1093/emboj/19.24.6778

Kim, E. K., and Choi, E.-J. (2010). Pathological roles of MAPK signaling pathways in human diseases. *Biochim. Biophys. Acta* 1802, 396–405. doi: 10.1016/j.bbadis.2009.12.009

Kim, J. W., Botvinnik, O. B., Abudayyeh, O., Birger, C., Rosenbluh, J., Shrestha, Y., et al. (2016). Characterizing genomic alterations in cancer by complementary functional associations. *Nat. Biotechnol.* 34, 3–5. doi: 10.1038/nbt.3527

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Leiserson, M. D. M., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9:e1003054. doi: 10.1371/journal.pcbi.1003054

Leiserson, M. D. M., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Liu, J., Cho, S. N., Akkanti, B., Jin, N., Mao, J., Long, W., et al. (2015). ErbB2 pathway activation upon smad4 loss promotes lung tumor growth and metastasis. *Cell Rep.* 10, 1599–1613. doi: 10.1016/j.celrep.2015.02.014

Mascaux, C., Wynes, M. W., Kato, Y., Tran, C., Asuncion, B. R., Zhao, J. M., et al. (2011). EGFR protein expression in non-small cell lung cancer predicts response to an EGFR tyrosine kinase inhibitor - a novel antibody for immunohistochemistry or AQUA technology. *Clin. Cancer Res.* 17, 7796–7807. doi: 10.1158/1078-0432.CCR-11-0209

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Monti, S., Chapuy, B., Takeyama, K., Rodig, S. J., Hao, Y., Yeda, K. T., et al. (2012). Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* 22, 359–372. doi: 10.1016/j.ccr.2012.07.014

Moon, Y. W., Rao, G., Kim, J. J., Shim, H. S., Park, K. S., An, S. S., et al. (2015). LAMC2 enhances the metastatic potential of lung adenocarcinoma. *Cell Death Differ.* 22, 1341–1352. doi: 10.1038/cdd.2014.228

Moroishi, T., Hansen, C. G., and Guan, K.-L. (2015). The emerging roles of YAP and TAZ in cancer. *Nat. Rev. Cancer* 15, 73–79. doi: 10.1038/nrc3876

Moustakas, A., and Heldin, C. H. (2005). Non-Smad TGF-beta signals. *J. Cell Sci.* 118, 3573–3584. doi: 10.1242/jcs.02554

Ng, S., Collisson, E. A., Sokolov, A., Goldstein, T., Lopez-bigas, N., Benz, C., et al. (2012). PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* 28, 640–646. doi: 10.1093/bioinformatics/bts402

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., et al. (2004). EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13306–13311. doi: 10.1073/pnas.0405220101

Piccolo, S., Dupont, S., and Cordenonsi, M. (2014). The biology of YAP/TAZ: hippo signaling and beyond. *Physiol. Rev.* 94, 1287–1312. doi: 10.1152/physrev.00005.2014

Roberts, P. J., and Der, C. J. (2007). Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* 26, 3291–3310. doi: 10.1038/sj.onc.1210422

Rojas, A., Padidam, M., Cress, D., and Grady, W. M. (2009). TGF-B receptor levels regulate the specificity of signaling pathway activation and biological effects of TGF-B. *Biochim. Biophys. Acta* 1793, 1165–1173. doi: 10.1016/j.bbamcr.2009.02.001

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321.e10–337.e10. doi: 10.1016/j.cell.2018.03.035

Savage, K. J., Monti, S., Kutok, J. L., Cattoretti, G., Neuberg, D., De Leval, L., et al. (2003). The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma. *Blood* 102, 3871–3879. doi: 10.1182/blood-2003-06-1841

Sensi, M., Nicolini, G., Petti, C., Bersani, I., Lozupone, F., Molla, A., et al. (2006). Mutually exclusive NRASQ61R and BRAFV600E mutations at the single-cell level in the same human melanoma. *Oncogene* 25, 3357–3364. doi: 10.1038/sj.onc.1209379

Shen, Y., Rahman, M., Piccolo, S. R., Gusenleitner, D., El-Chaar, N. N., Cheng, L., et al. (2015). ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. *Bioinformatics* 31, 1745–1753. doi: 10.1093/bioinformatics/btv031

Soengas, M. S., Gerald, W. L., Cordon-Cardo, C., Lazebnik, Y., and Lowe, S. W. (2006). Apaf-1 expression in malignant melanoma. *Cell Death Differ.* 13, 352–353. doi: 10.1038/sj.cdd.4401755

Stein, T., Cosimo, E., Yu, X., Smith, P. R., Simon, R., Cottrell, L., et al. (2010). Loss of reelin expression in breast cancer is epigenetically controlled and associated with poor prognosis. *Am. J. Pathol.* 177, 2323–2333. doi: 10.2353/ajpath.2010.100209

Stone, A. V., Vanderman, K. S., Willey, J. S., David, L., Register, T. C., Shively, C. A., et al. (2016). Anti-Müllerian hormone signaling regulates epithelial plasticity and chemoresistance in lung cancer. *Cell Rep.* 23, 1780–1789. doi: 10.1016/j.joca.2015.05.020

Sudol, M. (1994). Yes-associated protein (YAP65) is a proline-rich phosphoprotein that binds to the SH3 domain of the Yes proto-oncogene product. *Oncogene* 9, 2145–2152.

Tsai, J., Lee, J. T., Wang, W., Zhang, J., Cho, H., Mamo, S., et al. (2008). Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3041–3046. doi: 10.1073/pnas.0711741105

Vandin, F., Upfal, E., and Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res.* 22, 375–385. doi: 10.1101/gr.120477.111

Varelas, X. (2014). The Hippo pathway effectors TAZ and YAP in development, homeostasis and disease. *Development* 141, 1614–1626. doi: 10.1242/dev.102376

Xi, J., Wang, M., and Li, A. (2017). Discovering potential driver genes through an integrated model of somatic mutation profiles and gene functional information. *Mol. Biosyst.* 13, 2135–2144. doi: 10.1039/c7mb00303j

Yeh, T. C., Marsh, V., Bernat, B. A., Ballard, J., Colwell, H., Evans, R. J., et al. (2007). Biological characterization of ARRY-142886 (AZD6244), a potent, highly selective mitogen-activated protein kinase kinase 1/2 inhibitor. *Clin. Cancer Res.* 13, 1576–1583. doi: 10.1158/1078-0432.CCR-06-1150

Youn, A., and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 27, 175–181. doi: 10.1093/bioinformatics/btq630

Yuan, Y., Chen, H., Ma, G., Cao, X., and Liu, Z. (2012). Reelin is involved in transforming growth factor-B1-induced cell migration in esophageal carcinoma cells. *PLoS One* 7:e31802. doi: 10.1371/journal.pone.0031802

Zanconato, F., Cordenonsi, M., and Piccolo, S. (2016). YAP/TAZ at the roots of cancer. *Cancer Cell* 29, 783–803. doi: 10.1016/j.ccell.2016.05.005

Zanconato, F., Forcato, M., Battilana, G., Azzolin, L., Quaranta, E., Bodega, B., et al. (2015). Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.* 17, 1218–1227. doi: 10.1038/ncb3216

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 March 2019
doi: 10.3389/fgene.2019.00156

Gene Expression-Based Predictive Markers for Paclitaxel Treatment in ER+ and ER– Breast Cancer

Xiaowen Feng[1,2], Edwin Wang[2,3] and Qinghua Cui[1]**

[1]Department of Biomedical Informatics, School of Basic Medical Sciences, MOE Key Lab of Cardiovascular Sciences, Peking University, Beijing, China, [2]Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, [3]Faculty of Medicine, McGill University, Montreal, QC, Canada

OPEN ACCESS

Edited by:
Victor Jin,
The University of Texas Health Science Center at San Antonio, United States

Reviewed by:
Ao Li,
University of Science and Technology of China, China
Tianbao Li,
The University of Texas Health Science Center at San Antonio, United States

***Correspondence:**
Edwin Wang
edwin.wang@ucalgary.ca
Qinghua Cui
cuiqinghua@hsc.pku.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 15 November 2018
Accepted: 13 February 2019
Published: 01 March 2019

Citation:
Feng X, Wang E and Cui Q (2019) Gene Expression-Based Predictive Markers for Paclitaxel Treatment in ER+ and ER– Breast Cancer. Front. Genet. 10:156. doi: 10.3389/fgene.2019.00156

One of the objectives of precision oncology is to identify patient's responsiveness to a given treatment and prevent potential overtreatments through molecular profiling. Predictive gene expression biomarkers are a promising and practical means to this purpose. The overall response rate of paclitaxel drugs in breast cancer has been reported to be in the range of 20–60% and is in the even lower range for ER-positive patients. Predicting responsiveness of breast cancer patients, either ER-positive or ER-negative, to paclitaxel treatment could prevent individuals with poor response to the therapy from undergoing excess exposure to the agent. In this study, we identified six sets of gene signatures whose gene expression profiles could robustly predict nonresponding patients with precisions more than 94% and recalls more than 93% on various discovery datasets (n = 469 for the largest set) and independent validation datasets (n = 278), using the previously developed Multiple Survival Screening algorithm, a random-sampling-based methodology. The gene signatures reported were stable regardless of half of the discovery datasets being swapped, demonstrating their robustness. We also reported a set of optimizations that enabled the algorithm to train on small-scale computational resources. The gene signatures and optimized methodology described in this study could be used for identifying unresponsiveness in patients of ER-positive or ER-negative breast cancers.

Keywords: microarray gene expression profile, breast cancer, signature genes, drug resistance, predictor

INTRODUCTION

Predicting if a given patient would not respond to a specific treatment could save enormous health care resources and potentially make it possible to reallocate the individual to better suited medication programs earlier (Garraway et al., 2013; Collins and Varmus, 2015). Paclitaxel treatment, which targets at cell cycle processes through stabilizing microtubules, is a prevalent medication used in various cancer types including breast, ovarian, and prostate cancer. Up to 20% of the ER-positive (ER+) breast cancer patients, who represent 80% of breast cancer population, could gain survival benefit from paclitaxel treatment. With high-confident prediction, it would be made possible to prevent nearly 20,000 women from ineffective paclitaxel treatment, which might cause additional neurotoxicity and adverse effects, in the United States alone. Network representation learning as well as integration of somatic mutation profile and gene functional annotation

information were utilized to discovery driver genes related to drug treatment responsiveness (Xi et al., 2017, 2018; Yang et al., 2018; Zhang et al., 2018). Existing studies either focused on triple-negative cases, or provided insights on a small number of tipping point genes more biologically other than computationally. For example, ABCB1/PgP and ABCC3/MRP3 were reported to be closely associated with resistance to paclitaxel (Němcová-Fürstová et al., 2016; Delou et al., 2017), while the resistance might be driven by hundreds of genes (Duan et al., 2004). Xu et al. collected 22 key genes involved in paclitaxel treatment resistance for miscellaneous cancer types by analyzing literatures (Xu et al., 2016) with the assistance of GeneMANIA (Warde-Farley et al., 2010), a gene/protein function predicting tool.

In this study, we improved the Multiple Survival Screening (MSS), a methodology developed by Li et al. (2010). for identifying cancer prognostic markers with high robustness and prediction power (Li et al., 2010), and employed it to five microarray gene expression datasets [GSE20194 (MAQC Consortium, 2010; Popovici et al., 2010), GSE20271 (Tabchy et al., 2010), GSE22093 (Iwamoto et al., 2010), GSE23988 (Iwamoto et al., 2010), and GSE25066 (Hatzis, 2011; Itoh et al., 2013)], which were partitioned into discovery set and independent validation set, in search of signature genes of nonresponsiveness in ER+ breast cancer. We discovered sets of such genes that gave precision up to 94.6% and recall rate up to 93.3%, and performed consistently in cross validation inside discovery datasets, and different discovery datasets against their corresponding independent validation datasets. Similar results were obtained for ER-negative patients, demonstrating the prediction power and potential of real-life applications of the optimized methodology and reported gene sets.

RESULTS

Gene Signatures for Unresponsiveness of Paclitaxel Treatment in ER-Positive Breast Cancer

To explore efficient and generalizable gene signatures for predicting of whether a given breast cancer patient should be admitted to paclitaxel treatment, we constructed a discovery dataset comprised of microarray data generated by four cohorts (GSE20271, GSE22093, GSE23988, and GSE25066; referred to as $T1_{pos}$; see Methods for details), where in total 469 patients were acquired (n_{RD} = 418, n_{CR} = 51; RD, residual disease; CR, complete response). Similarly, an independent validation dataset was formed using microarray data from the cohort of GSE20194 (n_{RD} = 213, n_{CR} = 65; referred to as $V1_{pos}$). MAS5 normalization was employed for both $T1_{pos}$ and $V1_{pos}$, respectively. Both expression profile matrices then underwent additional normalizations to address batch effects between the cohorts as well as merging of multiple probes that represented same gene on the gene expression microarray (see Methods).

Implementing a methodology based on Multiple Survival Screening (MSS) (Li et al., 2010), which as a random search computational scheme that could identify reliable signature genes, we obtained six gene signatures ("Signatures," A_1–F_1) from $T1_{pos}$ corresponding to six groups of Gene Ontology (GO) terms closely associated with carcinogenesis (**Figure 1**): cell adhesion, apoptosis, cell cycle, immune response, phosphorylation, and DNA damage & repair. Each signature gene set contained 30 unique genes and was used to translate a given expression profile into a feature vector. Testing the six signatures against $V1_{pos}$, we observed that the prediction achieved precision of 94.4% and recall rate of 90.1% for RD (residual disease; mutually exclusive to CR, complete response) subgroup, where a true positive prediction was defined as predicting a nonresponding patient to be so, and a false positive prediction to be predicting a patient that responded to the treatment as a nonresponding one. Precision and recall rate aligned with convention definition. Comparing to the genes with most significantly differential expression profiles (see Method), less than 50% of the most significant genes were selected (i.e., if selecting 130 genes, less than 65 genes were among the 130 top listed genes). Simply using the most significant genes gave inferior prediction power in the independent validation dataset (recall rate of 88%), implying that most prominent differential expression patterns contained cohort-specific features and might not be feasible to be utilized directly.

Further, we examined the predicting performances of all possible combinations of six signatures (k = 2, 3, 4, 5) (**Figures 2–4**) through 10-fold cross validation tests in $T1_{pos}$. While all choices gave precisions more than 94%, recall rates varied between 80 and 95%, exhibiting differences in prediction power. The combination of Signature B_1 (apoptosis), C_1 (cell cycle), and F_1 (DNA damage and repair) provided the best-balanced precision and recall rate (using the average values of 10-fold cross validations), of 94.0 and 93.4%, respectively. Predictor comprised of the selected combination of signatures had a better performance on the independent validation (precision of 93.1% and recall rate of 92.7%). We considered the recall rate to be the most important metric, as the methodology was intended to reliably predict whether an individual can skip a treatment without adverse consequences. In comparison, we tested seven signature genes (BRCA1, APC, p16/CDKN2A, FRMD6/hEx, YAP, BAX, and LZTS1/FEZ1) related to drug resistance in breast cancer, collected by Xu et al. (2016), for their prediction power. In the four-cohort discovery dataset, two-cohort discovery dataset and validation dataset, the signature gave precision rates of 92.3, 89.5, and 94.0% and recall rates of 82.7, 78.9, and 85.2%, respectively. Overall, our proposed signature genes provided better prediction power, and the methodology allowed the aggregation of accumulating datasets to discover potential better gene combinations.

To demonstrate the contribution of the signature genes against drug resistance, we calculated their relative contribution scores (RCS) based on randomization tests. Similar to the signature selection process but with reduced randomization count per iteration (50,000) and higher total iteration counts (200 for each of the six GO terms), fuzzy K-means clustering combined with Fisher's test was performed to measure randomized gene

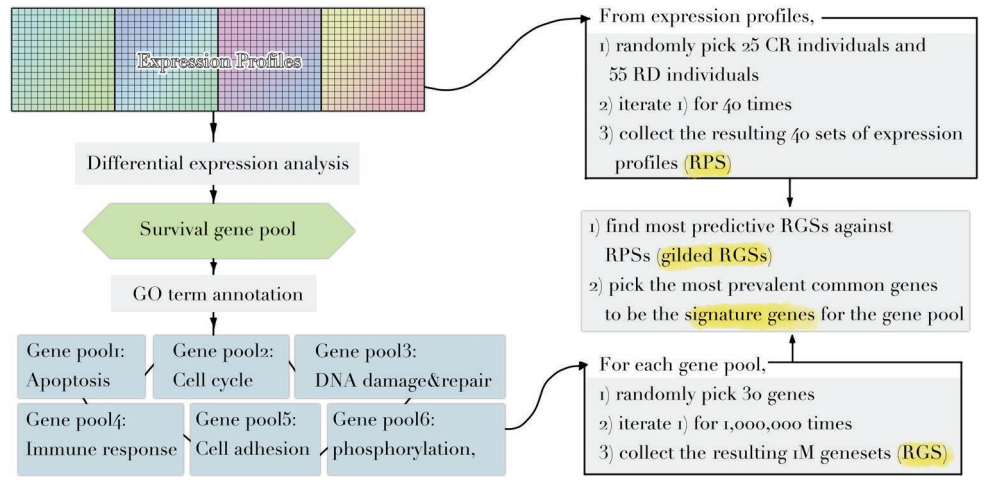

FIGURE 1 | Diagram illustrating the workflow of methodology used. Refer to Methods for dataset information and details in each step.

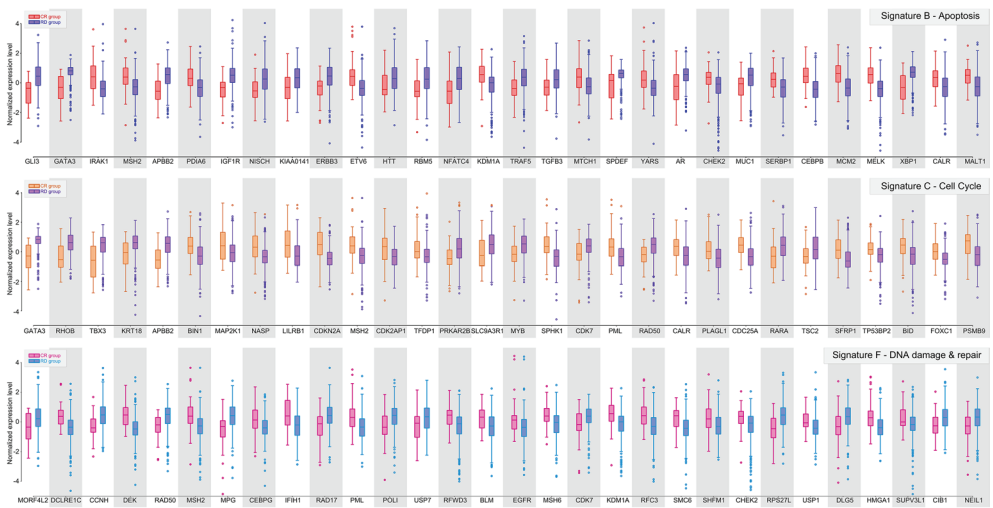

FIGURE 2 | Gene signature B, C, and F of ER-positive breast cancer. Box plots showing the distributions of normalized expression levels of the signature genes, whose centroids were further used to construct the predictor.

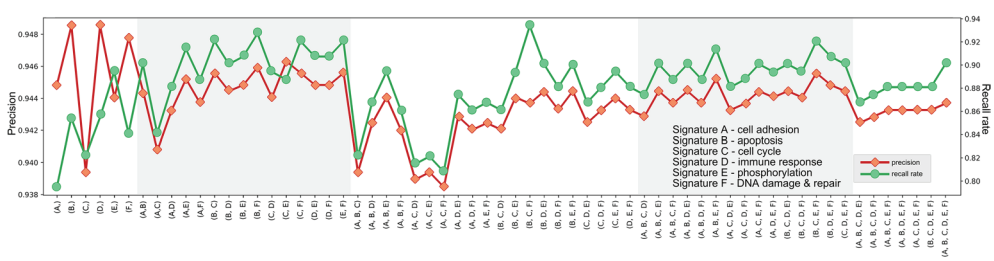

FIGURE 3 | Precisions and recall rates of predictor comprised of potential signature combinations, trained on $T1_{pos}$, tested using 10-fold validation. Although the combination of Signature B, C, and F provided not the best precision, its recall rate was finest.

Apoptosis signature genes of ER+

Entrez gene ID	Gene description	Gene symbol
2065	erb-b2 receptor tyrosine kinase 3	ERBB3
9833	maternal embryonic leucine zipper kinase	MELK
25803	SAM pointed domain containing ETS transcription factor	SPDEF
7043	transforming growth factor beta 3	TGFB3
2737	GLI family zinc finger 3	GLI3
3064	huntingtin	HTT
1051	CCAAT/enhancer binding protein beta	CEBPB
323	amyloid beta precursor protein binding family B member 2	APBB2
23767	mitochondrial carrier 1	MTCH1
11188	nischarin	NISCH
3480	insulin like growth factor 1 receptor	IGF1R
4582	mucin 1, cell surface associated	MUC1
2120	ETS variant 6	ETV6
23028	lysine demethylase 1A	KDM1A
811	calreticulin	CALR
4436	mutS homolog 2	MSH2
8565	tyrosyl-tRNA synthetase	YARS
4776	nuclear factor of activated T-cells 4	NFATC4
10130	protein disulfide isomerase family A member 6	PDIA6
7494	X-box binding protein 1	XBP1
11200	checkpoint kinase 2	CHEK2
7188	TNF receptor associated factor 5	TRAF5
4171	minichromosome maintenance complex component 2	MCM2
2625	GATA binding protein 3	GATA3
10892	MALT1 paracaspase	MALT1
10181	RNA binding motif protein 5	RBM5
367	androgen receptor	AR
26135	SERPINE1 mRNA binding protein 1	SERBP1
3654	interleukin 1 receptor associated kinase 1	IRAK1
9812	KIAA0141	KIAA0141

Cell cycle signature genes of ER+

Entrez gene ID	Gene description	Gene symbol
8877	sphingosine kinase 1	SPHK1
274	bridging integrator 1	BIN1
10111	RAD50 double strand break repair protein	RAD50
5914	retinoic acid receptor alpha	RARA
1022	cyclin dependent kinase 7	CDK7
4678	nuclear autoantigenic sperm protein	NASP
4602	MYB proto-oncogene, transcription factor	MYB
6926	T-box 3	TBX3
5698	proteasome subunit beta 9	PSMB9
1029	cyclin dependent kinase inhibitor 2A	CDKN2A
3875	keratin 18	KRT18
5604	mitogen-activated protein kinase kinase 1	MAP2K1
323	amyloid beta precursor protein binding family B member 2	APBB2
993	cell division cycle 25A	CDC25A
9368	SLC9A3 regulator 1	SLC9A3R1
10859	leukocyte immunoglobulin like receptor B1	LILRB1
5371	promyelocytic leukemia	PML
637	BH3 interacting domain death agonist	BID
811	calreticulin	CALR
388	ras homolog family member B	RHOB
4436	mutS homolog 2	MSH2
2296	forkhead box C1	FOXC1
8099	cyclin dependent kinase 2 associated protein 1	CDK2AP1
5325	PLAG1 like zinc finger 1	PLAGL1
7249	tuberous sclerosis 2	TSC2
2625	GATA binding protein 3	GATA3
7027	transcription factor Dp-1	TFDP1
6422	secreted frizzled related protein 1	SFRP1
7159	tumor protein p53 binding protein 2	TP53BP2
5577	protein kinase cAMP-dependent type II regulatory subunit beta	PRKAR2B

DNA damage & repair signature genes of ER+

Entrez gene ID	Gene description	Gene symbol
10111	RAD50 double strand break repair protein	RAD50
5983	replication factor C subunit 3	RFC3
9231	discs large MAGUK scaffold protein 5	DLG5
1022	cyclin dependent kinase 7	CDK7
3159	high mobility group AT-hook 1	HMGA1
51065	ribosomal protein S27 like	RPS27L
7396	ubiquitin specific peptidase 1	USP1
4350	N-methylpurine DNA glycosylase	MPG
1956	epidermal growth factor receptor	EGFR
7874	ubiquitin specific peptidase 7	USP7
64135	interferon induced with helicase C domain 1	IFIH1
10519	calcium and integrin binding 1	CIB1
1054	CCAAT/enhancer binding protein gamma	CEBPG
5371	promyelocytic leukemia	PML
23028	lysine demethylase 1A	KDM1A
4436	mutS homolog 2	MSH2
79681	nei like DNA glycosylase 1	NEIL1
2956	mutS homolog 6	MSH6
55159	ring finger and WD repeat domain 3	RFWD3
7979	split hand/foot malformation (ectrodactyly) type 1	SHFM1
64421	DNA cross-link repair 1C	DCLRE1C
11201	DNA polymerase iota	POLI
11200	checkpoint kinase 2	CHEK2
641	Bloom syndrome RecQ like helicase	BLM
6832	Suv3 like RNA helicase	SUPV3L1
79677	structural maintenance of chromosomes 6	SMC6
9643	mortality factor 4 like 2	MORF4L2
902	cyclin H	CCNH
7913	DEK proto-oncogene	DEK
5884	RAD17 checkpoint clamp loader component	RAD17

FIGURE 4 | List of gene signatures of ER-positive breast cancer.

sets' partition power over responsiveness, where gene set that exhibited statistical significance stronger than $p < 0.001$ was collected as "candidate geneset." Relative prevalence of a given signature gene was then obtained by measuring its presence amongst the candidate gene sets and normalizing the value through dividing the largest absolute prevalence value.

Robustness and Generalizability of Signature Gene Sets

To examine whether the identified gene signatures were not impacted by random factors, we performed another round of signature discovery process on $T1_{pos}$ with same set of hyperparameters and a new initial random state. We found that 99.2% (129 out of 130) gene selections remained the same in the new iteration, with the only altered gene selection resided in the Signature A_1 (adhesion). Expanding the number of random gene sets or iterations of the algorithm (see Methods) would not significantly impact on the gene signatures.

Further, the same gene signature discovery methodology was employed to $T2_{pos}$, a discovery dataset comprised of two cohorts (GSE22093 and GSE25066) and validated against the remaining three cohorts (GSE22093, GSE23988, and GSE20194) to prove the generalizability of the signatures. Regardless of shrank dataset size, the identified Signature B_2 (apoptosis), C_2 (cell cycle), and F_2 (DNA damage & repair) were exactly the same as the above Signature B_1, C_1, and F_1. This signature combination achieved best precisions and recall rates in GSE20194 (a.k.a. $V1_{pos}$; 94.6 and 93.4%, respectively), GSE20271 (95.4 and 91.2%, respectively), and GSE23988 (95.7 and 96.0%, respectively). Swapping the components of the discovery dataset did not significantly impact on signature discovery (none or less than two gene selections altered in each GO term signature) and the above reported prediction power. These results demonstrated that Signature C and E were generic and stable for nonresponsive ER-positive breast cancer cases and might be applied to new incoming datasets.

Gene Signatures for Unresponsiveness of Paclitaxel Treatment in ER-Negative Breast Cancer

We further demonstrated that the methodology may work equally well for ER-negative population. To obtain signature genes for ER-negative (ER−) group, we constructed a discovery dataset comprised of the four cohorts described above (see Methods (GSE20271, GSE22093, GSE23988, and GSE25066; referred to as T_{neg}; $n_{RD\text{-}and\text{-}ERneg} = 152$, $n_{CR\text{-}and\text{-}ERneg} = 217$). Similarly, GSE20194 ($n_{RD\text{-}and\text{-}ERneg} = 62$, $n_{CR\text{-}and\text{-}ERneg} = 45$; referred to as V_{neg}) was utilized as an independent validation dataset. MAS5 normalization and further regularizations addressing batch effects were performed as mentioned previously. We obtained five sets of signature genes ("Signatures," a–e) corresponding to five groups of GO terms which were closely associated with carcinogenesis: phosphorylation, immune response, apoptosis, DNA damage and repair, and cell cycle. Regardless of distinct ratio of sample size of RD and CR subgroup (ratios in range 0.7–1.4), compared to ER+ datasets (ratios in range 3–10), the prediction power of the signature gene sets was similarly steady. Validating in V_{neg}, the combination of Signature b (immune response), c (apoptosis), and d (DNA damage and repair) (**Figure 5**) achieved precision of 94.8% and recall rate of 92.0%.

Optimizing Methodology to Use 50-Fold Less Computation Resources

The original MSS methodology essentially relied on random searching, which was implemented through randomly generating sets of genes, ranking their ability to represent nonresponding patients, and selecting consensus genes from top-ranked gene sets to serve as gene signatures in the predictor. This process was computationally expensive, where training a model distributed on 672 cores (2.60 GHz) would cost 30–60 min to finish the 6 million iterations for six GO subsets (see Methods), and had also undefined hyperparameters that accounted for the number of total iterations as well as ranking criteria.

We found that the signature genes were prominent enough in most discovery datasets, as long as the overall sample size was reasonable, to allow optimization of signature discovery processes. First, hyperparameters that determine the base "gene pool" of random sampling could be replaced by simply picking the 500 most significantly differentially expressed genes, trivializing parameter tuning. Then, through introducing one single threshold and an ensemble method (see Methods), we were able to reduce the 1 million iterations required by the original methodology to 20,000 iterations while retaining same prediction power. While signatures reported above could

Apoptosis signature genes of ER-

Entrez gene ID	Gene description	Gene symbol
4776	nuclear factor of activated T-cells 4	NFATC4
7188	TNF receptor associated factor 5	TRAF5
25803	SAM pointed domain containing ETS transcription factor	SPDEF
7043	transforming growth factor beta 3	TGFB3
23028	lysine demethylase 1A	KDM1A
1051	CCAAT/enhancer binding protein beta	CEBPB
323	amyloid beta precursor protein binding family B member 2	APBB2
23787	mitochondrial carrier 1	MTCH1
11188	nischarin	NISCH
3480	insulin like growth factor 1 receptor	IGF1R
4582	mucin 1, cell surface associated	MUC1
2120	ETS variant 6	ETV6
367	androgen receptor	AR
811	calreticulin	CALR
4436	mutS homolog 2	MSH2
8565	tyrosyl-tRNA synthetase	YARS
26135	SERPINE1 mRNA binding protein 1	SERBP1
10130	protein disulfide isomerase family A member 6	PDIA6
7494	X-box binding protein 1	XBP1
11200	checkpoint kinase 2	CHEK2
2737	GLI family zinc finger 3	GLI3
4171	minichromosome maintenance complex component 2	MCM2
2625	GATA binding protein 3	GATA3
10892	MALT1 paracaspase	MALT1
10181	RNA binding motif protein 5	RBM5
3654	interleukin 1 receptor associated kinase 1	IRAK1
9812	KIAA0141	KIAA0141
3064	huntingtin	HTT
2065	erb-b2 receptor tyrosine kinase 3	ERBB3
9833	maternal embryonic leucine zipper kinase	MELK

DNA damage & repair signature genes of ER-

Entrez gene ID	Gene description	Gene symbol
11200	checkpoint kinase 2	CHEK2
5983	replication factor C subunit 3	RFC3
4436	mutS homolog 2	MSH2
79661	nei like DNA glycosylase 1	NEIL1
2956	mutS homolog 6	MSH6
9231	discs large MAGUK scaffold protein 5	DLG5
7398	ubiquitin specific peptidase 1	USP1
4350	N-methylpurine DNA glycosylase	MPG
1956	epidermal growth factor receptor	EGFR
55159	ring finger and WD repeat domain 3	RFWD3
64135	interferon induced with helicase C domain 1	IFIH1
10519	calcium and integrin binding 1	CIB1
1054	CCAAT/enhancer binding protein gamma	CEBPG
5371	promyelocytic leukemia	PML
23028	lysine demethylase 1A	KDM1A
7874	ubiquitin specific peptidase 7	USP7
7979	split hand/foot malformation (ectrodactyly) type1	SHFM1
64421	DNA cross-link repair 1C	DCLRE1C
11201	DNA polymerase iota	POLI
51065	ribosomal protein S27 like	RPS27L
641	Bloom syndrome RecQ like helicase	BLM
6832	Suv3 like RNA helicase	SUPV3L1
79677	structural maintenance of chromosomes 6	SMC6
9643	mortality factor 4 like 2	MORF4L2
902	cyclin H	CCNH
7913	DEK proto-oncogene	DEK
5884	RAD17 checkpoint clamp loader component	RAD17
1022	cyclin dependent kinase 7	CDK7
3159	high mobility group AT-hook 1	HMGA1
10111	RAD50 double strand break repair protein	RAD50

Immune response signature genes of ER-

Entrez gene ID	Gene description	Gene symbol
81603	tripartite motif containing 8	TRIM8
5167	ectonucleotide pyrophosphatase/phosphodiesterase 1	ENPP1
10563	C-X-C motif chemokine ligand 13	CXCL13
3148	high mobility group box 2	HMGB2
10512	semaphorin 3C	SEMA3C
1051	CCAAT/enhancer binding protein beta	CEBPB
3627	C-X-C motif chemokine ligand 10	CXCL10
6590	secretory leukocyte peptidase inhibitor	SLPI
1054	CCAAT/enhancer binding protein gamma	CEBPG
3480	insulin like growth factor 1 receptor	IGF1R
6373	C-X-C motif chemokine ligand 11	CXCL11
10087	collagen type IV alpha 3 binding protein	COL4A3BP
6732	SRSF protein kinase 1	SRPK1
6364	C-C motif chemokine ligand 20	CCL20
64421	DNA cross-link repair 1C	DCLRE1C
7494	X-box binding protein 1	XBP1
80762	Nedd4 family interacting protein 1	NDFIP1
1672	defensin beta 1	DEFB1
9582	apolipoprotein B mRNA editing enzyme catalytic subunit 3B	APOBEC3B
3572	interleukin 6 signal transducer	IL6ST
2625	GATA binding protein 3	GATA3
3929	lipopolysaccharide binding protein	LBP
10892	MALT1 paracaspase	MALT1
3934	lipocalin 2	LCN2
6347	C-C motif chemokine ligand 2	CCL2
720	complement C4A (Rodgers blood group)	C4A
1075	cathepsin C	CTSC
5819	nectin cell adhesion molecule 2	NECTIN2
9156	exonuclease 1	EXO1
3654	interleukin 1 receptor associated kinase 1	IRAK1

FIGURE 5 | List of gene signatures of ER-negative breast cancer.

be used for potential application in breast cancer nonresponsive screening without redoing the discovery processes, the optimization was suitable for implementations of the methodology on small computation resource, e.g., personal computer.

DISCUSSION

Precision oncology addresses the following aspects of targeted therapies: for example, developing medications that would benefit patients with a certain phenotype or symptom helps improve overall survival, finding means to confidently suggest patients to opt-out treatments that provide little benefit to them is as important. Paclitaxel, a drug which targets microtube components (β subunit of tubulin) of cell cycle regulatory to oppress expansion of cancer cells, has been considered as an important agent for treating breast cancer, providing valid efficacy and tolerability while low in cross-resistance with other drugs. However, paclitaxel's response rate among breast cancer patients resides in a loose range of 10–60%. Only 20% ER-positive patients would respond or partially respond to the drug. Accurately predicting whether a given patient will respond to paclitaxel treatment with confident would help preventing enormous breast cancer patients from undergoing excess effectless treatment and adverse effects. Gene expression profile was reported to be the strongest indicator of paclitaxel sensitivity in breast cancer patients (Dorman et al., 2015). Although resistance to paclitaxel has been reported to be associated with the expression levels of hundreds of transcripts and studied for the underlying molecular mechanisms as well as key pathways, existing signature genes did not perform well in predicting the lack of response in breast cancer patients.

While microarray and RNA-seq are becoming more applicable and affordable for clinical diagnostics, preventing patients from excessive treatments is desirable. In this study, we reported six sets of robust and generalizable gene signatures for the prediction of nonresponding individuals in ER+ and ER− groups of breast cancer, where combination of Signature B (30 genes related to apoptosis), C (30 genes related to cell cycle), and F (30 genes related to DNA damage and repair) achieved the best precision (>94%) and recall (>93%) predicting nonresponding patients in independent validation datasets, which were significant improvements compared to previous studies [e.g., 82% accuracy in cell lines, using expression profile of 15 genes and SVM model (Dorman et al., 2015)]. Signature genes were given relative contribution scores (RCS) based on randomization tests to demonstrate their contribution to the predictor, or relatively to what extent they contributed to the resistance. Moreover, we described a potential optimization of the methodology that rendered the algorithm less computational demanding, and therefore enabling faster gene signature discovery in new datasets.

MATERIALS AND METHODS

Data Processing and Normalization

The following five microarray-based gene expression profiles (samples examined before treatments) were collected from the repository of Gene Expression Omnibus (GEO): (1) GSE20194, comprised of 278 samples using Affymetrix Human Genome U133A Array (GPL96), where 161 samples were labeled as ER+. Of the 161 samples, 151 samples were marked as residual disease (RD) and 10 samples as partial complete response (pCR) or complete response (CR); (2) GSE 20271, comprised of 178 samples using Affymetrix Human Genome U133A Array (GPL96). In total, 98 samples were labeled as ER+, where 91 samples were marked as RD and 7 samples as pCR or CR; (3) GSE22093, comprised of 103 samples using Affymetrix Human Genome U133A Array (GPL96). In total, 42 samples were labeled as ER+, where 32 samples were marked as RD and 10 samples as pCR or CR; (4) GSE23988, comprised of 61 samples using Affymetrix Human Genome U133A Array (GPL96). In total, 32 samples were labeled as ER+, where 25 samples were marked as RD and 7 samples as pCR or CR; (5) GSE25066, comprised of 508 samples using Affymetrix Human Genome U133A Array (GPL96). In total, 297 samples were labeled as ER+, where 270 samples were marked as RD and 27 samples as pCR or CR.

We retrieved all five cohorts in their raw data format (CEL files) along with clinical data records. Expression profiles of each cohort were then normalized through MAS5.0 normalization (using RMA normalization instead in this step did not demonstrate visible impact on the results reported). After log2

transformation, we mapped the probes to Entrez Gene IDs (mapping provided by GEO) and removed duplicated reads of a given gene by retaining their average read. In total 4,075 unique genes were preserved. Probes pointed to unidentified genes (i.e., genes without Entrez ID) were not removed deliberately. They were practically invisible during the downstream analysis (see below), however. Data were further median-centered and z-scored across cohorts to address batch effects.

The four-cohort discovery datasets comprised of GSE20271, GSE22093, GSE23988, and GSE25066, utilizing GSE20194 as independent validation dataset. The two-cohort discovery dataset comprised of GSE22093 and GSE25066, utilizing GSE20194, GSE20271, and GSE23988 as validation set.

MSS Methodology and Optimization

Based on the study of Li et al., we utilized the following random-sampling-focused methodology in a given pair of discovery dataset and independent validation dataset.

1. In discovery dataset, genes that demonstrated significant differential expression profiles between subgroup of responsive patients (i.e., samples marked as pCR or CR) and subgroup of nonresponsive patients (samples marked as RD) were selected to form a gene pool. Significance was defined by the criteria that in more than 80 of 100 iterations of randomly drawing 30 responsive samples and 70 nonresponsive samples, *t*-test between such randomly drew subgroups showed $p < 0.05$. The 30–70 ratio can be relaxed to up to 30–120 without altering downstream results; in fact, only half of the differentially expressed genes that made to the final collections were at the top of this list, implying the following feature selection steps were of more importance. For the four-cohort discovery dataset, we obtained 389 unique genes to form the pool; for the two-cohort discovery dataset, 593 genes were selected. The two pools shared 369 unique genes, implying that although more significantly differentially expressed genes were found in two-cohort discovery dataset, many of which might be cohort-specific or at least not generic. Gene pools were annotated for GO terms by DAVID (Huang et al., 2008, 2009) (v6.8). In original MSS methodology, criteria of significance were considered to be hyperparameters, ideally controlling the number of selected genes during the corresponding step. However, training on the discovery dataset, we noticed that none of the signature genes came from the less significant ones, i.e., the bottom of the ranking list, therefore simply performing the *t*-tests and selecting the most significant 300–500 genes would serve the same objective. We discarded the hyperparameter in favor of this optimization and observed same results as reported, with less tuning attempts.
2. For a given gene pool, we partitioned genes with replacement into GO-defined subgroups (or, "subpool"). One gene could appear in more than one such subgroup according to its annotations. For the four-cohort discovery dataset, subgroup of apoptosis-related functions comprised of 186 unique genes; similarly, the numbers of genes were as the following for other subgroups: DNA damage & repair (56), immune response (104), cell adhesion (56), cell cycle (84), and phosphorylation (77). For the two-cohort discovery dataset, the numbers of genes were as the following for subgroups: apoptosis (290), DNA damage & repair (81), immune response (142), cell adhesion (93), cell cycle (115), and phosphorylation (111).
3. Following the original MSS methodology, for a given GO-defined subpool, 30 genes were randomly drew without replacement to form a random gene set (RGS) for 1,000,000 iterations, yielding 1 million RGSs. For a given discovery dataset, 25 CR individuals and 55 RD individuals were randomly drew without replacement to form a random patient set (RPS) for 40 iterations, yielding 40 RPSs. We optimized this step computationally through the following, without significant impact on the outputs:

    a. The number of RGSs can be reduced to up to 20-fold less by monitoring the list of most frequently appeared genes of the RGSs, without affecting the reported results. In original MSS, arbitrary 1 or 2 millions of iterations were performed to obtain the "gilded RGSs" and then the signature genes (see below). Instead we observed that, combinations of signature genes were prominent enough that it was possible to set a stopping criterion T, such that if after T iterations, the top 30 most frequently appeared genes of the "gilded RGSs" had no change, terminate this step and accept the "gilded RGSs" along with the list of top 30 most frequent genes as the final results. It was safe to assume such a parameter T in the range of 100–500, where a lesser T implied more tradeoff of robustness of the gene list in favor of computational complexity.
    b. Computational complexity could be further reduced by using an ensemble model. Instead of allowing each signature gene set to claim one vote in the predicting (see below), we lowered the parameter T to as less as 30 and obtained five gene lists for each GO-defined subpool. Each gene list was then treated as one independent voter during voting.

    Combining a and b, the number of total executed iterations could be reduced to 50-fold less. In this study, we implemented the original MSS methodology distributed on a cluster with 672 CPUs, paralleling all 1 million iterations for each GO-defined subpool, and the runtime was around half an hour. Using the optimization, it was possible to calculate the predictor of desire at regular PCs or workstations in reasonable time frame.

    Altering the proportion of CR and RD cases in RPSs would not significantly affect reported results, as long as the ratio was kept around 1:2 to 1:5.

4. Each RGS was tested against all 40 RPSs (if not using optimized version): patients in a RPS were partitioned into two clusters through K-means (Euclidean distance; using fuzzy K-means that implemented by sklearn-extension with fuzzy factor as 2 would not significantly alter the reported results, but with much less efficiency). Fisher's test was used to determine if the clusters enriched CR or RD individuals, respectively. The p's yielded by Fisher's tests were recorded, and the reciprocal of their average was considered as the enrichment score of

the RGS. For each GO term, top 3,000 most significant RGSs were selected to be "gilded RGSs" based on the enrichment score. This threshold could be chosen freely between 1,000 and 3,000 and did not significantly affect the report results.
5. The unique 30 most frequently picked genes across gilded RGSs of a GO term were drew as the set of signature genes for the corresponding GO term.

Gene Sets Selection

Combinations of gene sets were tested using 10-fold cross validation and independent validation dataset. Prediction of labels (either the given individual being nonresponsive or responsive to paclitaxel treatment) was made through voting: (1) for each GO term, we used their 30 signature genes to translate expression profiles of patients in the training dataset into 1D vectors of shape (30, 1). (The expression profile of the individual being predicted underwent the same transformation.) Centroids of the feature vectors were calculated for RD subgroup and CR subgroup, respectively. If cosine distance between feature vectors of an individual and RD subgroups' centroid was smaller than such cosine distance between feature vectors and CR's centroid, the individual would gain one point on belonging to RD; one point be given to CR otherwise. (2) After all signature genesets had their votes assigned, the individual was labeled as the prediction with most votes. Having even number of signature genesets rarely was a problem in this study; we observed that predictions of nonresponsive labels were mostly being consented by majority or all genesets. If being of concern, cosine-distances-based fuzzy votes could be used in place of the binary votes.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

QC, EW, and XF designed the study. XF performed data preparation, coding, signature extraction, optimization, and downstream analysis.

FUNDING

This work was supported by Natural Science Foundation of China (81670462).

REFERENCES

Collins, F. S., and Varmus, H. (2015). A new Initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795. doi: 10.1056/nejmp1500523

de Delou, J. M. A., Vignal, G. M., Índio-do-Brasil, V., de Accioly, M. T. S., da Silva, T. S. L., Piranda, D. N., et al. (2017). Loss of constitutive ABCB1 expression in breast cancer associated with worse prognosis. *BCTT* 9, 415–428. doi: 10.2147/BCTT.S131284

Dorman, S. N., Baranova, K., Knoll, J. H. M., Urquhart, B. L., Mariani, G., Carcangiu, M. L., et al. (2015). Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol. Oncol.* 10, 85–100. doi: 10.1016/j.molonc.2015.07.006

Duan, Z., Lamendola, D. E., Duan, Y., Yusuf, R. Z., and Seiden, M. V. (2004). Description of paclitaxel resistance-associated genes in ovarian and breast cancer cell lines. *Cancer Chemother. Pharmacol.* 55, 277–285. doi: 10.1007/s00280-004-0878-y

Garraway, L. A., Verweij, J., and Ballman, K. V. (2013). Precision oncology: an overview. *J. Clin. Oncol.* 31, 1803–1805. doi: 10.1200/jco.2013.49.4799

Hatzis, C. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 305, 1873–1823. doi: 10.1001/jama.2011.593

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Itoh, M., Iwamoto, T., Matsuoka, J., Nogami, T., Motoki, T., Shien, T., et al. (2013). Estrogen receptor (ER) mRNA expression and molecular subtype distribution in ER-negative/progesterone receptor-positive breast cancers. *Breast Cancer Res. Treat.* 143, 403–409. doi: 10.1007/s10549-013-2763-z

Iwamoto, T., Bianchini, G., Booser, D., Qi, Y., Coutant, C., Ya-Hui Shiang, C., et al. (2010). Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. *J. Natl. Cancer Inst.* 103, 264–272. doi: 10.1093/jnci/djq524

Li, J., Lenferink, A. E. G., Deng, Y., Collins, C., Cui, Q., Purisima, E. O., et al. (2010). Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat. Commun.* 1, 34–38. doi: 10.1038/ncomms1033

MAQC Consortium. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* 28, 827–838. doi: 10.1038/nbt.1665

Němcová-Fürstová, V., Kopperová, D., Balušíková, K., Ehrlichová, M., Brynychová, V., Václavíková, R., et al. (2016). Characterization of acquired paclitaxel resistance of breast cancer cells and involvement of ABC transporters. *Toxicol. Appl. Pharmacol.* 310, 215–228. doi: 10.1016/j.taap.2016.09.020

Popovici, V., Chen, W., Gallas, B. D., Hatzis, C., Shi, W., Samuelson, F. W., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* 12, 1999–2013. doi: 10.1186/bcr2468

Tabchy, A., Valero, V., Vidaurre, T., Lluch, A., Gomez, H., Martin, M., et al. (2010). Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clin. Cancer Res.* 16, 5351–5361. doi: 10.1158/1078-0432.CCR-10-1265

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Xi, J., Wang, M., and Li, A. (2017). Discovering potential driver genes through an integrated model of somatic mutation profiles and gene functional information. *Mol. BioSyst.* 13, 2135–2144. doi: 10.1039/C7MB00303J

Xi, J., Wang, M., and Li, A. (2018). Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. *BMC Bioinf.* 19:1546. doi: 10.1186/s12859-018-2218-y

Xu, J. -H., Hu, S. -L., Shen, G. -D., and Shen, G. (2016). Tumor suppressor genes and their underlying interactions in paclitaxel resistance in cancer therapy. *Cancer Cell international* 16:13. doi: 10.1186/s12935-016-0290-9

Yang, J., Li, A., Li, Y., Guo, X., and Bioinformatics, M. W. (2018). A novel approach for drug response prediction in cancer cell lines via network representation learning. *Bioinformatics* [Epub ahead of print]. 1–9.

Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. (2018). A novel heterogeneous network- based method for drug response prediction in cancer cell lines. *Sci. Rep.* 1–9. doi: 10.1038/s41598-018-21622-4

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 March 2019
doi: 10.3389/fgene.2019.00166

SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer

Zhi Huang[1,2,3], *Xiaohui Zhan*[2,4], *Shunian Xiang*[4,5], *Travis S. Johnson*[2,6], *Bryan Helm*[2], *Christina Y. Yu*[2,6], *Jie Zhang*[5], *Paul Salama*[3], *Maher Rizkalla*[3], *Zhi Han*[2,7]* and *Kun Huang*[2,3,7]*

[1] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States, [2] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, United States, [3] Department of Electrical and Computer Engineering, Indiana University-Purdue University Indianapolis, Indianapolis, IN, United States, [4] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China, [5] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, United States, [6] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, United States, [7] Regenstrief Institute, Indianapolis, IN, United States

OPEN ACCESS

Edited by:
Victor Jin,
The University of Texas Health Science Center at San Antonio, United States

Reviewed by:
Long Gao,
University of Pennsylvania, United States
Dong Xu,
University of Missouri, United States

***Correspondence:**
Kun Huang
kunhuang@iu.edu
Zhi Han
zhihan@iu.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 01 December 2018
Accepted: 14 February 2019
Published: 08 March 2019

Citation:
Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, Zhang J, Salama P, Rizkalla M, Han Z and Huang K (2019) SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. Front. Genet. 10:166. doi: 10.3389/fgene.2019.00166

Improved cancer prognosis is a central goal for precision health medicine. Though many models can predict differential survival from data, there is a strong need for sophisticated algorithms that can aggregate and filter relevant predictors from increasingly complex data inputs. In turn, these models should provide deeper insight into which types of data are most relevant to improve prognosis. Deep Learning-based neural networks offer a potential solution for both problems because they are highly flexible and account for data complexity in a non-linear fashion. In this study, we implement Deep Learning-based networks to determine how gene expression data predicts Cox regression survival in breast cancer. We accomplish this through an algorithm called SALMON (Survival Analysis Learning with Multi-Omics Neural Networks), which aggregates and simplifies gene expression data and cancer biomarkers to enable prognosis prediction. The results revealed improved performance when more omics data were used in model construction. Rather than use raw gene expression values as model inputs, we innovatively use eigengene modules from the result of gene co-expression network analysis. The corresponding high impact co-expression modules and other omics data are identified by feature selection technique, then examined by conducting enrichment analysis and exploiting biological functions, escalated the interpretation of input feature from gene level to co-expression modules level. Our study shows the feasibility of discovering breast cancer related co-expression modules, sketch a blueprint of future endeavors on Deep Learning-based survival analysis. SALMON source code is available at https://github.com/huangzhii/SALMON/.

Keywords: deep Learning, co-expression analysis, survival prognosis, breast cancer, multi-omics, neural networks, cox regression

BACKGROUND AND INTRODUCTION

There is a strong need to identify effective prognostic biomarkers to help optimize and personalize treatment (Liu et al., 2016). Among cancers, breast invasive carcinoma is one of the most heterogeneous cancers with distinct prognoses based on morphological, phenological, and molecular stratifications (Nagini, 2017; Wu et al., 2017). Breast invasive carcinoma patients have a 77% survival rate after 5 years and 44% survival rate after 15 years (Pereira et al., 2016), so developing accurate prognostic models could significantly improve risk stratification after diagnosis.

Recent Deep Learning-based approaches have been widely applied to Computational Biology and Bioinformatics (Huang et al., 2017; Zhang et al., 2018b). The advantages of learning non-linear functions and retrieving lower dimensional representation (Ching et al., 2018) reveal advances of Deep Learning models. The application of survival prognosis that incorporates Cox proportional hazards regression with a single transcriptomic dataset (Ching et al., 2018; Katzman et al., 2018; Shao et al., 2018) and with multi-omics data (Chaudhary et al., 2018; Poirion et al., 2018; Ramazzotti et al., 2018; Sun et al., 2018; Zhang et al., 2018a) is of major interest in precision health.

For these reasons, we integrate multi-omics data with Deep Learning-based survival prognosis models. While most contemporary approaches incorporate one or few types of omics data, such as mRNA-seq data and miRNA-seq data (Gupta et al., 2015; Nassar et al., 2017), we propose that integrating more diverse data may lead to improved modeling—especially when driven by machine learning. Moreover, classic cancer biomarkers can often stratify patients into risk groups, and these too should be integrated when available. Specifically, copy number burden (CNB) and tumor mutation burden (TMB) are important for predicting tumor progression (Marshall et al., 2017; Thomas et al., 2018) and immunotherapy (Birkbak et al., 2013; Chalmers et al., 2017; Goodman et al., 2017). Other demographical and clinical information such as diagnosis age, estrogen receptors (ER) status, progesterone receptors (PR) status should also be considered during model construction. One of the challenges for such diverse data is high-dimensionality.

Most Deep Learning approaches employ neural networks (multilayer perceptron) with huge numbers of parameters to be optimized. Optimizing such large sets of parameters with limited patient samples tends to introduce overfitting that renders the models ineffective. In this paper, we advocate the use of eigengene matrices instead of original mRNA-seq and miRNA-seq data derived from co-expression analysis with R package "lmQCM." Using neural network architecture, multi-omics data, and the Cox proportional hazards model, we develop our model called SALMON (Survival Analysis Learning with Multi-Omics Neural Networks). SALMON adopts co-expression modules as input, namely, the eigengene matrix derived from co-expression network analysis. It greatly reduces the dimension of the original feature space addressing the "curse of dimensionality" and increases the robustness and learnability of the model. This novel technique was not adopted by any other Deep Learning-based survival prognosis model such as Cox-nnet (Ching et al., 2018).

SALMON is trained on co-expression module eigengenes instead of gene expressions and thus we were able to investigate co-expression modules contribution to the hazard ratio (**Figure 1**). These gene co-expression modules contained individual genes from the initial lmQCM gene co-expression network analysis. Genes from modules that highly contributed to the hazard ratio were further evaluated with gene enrichment analysis to confirm certain gene regulations and biological processes. These biological findings confirm the validity of our models and provide insight into the complex regulatory relationships at work in breast invasive carcinoma.

MATERIALS AND METHODS

Datasets and Study Design

In this experiment, we analyzed 583 female breast invasive carcinoma (BRCA) patients which had five omics data types including gene expression data (illuminahiseq_rnaseqv2-RSEM_genes_normalized) and miRNA data (illuminahiseq_mirnaseq-miR_gene_expression) from Broad GDAC Firehose (https://gdac.broadinstitute.org/), copy number burden (CNB) was measured by total Kb length and the data (broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg) was provided from Pan-Cancer Atlas (PanCanAtlas) Initiative (https://gdc.cancer.gov/about-data/publications/pancanatlas). Tumor mutation burden (TMB) was calculated by the total number of mutated genes based on MAF files (Mutation_Packager_Oncotated_Calls) from Broad GDAC Firehose. Demographical and clinical information (diagnosis age, Estrogen Receptor (ER) status, Progesterone Receptor (PR) status) and overall survival (OS) events and months were collected from cBioPortal (http://www.cbioportal.org/). HER2 status was not considered in this article because of insufficient data. **Table 1** shows the statistical information of this patient cohort.

We performed 5-fold cross-validation on the dataset. In each fold, 80% of the data were used for model training and 20% of the data were used for model testing. mRNA and miRNA data were pre-processed by TSUNAMI online analysis suite (https://apps.medgen.iupui.edu/rsc/tsunami/). The pre-processing steps are 2-fold: It firstly removed genes with lowest 20% of mean expression values shared by all patients. Then it removed genes with lowest 20% of expression values' variance. These pre-processing steps were necessary to ensure the robustness for the downstream correlational computation in gene co-expression module analysis step.

Gene Co-expression Module Analysis

Instead of feeding mRNA-seq and miRNA-seq data to the neural networks and analyzing results at the gene level, we used eigengene matrices of gene co-expression modules obtained from lmQCM algorithm (Zhang and Huang, 2014) as the input to the SALMON algorithm. This reduced 99.46% of input features and greatly reduced the number of parameters in the neural networks. Using eigengenes as features can be considered as bias/variance (error/complexity) trade-off in machine learning (Weigend et al., 1990; Geman et al., 1992), which simplifies

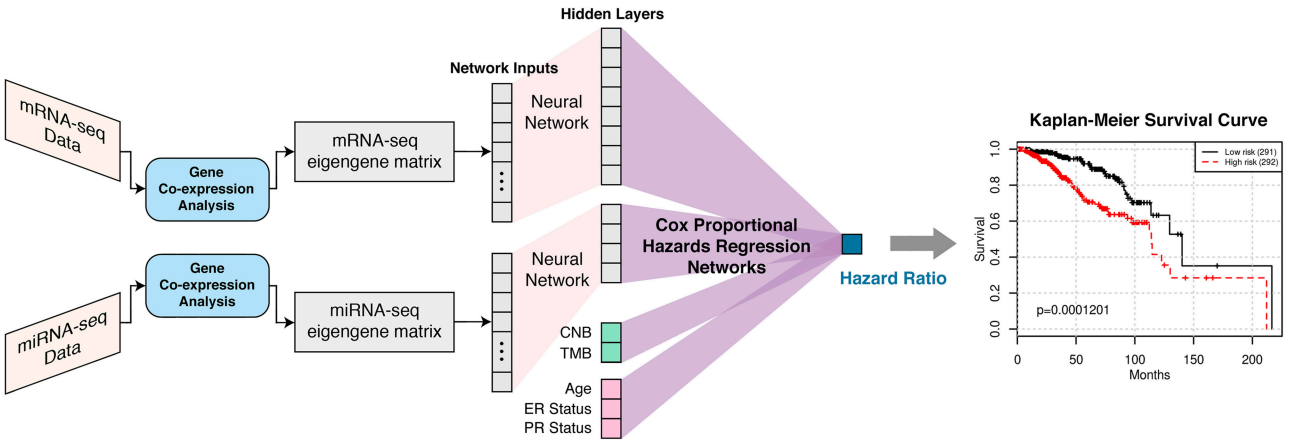

FIGURE 1 | SALMON (Survival Analysis Learning with Multi-Omics Neural Networks) architecture with the implementation of Cox proportional hazards regression networks. Co-expression modules (eigengene matrices) are the inputs to the SALMON. Number of the hidden layers and dimensions of hidden layers can also be fine-tuned (not included in this paper). The output is the hazard ratios which can be interpreted as the relative risks of patients.

TABLE 1 | Demographical and clinical characteristics of 583 female breast invasive carcinoma (BRCA) patients.

mRNA size		miRNA size		OS Months		Diagnostic age		ER positive ratio	PR positive ratio
Original	Co-expression module	Original	Co-expression module	Median	Range	Median	Range		
13,132	57	530	12	31.70	0.00–216.59	57	26–90	76.16%	67.41%

mRNA and miRNA stand for mRNA-seq data and miRNA-seq data. OS stands for overall survival. The status of ER and PR were derived from IHC (immunohistochemistry). All clinical information was collected from cBioPortal.

the networks significantly. The total number of neural network weights to be learned was then narrowed down from 107193 to 521, ensuring the robustness of the learning process and alleviate the overfitting issue (Caruana et al., 2001; Schmidhuber, 2015).

There are many gene co-expression network analysis packages, such as the R package for weighted correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) and local maximal Quasi-Clique Merger (lmQCM) (Zhang and Huang, 2014), which can discover densely connected gene modules across samples/patients. Co-expression network analyses are used increasingly to reveal latent gene-gene interactions, biomarkers and novel gene functions (Horvath et al., 2012; Chandran et al., 2016; Han et al., 2016, 2017; Zhang and Huang, 2017; Xiang et al., 2018). Comparing to WGCNA, weight normalization process in lmQCM was inspired by the spectral clustering (Ng et al., 2002) in machine learning. With efficient implementation of the revision from eQCM (edge-covering quasi-clique merger) algorithm (Xiang et al., 2012), lmQCM allowed module overlap, mining smaller densely co-expressed modules, and thus was adopted in this article. The generally smaller size of mined modules can also generate more meaningful gene ontology (GO) enrichment results (Zhang et al., 2012, 2013, 2016; Shroff et al., 2016; Cheng et al., 2017). The implementation was performed on TSUNAMI. For mRNA-seq data, we set lmQCM parameters $\gamma = 0.7$, $\lambda = 1$, $t = 1$, $\beta = 0.4$, minimum size of cluster $= 10$, and adopted Spearman's rank correlation coefficient (Mukaka, 2012) to calculate gene-wised correlations. The parameters setting of miRNA-seq data were the same except $\gamma = 0.4$, $\beta = 0.6$, and minimum size of cluster $= 4$.

After calculating gene co-expression modules with lmQCM, eigengene matrices were then determined. The eigengene matrix is the expression values of each gene co-expression module summarized into the first principal component using singular value decomposition (SVD) (Golub and Reinsch, 1970). With the first right-singular vector of each module as the summarized expression values, it projects co-expressed genes to 1-D space and thus can be treated as the "super gene." In our experiment with breast invasive carcinoma, an eigengene matrix with 57 dimensions was derived from mRNA-seq data and an eigengene matrix with 12 dimensions was also derived from miRNA-seq data. Details of co-expression modules and eigengene matrices we derived for this paper are available in **Supplementary files**. These eigengene matrices were treated as the substitution of the original expression inputs.

Neural Networks Design, Architecture, and Evaluation Metric

SALMON was designed and implemented in PyTorch 1.0. mRNA-seq and miRNA-seq eigengene matrices were firstly connected to hidden layers with dimensions 8 and 4, respectively, then connected to the final output (hazard ratio) with Cox proportional hazards regression networks. Alternatively, CNB, TMB, and demographical and clinical information (diagnosis

age, ER status, PR status) had no hidden layer and were connected to final output directly as covariates. This architecture was explained graphically in **Figure 1**. The rationale behind this network architecture instead of using simple fully connected networks such as Cox-nnet (Ching et al., 2018) was by assuming (1) each omics type affects the hazard ratio independently; (2) downscale eigengene matrices by hidden layers can force multi-omics data contributed to hazard ratios in a relatively equal scale at Cox proportional hazards regression networks part.

SALMON adopts Adaptive moment estimation (Adam) optimizer (Kingma and Ba, 2015). We set the number of epochs = 100 with fine-tuned learning rates for each 5-folds cross-validation experiments. LASSO (least absolute shrinkage and selection operator) regularization (Santosa and Symes, 1986) is applied to the networks. Sigmoid activation function is also applied right after each forward propagation and Cox proportional hazards regression networks. The Sigmoid function

$$sigmoid\,(x) = \frac{1}{1+e^{-x}} \tag{1}$$

forces the output range be within 0 to 1, introduces non-linearity to the system. In this model, we set the batch size = 64, and the batch normalization was not adopted. The number of the hidden layers and dimensions of hidden layers can be fine-tuned, in this paper, single hidden layers were attached to transcriptomic data with size = 8 for mRNA-seq modules, and size = 4 for miRNA-seq modules.

Cox Proportional Hazards Regression Networks

Our algorithm SALMON, integrated Cox proportional hazards model, differs from previous work (Ma and Zhang, 2018; Sun et al., 2018) which use survival status (living or deceased) in a binary classification problem. In contrast, we also took survival times (overall survival months) into account denoted as Y_i and made our neural networks into a Cox regression learning task. Maximum likelihood estimation (MLE) is then applied to the log partial likelihood

$$\ell\,(\beta) = \sum_{i\,:\,C_i=1}\left(\sum_{k=1}^{K}\beta_k X_{ik} - \log\left(\sum_{j\,:\,Y_j\geq Y_i}\exp(\sum_{k=1}^{K}\beta_k X_{ik})\right)\right) \tag{2}$$

where β are the parameters to be estimated. $C_i = 1$ indicates the occurrence of the death events for patient i with K-dimensional input vector X_i.

Objective Function

Based on Cox proportional hazards regression networks we formulized the objective function of neural networks as:

$$\hat{\Theta} = \mathbf{argmin}_{\Theta}\left\{\sum_{i\,:\,C_i=1}\left(\sum_{k=1}^{K}\beta_k X_{ik} - \log\left(\sum_{j\,:\,Y_j\geq Y_i}\exp(\sum_{k=1}^{K}\beta_k X_{ik})\right)\right) + \lambda\,\|\Theta\|_1\right\} \tag{3}$$

where Θ are the entire network weights (including β) to be optimized via back-propagation, λ is the weight multiplier of LASSO regularization. We set $\lambda = 1 \times 10^{-5}$ in the experiments.

Evaluation Metric

Concordance index (Steck et al., 2007), valued from 0 to 1, is used in this article as the evaluation metric of survival prognosis. It is widely adopted to evaluate the performances of survival prognosis models (Ching et al., 2018; Katzman et al., 2018) and is equivalent to the area under the ROC curve (AUC) (Bradley, 1997), which measures the model's distinguishability between living and deceased groups. A concordance index = 0.5 indicates the model makes ineffective prediction. A higher concordance index > 0.5 indicates a better survival prognosis model. For breast invasive carcinoma cancer, we consider a concordance index > 0.7 indicates a good model performance.

Survival Analysis

Survival analysis with log-rank test (Mantel, 1966) is used to inspect the performances of SALMON on 5-folds cross-validation testing sets. The Kaplan-Meier survival curves are generated by dichotomizing all testing patients to low risk and high risk groups via the median hazard ratio. The corresponding log-rank p-value implies the ability of the model to differentiate two risk groups. Lower p-values convey better model performances.

Gene Ontology and Functional Enrichment Analysis

Co-expression modules generated by lmQCM are then exported to ToppGene Suite (Chen et al., 2009) (https://toppgene.cchmc.org/) and Enrichr (Kuleshov et al., 2016) (http://amp.pharm.mssm.edu/Enrichr/). Using ToppGene, we performed functional analysis including Gene ontology (GO) and cytoband analysis. The false discovery rate (FDR) <0.05 and FDR <1.0 were considered to be significantly enriched for GO analysis and cytoband analysis, respectively. Human Gene Atlas [up regulated genes in human tissues from BioGPS (http://biogps.org)] and ARCHS4 tissues were also investigated for some certain co-expression modules by Enrichr.

RESULTS

The experiments were performed with six different combinations of multi-omics data as input sources, they are: (i) mRNA-seq data (mRNA) (57 features); (ii) miRNA-seq data (miRNA) (12 features); (iii) integration of mRNA and miRNA (69 features); (iv) integration of mRNA, miRNA, copy number burden (CNB), and tumor mutation burden (TMB) (71 features); (v) integration of mRNA, miRNA, and demographical and clinical (diagnosis age, ER status, PR status) data (72 features); (vi) integration of mRNA, miRNA, CNB, TMB, and demographical and clinical (diagnosis age, ER status, PR status) data (74 features). Where both RNA-seq co-expression modules are required for all integrative combinations. The SALMON model architecture from **Figure 1** removed certain network substructures which not been used and performed 5-folds cross-validation with

583 patients. Concordance index was used to evaluate the performances. SALMON was then compared to several other survival prognosis algorithm Cox-nnet (Ching et al., 2018), DeepSurv (Katzman et al., 2018), generalized linear model with Cox regression (GLMNET) (Friedman et al., 2010), and RSF (Ishwaran et al., 2008) with all omics data fed in. Since their Cox regression model didn't take multi-omics data sources into consideration, we modified their original framework to integrate multi-omics data (with co-expression modules) altogether as single input vector. The feature importance of all 74 covariates were also investigated by repeated feature deletion, then ranked by the median of decreased concordance index, proved and revealed certain biological interpretations.

Integrating Multi-Omics Features Increased the Performances

From **Figure 2A**, we observed an upward trend on median/mean concordance indices with more omics data are integrated. Integrating all omics data (74 features) gave the optimal performances (concordance index: median = 0.7285; mean = 0.6918). Next, all hazard ratios from 5-folds testing sets were concatenated and performed the log-rank test (Mantel, 1966) as shown in **Figures 2C–E** and **Figure S1**. Another feature set without transcriptomics data was also considered as reference (5 features containing CNB, TMB, and demographical and clinical features) with median concordance index = 0.6949 and the Kaplan-Meier plot was shown in **Figure S1F** (log-rank test p-value = 3.67E-03). We found that integrating all omics data (**Figure 2E**) gave the most significant p-value (1.201E-04) with respect to the log-rank test, proving that integrating more multi-omics data to SALMON can enhance the prediction.

We further performed pairwise paired t-test to the resulting concordance indices. As shown in **Table 2**, a negative t-statistic implied that the set 1 is lower than set 2. This concludes that integrating more omics data can generally increase the performance of survival prognosis in breast cancer.

Next, we compared SALMON to the state-of-the-art Deep Learning-based cancer survival prognosis model Cox-nnet (Ching et al., 2018), as well as another recently proposed DeepSurv (Katzman et al., 2018), and two traditional models generalized linear model with Cox regression (GLMNET) (Friedman et al., 2010) and RSF (Ishwaran et al., 2008).We further modified their original implementation with all omics data as inputs. As shown in **Figure 2B**, the median concordance index of SALMON (0.7285) was reported higher than the modified Cox-nnet (0.7234), DeepSurv (0.6563), GLMNET (0.6490), and RSF (0.6229). Compare to the modified Cox-nnet with similar performance in terms of concordance index, SALMON has a more significant result in log-rank test (p-value = 1.201E-04) than the modified Cox-nnet (p-value = 2.282E-04) with all testing sets and all 74 features as inputs (**Figure S2**). Between SALMON and the modified Cox-nnet the performance is insignificant (paired t-test statistic = −2.105, p-value = 0.103) suggesting these two methods are comparable. But from the neural network structure perspective, SALMON is more flexible since it separates forward propagation for each omics data, which enable a scalable integration of multi-omics data.

Interpreting and Ranking the Importance of Co-expression Modules

Interpreting feature importance for neural networks has been studied over years. One way is to assign each feature be zero repeatedly, then the feature with lowest change of the resulting accuracy implies the least importance that affects to the prediction model. This approach is widely adopted for feature selection and ranking the importance of features in neural network (Setiono and Liu, 1997; Zhang, 2000; Sung and Mukkamala, 2003). Based on this approach, we analyzed the contribution of each eigengene's module to the final hazard ratio by forcing each input feature of the testing sets be zero. By feeding the modified testing sets to the pre-trained SALMON networks, we rank the importance of features by inspecting how much the concordance indices decreased. Features that decrease the testing concordance indices more are considered to be more important. At this moment, we integrated all omics data for training and testing. **Table 3** presented top features that mostly reduced the concordance index. The leading two features are the diagnosis age and PR status, then five mRNA-seq co-expression modules are followed.

Next, we selected those features (33 in total) of which their median values < 0 in **Figure 3** and re-performed the training testing in SALMON. Results showed that before and after feature selection, the performances are insignificant in terms of concordance index (before feature selection: mean = 0.6918, median = 0.7285; after feature selection: mean = 0.7108, median = 0.7200; paired t-test statistic = −0.861, p-value = 0.438) (**Figure S3**). This implying that training with selected "important" multi-omics features instead of all can still preserve the prognosis performances.

Identification of Breast Cancer Related Genes and Cytobands Associated With Important Modules

To inference the biological implication from the feature ranking, we performed Gene Ontology (GO) and cytoband enrichment from ToppGene Suite (https://toppgene.cchmc.org/) (Chen et al., 2009). Specifically, we focused on analyzing top five mRNA co-expression modules (**Table 3**). Totally we identified 10 genes such as MST1, CPT1B, MAP3K7, CCNC, etc. We also identified various enriched cytoband and other biological functions. **Table 3** is further discussed and explained in Discussion section. Genes list within each mRNA-seq, miRNA-seq module is provided in **Supplementary Material**.

Investigating Feature Importance With Different Age Groups

As shown in **Figure 3**, we observed the strong predictive power of diagnosis age, which is consistent with previous studies demonstrating age as one of the most prominent cancer risk factors (Adami et al., 1986). Thus, it is crucial to further investigate if patients in different groups can be stratified using

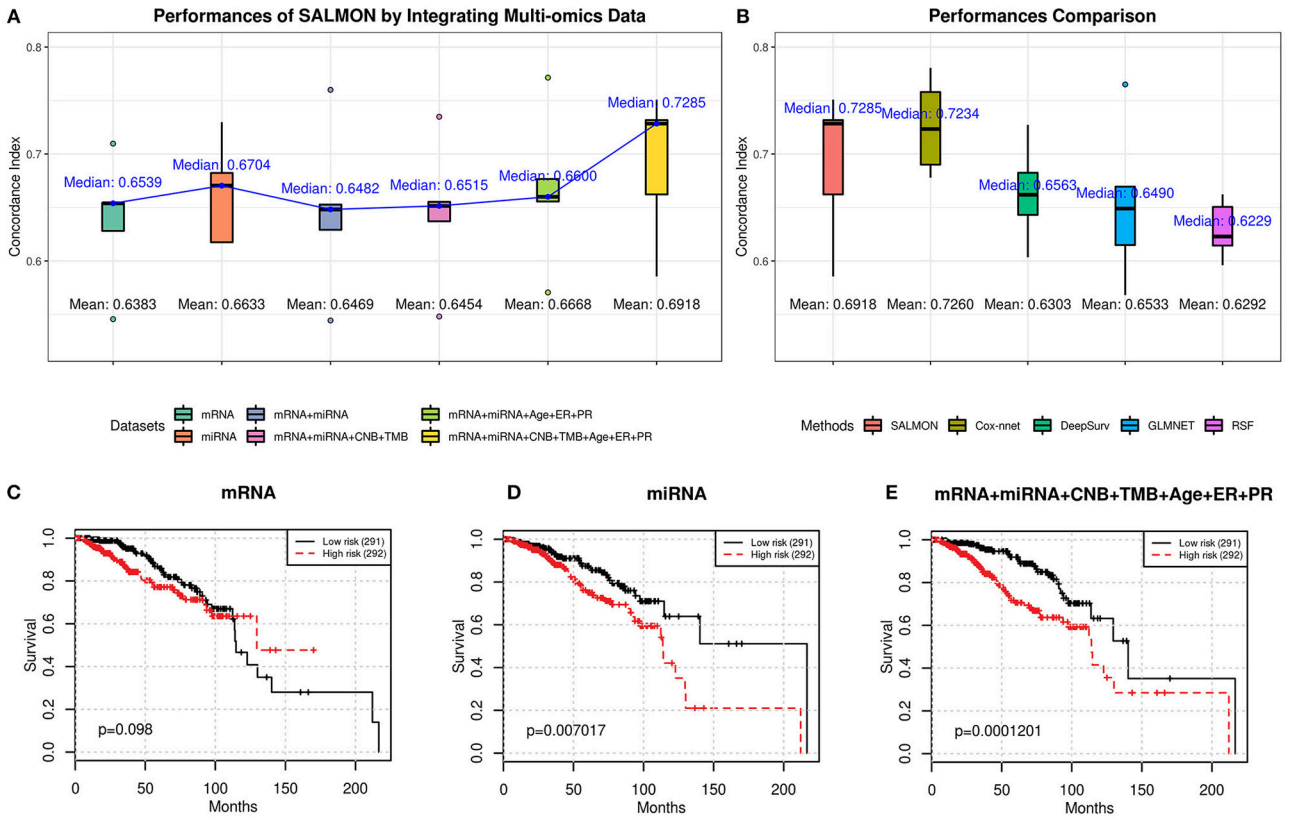

FIGURE 2 | (A) Performances of SALMON with multi-omics data integrated in terms of concordance index. **(B)** Performance comparison between SALMON and the modified Cox-nnet, DeepSurv, GLMNET, and RSF in terms of concordance index with all omics data used for learning. **(C–E)** Kaplan-Meier plot of survival prognosis. Hazard ratios were derived from all five testing sets. Log-rank test was used to find the corresponding p-value with low risk and high risk groups dichotomized by the median hazard ratio. Omics data used for training and testing: **(C)** mRNA-seq data (mRNA); **(D)** miRNA-seq data (miRNA); **(E)** integration of mRNA, miRNA, CNB, TMB, and demographical & clinical (diagnosis age, ER status, PR status) data. All other combinations of multi-omics results are in **Figure S1**.

TABLE 2 | Performances comparison with different combinations of multi-omics data by pairwise paired *t*-test, according to concordance index among 5-folds cross-validation results.

Pairwise paired *T*-test

		Set 2									
		ii		iii		iv		v		vi	
		t	***P***	***t***	***P***	***t***	***P***	***t***	***P***	***t***	***P***
Set 1	**i**	−0.784	0.477	−0.676	0.536	−0.832	0.452	−2.928	0.043*	−3.315	0.030*
	ii	-	-	0.406	0.705	−0.487	0.652	−0.092	0.931	−0.652	0.550
	iii	–	–	–	–	0.247	0.817	−5.804	0.004*	−2.710	0.054
	iv	–	–	–	–	–	–	−4.168	0.014*	−3.603	0.023*
	v	–	–	–	–	–	–	–	–	−1.529	0.201

Note that a negative t-statistic indicated set 1 worse than set 2 in terms of performances. Multi-omics dataset applied as inputs: (i) mRNA-seq data (mRNA) (57 features); (ii) miRNA-seq data (miRNA) (12 features); (iii) integration of mRNA and miRNA (69 features); (iv) integration of mRNA, miRNA, copy number burden (CNB), and tumor mutation burden (TMB) (71 features); (v) integration of mRNA, miRNA, and demographical and clinical (diagnosis age, ER status, PR status) data (72 features); (vi) integration of mRNA, miRNA, CNB, TMB, and demographical and clinical (diagnosis age, ER status, PR status) data (74 features).

*t-denotes the pairwise paired Student's t-test statistic, P denotes the p-value obtained. P-value < 0.05 are considered to be significant and indicated with * symbol.*

the same set of features. In this paper, we define three age groups: (1) age in range 26–50 (191 patients), (2) age in range 51–70 (280 patients), (3) age in range 71–90 (112 patients) to represent younger, middle aged, and elderly patients. By training and testing these three distinct groups with SALMON algorithm, we aim to answer two questions: (1) whether the diagnosis age still be a strong factor that affect prognosis performance; (2) what are the differences of feature rankings between these three distinct groups.

The performances in terms of concordance index by integrating all omics and clinical data (including mRNA, miRNA, CNB, TMB, diagnosis age, ER status, PR status) are shown in **Figure 4**. As expected they are all slightly inferior than the performance when not stratifying patients by age (median = 0.7285; mean = 0.6918), there is not a statistical significant difference. When inspecting the feature rankings, as shown in **Table 4**, we observed that in the age group 26–50, PR status (Progesterone Receptors status) plays a pivotal role in prognosis, while other features do not have substantial contributions to the prognosis including the diagnosis age (we still listed some modules). This situation changed in the age group 51–70 as ER status (Estrogen Receptors status) becomes the most important feature, while diagnosis age ranked at #5 with only marginal contribution. In age group 71–90, neither ER, PR status nor diagnosis age ranked in the front, instead mRNA-seq co-expression modules appeared to have the major influence on prognosis. The top ranked modules are #11, #1, #29, #35, and #4. By performing enrichment analysis, we found that the module #11 is significantly enriched with epithelium development genes (GO:0060429, p = 2.253E-9); module #1 is significantly enriched with chromosome organization genes (GO:0051276, p = 5.344E-17) and two well-known breast cancer genes NCOA3 (Burwinkel et al., 2005) and FOXA1 (Meyer and Carroll, 2012; Rangel et al., 2018) were identified in module 1; module #29 was enriched on cytoband 19q13.41 (p = 1.517E-25) and are exclusively zinc-finger proteins; module #35 was enriched on cytoband 1q34 (p = 1.252E-15) and contains multiple genes which have been previously detected in multiple breast cancer studies including UQCRH, PSMB2, PPIH, and YBX1 (Miller et al., 2005; Pujana et al., 2007; Barry et al., 2010); and module #4 is highly enriched with mitotic cell cycle genes (GO:1903047, p = 2.183E-70) including well-known breast cancer genes such as MKI67 (Gyorffy et al., 2010) and AURKA (Cox et al., 2006). Detailed feature rankings are in **Figures S5**–**S7**.

TABLE 3 | Top features that reduced the concordance index, including two demographical and clinical features, and five mRNA-seq co-expression modules (eigengene matrices as inputs to the SALMON).

Ranks	Feature names	Concordance index changed (median)	Highlighted genes/interpretations/enrichments or notes
1	Diagnosis age	−0.1257	Age
2	PR status	−0.0343	Progesterone receptors status
3	Module 13	−0.0150	Genes MST1, CPT1B. CD8+, CD4+, Breast bulk tissue.
4	Module 47	−0.0071	Genes MAP3K7, CCNC. Cytoband chr6q14-q16 and chr6q21.
5	Module 5	−0.0059	Genes DDR2, FLNA, TCF4. Associated with extracellular matrix (ECM), cell adhesion, and cell migration.
6	Module 36	−0.0053	Gene SNW1. Cytoband chr14q23-q24 and chr14q31-q32.
7	Module 51	−0.0047	Genes TCP1, HDAC2. Cytoband chr6q14-q15and chr6q21-q26.

DISCUSSION

In this work, we demonstrated the feasibility of breast cancer survival prognosis by integrating multi-omics data using Deep Learning-based approaches and opened up a new avenue for deriving new prognostic biomarkers in breast cancer. We introduced our SALMON (Survival Analysis Learning with Multi-Omics Neural Networks) algorithm with the implementation of Cox proportional hazards regression networks in breast invasive carcinoma. Instead of using gene level mRNA-seq or miRNA-seq data directly, SALMON adopts eigengene matrices as the network input derived from weighted gene co-expression network analysis. Unlike other algorithms, SALMON performs forward propagation separately with respect to each type of omics or clinical data in contrast with some other models such as Cox-nnet [which originally did not integrate multi-omics data nor use the co-expression modules as inputs (Ching et al., 2018)]. The separation of forward propagation prevents the interactions across omics data types thus enable easier examination of the module/feature importance for interpretability. It showed good prognosis results in terms of concordance index and log-rank test. Though experiments showed that SALMON has the competitive yet insignificantly superior performance compared to the state-of-the-art Cox-nnet (Ching et al., 2018), we have different paradigm in investigating how prognosis performance increases when integrating more omics and clinical data types, since other models such as Cox-nnet (Ching et al., 2018), DeepSurv (Katzman et al., 2018), etc. do not handle multi-omics data as input. The improved performances (concordance index) by integrating more omics data validates the hypothesis that integrative analysis enhances the survival prognosis accuracy for breast cancer. Moreover, using gene co-expression modules than gene expressions to reduce features upfront is the feature engineering technique we introduced based on bioinformatics techniques. By bridging the gap between gene co-expression analysis and Deep Learning, the advantages can be observed when we backtrack to identify the module/feature can affect the performances. The detected modules reveal clear cancer related biological processes, functions or structural variations allowing further biomedical investigations.

As feature importance has been conveyed and ranked from SALMON, we discovered that keeping only top important

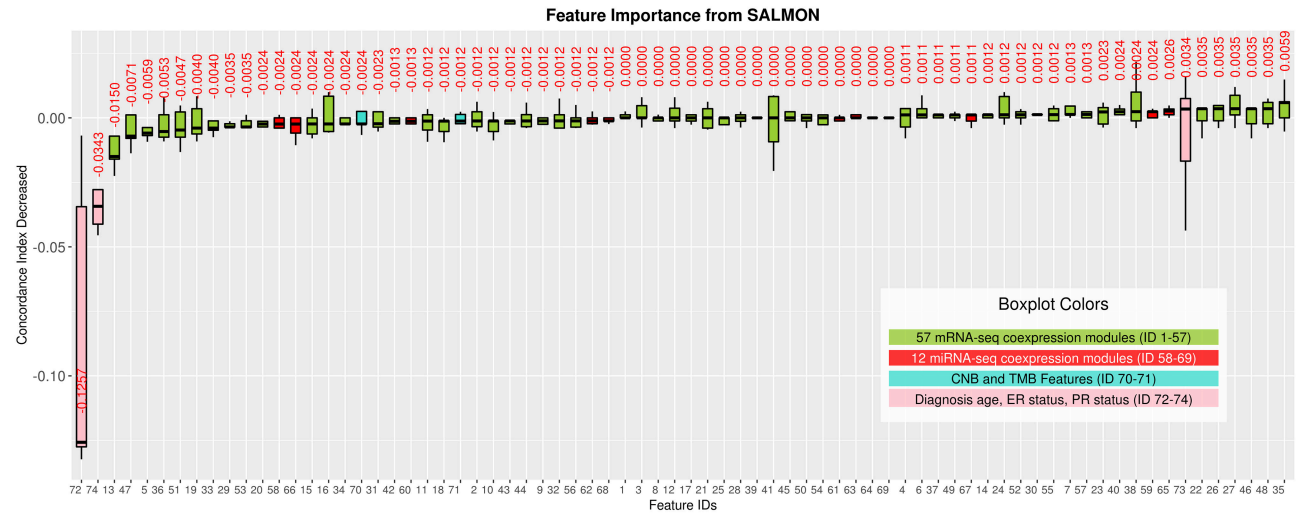

FIGURE 3 | Features importance evaluated by the decrease of concordance index, sorted based on median values. Boxplots in Green: 57 mRNA co-expression module features (ID from 1 to 57); boxplots in red: 12 miRNA co-expression module features (ID from 58 to 69); boxplots in turquoise: copy number burden (CNB) and tumor mutation burden (TMB) features (ID from 70 to 71); boxplots in pink: demographical and clinical features (ID from 72 to 74).

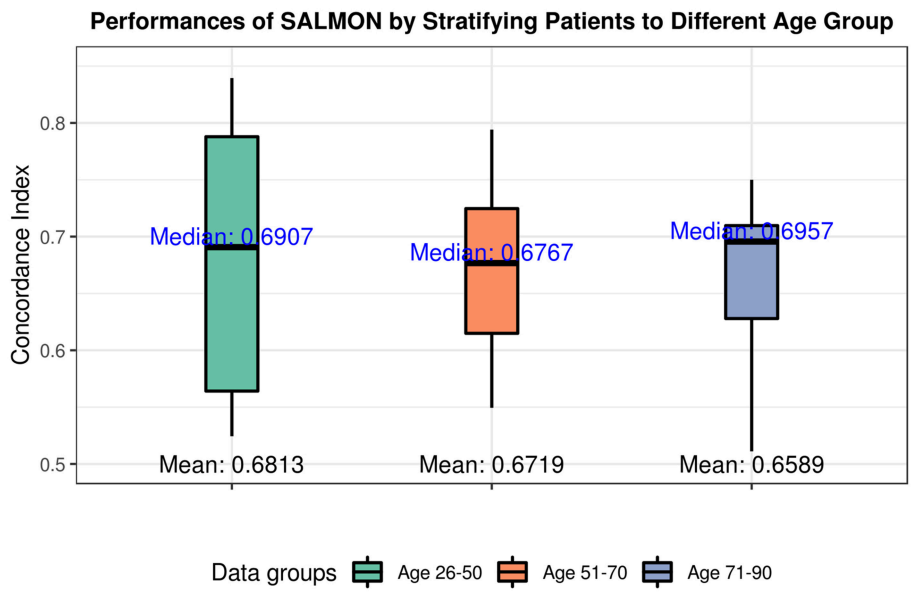

FIGURE 4 | Performances of SALMON algorithm stratified by three age groups: 26–50 group; 51–70 group; 71–90 group with integrating all omics data (integration of mRNA, miRNA, CNB, TMB, diagnosis age, ER status, PR status).

features can still preserve the testing performances. Based on features ranking, we also investigated the biological interpretation behind each demographical feature, clinical feature, and co-expression module. For the leading two features, since the importance of diagnosis age and PR status have been widely examined and recognized in breast cancer (Adami et al., 1986; Boyd et al., 1995; Huang et al., 2000; Bauer et al., 2007) and further confirmed by our results (**Figure 2C**), we focused on the top five mRNA-seq data co-expression modules ranked from 3 to 7. Those top five mRNA-seq data co-expression modules are: module #13, #47, #5, #36, #51.

In module #13, appears to be significantly associated with CD8+ T Cells (p-value = 6.54E-06) and CD4+ T Cells (p-value = 1.50E-02) based on Human Gene Atlas analysis. CD8+ and CD4+ T cells are important components of the immune system, which has been proved to have strong correlation with cancers (Hung et al., 1998; Hadrup et al., 2013). It contains multiple breast cancer related genes: (1) MST1 kinase, a core

TABLE 4 | Top features that reduced the concordance indices.

Ranks	Age group 26–50		Age group 51–70		Age group 71–90	
	Feature names	Concordance index changed (median)	Feature names	Concordance index changed (median)	Feature names	Concordance index changed (median)
1	**PR status**	**–0.0247**	**ER status**	**–0.0807**	**Module 11**	**–0.0323**
2	Module 1	0	Module 13	–0.0221	**Module 1**	**–0.0233**
3	Module 2	0	Module 4	–0.0185	**Module 29**	**–0.0233**
4	Module 3	0	Module 5	–0.0150	**Module 35**	**–0.0233**
5	Module 4	0	Diagnosis age	–0.0150	**Module 4**	**–0.0222**

*Experiments performed separately with three age groups: 26–50 group; 51–70 group; 71–90 group, with integrating all omics data (integration of mRNA, miRNA, CNB, TMB, diagnosis age, ER status, PR status). Detailed feature rankings are in **Figures S5–S7**. The bold values are of our interests and are being discussed.*

component of Hippo pathway, its phosphorylation can inhibit oncoproteins TAZ/YAP and regulate T-cell function. (Arash et al., 2017; Ercolani et al., 2017); (2) CPT1B, which encodes the critical enzyme for fatty acid beta-oxidation (FAO), the inhibition of FAO can inhibit breast cancer stem cells, chemoresistance, and breast tumor growth (Wang et al., 2018). In addition, tissues enrichment analysis using ARCHS4 (https://amp.pharm.mssm.edu/archs4/) also revealed that nearly one third of genes (11 out of 36) in this module were associated with breast cancer bulk tissue (p-value = 1.867E-03) (**Figure S4**).

In module #47, two genes are related to breast cancer have been identified: (1) MAP3K7, also known as TAK1, is a key mediator between survival and cell death in TNF-α-mediated signaling (Totzke et al., 2017); and (2) CCNC, an important transcriptional regulator whose higher expression is associated with shorter relapse-free survival (RFS) and impact the response to adjuvant therapy in breast cancer. Gene amplification of CCNC is also the most frequent type of genetic alterations in breast cancers (Broude et al., 2015). Module #47 was also enriched in cytoband chr6q.

In module #5, genes are highly enriched on tumor microenvironment (TME) related processes such as extracellular matrix (ECM), cell adhesion, and cell migration. Among them, DDR2 plays an indispensable role in a series of hypoxia-induced behaviors of breast cancer cells, such as migration, invasion, and epithelial-mesenchymal transition (EMT), the activated DDR2 can promote the metastasis of breast cancer (Ren et al., 2014). In addition, FLNA, whose overexpression is associated with the advanced stage, lymph node metastasis, and vascular or neural invasion of breast cancer (Feng et al., 2006). It also contributes the development of breast cancer (Tian et al., 2013). Finally, TCF4 is an important transcription factor, its loss is related with breast cancer chemoresistance (Ruiz de Garibay et al., 2018).

In module #36, SNW1 is a component of spliceosome in RNA splicing, its deletion can induce apoptosis, where the inhibition of SNW1 or its associating proteins may be a novel therapeutic strategy for cancer treatment (Sato et al., 2015). Module #36 was also enriched in cytoband chr14q23-q24 and chr14q31-q32.

In module #51, TCP1 functioned as a cytosolic chaperone in the biogenesis of tubulin (Yaffe et al., 1992), which has been proved to have an association with breast cancer (Bassiouni et al., 2016). HDAC2, its overexpression has a correlation with DNA-damage response and promote tumor progression (Shan et al., 2017). Module #51 was also enriched on cytoband chr6q.

Instead of identified breast cancer related genes, the Enrichment analysis in selected modules also revealed important biological functions. Module 47 and 51 were enriched in chr6q. Not surprisingly, previous studies have identified the frequent alterations at chr6q in archival breast cancer specimens (Shadeo and Lam, 2006), while chr6q21 is hotspots copy number alteration region (Chin et al., 2007). The copy number alterations at chr6q26 can affect MAP3K4, plays an important role of epidermal growth factor receptor pathway (Shadeo and Lam, 2006). Module 36 was enriched in chr14q, the cytoband where the high-level alterations at 14q31.3-32.12 were found in breast cancer from Shadeo and Lam (2006). Besides, the deletion of chr14q is a common feature of tumors with BRCA2 mutations (Rouault et al., 2012). Modules 5 was specifically associated with TME related biological process such as extracellular matrix (ECM), cell adhesion and cell migration. All these GO Biological Processes (BPs) have been shown to play pivotal roles in TME development in cancers while TME has now been widely recognized as a critical participant in tumor progression (Quail and Joyce, 2013). Abnormal ECM in tumors can promote the aggressiveness of breast cancer (Robertson, 2016). Cell adhesion as a common event in cancer can promote cell growth as well as tumor dissemination (Moh and Shen, 2009; Saadatmand et al., 2013). All these discoveries not only confirmed the existed literatures for breast cancer, but also justified the feature importance that SALMON generated.

Another interesting finding is that no miRNA-seq module was ranked in top features although miRNA-seq modules show a better prognosis performance than mRNA-seq modules. This could due to the modules within miRNA-seq are more dependent with each other than the modules within mRNA-seq, thus simply knock out one module/feature may not reduce the performance too much. Indeed, by performing pair-wised Pearson correlation analysis, we found 3.03% miRNA-seq modules has strong correlations (Pearson $\rho > 0.8$), while in mRNA-seq modules this ratio is down to 0.94%. It leads us a new perspective to inspect modules dependency in the future.

Since we confirmed that diagnosis age is the most powerful predictor, we examined the feature rankings with three different

age groups, namely, younger group (age 26–50), middle aged group (age 51–70), and elderly group (age 71–90). We confirmed that by separating the 583 patients to three distinct age groups, the diagnosis age becomes unimportant to the prognosis outcome. While in younger group, PR status is the most important feature. In middle aged group, ER status is the most important feature. When we inspected the elderly group with age in range 71–90, we found that only mRNA-seq co-expression modules were ranked at the top and the five most conspicuous ones are modules #11, #1, #29, #35, and #4. These observations suggest that specific biological processes may play different roles in breast cancer patients of different ages while different biomarkers and predictive models may be needed for different age groups. Further inspection of the modules found that three of these modules are related to known breast cancer related processes such as epithelium development (Vincent-Salomon and Thiery, 2003), chromosome organization (Muleris et al., 1995), and mitotic cell cycle (Kastan and Bartek, 2004) including well-known breast cancers genes such as NCOA3, AURKA, MKI67, and FOXA1. The other two modules are highly enriched on specific cytobands on different chromosomes, implying potential copy number variations on these regions. Indeed, both cytobands (19q13.41 and 1q34) are known to be associated with breast cancer outcomes (Han et al., 2006; Ton et al., 2009). For module #35, while most of the genes locate on 1q34, many of the genes such as UQCRH, PSMB2, PPIH, and YBX1 are involved in RNA processing and have been identified with breast cancer in multiple studies (Miller et al., 2005; Pujana et al., 2007; Barry et al., 2010). Interestingly, all genes identified from module #29 are zinc finger transcription factors. While it is not clear if any of them are specifically related to breast cancer, it is of great interest to further investigate the roles of the ZNF family genes in breast cancer development.

CONCLUSION

We performed survival prognosis on breast cancer, proposed a Deep Learning-based algorithm SALMON (Survival Analysis Learning with Multi-Omics Network) by integrating Cox proportional hazards model and adopting gene co-expression network analysis results as input, and predict patient hazard ratios precisely. Performances (concordance index and log-rank test *p*-value) improved when more omics data integrated to the input of SALMON. SALMON also showed a competitive performance compared to other Deep Learning survival prognosis model. By inspecting how each feature contributes to the hazard ratios, SALMON confirmed certain mRNA-seq co-expression modules and clinical information, which play pivotal roles in breast cancer prognosis, revealed several biological functions. By further stratifying patients with diagnosis age, SALMON confirmed that different age groups have different main features that controls survival prognosis performance. To sum up, SALMON fuses the gene co-expression network analysis, Deep Learning technique, feature selection, Cox proportional hazard model, integrative analysis, and module-level enrichment analysis altogether, offers a new avenue for the future integrative analysis and Deep Learning-based cancer survival prognosis.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or the supplementary files.

AUTHOR CONTRIBUTIONS

ZHu conceived and designed the algorithm and analysis, conducted the experiments, and wrote the paper. XZ, KH, ZHu performed the biological analysis and wrote the paper. SX, JZ performed the biological analysis. TJ, CY, ZHa collected the data. TJ, BH, KH edited the paper. JZ, PS, MR, ZHa, KH provided the research guide. PS, ZHa, KH supervised this project.

FUNDING

This work was partially supported by Indiana University School of Medicine (IUSM) start-up fund (JZ), the National Cancer Institute Informatics Technology for Cancer Research (NCI ITCR) U01 [CA188547] (KH, JZ), Indiana University Precision Health Initiative (KH, JZ, ZHu, ZHa, TJ, BH, CY), and Shenzhen Peacock Plan [KQTD2016053112051497] (XZ, SX).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00166/full#supplementary-material

REFERENCES

Adami, H. O., Malker, B., Holmberg, L., Persson, I., and Stone, B. (1986). The relation between survival and age at diagnosis in breast cancer. *N. Engl. J. Med.* 315, 559–563. doi: 10.1056/NEJM198608283150906

Arash, E. H., Shiban, A., Song, S. Y., and Attisano, L. (2017). MARK4 inhibits Hippo signaling to promote proliferation and migration of breast cancer cells. *EMBO Rep.* 18, 420–436. doi: 10.15252/embr.201642455

Barry, W. T., Kernagis, D. N., Dressman, H. K., Griffis, R. J., Hunter, J. D., Olson, J. A., et al. (2010). Intratumor heterogeneity and precision of microarray-based predictors of breast cancer biology and clinical outcome. *J. Clin. Oncol.* 28, 2198–2206. doi: 10.1200/JCO.2009.26.7245

Bassiouni, R., Nemec, K. N., Iketani, A., Flores, O., Showalter, A., Khaled, A. S., et al. (2016). Chaperonin containing TCP-1 protein level in breast cancer cells predicts therapeutic application of a cytotoxic peptide. *Clin. Cancer Res.* 22, 4366–4379. doi: 10.1158/1078-0432.CCR-15-2502

Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A., and Caggiano, V. (2007). Descriptive analysis of estrogen receptor (ER)negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype - a population-based study from the California Cancer Registry. *Cancer* 109, 1721–1728. doi: 10.1002/cncr.22618

Birkbak, N. J., Kochupurakkal, B., Izarzugaza, J. M., Eklund, A. C., Li, Y., Liu, J., et al. (2013). Tumor mutation burden forecasts outcome in ovarian cancer with BRCA1 or BRCA2 mutations. *PLoS ONE* 8:e80023. doi: 10.1371/journal.pone.0080023

Boyd, N. F., Byng, J. W., Jong, R. A., Fishell, E. K., Little, L. E., Miller, A. B., et al. (1995). Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. *J. Natl. Cancer Inst.* 87, 670–675. doi: 10.1093/jnci/87.9.670

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

Broude, E. V., Gyorffy, B., Chumanevich, A. A., Chen, M. Q., McDermott, M. S. J., Shtutman, M., et al. (2015). Expression of CDK8 and CDK8-interacting genes as potential biomarkers in breast cancer. *Curr. Cancer Drug Targets* 15, 739–749. doi: 10.2174/1568009615081510011058l4

Burwinkel, B., Wirtenberger, M., Klaes, R., Schmutzler, R. K., Grzybowska, E., Forsti, A., et al. (2005). Association of NCOA3 polymorphisms with breast cancer risk. *Clin. Cancer Res.* 11, 2169–2174. doi: 10.1158/1078-0432.CCR-04-1621

Caruana, R., Lawrence, S., and Giles, L. (2001). "Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping," in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000* (Denver, CO), 402–408.

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., et al. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 9:34. doi: 10.1186/s13073-017-0424-2

Chandran, V., Coppola, G., Nawabi, H., Omura, T., Versano, R., Huebner, E. A., et al. (2016). A systems-level analysis of the peripheral nerve intrinsic axonal growth program. *Neuron* 89, 956–970. doi: 10.1016/j.neuron.2016.01.034

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* 24, 1248–1259. doi: 10.1158/1078-0432.CCR-17-0853

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427

Cheng, J., Zhang, J., Han, Y., Wang, X., Ye, X., Meng, Y., et al. (2017). Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Res.* 77, e91–e100. doi: 10.1158/0008-5472.CAN-17-0313

Chin, S. F., Teschendorff, A. E., Marioni, J. C., Wang, Y., Barbosa-Morais, N. L., Thorne, N. P., et al. (2007). High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.* 8: R215 . doi: 10.1186/gb-2007-8-10-r215

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* 14:e1006076 . doi: 10.1371/journal.pcbi.1006076

Cox, D. G., Hankinson, S. E., and Hunter, D. J. (2006). Polymorphisms of the AURKA (STK15/Aurora Kinase) gene and breast cancer risk (United States). *Cancer Causes Control* 17, 81–83. doi: 10.1007/s10552-005-0429-9

Ercolani, C., Di Benedetto, A., Terrenato, I., Pizzuti, L., Di Lauro, L., Sergi, D., et al. (2017). Expression of phosphorylated Hippo pathway kinases (MST1/2 and LATS1/2) in HER2-positive and triple-negative breast cancer patients treated with neoadjuvant therapy. *Cancer Biol. Ther.* 18, 339–346. doi: 10.1080/15384047.2017.1312230

Feng, Y., Chen, M. H., Moskowitz, I. P., Mendonza, A. M., Vidali, L., Nakamura, F., et al. (2006). Filamin A (FLNA) is required for cell-cell contact in vascular development and cardiac morphogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19836–19841. doi: 10.1073/pnas.0609628104

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias variance dilemma. *Neural Comput.* 4, 1–58. doi: 10.1162/neco.1992.4.1.1

Golub, G. H., and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Math.* 14, 403–420. doi: 10.1007/BF02163027

Goodman, A. M., Kato, S., Bazhenova, L., Patel, S. P., Frampton, G. M., Miller, V., et al. (2017). Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* 16, 2598–2608. doi: 10.1158/1535-7163.MCT-17-0386

Gupta, A., Mutebi, M., and Bardia, A. (2015). Gene-expression-based predictors for breast cancer. *Ann. Surg. Oncol.* 22, 3418–3432. doi: 10.1245/s10434-015-4703-0

Gyorffy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., et al. (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* 123, 725–731. doi: 10.1007/s10549-009-0674-9

Hadrup, S., Donia, M., and Thor Straten, P. (2013). Effector CD4 and CD8 T cells and their role in the tumor microenvironment. *Cancer Microenviron.* 6, 123–133. doi: 10.1007/s12307-012-0127-6

Han, W., Han, M. R., Kang, J. J., Bae, J. Y., Lee, J. H., Bae, Y. J., et al. (2006). Genomic alterations identified by array comparative genomic hybridization as prognostic markers in tamoxifen-treated estrogen receptor-positive breast cancer. *BMC Cancer* 6:92. doi: 10.1186/1471-2407-6-92

Han, Z., Johnson, T., Zhang, J., Zhang, X., and Huang, K. (2017). Functional virtual flow cytometry: a visual analytic approach for characterizing single-cell gene expression patterns. *Biomed. Res. Int.* 2017:3035481. doi: 10.1155/2017/303548

Han, Z., Zhang, J., Sun, G., Liu, G., and Huang, K. (2016). A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules. *BMC Genomics* 17(Suppl. 7):519. doi: 10.1186/s12864-016-2912-y

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., et al. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 13:R97. doi: 10.1186/gb-2012-13-10-r97

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084

Huang, W. Y., Newman, B., Millikan, R. C., Schell, M. J., Hulka, B. S., and Moorman, P. G. (2000). Hormone-related factors and risk of breast cancer in relation to estrogen receptor and progesterone receptor status. *Am. J. Epidemiol.* 151, 703–714. doi: 10.1093/oxfordjournals.aje.a010265

Hung, K., Hayashi, R., Lafond-Walker, A., Lowenstein, C., Pardoll, D., and Levitsky, H. (1998). The central role of CD4(+) T cells in the antitumor immune response. *J. Exp. Med.* 188, 2357–2368. doi: 10.1084/jem.188.12.2357

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* 2, 841–860. doi: 10.1214/08-AOAS169

Kastan, M. B., and Bartek, J. (2004). Cell-cycle checkpoints and cancer. *Nature* 432, 316–323. doi: 10.1038/nature03097

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18:24. doi: 10.1186/s12874-018-0482-1

Kingma, D., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Liu, G., Dong, C., and Liu, L. (2016). Integrated multiple "-omics" data reveal subtypes of hepatocellular carcinoma. *PLoS ONE* 11:e0165457. doi: 10.1371/journal.pone.0165457

Ma, T., and Zhang, A. (2018). "Multi-view factorization AutoEncoder with network constraints for multi-omic integrative analysis," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE). doi: 10.1109/BIBM.2018.8621379

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* 50, 163–170.

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., et al. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* 49, 27–35. doi: 10.1038/ng.3725

Meyer, K. B., and Carroll, J. S. (2012). FOXA1 and breast cancer risk. *Nat. Genet.* 44, 1176–1177. doi: 10.1038/ng.2449

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., et al. (2005). An expression signature for p53 status in human

breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13550–13555. doi: 10.1073/pnas.0506230102

Moh, M. C., and Shen, S. L. (2009). The roles of cell adhesion molecules in tumor suppression and cell migration a new paradox. *Cell Adh. Migr.* 3, 334–336. doi: 10.4161/cam.3.4.9246

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24, 69–71.

Muleris, M., Almeida, A., Gerbault-Seureau, M., Malfoy, B., and Dutrillaux, B. (1995). Identification of amplified DNA sequences in breast cancer and their organization within homogeneously staining regions. *Genes Chromosomes Cancer* 14, 155–163. doi: 10.1002/gcc.2870140302

Nagini, S. (2017). Breast cancer: current molecular therapeutic targets and new players. *Anticancer. Agents Med. Chem.* 17, 152–163. doi: 10.2174/1871520616666160502122724

Nassar, F. J., Nasr, R., and Talhouk, R. (2017). MicroRNAs as biomarkers for early breast cancer diagnosis, prognosis and therapy prediction. *Pharmacol. Ther.* 172, 34–49. doi: 10.1016/j.pharmthera.2016.11.012

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). "On spectral clustering: analysis and an algorithm," in *Advances in Neural Information Processing Systems* (MIT Press), 849–856.

Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K., Provenzano, E., Bardwell, H. A., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* 7:11479. doi: 10.1038/ncomms11479

Poirion, O. B., Chaudhary, K., and Garmire, L. X. (2018). Deep learning data integration for better risk stratification models of bladder cancer. *AMIA Jt. Summits Transl. Sci. Proc.* 2017, 197–206.

Pujana, M. A., Han, J. D., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., et al. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.* 39, 1338–1349. doi: 10.1038/ng.2007.2

Quail, D. F., and Joyce, J. A. (2013). Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* 19, 1423–1437. doi: 10.1038/nm.3394

Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S., and Sidow, A. (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. Commun.* 9:4453. doi: 10.1038/s41467-018-06921-8

Rangel, N., Fortunati, N., Osella-Abate, S., Annaratone, L., Isella, C., Catalano, M. G., et al. (2018). FOXA1 and AR in invasive breast cancer: new findings on their co-expression and impact on prognosis in ER-positive patients. *BMC Cancer* 18:703. doi: 10.1186/s12885-018-4624-y

Ren, T., Zhang, W., Liu, X., Zhao, H., Zhang, J., Zhang, J., et al. (2014). Discoidin domain receptor 2 (DDR2) promotes breast cancer cell metastasis and the mechanism implicates epithelial-mesenchymal transition programme under hypoxia. *J. Pathol.* 234, 526–537. doi: 10.1002/path.4415

Robertson, C. (2016). The extracellular matrix in breast cancer predicts prognosis through composition, splicing, and crosslinking. *Exp. Cell Res.* 343, 73–81. doi: 10.1016/j.yexcr.2015.11.009

Rouault, A., Banneau, G., MacGrogan, G., Jones, N., Elarouci, N., Barouk-Simonet, E., et al. (2012). Deletion of chromosomes 13q and 14q is a common feature of tumors with BRCA2 mutations. *PLoS ONE* 7:e52079. doi: 10.1371/journal.pone.0052079

Ruiz de Garibay, G., Mateo, F., Stradella, A., Valdes-Mas, R., Palomero, L., Serra-Musach, J., et al. (2018). Tumor xenograft modeling identifies an association between TCF4 loss and breast cancer chemoresistance. *Dis. Model. Mech.* 11:dmm032292. doi: 10.1242/dmm.032292

Saadatmand, S., de Kruijf, E. M., Sajet, A., Dekker-Ensink, N. G., van Nes, J. G. H., Putter, H., et al. (2013). Expression of cell adhesion molecules and prognosis in breast cancer. *Br. J. Surg.* 100, 252–260. doi: 10.1002/bjs.8980

Santosa, F., and Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* 7, 1307–1330. doi: 10.1137/0907087

Sato, N., Maeda, M., Sugiyama, M., Ito, S., Hyodo, T., Masuda, A., et al. (2015). Inhibition of SNW1 association with spliceosomal proteins promotes apoptosis in breast cancer cells. *Cancer Med.* 4, 268–277. doi: 10.1002/cam4.366

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Setiono, R., and Liu, H. (1997). Neural-network feature selector. *IEEE Trans. Neural Netw.* 8, 654–662. doi: 10.1109/72.572104

Shadeo, A., and Lam, W. L. (2006). Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res.* 8:R9. doi: 10.1186/bcr1370

Shan, W., Jiang, Y., Yu, H., Huang, Q., Liu, L., Guo, X., et al. (2017). HDAC2 overexpression correlates with aggressive clinicopathological features and DNA-damage response pathway of breast cancer. *Am. J. Cancer Res.* 7, 1213–1226.

Shao, W., Cheng, J., Sun, L., Han, Z., Feng, Q., Zhang, D., et al. (2018). "Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Granada: Springer), 648–656.

Shroff, S., Zhang, J., and Huang, K. (2016). Gene co-expression analysis predicts genetic variants associated with drug responsiveness in lung cancer. *AMIA Jt Summits Transl. Sci. Proc.* 2016, 32–41.

Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. (2007). "On ranking in survival analysis: bounds on the concordance index," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 1209–1216.

Sun, D., Wang, M., and Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 99:1. doi: 10.1109/TCBB.2018.2806438

Sung, A. H., and Mukkamala, S. (2003). "Identifying important features for intrusion detection using support vector machines and neural networks," in *2003 Symposium on Applications and the Internet, Proceedings* (Orlando, FL), 209–216. doi: 10.1109/SAINT.2003.1183050

Thomas, A., Routh, E. D., Pullikuth, A., Jin, G., Su, J., Chou, J. W., et al. (2018). Tumor mutational burden is a determinant of immune-mediated survival in breast cancer. *Oncoimmunology* 7:e1490854. doi: 10.1080/2162402X.2018.1490854

Tian, H. M., Liu, X. H., Han, W., Zhao, L. L., Yuan, B., and Yuan, C. J. (2013). Differential expression of filamin A and its clinical significance in breast cancer. *Oncol. Lett.* 6, 681–686. doi: 10.3892/ol.2013.1454

Ton, C., Guenthoer, J., and Porter, P. L. (2009). "Somatic alterations and implications in breast cancer," in *Role of Genetics in Breast and Productive Cancers*, ed P. Welcsh (Seattle, WA: Springer), 183–213. doi: 10.1007/978-1-4419-0477-5_9

Totzke, J., Gurbani, D., Raphemot, R., Hughes, P. F., Bodoor, K., Carlson, D. A., et al. (2017). Takinib, a selective TAK1 inhibitor, broadens the therapeutic efficacy of TNF-alpha inhibition for cancer and autoimmune disease. *Cell Chem. Biol.* 24, 1029–1039 e1027. doi: 10.1016/j.chembiol.2017.07.011

Vincent-Salomon, A., and Thiery, J. P. (2003). Host microenvironment in breast cancer development: epithelial-mesenchymal transition in breast cancer development. *Breast Cancer Res.* 5, 101–106. doi: 10.1186/bcr578

Wang, T., Fahrmann, J. F., Lee, H., Li, Y. J., Tripathi, S. C., Yue, C., et al. (2018). JAK/STAT3-regulated fatty acid beta-oxidation is critical for breast cancer stem cell self-renewal and chemoresistance. *Cell Metab.* 27, 136–150 e135. doi: 10.1016/j.cmet.2018.04.018

Weigend, A. S., Rumelhart, D. E., and Huberman, B. A. (1990). "Generalization by weight-elimination with application to forecasting," in *Advances in Neural Information Processing Systems* (Denver, CO), 875–882

Wu, X. F., Ye, Y. Q., Barcenas, C. H., Chow, W. H., Meng, Q. H., Chavez-MacGregor, M., et al. (2017). Personalized prognostic prediction models for breast cancer recurrence and survival incorporating multidimensional data. *J. Natl. Cancer Inst.* 109. doi: 10.1093/jnci/djw314

Xiang, S., Huang, Z., Wang, T., Han, Z., Yu, C. Y., Ni, D., et al. (2018). Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. 11:115. doi: 10.1186/s12920-018-0431-1

Xiang, Y., Zhang, C. Q., and Huang, K. (2012). Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. *BMC Bioinformatics* 13:S12. doi: 10.1186/1471-2105-13-S2-S12

Yaffe, M. B., Farr, G. W., Miklos, D., Horwich, A. L., Sternlicht, M. L., and Sternlicht, H. (1992). TCP1 complex is a molecular chaperone in tubulin biogenesis. *Nature* 358, 245–248. doi: 10.1038/358245a0

Zhang, G. Q. P. (2000). Neural networks for classification: a survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 30, 451–462. doi: 10.1109/5326.897072

Zhang, J., Abrams, Z., Parvin, J. D., and Huang, K. (2016). Integrative analysis of somatic mutations and transcriptomic data to functionally stratify breast cancer patients. *BMC Genomics* 17(Suppl. 7):513. doi: 10.1186/s12864-016-2902-0

Zhang, J., and Huang, K. (2014). Normalized lmQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. *Cancer Inform.* 13(Suppl. 3), 137–146. doi: 10.4137/CIN.S14021

Zhang, J., and Huang, K. (2017). Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *BMC Genomics* 18(Suppl. 1):1045. doi: 10.1186/s12864-016-3259-0

Zhang, J., Lu, K. W., Xiang, Y., Islam, M., Kotian, S., Kais, Z., et al. (2012). Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Comput. Biol.* 8:e1002656. doi: 10.1371/journal.pcbi.1002656

Zhang, J., Ni, S., Xiang, Y., Parvin, J. D., Yang, Y., Zhou, Y., et al. (2013). Gene co-expression analysis predicts genetic aberration loci associated with colon cancer metastasis. *Int. J. Comput. Biol. Drug Des.* 6, 60–71. doi: 10.1504/IJCBDD.2013.052202

Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., et al. (2018a). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front. Genet.* 9:477. doi: 10.3389/fgene.2018.00477

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., et al. (2018b). Deep learning in omics: a survey and guideline. *Brief. Funct. Genomics.* 18, 41–57. doi: 10.1093/bfgp/ely030

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

frontiers in Genetics

REVIEW
published: 26 March 2019
doi: 10.3389/fgene.2019.00214

Recent Advances of Deep Learning in Bioinformatics and Computational Biology

Binhua Tang[1,2*†], *Zixiang Pan*[1†], *Kang Yin*[1] *and Asif Khateeb*[1]

[1] Epigenetics & Function Group, Hohai University, Nanjing, China, [2] School of Public Health, Shanghai Jiao Tong University, Shanghai, China

Extracting inherent valuable knowledge from omics big data remains as a daunting problem in bioinformatics and computational biology. Deep learning, as an emerging branch from machine learning, has exhibited unprecedented performance in quite a few applications from academia and industry. We highlight the difference and similarity in widely utilized models in deep learning studies, through discussing their basic structures, and reviewing diverse applications and disadvantages. We anticipate the work can serve as a meaningful perspective for further development of its theory, algorithm and application in bioinformatic and computational biology.

Keywords: computational biology, bioinformatics, application, algorithm, deep learning

OPEN ACCESS

Edited by:
Juan Caballero, Universidad Autónoma de Querétaro, Mexico

Reviewed by:
Wenhai Zhang, Hengyang Normal University, China
Zhuliang Yu, South China University of Technology, China

***Correspondence:**
Binhua Tang
bh.tang@hhu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *20 August 2018*
Accepted: *27 February 2019*
Published: *26 March 2019*

Citation:
Tang B, Pan Z, Yin K and Khateeb A (2019) Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Front. Genet. 10:214. doi: 10.3389/fgene.2019.00214

INTRODUCTION

Deep learning is the emerging generation of the artificial intelligence techniques, specifically in machine learning. The earliest artificial intelligence was firstly implemented on hardware system in the 1950s. The newer concept with the more systematic theorems, named machine learning, appeared in the 1960s. And its newly-evolved branch, deep learning, was first brought up around the 2000s, and soon led to rapid applications in different fields, due to its unprecedented prediction performance on big data (Hinton and Salakhutdinov, 2006; LeCun et al., 2015; Nussinov, 2015).

The basic concepts and models in deep learning have derived from the artificial neural network, which mimic human brain's activity pattern to intelligentize the algorithms and save tedious human labor (Mnih et al., 2015; Schmidhuber, 2015; Mamoshina et al., 2016). Although deep learning is an emerging subfield recently from machine learning, it has immense utilizations spreading from machine vision, voice, and signal processing, sequence and text prediction, and computational biology topics, altogether shaping the productive AI fields (Bengio and LeCun, 2007; Alipanahi et al., 2015; Libbrecht and Noble, 2015; Zhang et al., 2016; Esteva et al., 2017; Ching et al., 2018). Deep learning has several implementation models as artificial neural network, deep structured learning, and hierarchical learning, which commonly apply a class of structured networks to infer the quantitative properties between responses and causes within a group of data (Ditzler et al., 2015; Liang et al., 2015; Xu J. et al., 2016; Giorgi and Bader, 2018).

The subsequent paragraphs mainly summarize the essential concepts and recent applications of deep learning, together highlight the key achievements and future directions of deep learning, especially from the perspectives of bioinformatics and computational biology.

ESSENTIAL CONCEPTS IN DEEP NEURAL NETWORK

Basic Structure of Neural Network

Neural network is a class of information processing modules, frequently utilized in machine learning. Within a multi-layer context, the basic building units, namely neurons, are connected to each other among the adjacent layers via internal links, but the neurons belonging to the same layer have no connection, as depicted in **Figure 1**.

In **Figure 1**, each hidden layer processes its inputs via a connection function denoted as below,

$$h_{W,b}(X) = f(W^T X + b) \tag{1}$$

where W refers to the weight and b for bias. When all input layer neurons are active, each input neuron will multiply their respective weight matrix and the output will be summed up with a bias, which then will be fed into an adjacent hidden layer. Although the input-output formalization may repeat similarly among hidden layers, there is usually no direct connection between neurons within the same layer. And activation function is to quantify the connection between two neighboring neurons across two (hidden) layers.

Specifically, the input of the activation function is the combination $W^T X+b$ denoted in Equation (1), and the function output is then fed into the next neuron as a new input. Following the connection formula, the former input feature can be extracted to the next layer; by this means the features can be well-extracted and refined further. And the performance of the feature extraction depends significantly on the selection of the activation function.

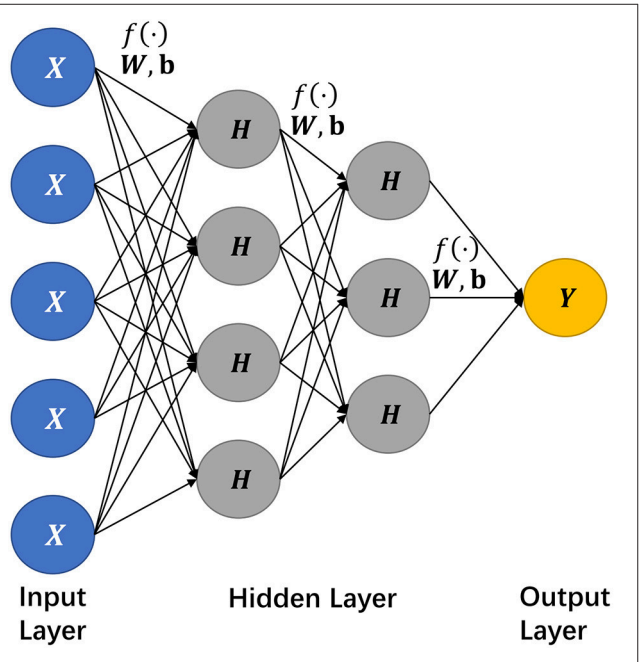

FIGURE 1 | The network structure of a deep learning model. Here we select a network structure with two hidden layers as an illustration, where *X* nodes constitute the input layer, *H*s for the hidden layers, *Y* for the output layer, and *f*(·) denotes an activation function.

Before training the network structure, the input raw datasets are usually separated into two or three groups, namely a training set and a test set, sometimes a validation set to examine the performance of previously trained network models, as depicted in **Figure 2**. In practice, the original datasets are separated stochastically to avoid the potential local tendency, but the proportion of each set can be determined manually.

Learning by Training, Validation, and Testing

Normally, training a neural network refers to a process the network self-tunes its parameters or weights to meet the prespecified performance criteria, thus the trained model can be further used in regression or classification purposes. As depicted in **Figure 2**, generally a complete dataset collected from a specific experiment beforehand can be split into the training and testing, and even validation sets, then followed by conventional tasks as model training, validation and performance comparison.

During training with initial batches of data samples, model parameters and their characteristics normally can be tuned by various learning paradigms, including appropriate activation and rectification functions. Then the trained network should be further tested or even validated with the other batch of samples, to acquire high robustness and satisfactory predictability, the processes of which are often referred as model testing and validation.

Usually, the three procedures above are faithfully implemented in conventional machine learning studies; and even in its quickly-evolving subfield, deep learning, the similar paradigm is always observed (LeCun et al., 2015; Schmidhuber, 2015).

Activation and Loss Function

After training completed, the neural network can perform regression or classification task on testing data, while there usually exists the difference between the predicted outputs and actual values. And the difference should be minimized to acquire optimal model performance.

Within a certain layer, error reduction requires scaling it back within a preset range before passing it onto the next layer of neurons. Activation herein is defined to control neurons' outputs in "active" or "inactive" status, using those non-linear functions as rectified linear unit (ReLU), tanh, and logistic (Sigmoid or soft step) (LeCun et al., 2015).

Besides, a loss function herein is to measure the total difference between the predicted and accurate values, through fine-tuning in backpropagation process. And it acts as an ending threshold for parameter optimization by means of iteratively evaluating the trained models.

With activation function in each neuron throughout diverse layers, a training procedure will continue searching a whole hyperparameter space till the ending threshold, compare and detect an optimal parameter combination by minimizing the preset loss function.

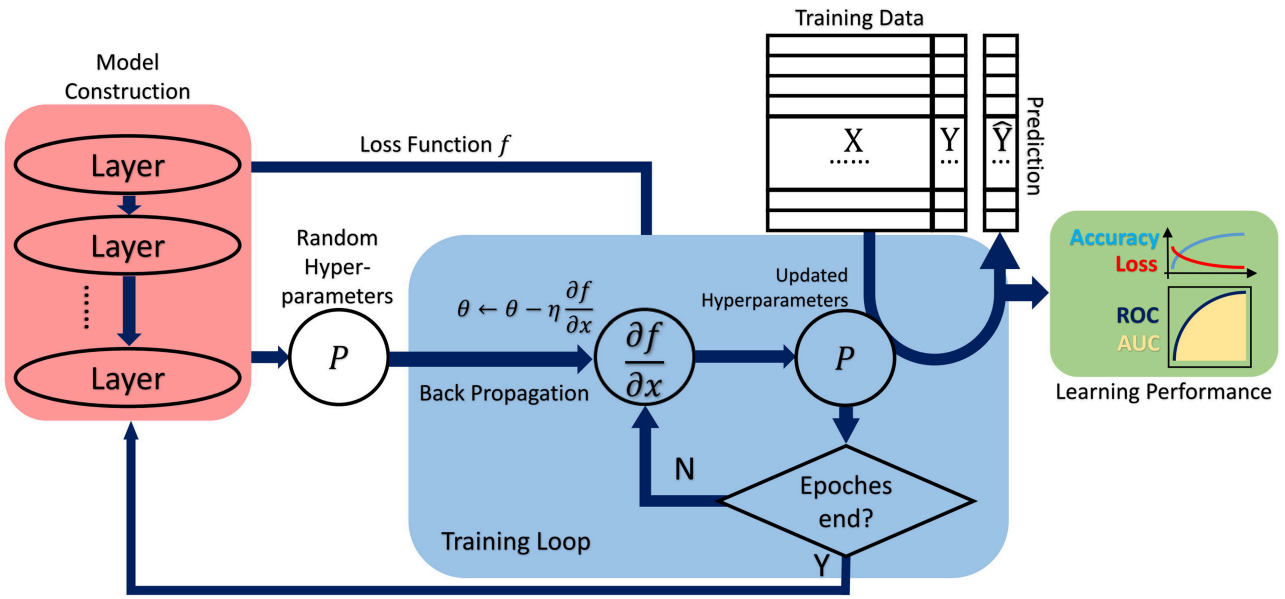

FIGURE 2 | The general analysis procedure commonly adopted in deep learning, which covers training data preparation, model construction, hyperparameter fine-tuning (in training loop), prediction and performance evaluation. Basically, it still follows the requisite schema in machine learning.

TYPICAL ALGORITHMS AND APPLICATIONS

With the substantial progresses in advanced computation and Graphic Processing Unit (GPU) technologies, systematic interrogation into massive data to understand its inherent mechanisms becomes possible, especially through deep learning approaches. Hereinafter, we illustrated several frequently utilized models in deep learning literatures, in both recent computation theories and diverse applications.

Recurrent Neural Network

Recurrent Neural Network (RNN) is a deep learning model different from traditional neural networks, since the former can integrate the previously learned status through a recurrent approach, namely backpropagation; while traditional neural network usually outputs prediction based on the status of the current layer.

Compared with traditional network models, RNN only has one hidden layer but it can unfold horizontally, and multi-vertical-groups are enabled to utilize most of the previous results, namely "using memory".

As depicted in **Figure 3**, the hidden layer neuron H_n is defined by Equation (2),

$$H_n = \sigma_1(W_{1,n}^T H_{n-1} + W_{2,n}^T X_n + b_{1,n}) \tag{2}$$

where $W_{1,n}$ and $W_{2,n}$ represent weight matrix, $b_{1,n}$ is a bias matrix, and $\sigma(\cdot)$ (usually $tanh(\cdot)$) for an activation function. Thus, each layer will generate a partial of output from the current hidden layer neuron with a weight matrix $W_{3,n}$ and bias $b_{2,n}$, defined by Equation (3),

$$\hat{Y}_n = \sigma_2(W_{3,n} H_n + b_{2,n}) \tag{3}$$

And the total loss L_{total} will be the sum of the loss functions from each hidden layer, defined as below,

$$L_{total} = \sum_{n=1}^{N} L_n = \sum_{n=1}^{N} L(\hat{Y}, Y) \tag{4}$$

Thus, fine tuning of RNN backpropagation is based on three weights, $W_{1,n}$, $W_{2,n}$, and $W_{3,n}$. Since the multi-parameter setting in weights adds to the optimization burden, RNN usually performs worse than Convolutional Neural Network (CNN) in terms of fine-tuning. But frequently it is ensembled with CNN in diverse applications, such as dimension reduction, image, and video processing (Hinton and Salakhutdinov, 2006; Hu and Lu, 2018). Angermueller et al. proposed an ensembled RNN-CNN architecture, DeepCpG, on single-cell DNA methylation data, to better predict missing CpG status for genome-wide analysis; together the model's interpretable parameters shed light on the connection between sequence composition and methylation variability (Angermueller et al., 2017). Section Autoencoder will specifically discuss CNN and its typical applications.

Moreover, RNN outperforms those conventional models as logistic regression and SVM, and it can be implemented in various environments, accelerated by GPUs (Li et al., 2017). Due to its structural characteristics, RNN is suitable to deal with long and sequential data, such as DNA array and genomics sequence (Pan et al., 2008; Ray et al., 2009; Jolma et al., 2013; Lee and Young, 2013; Alipanahi et al., 2015; Xu T. et al., 2016).

But RNN cannot interact with hidden neurons far from the current one. To construct an efficient framework

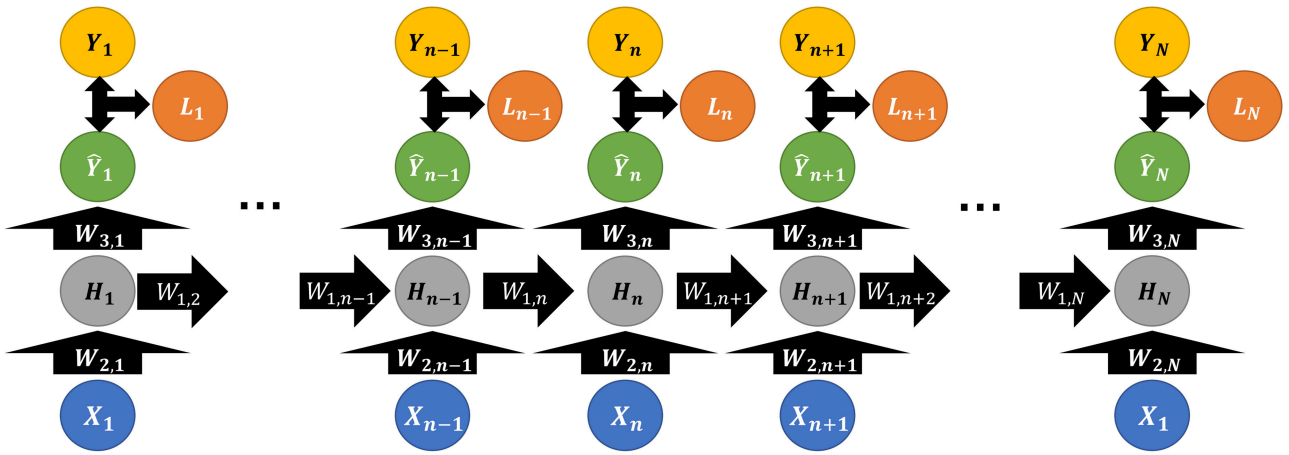

FIGURE 3 | Illustrative structure diagram of Recurrent Neural Network, where *X*, *Y*, and *W* are defined the same as above; L_i denotes the loss function between the actual Y_i and predicted $\hat{Y}_i$ ($i \in N$).

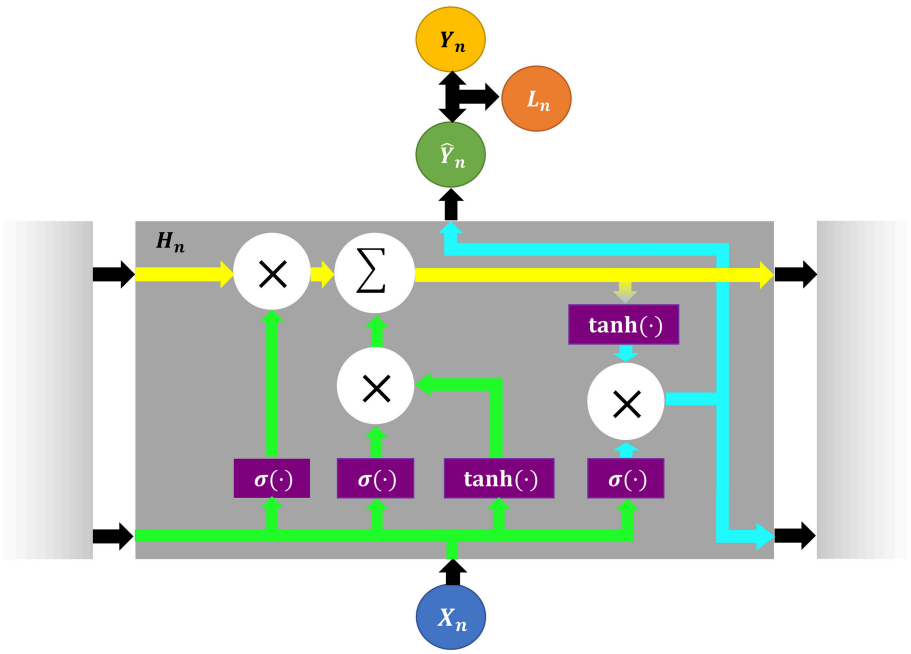

FIGURE 4 | The LSTM network structure and its general information flow chart, where *X*, *Y*, and *W* are defined the same as above.

of recalling deep memory, many improved algorithms have been proposed, like BRNN in protein secondary structure prediction (Baldi et al., 1999), and MD-RNN in analyzing electron microscopy and MRIs of breast cancer samples (Kim et al., 2018).

LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are two recently-improved derivatives of RNN to solve the long-time dependence issues. GRU shares a similar structure with LSTM, which has several gates used for modeling its memory center. The current memory output is jointly influenced by its current input feature, the context (namely the past influence), and the inner action toward the input, as shown in **Figure 4**.

In **Figure 4**, the yellow track refers to an input gate transfering its total past features, and is accessible for any new feature to be added. The green track is a mixture of an input gate and its former hidden layer neurons; and it decides what to omit, namely resetting activation function close to 0, and what to be updated into the yellow track. The blue track is the output gate integrating the inner influence from the yellow track, and it decides the output of the current hidden neurons and what to be passed to the next hidden neuron.

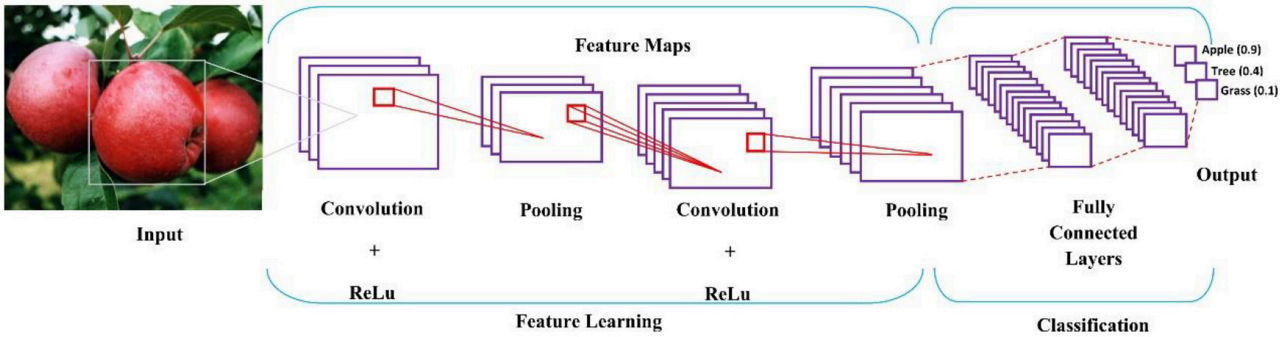

FIGURE 5 | The basic architecture and analysis procedure of a CNN model, which illustrates a classification procedure for an apple on a tree.

Recently an attention-based architecture, DeepDiff, utilizes a hierarchy of LSTM modules to characterize how various histone modifications cooperate simultaneously, and it can effectively predict cell-type-specific gene expression (Sekhon et al., 2018).

Convolutional Neural Network

Convolutional neural networks (CNN or ConvNet) are suitable to process information in the form of multiple arrays (LeCun et al., 2015; Esteva et al., 2017; Hu and Lu, 2018). To reduce the parameters without compromising its learning capacity is the general design principle of CNN (LeCun et al., 2015; Krizhevsky et al., 2017). And each convolution kernel's parameters in CNN are trained by the backpropagation algorithm.

Especially in image-related applications, CNN can cope with pixel scanning and processing, thus it greatly accelerates the implementation of optimized algorithms into practice (Esteva et al., 2017; Quang et al., 2018). Structurally, CNN consists of linear convolution operation, followed by nonlinear activators, pooling layers, and deep neural network classifier, depicted in **Figure 5**.

In **Figure 5**, several filters are applied to convolve an input image, and its output is subsampled as a new input into the next layer; and convolution and subsampling processes are repeated till high level features, namely shapes, can be extracted. The more layers a CNN model has, the higher-level features it will extract.

In feature learning, convolution operation is to scan a 2D image with a given pattern, and calculate the matching degree at each step, then pooling identifies the pattern presence in the scanned region (Angermueller et al., 2016). Activation function defines a neuron's output based on a set of given inputs. The weighted sum of inputs is passed through an activation function for non-linear transformation. A typical activation function returns a binary output, 0 or 1; when a neuron's accumulation exceeds a preset threshold, the neuron is activated and passes its information to the next layers; otherwise, the neuron is deactivated. Sigmoid, tanh, ReLU, leaky ReLU, and softmax are the commonly used activation functions (LeCun et al., 2015; Schmidhuber, 2015).

Through pooling layers, pixels are stretched to a single column vector. The vectorized and concatenated pixel information is fed into dense layers, known as fully connected layers for further classification. The fully-connected layer renders the final decision, where CNN returns a probability that an object in the image belongs to a specific type.

Following the fully-connected layer is a loss layer, which adjusts their weights across the network. A loss function is used to measure the model performance and inconsistency between the actual and predicted values. Model performance increases with decreasing of the loss function. For an output vector y_i and an input x=(x_1, x_2, …, x_n), the mapping loss function $L(\cdot)$ between x and y is defined as,

$$L(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=1,j=1}^{n,k} \varphi[y_i, f(x_i, \sigma_i, \omega_{ij}, b_i)] \quad (5)$$

where φ denotes an empirical risk for each output, $\hat{y}_i$ for the i-th prediction, n the total number of training samples, k the count of the weights ω_{ij} and b_i the bias for the activation function σ_i.

Recently, CNN has been adopted rapidly in biomedical imaging studies for its outstanding performance in computer vision and concurrent computation with GPUs (Ravi et al., 2017). Usually convolution-pooling structure can better learn imaging features from CT scans and MRI images from head trauma, stroke diagnosis and brain EPV (enlarged perivascular space) detection (Chilamkurthy et al., 2018; Dubost et al., 2019).

In recent computational biology, a discriminative CNN framework, DeepChrome, is proposed to predict gene expression by feature extraction from histone modification. And the deep learning model outperforms traditional Random Forests and SVM on 56 cell types from REMC database (Singh et al., 2016).

Furthermore, CNN can be combined with other deep learning models, such as RNN to predict imaging content, where CNN encodes an image and RNN generates the corresponding image description (Angermueller et al., 2016). Till now, quite a few variants of CNN have been also proposed in diverse classification applications, like AlexNet with GPU support and DQN in reinforcement learning (Mnih et al., 2015).

Autoencoder

Through an unsupervised manner, autoencoder is another typical artificial neural network, designed to precisely extract coding or

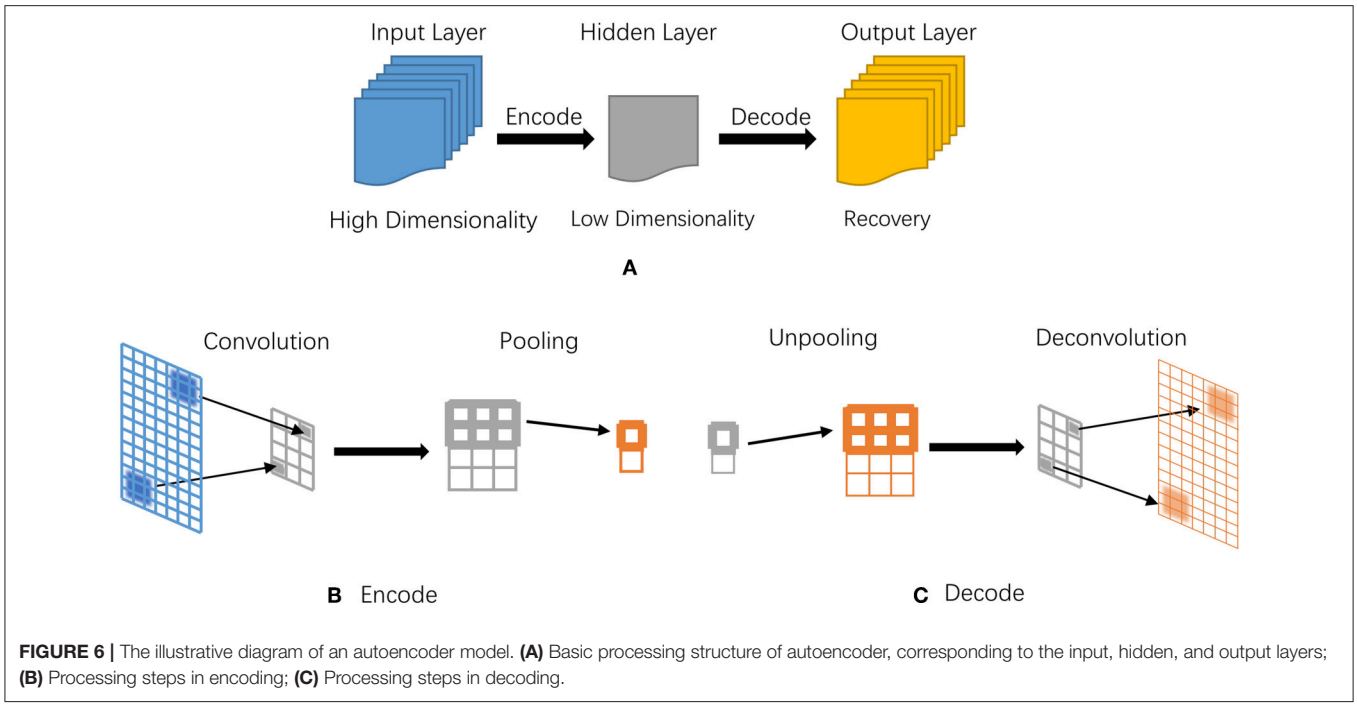

FIGURE 6 | The illustrative diagram of an autoencoder model. **(A)** Basic processing structure of autoencoder, corresponding to the input, hidden, and output layers; **(B)** Processing steps in encoding; **(C)** Processing steps in decoding.

representation features using data-driven learning (Min et al., 2017; Zeng et al., 2017; Yang et al., 2018). For high-dimensional data, it is time-consuming and infeasible to load all raw data into a network, thus dimension reduction or compression is a necessity in preprocessing of raw data.

Autoencoder can compress and encode information from the input layer into a short code, then after specific processing, it will decode into the output closely matching the original input. **Figure 6** illustrates its basic model structure and processing steps.

Convolution and pooling are two major steps in encoder, depcited in **Figure 6B**; while decoder has two complete opposite steps, namely unpooling and deconvolution in **Figure 6C**. Both convolution and pooling can compress data while preserving the most representative features in two different ways. Convolution involves continuously scanning data with a rectangle window, for example a 3×3 size; after each scanning, the window moves to a next position, namely pixel, by replacing the oldest elements with new ones, together with convolution operation. After the whole scanning and convolution, pooling is utilized to deeper compress on redundancy.

Similar to traditional PCA in dimension reduction to some extent, but autoencoder is more robust and effective in extracting data features for its non-linear transformation in hidden layers. Given an input x, the model extracts its main feature and generates $\hat{x} = Wb$, where W and b denote weighting and bias vectors, respectively. Commonly, the output cannot fit the input precisely, which can be measured with a loss function in mean squared error (MSE) defined in Equation (6),

$$L(W,b) = \frac{1}{m}\sum_{i=1}^{m}(\hat{x} - x)^2 \tag{6}$$

Thus, the learning process is to minimize the loss L after iterative optimization.

Recently, sparse autoencoder (SAE) is frequented discussed for its admirable performance in dimension reduction and denoising corrupted data. And the loss function in SAE is defined in Equation (7),

$$L_{SAE} = L(W,b) + \beta\sum_{k} KL(\rho||\widehat{\rho_k}) \tag{7}$$

where KL refers to KL-divergence in Equation (10), ρ for the activation level of neurons, usually set as 0.05 in condition of sigmoid, indicating most neurons are inactive, ρ_k for the average activation level of neuron k, and β for the regularization coefficient.

$$KL(\rho||\widehat{\rho_k}) = \rho log\frac{\rho}{\widehat{\rho_k}} + (1-\rho)log\frac{1-\rho}{1-\widehat{\rho_k}} \tag{8}$$

where $\widehat{\rho_k}$ represents the average activation level of test samples, and $x^{(i)}$ is the i-th test sample in Equation (9).

$$\widehat{\rho_k} = \frac{1}{m}\sum_{i}\left[a_j(x^{(i)})\right] \tag{9}$$

For high dimensional data, multiple autoencoders can be stacked to act as a deep autoencoder (Hinton and Salakhutdinov, 2006). And this architecture may lead to vanishing gradient, due to its gradient-based and backpropagation learning, and the current solutions include adopting ReLu activation and dropout (Szegedy et al., 2015; Krizhevsky et al., 2017). During configuration and pretraining, the model weights can be acquired by greedy layer-wise training, then the network can be fine-tuned with the backpropagation algorithm.

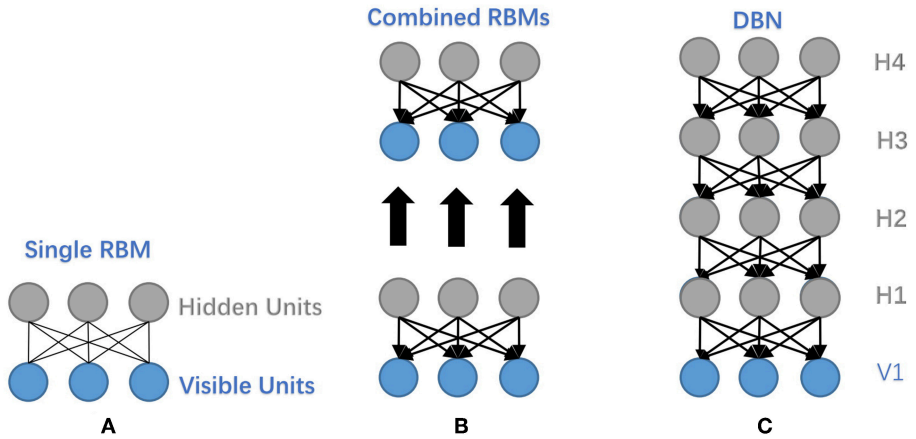

FIGURE 7 | Illustrative network structures of RBM and DBN. **(A)** The structure of RBM. **(B)**Take the hidden layer of the trained RBM to function as the visible layer of another RBM. **(C)** The structure of a DBN. It stacks several RBMs on top of each other to form a DBN.

Many variations of autoencoder have been proposed recently, such as sparse autoencoder (SAE), denoising autoencoder (DAE). Typically, stacked sparse autoencoder (SSAE) was proposed to analyze high-resolution histopathological images in breast cancer (Xu J. et al., 2016). By using SAE with three iterations, Heffernan et al. reported the successful prediction of protein secondary structure, local backbone angles, and solvent accessible surface area (Heffernan et al., 2015). Miotto et al. introduced a stack of DAEs to predict features from a large scale of electronic health records (EHR), via an unsupervised representation approach (Miotto et al., 2016). Ithapu et al. proposed a randomized denoising autoencoder marker (rDAm) to predict future cognitive and neural decline for Alzheimer diseases, with its performance surpassing the existing methods (Ithapu et al., 2015).

Deep Belief Network

As a generative graphical model, Deep Belief Network (DBN) is composed of multiple Restricted Boltzmann Machines (RBM) or autoencoders stacked on top of each other, where each hidden layer in subnetworks serves as a visible layer for the next layer (Hinton et al., 2006). The main network structures of RBM and DBN are depicted in **Figure 7**, where it manifests the construction relations between the two network models.

DBN trains layer by layer in an unsupervised greedy approach to initialize network weights, separately; then it can utilize the wake-sleep or backpropagation algorithm during fine-tuning. While for traditional backpropagation used in fine-tuning, DBN may encounter several problems: (1) requiring labeled data for training; (2) low learning rate; (3) inappropriate parameters tending to acquire local optimum.

Within recent applications, Plis et al. classified schizophrenia patients based on brain MRIs with DBN (Plis et al., 2014); in drug design based on high-throughput screening, DBN was exploited to perform quantitative structure activity relationship (QSAR) study. And the result showed that the optimization in parameter initialization highly improves the capability of DNN to provide high-quality model predictions (Ghasemi et al., 2018). DBN was also used to study the combination of resting-state fMRI (rs-fMRI), gray matter, and white matter data by exploiting the latent and abstract high-level features (Akhavan Aghdam et al., 2018). Meanwhile, DBN and CNN were compared to prove that deep learning has better discriminative results and holds promise in the medical image diagnosis (Hua et al., 2015).

Transfer Learning in Deep Learning

Besides the above deep learning models, transfer learning is frequently utilized in specific cases without sufficient labeling information or dimensionality (Pan and Yang, 2010). Although conceptually it does not belong to deep learning, due to its transferability of high-level semantic classification for deep neural network, transfer learning has gained emerging notices from deep learning fields (O'Shea et al., 2013; Anthimopoulos et al., 2016).

In quite a few deep learning studies, transfer learning enables a previously-trained model to transfer its optimized parameters to a new model, thus to implement the knowledge transmission and reduce repetitive training from scratch, as depicted in **Figure 8**.

Normally, source and target domains have certain statistical relationship or similarity that directly affects the transferability. The domain contains the original dataset, for example image matrix, and the task refers to certain processes, like classification or pattern recognition. The mission of transfer learning includes transferring not only the parameters like weight, but the concentrated small-size matrix from the origin data domain called knowledge distillation.

The knowledge distillation usually uses both "hard target" and "soft target" to train the model and obtain lower information entropy. The below softmax function is usually utilized to soften

the sparse data and excavate its inherent features,

$$f(\alpha_k) = \frac{e^{\frac{\alpha_k}{T}}}{\sum_k e^{\frac{\alpha_k}{T}}} \quad (10)$$

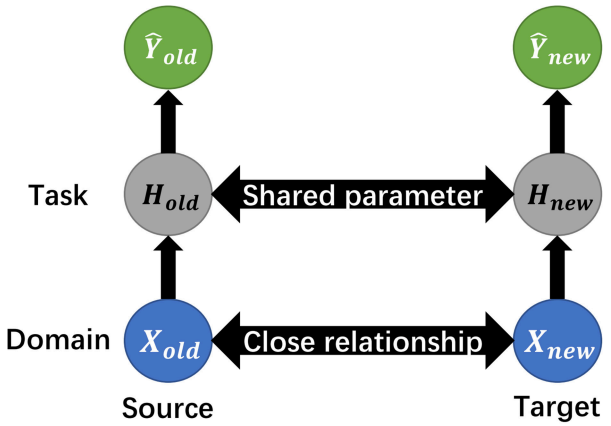

FIGURE 8 | The schematic illustration of transfer learning. Given source domain and its learning task, together with target domain and respective task, transfer learning aims to improve the learning of the target prediction function, with the knowledge in source domain and its task.

where the logical judger α_k is the input, $f(\cdot)$ is to soft target data and can offer smaller gradient variance, k denotes the k-th segmented data slice. The parameter T is called temperature and the larger T is, the softer the target is.

Furthermore, transfer learning is categorized into instance-based, feature-based, parameter-based and relation-based derivatives, depicted in **Figure 9**. Currently transfer learning is frequently discussed in the deep learning fields for its great applicability and performance. Ensembled with CNN, transfer learning can attain greater prediction performance of interstitial lung disease CT scans (Anthimopoulos et al., 2016). It was also used as a ligament between the multi-layer LSTM and conditional random field (CRF), and the result showed that the LSTM-CRF approach outperformed the baseline methods on the target datasets (Giorgi and Bader, 2018).

CONCLUSIONS

Within the work, we comprehensively summarized the basic but essential concepts and methods in deep learning, together with its recent applications in diverse biomedical studies. Through reviewing those typical deep learning models as RNN, CNN, autoencoder, and DBN, we highlight that the specific application scenario or context, such as data feature and model applicability,

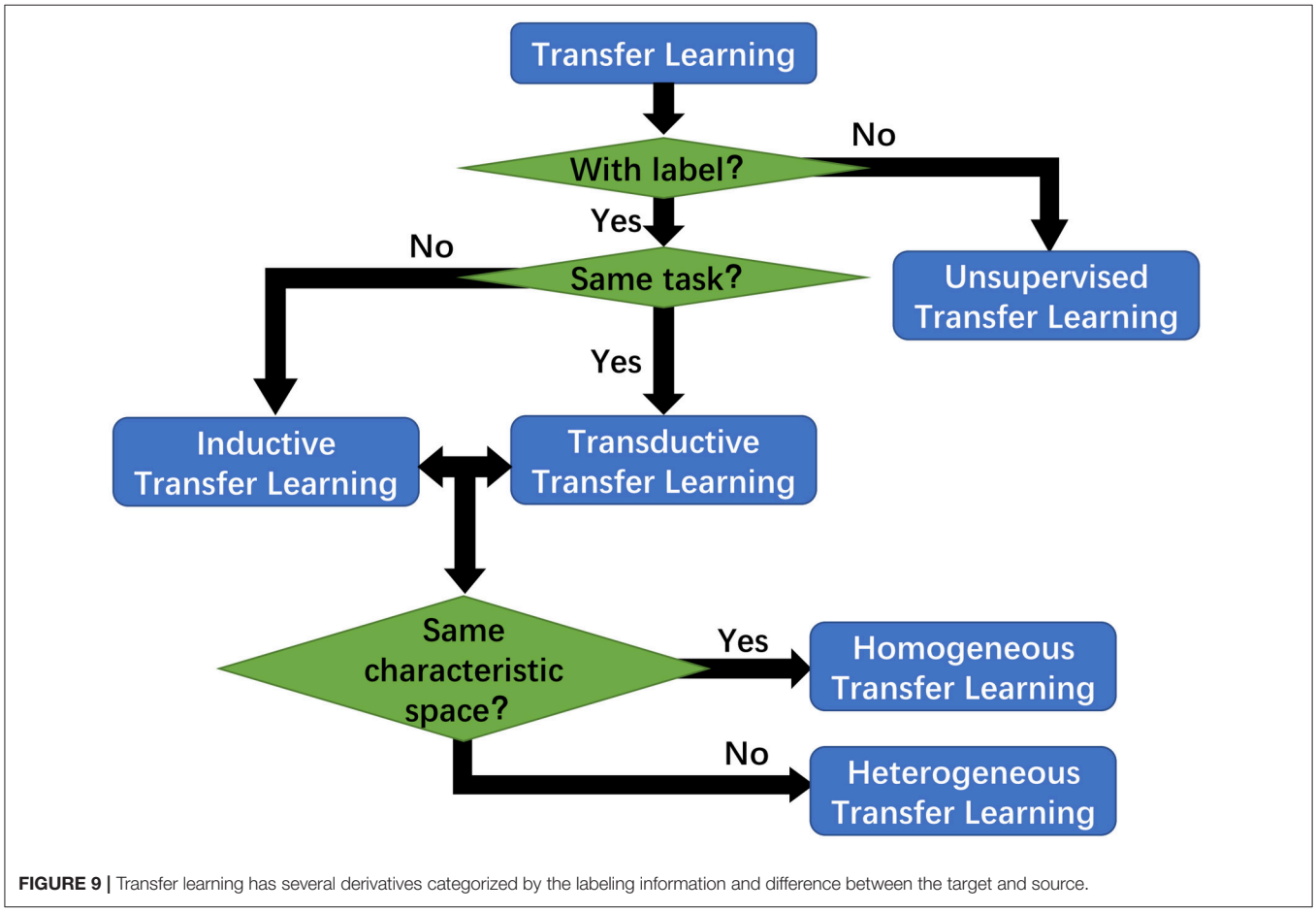

FIGURE 9 | Transfer learning has several derivatives categorized by the labeling information and difference between the target and source.

are the prominent factors in designing a suitable deep learning approach to extract knowledge from data; thus, how to decipher and characterize data feature is not a trivial work in deep-learning workflow yet. In recent deep learning studies, many derivatives from classic network models, including the network models depicted above, manifest that model selection affects the effectiveness of deep learning application.

Secondly, for its limitation and further improvement direction, we should revisit the nature of the method: deep learning is essentially a continuous manifold transformation among diverse vector spaces, but there exist quite a few tasks cannot be converted into a deep learning model, or in a learnable approach, due to the complex geometric transform. Moreover, deep learning is generally a big-data-driven technique, which has made it unique from conventional statistical learning or Bayesian approaches. Thus, it is a new direction for deep learning to integrate or embed with other conventional algorithms in tackling those complicated tasks.

Thirdly, when it comes to innovation in computational algorithm and hardware. As an inference technique driven by big data, deep learning demands parallel computation facilities of high performance, together with more algorithmic breakthroughs and fast accumulation of diverse perceptual data, it is achieving pervasive successes in many fields and applications. Particularly in bioinformatics and computational biology, which is a typical data-oriented field, it has witnessed the remarkable changes taken place in its research methods.

Finally, as unprecedented innovation and successes acquired with deep learning in diverse subfields, some even argued that deep learning could bring about another wave like the internet. In the long term, deep learning technique is shaping the future of our lives and societies to its full extent. But deep learning should not be misinterpreted or overestimated either in academia or AI industry, and actually it has lots of technical problems to solve due to its nature. In all, we anticipate this review work will provide a meaningful perspective to help our researchers gain comprehensive knowledge and make more progresses in this ever-faster developing field.

AUTHOR CONTRIBUTIONS

BT conceived the study. ZP, KY, AK, and BT drafted the application sections and revised and approved the final manuscript.

FUNDING

This work was supported by the Natural Science Foundation of Jiangsu, China (BE2016655 and BK20161196), and the Fundamental Research Funds for China Central Universities (2019B22414). This work made use of the resources supported by the NSFC-Guangdong Mutual Funds for Super Computing Program (2nd Phase), and the Open Cloud Consortium sponsored project resource, supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation (USA) and major contributions from OCC members.

REFERENCES

Akhavan Aghdam, M., Sharifi, A., and Pedram, M. M. (2018). Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network. *J. Digit. Imaging*. 31, 895–903. doi: 10.1007/s10278-018-0093-8

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33:831–838. doi: 10.1038/nbt.3300

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18:67. doi: 10.1186/s13059-017-1189-z

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12:878. doi: 10.15252/msb.20156651

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., and Mougiakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans. Med. Imag.* 35, 1207–1216. doi: 10.1109/TMI.2016.2535865

Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15:937. doi: 10.1093/bioinformatics/15.11.937

Bengio, Y., and LeCun, Y. (2007). "Scaling learning algorithms toward AI," in *Large-Scale Kernel Machines*, eds L. Bottou, O. Chapelle, D. DeCoste and J. Weston (Cambridge, MA: The MIT Press).

Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., et al. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 392, 2388–2396. doi: 10.1016/S0140-6736(18)31645-3

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15:20170387. doi: 10.1098/rsif.2017.0387

Ditzler, G., Polikar, R., Member, S., Rosen, G., and Member, S. (2015). Multi-layer and recursive neural networks for metagenomic classification. *IEEE. Trans. Nanobiosci.* 14:608. doi: 10.1109/TNB.2015.2461219

Dubost, F., Adams, H., Bortsova, G., Ikram, M. A., Niessen, W., Vernooij, M., et al. (2019). 3D regression neural network for the quantification of enlarged perivascular spaces in brain MRI. *Med. Image Anal.* 51, 89–100. doi: 10.1016/j.media.2018.10.008

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–118. doi: 10.1038/nature21056

Ghasemi, F., Mehridehnavi, A., Fassihi, A., and Pérez-Sánchez, H. (2018). Deep neural network in QSAR studies using deep belief network. *Appl. Soft Comput.* 62, 251–258. doi: 10.1016/j.asoc.2017.09.040

Giorgi, J. M., and Bader, G. D. (2018). Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics* 34, 4087–4094. doi: 10.1093/bioinformatics/bty449

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., et al. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* 5:11476. doi: 10.1038/srep11476

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural. Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hu, Y., and Lu, X. (2018). Learning spatial-temporal features for video copy detection by the combination of CNN and RNN. *J. Vis. Commun. Image Rep.* 55, 21–29. doi: 10.1016/j.jvcir.2018.05.013

Hua, K. L., Hsu, C. H., Hidayati, H. C., Cheng, W. H., and Chen, Y. J. (2015). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Oncotargets Ther.* 8:2015–2022. doi: 10.2147/OTT.S80733

Ithapu, V. K., Singh, V., Okonkwo, O. C., Chappell, R. J., Dowling, N. M., and Johnson, S. C. (2015). Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimer's Dement.* 11:1489–1499. doi: 10.1016/j.jalz.2015.01.010

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta Kazuhiro, R, Rastas, P., et al. (2013). DNA-binding specificities of human transcription factors. *Cell* 152, 327–339. doi: 10.1016/j.cell.2012.12.009

Kim, Y., Sim, S. H., Park, B., Lee, K. S., Chae, I. H., Park, I. H., et al. (2018). MRI assessment of residual breast cancer after neoadjuvant chemotherapy: relevance to tumor subtypes and MRI interpretation threshold. *Clin. Breast Cancer* 18, 459–467.e1 doi: 10.1016/j.clbc.2018.05.009

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436. doi: 10.1038/nature14539

Lee, T. I., and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251. doi: 10.1016/j.cell.2013.02.014

Li, A., Serban, R., and Negrut, D. (2017). Analysis of a splitting approach for the parallel solution of linear systems on GPU cards. *SIAM J. Sci. Comput.* 39, C215–C237. doi: 10.1137/15M1039523

Liang, M., Li, Z., Chen, T., and Zeng, J. (2015). Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 12, 928–937. doi: 10.1109/TCBB.2014.2377729

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16:321–322. doi: 10.1038/nrg3920

Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Mol. Pharmaceut.* 13, 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Brief Bioinform.* 18, 851–869. doi: 10.1093/bib/bbw068

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6:26094. doi: 10.1038/srep26094

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Nussinov, R. (2015). Advancements and challenges in computational biology. *PLoS Comput. Biol.* 11:e1004053. doi: 10.1371/journal.pcbi.1004053

O'Shea, J. P., Chou, M. F., Quader, S. A., Ryan, J. K., Church, G. M., and Schwartz, D. (2013). pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods* 10, 1211–1212. doi: 10.1038/nmeth.2646

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Quang, D., Guan, Y., and Parker, S. C. J. (2018). YAMDA thousandfold speedup of EM-based motif discovery using deep learning libraries and GPU. *Bioinformatics* 34, 3578–3580. doi: 10.1093/bioinformatics/bty396

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., et al. (2017). Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* 21, 4–21. doi: 10.1109/JBHI.2016.2636665

Ray, D., Kazan, H., Chan, E. T., Peña, L. C., Chaudhry, S., Talukder, S., et al. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* 27, 667–670. doi: 10.1038/nbt.1550

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. (2015). *Neural. Net.* 61:85. doi: 10.1016/j.neunet.2014.09.003

Sekhon, A., Singh, R., and Qi, Y. (2018). DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics* 34, i891–i900. doi: 10.1093/bioinformatics/bty612

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32, i639–i648. doi: 10.1093/bioinformatics/btw427

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., and Madabhushi, A. (2016). Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE Trans. Med. Imaging* 35, 119–130. doi: 10.1109/TMI.2015.2458702

Xu, T., Zhang, H., Huang, X., Zhang, S., and Metaxas, D. N. (2016). "Multimodal deep learning for cervical dysplasia diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Boston, MA), 115–123.

Yang, W., Liu, Q., Wang, S., Cui, Z., Chen, X., Chen, L., and Zhang, N. (2018). Down image recognition based on deep convolutional neural network. *Inform. Process. Agric.* 5, 246–252. doi: 10.1016/j.inpa.2018.01.004

Zeng, K., Yu, J., Wang, R., Li, C., and Tao, D. (2017). Coupled deep autoencoder for single image super-resolution. *IEEE Trans. Cybernet.* 47, 27–37. doi: 10.1109/TCYB.2015.2501373

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., and Zeng, J. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* 44:e32. doi: 10.1093/nar/gkv1025

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 March 2019
doi: 10.3389/fgene.2019.00233

Autoencoder Based Feature Selection Method for Classification of Anticancer Drug Response

Xiaolu Xu[1], *Hong Gu*[1], *Yang Wang*[2], *Jia Wang*[3]* *and Pan Qin*[1]*

[1] Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China, [2] Institute of Cancer Stem Cell, Dalian Medical University, Dalian, China, [3] Department of Breast Surgery, Institute of Breast Disease, Second Hospital of Dalian Medical University, Dalian, China

Anticancer drug responses can be varied for individual patients. This difference is mainly caused by genetic reasons, like mutations and RNA expression. Thus, these genetic features are often used to construct classification models to predict the drug response. This research focuses on the feature selection issue for the classification models. Because of the vast dimensions of the feature space for predicting drug response, the autoencoder network was first built, and a subset of inputs with the important contribution was selected. Then by using the Boruta algorithm, a further small set of features was determined for the random forest, which was used to predict drug response. Two datasets, GDSC and CCLE, were used to illustrate the efficiency of the proposed method.

Keywords: anticancer drug response, autoencoder, classification model, feature selection, random forest

OPEN ACCESS

Edited by:
Binhua Tang,
Hohai University, China

Reviewed by:
Sandeep Kumar Dhanda,
La Jolla Institute for Immunology (LJI), United States
Firoz Ahmed,
Jeddah University, Saudi Arabia

***Correspondence:**
Jia Wang
wangjia77@hotmail.com
Pan Qin
qp112cn@dlut.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *28 October 2018*
Accepted: *04 March 2019*
Published: *27 March 2019*

Citation:
Xu X, Gu H, Wang Y, Wang J and Qin P (2019) Autoencoder Based Feature Selection Method for Classification of Anticancer Drug Response. Front. Genet. 10:233. doi: 10.3389/fgene.2019.00233

1. INTRODUCTION

The prediction of drug responses for individual patients is an essential issue in the research of precision medicine. It is known that the drug response for various patients can be different (Wilkinson, 2005). Thus, there are different therapeutic effects when using the same anticancer drug for a cohort of patients (Dong et al., 2015). It has been suggested that the patients with similar response to an anticancer drug can have similar genetic features, like gene mutations and expressions (Wang et al., 2017). These features can be used as the biomarkers to predict the drug response (La Thangue and Kerr, 2011).

Because the clinical trials are of high time and economic costs, the researchers prefer to use the cell lines obtained from the cancer patients for investigating drug responses. These investigations lead to several drug response databases, like Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2012) and Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). By using these databases, constructing models for the prediction of drug response becomes feasible. Primarily, researchers always use IC50 (Barretina et al., 2012; Garnett et al., 2012), which indicates the concentration required for 50% inhibition *in vitro*, to measure the sensitivity of drug response. Taking IC50 as the dependent variable, linear regression models, including ridge regression, lasso, and elastic net, were developed to predict drug response (Barretina et al., 2012; Garnett et al., 2012; Basu et al., 2013; Iorio et al., 2016). Further complex models, like support vector regression, artificial neural network, and random forest (RF), were also constructed for this purpose (Riddick et al., 2010; Menden et al., 2013; Ammad-Ud-Din et al., 2014; Ammad-ud din et al., 2016; Costello et al., 2014; Ospina et al., 2014; Cichonska et al., 2015; Dong et al., 2015; Zhang et al., 2015). Neto et al. (2014) proposed the STREAM algorithm that combined a Bayesian inference strategy with ridge regression for the prediction of drug response. Besides the regressions, several network-based

models were also proposed (Wang et al., 2014; Fey et al., 2015; Zhang et al., 2015). Model ensembles have also been considered by some works (Wan and Pal, 2014; Cortés-Ciriano et al., 2015). Meanwhile, deciding whether an individual patient is sensitive or not to the anticancer drugs is meaningful for treatment. By setting a proper threshold value for IC50, drug response can be divided into two categories: sensitivity and non-sensitivity. In this case, classification models can be fitted for predicting drug response. To this end, the recommender system, naive Bayes classifier and support vector machine have been used (Barretina et al., 2012; Dong et al., 2015; Suphavilai et al., 2018).

Nilsson et al. (2007) indicated that the appropriate selection of small feature set gives the best possible classification results. Thus, selecting an appropriate feature set from a large number of genetic feature candidates is a crucial issue for classification models for predicting drug response. In this paper, we developed a drug response prediction model, called AutoBorutaRF, by using autoencoder (Liou et al., 2008) and Boruta algorithm (Kursa et al., 2010) for feature selection and RF for classification. We first constructed the autoencoder network (Liou et al., 2008), which is a type of artificial neural network, for the reduction of genetic features. By using the Gedeon method (Gedeon, 1997), we initially reduced the total number of features. We further selected a smaller feature set feasible for RF by using the Boruta algorithm. By applying AutoBorutaRF to GDSC and CCLE, we proved that our proposed method is of excellent prediction accuracy. We further analyzed the biomarkers obtained from the lung cell lines in GDSC by the proposed feature selection method.

2. MATERIALS AND METHODS

2.1. Datasets and Preprocessing

In this research, we used two datasets, including GDSC (Garnett et al., 2012) and CCLE (Barretina et al., 2012). The datasets were downloaded by using R package PharmacoGx (Smirnov et al., 2015). We used the sensitivity measure IC50 (Barretina et al., 2012; Garnett et al., 2012) as the response variable (denoted by $y_{rs,c}$) for cell line c. We used three types of genetic features as the explanatory variables, including the gene expression (denoted by $\boldsymbol{x}_{rna,g}$), the single-nucleotide mutation (denoted by $\boldsymbol{x}_{snv,g}$), and the copy number alternation (denoted by $\boldsymbol{x}_{cna,g}$) for gene g. Note that the elements in $\boldsymbol{x}_{rna,g}$ and $\boldsymbol{x}_{cna,g}$ are real-valued; the elements in $\boldsymbol{x}_{snv,g}$ are binary-valued, i.e., "1" for mutation and "0" for wild type. In the two datasets, some cell lines missed the values of the response variable, the single-nucleotide mutation features, and the copy number alteration features. There was no missing value in the gene expression features. We first removed the features with the cell lines missing values more than 50%. Then, we removed the cell lines with more than 50% features missing values from the datasets. For the remaining cell lines with missing values, we used a weight mean method to compensate the missing values as follows:

1. Let $z^*_{c,g}$ denote the missing value for the cell line c in the response variable or the genetic feature g. Let $\boldsymbol{x}_{rna,c}$ denote the vector of gene expression features for the cell line c.
2. Assume the cell line k has no missing data for the features involved in $z^*_{c,g}$. The diversity between the cell lines c and k is obtained by $d(c,k) = \left\|\boldsymbol{x}_{rna,c} - \boldsymbol{x}_{rna,k}\right\|_2^2$. Search K cell lines nearest to g with respect to $d(c,i)$.
3. If g is the response variable or the copy number alternation feature, $z^*_{c,g}$ is compensated by

$$\widehat{z}^*_{c,g} = \sum_{k=1}^{K} \frac{d(c,k)}{\sum_{k=1}^{K} d(c,k)} z_{k,g}$$

4. If g is the single-nucleotide mutation feature, $z_{c,g}$ is compensated by

$$\widehat{z}^*_{c,g} = \begin{cases} 1 & \sum_{k=1}^{K} \mathbf{1}(z_{k,g} = 1) > \sum_{k=1}^{K} \mathbf{1}(z_{k,g} = 0) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{1}() = 1$ for the true statement in the parenthesis and $\mathbf{1}() = 0$ for the negative statement in the parenthesis.

We set $K = 10$ for the preprocessing of GDSC and CCLE datasets.

2.2. Label Assignment for Cell Lines According to IC50

This research is to construct classification models for predicting how the cell lines respond to the drugs under study. The drug responses can be divided into two categories: "sensitivity" and "non-sensitivity" (Liu et al., 2016). So far, several works have used various threshold values of IC50 to classify the drug responses (Brubaker et al., 2014; Li et al., 2015). Brubaker et al. (2014) used a hard threshold 0.1 to label sensitivity for IC50$<$ 0.1 and to label non-sensitivity (i.e., resistance in this work) for IC50$\geq$ 0.1. However, by investigating the histograms of IC50, we found that the statistics of drugs are various. It can be supposed that the decision of labels should be driven by the data of individual drugs. To this end, we adopted the strategy introduced in Li et al. (2015), which used the median of the observed IC50 values as a data-driven threshold. We labeled a cell line as "sensitivity" if its IC50 is smaller than the median overall the cell lines for an individual drug. We labeled a cell line "non-sensitivity" if its IC50 is equal to or larger than the median overall the cell lines for an individual drug.

2.3. Classification Model and Feature Selection for Predicting Drug Response

2.3.1. Classification Model

The drug response data are often of imbalanced classifications. Because RF is outstanding for the imbalanced classification problem, we used it as the classification model. In RF, we used classification and regression trees (CART) algorithm as

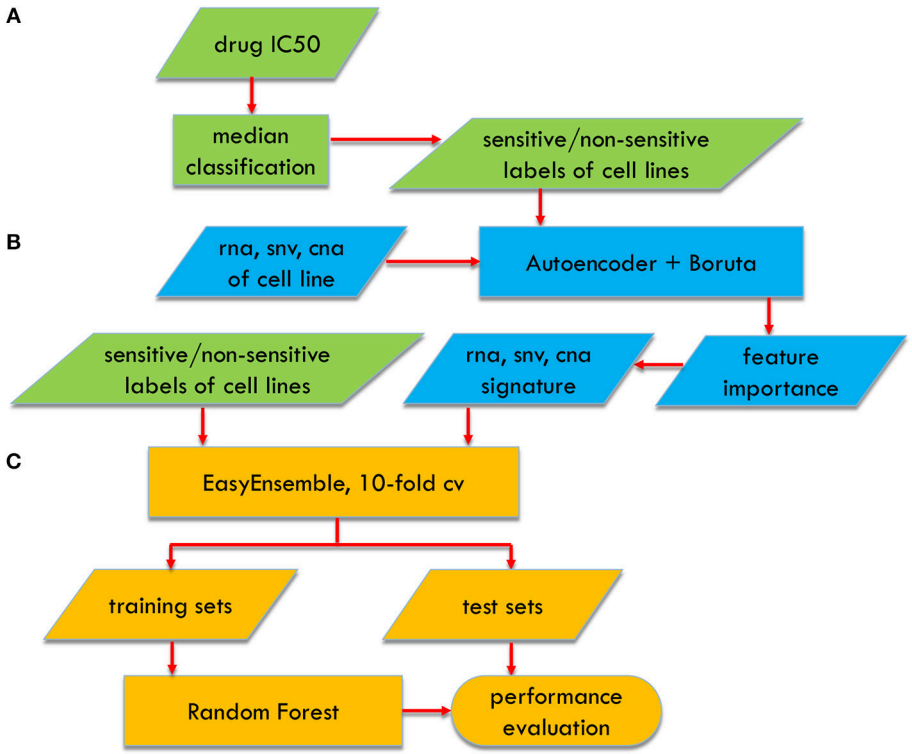

FIGURE 1 | Flowchart of AutoBorutaRF for predicting anticancer drug response, which includes three parts: **(A)** data preprocessing, **(B)** feature selection, and **(C)** classifier constructing.

the basic classifier. RF randomly generalizes 1,000 CARTs. Each CART is trained by using $\lceil 0.632 \times N_{sample} \rceil$ bootstrapping samples, where N_{sample} is a total of cell lines. The ultimate results were determined through voting with the prediction results of all CARTs.

2.3.2. Feature Selection With the Autoencoder and Boruta Algorithm

Feature selection is crucial for improving the prediction performance of the classification models. We used the Boruta algorithm, which aims to the feature selection problem for RF (Kursa et al., 2010) (**Figure 1**). The considerable cardinality of the feature candidate set leads to the curse of dimensionality for the Boruta algorithm. Thus, we first used the autoencoder network, to roughly screen out the features to a proper dimension. The detailed two-stepwise feature selection procedure is described as follows:

Step 1: We trained two single-hidden-layer autoencoder networks, with hyperbolic tangent being the activation functions, for screening out the features of the gene expression and the features of the copy number alteration, respectively. Different from the straight application of the hidden layers of the autoencoder, we used Gedeon method (Gedeon, 1997) to calculate the proportional contributions to select the significant genes. The contribution of the ith input (gene) to the jth output (gene) is calculated as

$$Q_{ij} = \sum_{k=1}^{K} (P_{ik} \times P_{kj})$$

Here K denotes the total number of the neurons of the hidden layer. P_{ik} is the contribution of the ith input to the kth neuron of the hidden layer calculated by

$$P_{ik} = \frac{|W_{ik}|}{\sum_{i^*=1}^{G} |W_{i^*k}|}$$

with G being the total number of the inputs and W_{i^*k}s being the weights linking the corresponding neuron couples. P_{kj} is the contribution of the kth neuron of the hidden layer to the jth output, whose calculation is similar to that of P_{ik}. The total contribution of the ith input is calculated by

$$q_i = \sum_{j=1}^{G} \frac{Q_{ij}}{\sum_{i^*=1}^{G} Q_{i^*j}}$$

We ranked the inputs of the autoencoder in the descending order with respect to q_i and removed the last

50% features. We also removed the features, whose means of correlation coefficients with other features were more than 0.95.

Step 2: From the features obtained by Step 2, the Boruta algorithm was used to select features for RF as follows:

2-1. Extend the dataset by adding copies of all the features obtained by Step 1.
2-2. Shuffle the values of the copied features, called shadow features, to remove their correlations with the response variable, i.e., IC50.
2-3. The shadow features are combined with the original ones.
2-4. Run a random forest classifier on the combined dataset and perform a variable importance measure, in which the mean decrease accuracy (MDA) is used.
2-5. Z score is calculated by dividing MDA with the standard deviation of accuracy loss.
2-6. Find the maximum Z score among shadow attributes (MZSA).
2-7. The features with importance significantly lower than MZSA are permanently removed from the dataset. The features with importance significantly higher than MZSA are retained as important features.
2-8. The shadow features are removed from the dataset.
2-9. Repeat the above steps until for the prefixed iterations (200 was prefixed in our study), or all the retained features are important features.

2.4. EasyEnsemble for Imbalanced Datasets

The total number of cell lines sensitive to drugs is much smaller than that of cell lines non-sensitive to drugs. Thus, the datasets in this research are the class imbalance. Let $\mathcal{N}$ and $\mathcal{R}$ denote the sample set of majority class (non-sensitivity) and that of minority class (sensitivity), respectively. The imbalance ratio $\mathrm{IR} = |\mathcal{N}|/|\mathcal{R}|$ is used to measure the class imbalance, with $|\cdot|$ being the cardinality of a set. For the various drugs under study, the values of IR are different. In this research, for the drugs with IR$\leq$ 2, the feature selection and classification method were directly used; for the drugs with IR$>$ 2, we used EasyEnsemble (Liu et al., 2009) resampling strategy to deal with the imbalance class problem. The core procedure of EasyEnsemble used here is described as follows:

1. Equally divide $\mathcal{N}$ into T subsets $\{\mathcal{N}_i | i = 1, 2, \cdots, T\}$, with $T = \lfloor \mathrm{IR} \rfloor$. Such that $|\mathcal{N}_i| \approx |\mathcal{R}|$.
2. The RF classifier $F_i(x)$ is constructed on each training subsets $\{\mathcal{N}_i, \mathcal{R}\}$ for $i = 1, 2, \cdots, T$.
3. Take the majority vote according to the T predictions of $\{F_i(x) | i = 1, 2, \cdots, T\}$.

2.5. Evaluation Criteria

We used the following metrics to evaluate the performance of the classification models:

Accuracy: $$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

Recall: $$REC = \frac{TP}{TP + FN}$$

Specificity: $$SPC = \frac{TN}{TN + FP}$$

F_1 score: $$F_1 = \frac{2TP}{2TP + FP + FN}$$

Matthews correlation coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(FP + TN)(FN + TN)}}$$

where

1. TP (true positive) is the number of cell lines labeled with sensitivity and predicted as sensitivity;
2. FP (false positive) is the number of cell lines labeled with resistance and predicted as sensitivity;
3. FN (false negative) is the number of cell lines labeled with sensitivity and predicted as non-sensivity;
4. TN (true negative) is the number of cell lines labeled with resistance and predicted as non-sensivity.

Besides the metrics above, AUC was also obtained.

Because the total number of samples was much smaller than that of the features, the above evaluation criteria were obtained by using 10-fold cross validation (CV). The dataset was randomly partitioned into 10 equal sized subsets. Of the ten subsets, a single subset was used as the test set to calculate the evaluation criteria of the models trained by the remaining nine subsets. The above process was then repeated 10 times, and the mean of the evaluation criteria obtained in the 10 times was used as the final criteria. In this way, the test datasets can be ensured to be independent of the training datasets.

3. RESULTS

3.1. Data Description

There are missing data in both datasets. These missing data were compensated by using the weighted mean method described in the section Materials and Methods. The total numbers of samples for each variable are listed in **Table 1**.

According to their histograms, the most of distributions of drug responses of cell lines in two datasets can be approximated by the Gauss distribution (**Figure 2**). t-hypothesis test showed that the significance of two groups divided by median of IC50 in GDSC is of p-values from 4.27×10^{-160} to 6.89×10^{-46}; such significance in CCLE is of p-value from 7.14×10^{-95} to 4.05×10^{-4}.

3.2. Prediction Performance of AutoBorutaRF

To illustrate the effectiveness of our AutoBorutaRF method, we demonstrated its prediction performance on GDSC and CCLE datasets. Meanwhile, we compared it with other four algorithms,

TABLE 1 | Total numbers of samples for three features.

Dataset	State	Drugs	Cell lines	x_{rna}	x_{snv}	x_{cna}
GDSC	Raw	139	1,124	11,833 (789)	70 (778)	24,960 (936)
	Preprocessed	98	555	11,712 (555)	54 (555)	24,959 (555)
CCLE	Raw	24	1,061	20,049 (1,028)	1,667 (1,044)	24,960 (742)
	Preprocessed	24	363	19,389 (363)	1,667 (363)	24,960 (363)

The number in the parenthesis means a total of cell lines corresponding to the features.

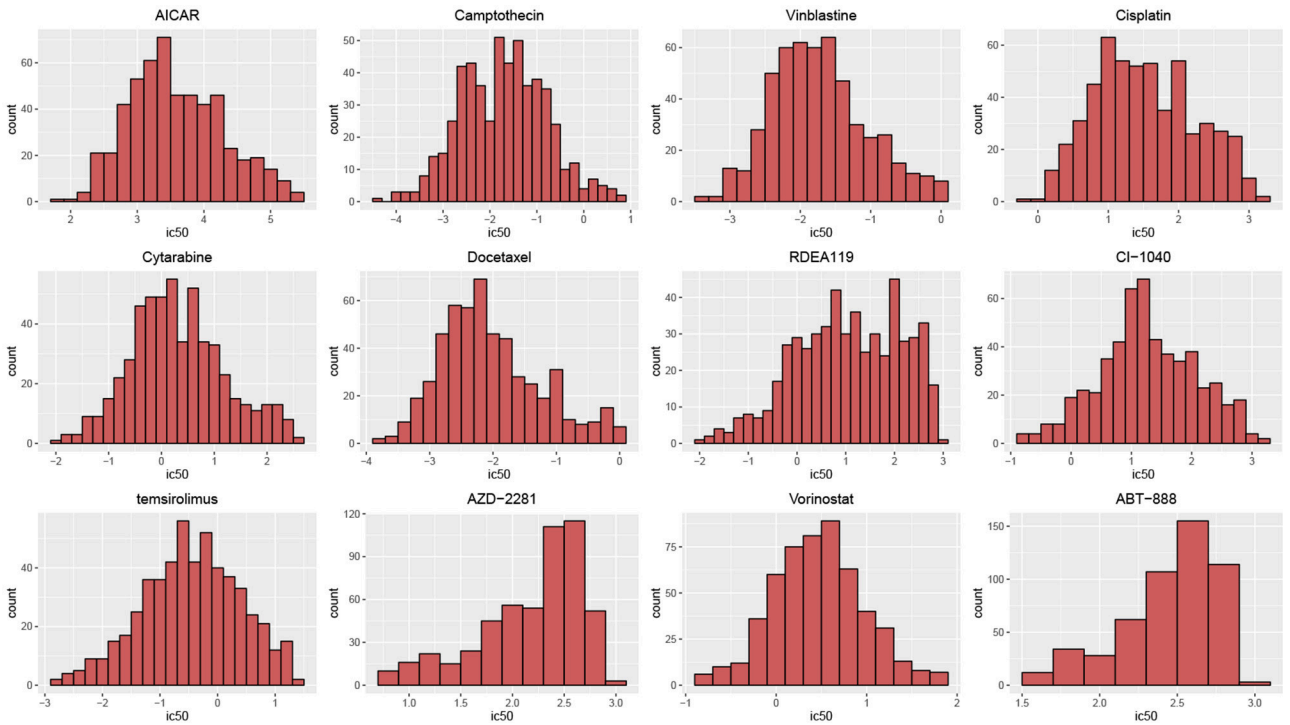

FIGURE 2 | Histograms of drug responses for 12 drugs in GDSC. The distributions of drug responses were different for various drugs.

including naive Bayes classifier (Barretina et al., 2012), SVM-RFE (Dong et al., 2015), FSelector for *k*-nearest-neighbors (KNN) algorithm (Soufan et al., 2015), and AutoHidden. The naive Bayes method first selected the top 30 features using either non-parametric Wilcoxon Sum Rank Test (for the gene expression features) or Fisher Exact Test (for the gene mutations). Then, the remaining significant features ($p < 0.25$) were clustered using a message-passing algorithm for each type of features. Then, they combined these two-part features and used a naive Bayes classifier for the drug response classification prediction. SVM-RFE is a wrapper method using a recursive feature selection and SVM classifier. The parameters of feature number, gamma and cost were set to be 10, 0.5, and 10, which were the optimal parameters selected by SVM-RFE. FSelector selected features using FSelector based on the information entropy and applied to the KNN algorithm. In AutoHidden, we directly use the hidden layer of the autoencoder constructed in our AutoBorutaRF, as the features.

TABLE 2 | Mean values of six evaluation metrics obtained from GDSC.

Method	AUC	ACC	REC	SPC	*F_1*	MCC
AutoBorutaRF	**0.7116**	**0.6534**	**0.6527**	0.6542	**0.6501**	**0.3109**
Naive Bayes	0.6792	0.6109	0.4242	**0.7969**	0.4947	0.2475
SVM-RFE	0.5159	0.5945	0.5797	0.6092	0.5855	0.1915
FSelector	0.6477	0.6061	0.6171	0.5952	0.6068	0.2155
AutoHidden	0.6095	0.5780	0.5576	0.5984	0.5651	0.1584

The bold number indicates the best result.

The overall prediction performance of the five methods for the two datasets is illustrated in **Tables 2**, **3** and **Figure 3**. All the metrics in the figure were obtained by using 10-fold CV. **Figure 3** showed that our method was of the best performance with respect to AUC, accuracy, recall, specificity, F_1 score, and Matthews correlation coefficient.

Among the 98 drugs in GDSC, ABT-888 presented the worst prediction with AUC being 0.5935, and the best prediction is for RDEA119 with AUC being 0.8282. Meanwhile, RDEA119, PD-0325901, 17-AAG, and Vorinostat were the only four drugs with AUC >0.8. However, there were 59 drugs, whose AUCs were higher than 0.7. Among the 24 drugs in CCLE, the worst prediction is for AEW541 with AUC being 0.6509. The best three predictions are for Nutlin-3, LBW242, and AZD6244, with AUC being 0.9633, 0.9300, and 0.9079, respectively. The AUCs of Irinotecan, Panobinostat, PD-0332991, PD-0325901, PHA-665752, PLX4720, and Topotecan are higher than 0.85. The receiver operating characteristic (ROC) curves are listed in **Supplementary File 1**.

3.3. Identified Biomarkers Are Associated With Cancer and Drug Target Pathway

We used 95 lung cell lines in the GDSC database to illustrate the biological significance of the identified biomarkers. **Figure 4A** shows the prediction performance of AutoBorutaRF for the lung cell lines. AutoBorutaRF showed satisfying prediction performance for predicting the drug responses for the lung cell lines. We used the non-parametric Wilcoxon sum rank test for the genetic features of gene expression and copy number alternation and a Fisher exact test for the genetic feature of single-nucleotide mutation, to test the significant difference of the genetic features between the sensitive and non-sensitive populations. Among all the identified 1,087 features (**Supplementary File 2**), a total of features with $p < 0.05$ was 1029, shown by **Figure 4B**. These results showed that most of the identified features were of significantly different genetic profiles between two classes (**Supplementary File 3**).

TABLE 3 | Mean values of six evaluation metrics obtained from CCLE.

Method	AUC	ACC	REC	SPC	*F_1*	MCC
AutoBorutaRF	**0.8210**	**0.7638**	**0.6560**	0.8137	**0.6248**	**0.4520**
Naive Bayes	0.7793	0.6838	0.3325	0.9194	0.3662	0.2759
SVM-RFE	0.5516	0.7287	0.4286	0.8129	0.5239	0.2961
FSelector	0.7372	0.7430	0.5061	0.8058	0.5639	0.3535
AutoHidden	0.7063	0.6970	0.1338	**0.9501**	0.3567	0.2198

The bold number indicates the best result.

We further use PLX4720 and BIBW2992 as two examples to illustrate the biological significance of the features selected for the lung cell lines. Prediction metrics of these two drugs are shown in **Figure 5**. PLX4720 is the inhibitor for B-raf and targets at MAPK signaling pathway (Michaelis et al., 2014). The selected significant features for PLX4720 were *CCL19, CCRL2, CST7, GPR143, HDAC5*, and *IDO1*. *CCRL2* inhibits p38 MAPK phosphorylation and up-regulates the expression of E-cadherin (Wang et al., 2015). Besides, *CCR7, CST7, GPR143, HDAC5*, and *IDO1* are also related to lung cancer or the MAPK pathway (Liu et al., 2014, 2018; Li and Seto, 2016; Matthews et al., 2016; Rose et al., 2016).

BIBW2992 inhibits *ERBB2* and *EGFR* and targets at EGFR signaling pathway (Iorio et al., 2016) and has been widely investigated for cancers, like lung cancer and melanoma (Rinehart et al., 2004; Nehs et al., 2010; Varmeh et al., 2016). The selected significant features were *FYN, KCNH2, REST, CDH12,*

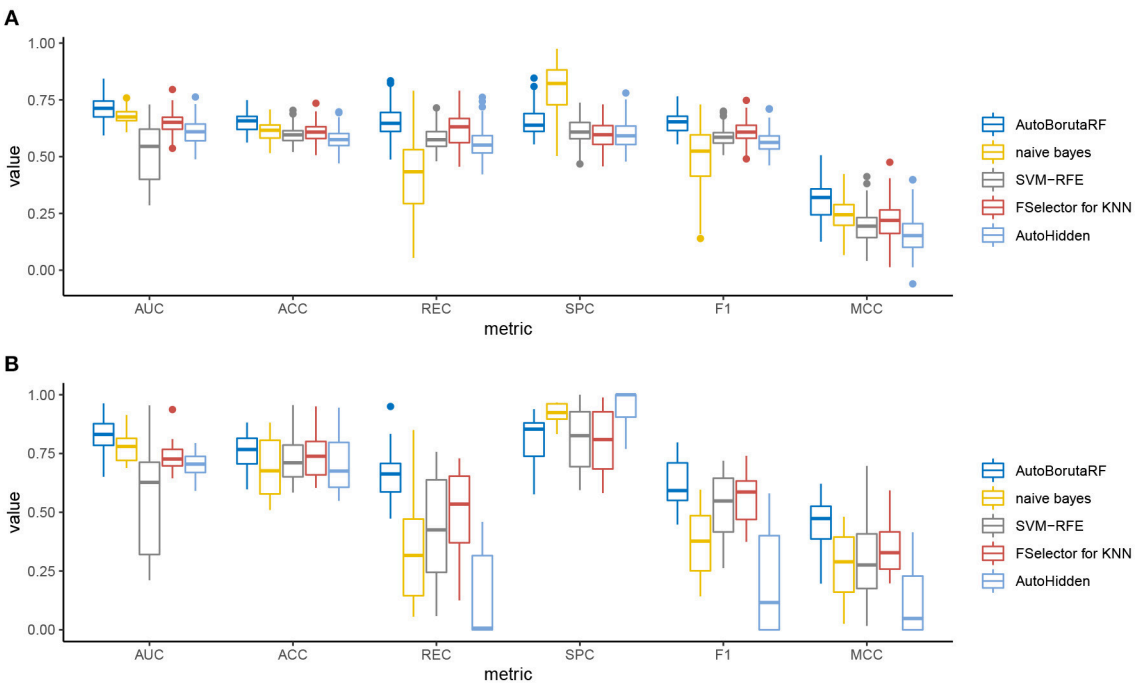

FIGURE 3 | Box plots of the six evaluation metrics overall the cell lines in the **(A)** GDSC and **(B)** CCLE datasets. Our method was of the best performance with respect to AUC, accuracy, recall, specificity, F_1 score, and Matthews correlation coefficient. The naive Bayes classifier and SVM-RFE outperformed at specificity.

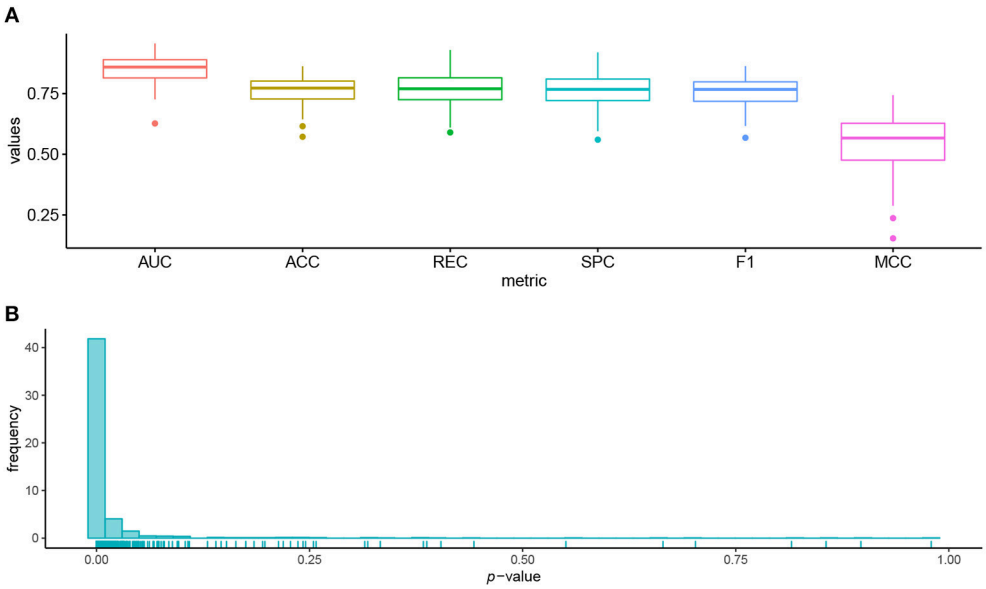

FIGURE 4 | Prediction performance for the lung cell lines in GDSC. **(A)** Box plots of six metrics overall the lung cells showed the satisfying prediction performance. **(B)** Histogram of *p*-values obtained by the statistical significance test for the identified features proved that most of the identified features were of significantly different genetic profiles between the sensitive and non-sensitive populations.

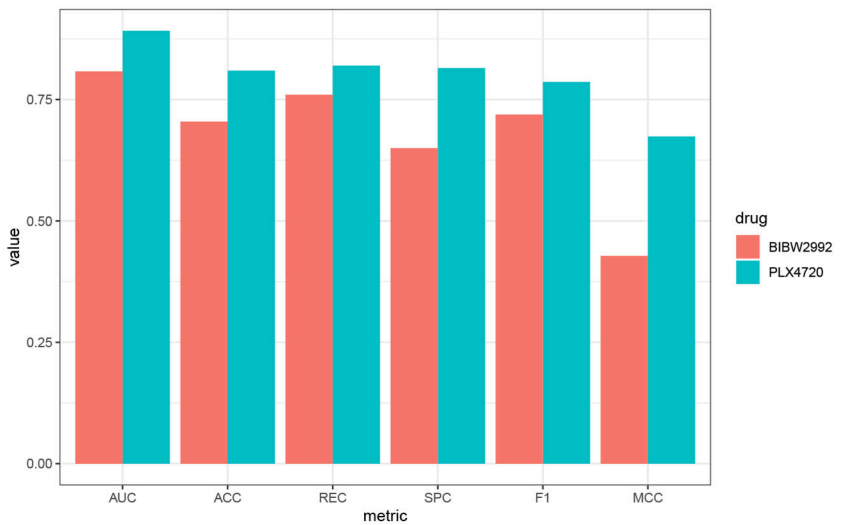

FIGURE 5 | Performance metrics of AutoBorutaRF overall the lung cell lines in GDSC for PLX4720 and BIBW2992.

LRRC8E, SCG2, PHF8, PCSK1, ANXA2, and *MIR6730*. *FYN* was an authentic Effector of oncogenic EGFR signaling, by limiting EGFR tumor cell motility (Lu et al., 2009). *CDH12* plays an important role in non-small-cell lung cancer(NSCLC) geneses, resulting from that the mutations of *CDH12* and other PRAME family members were equally distributed among tumors of different grades and stages (Bankovic et al., 2010). *SCG2* is in connection with the alteration of miRNA profiles in A549 human non-small-cell lung cancer cells (Shin et al., 2009). *KCNH2, REST, LRRC8E, PHF8, PCSK1, ANXA2, and MIR6730* have been also proved to be related to signaling pathway EGFR and lung cancer (Bonilla and Geha, 2006; de Castro et al., 2006; Kreisler et al., 2010; Wang et al., 2012; Demidyuk et al., 2013; Shen et al., 2014; Díaz-Rodríguez et al., 2018). The function descriptions and interaction networks of the identified features for PLX4720 and BIBW2992 are included in **Supplementary File 4**.

DISCUSSION

The prediction of anticancer drug response is crucial for many applications, like the preclinical setting and clinical trial design.

The prediction models for drug response include regression models and classification models. This research developed AutoBorutaRF for predicting the drug response for a two-fold aim: achieving proper features for RF and investigating biologically significant biomarkers for the explaining drug response. Because the genetic feature candidates are a vast set, we cannot directly apply the well developed Boruta algorithm for feature selection. We first drastically reduced the dimension by constructing the autoencoder network. Different from the typical application of a hidden layer of the autoencoder, we extracted the inputs with large contributions evaluated by the Gedeon method.

Considering AUC= 0.7 as a pass mark, 22 of 24 drugs in CCLE were of qualified prediction performance; 59 of 98 drugs in GDSC were of qualified prediction performance. Further analysis should be conducted to investigate the reasons leading to the prediction difference between two datasets.

We further investigated the biological significance. We proved that most of the identified genetic features between the sensitive and non-sensitive cell lines were significantly different. By using PLX4720 and BIBW2992 as two examples, we illustrated that many genes identified by AutoBorutaRF were reported to have close relationship with tumorigenesis or cancer progression. The detailed function explanations and interaction networks of the selected features can be referred to **Supplementary File 4**. Thus, AutoBorutaRF can be considered to be a capable machine learning method for determining the biomarkers for predicting the drug response for the preclinical and clinical purposes.

Note that our proposed method used no prior information to obtain the optimal feature set in the sense of prediction performance. In future research, the pre-determined information, like pathway knowledge, and the prior distribution describing the uncertainties of anticancer drugs can be considered to be embedded in our method.

DATA AVAILABILITY

The source code and datasets for this study can be downloaded from https://github.com/bioinformatics-xu/AutoBorutaRF.

AUTHOR CONTRIBUTIONS

XX and PQ processed the data, designed the algorithm, and the programming codes, and wrote the manuscript. YW supported result interpretation and manuscript writing. JW and HG supervised the project and contributed to writing the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61633006, 61502074, 81602309, 81422038, 81872247, 91540110, and 31471235).

ACKNOWLEDGMENTS

We thank Pi Xu Liu and Hailing Cheng for useful discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00233/full#supplementary-material

Supplementary File 1 | ROC curve of ten-fold cross validation.

Supplementary File 2 | Selected features.

Supplementary File 3 | Results of feature significance test.

Supplementary File 4 | Function descriptions and interaction networks for PLX4720 and BIBW2992.

REFERENCES

Ammad-ud din, M., Khan, S. A., Malani, D., Murumägi, A., Kallioniemi, O., Aittokallio, T., et al. (2016). Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics* 32, i455–i463. doi: 10.1093/bioinformatics/btw433

Ammad-Ud-Din, M., Georgii, E., Gonen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., et al. (2014). Integrative and personalized QSAR analysis in cancer by kernelized bayesian matrix factorization. *J. Chem. Inform. Model.* 54, 2347–2359. doi: 10.1021/ci500152b

Bankovic, J., Stojsic, J., Jovanovic, D., Andjelkovic, T., Milinkovic, V., Ruzdijic, S., et al. (2010). Identification of genes associated with non-small-cell lung cancer promotion and progression. *Lung Cancer* 67, 151–159. doi: 10.1016/j.lungcan.2009.04.010

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., et al. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 154, 1151–1161. doi: 10.1016/j.cell.2013.08.003

Bonilla, F. A., and Geha, R. S. (2006). 2. update on primary immunodeficiency diseases. *J. Allergy Clin. Immunol.* 117, S435–S441. doi: 10.1016/j.jaci.2005.09.051

Brubaker, D., Difeo, A., Chen, Y., Pearl, T., Zhai, K., Bebek, G., et al. (2014). "Drug intervention response predictions with paradigm (dirpp) identifies drug resistant cancer cell lines and pathway mechanisms of resistance," in *Biocomputing 2014*, eds R. B. Altman, A. K. Dunker, L. Hunter, T. A. Murray, T. E. Klein, and M. D. Ritchie (Hawaii, HI: World Scientific), 125–135.

Cichonska, A., Rousu, J., and Aittokallio, T. (2015). Identification of drug candidates and repurposing opportunities through compound–target interaction networks. *Expert Opin. Drug Discov.* 10, 1333–1345. doi: 10.1517/17460441.2015.1096926

Cortés-Ciriano, I., van Westen, G. J., Bouvier, G., Nilges, M., Overington, J. P., Bender, A., et al. (2015). Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* 32, 85–95. doi: 10.1093/bioinformatics/btv529

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 32, 1202–1212. doi: 10.1038/nbt.2877

de Castro, M. P., Aránega, A., and Franco, D. (2006). Protein distribution of Kcnq1, Kcnh2, and Kcne3 potassium channel subunits during mouse embryonic development. *Anat. Rec. Part A* 288, 304–315. doi: 10.1002/ar.a.20312

Demidyuk, I. V., Shubin, A. V., Gasanov, E. V., Kurinov, A. M., Demkin, V. V., Vinogradova, T. V., et al. (2013). Alterations in gene expression of proprotein convertases in human lung cancer have a limited number of scenarios. *PLoS ONE* 8:e55752. doi: 10.1371/journal.pone.0055752

Díaz-Rodríguez, E., Sanz, E., and Pandiella, A. (2018). Antitumoral effect of ocoxin, a natural compound-containing nutritional supplement, in small cell lung cancer. *Int. J. Oncol.* 53, 113–123. doi: 10.3892/ijo.2018.4373

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., et al. (2015). Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer* 15:489. doi: 10.1186/s12885-015-1492-6

Fey, D., Halasz, M., Dreidax, D., Kennedy, S. P., Hastings, J. F., Rauch, N., et al. (2015). Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.* 8, ra130–ra130. doi: 10.1126/scisignal.aab0990

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 570–575. doi: 10.1038/nature11005

Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *Int. J. Neural Syst.* 8, 209–218. doi: 10.1142/S0129065797000227

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell* 166, 740–754. doi: 10.1016/j.cell.2016.06.017

Kreisler, A., Strissel, P., Strick, R., Neumann, S., Schumacher, U., and Becker, C. (2010). Regulation of the NRSF/REST gene by methylation and CREB affects the cellular phenotype of small-cell lung cancer. *Oncogene* 29, 5828–5838. doi: 10.1038/onc.2010.321

Kursa, M. B., Rudnicki, W. R., et al. (2010). Feature selection with the boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

La Thangue, N. B., and Kerr, D. J. (2011). Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nat. Rev. Clin. Oncol.* 8, 587–596. doi: 10.1038/nrclinonc.2011.121

Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., et al. (2015). Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. *PLoS ONE* 10:e0130700. doi: 10.1371/journal.pone.0130700

Li, Y., and Seto, E. (2016). HDACs and HDAC inhibitors in cancer development and therapy. *Cold Spring Harb. Perspect. Med.* 6:a026831. doi: 10.1101/cshperspect.a026831

Liou, C.-Y., Huang, J.-C., and Yang, W.-C. (2008). Modeling word perception using the Elman network. *Neurocomputing* 71, 3150–3157. doi: 10.1016/j.neucom.2008.04.030

Liu, F.-Y., Safdar, J., Li, Z.-N., Fang, Q.-G., Zhang, X., Xu, Z.-F., et al. (2014). CCR7 regulates cell migration and invasion through MAPKs in metastatic squamous cell carcinoma of head and neck. *Int. J. Oncol.* 45, 2502–2510. doi: 10.3892/ijo.2014.2674

Liu, M., Wang, X., Wang, L., Ma, X., Gong, Z., Zhang, S., et al. (2018). Targeting the IDO1 pathway in cancer: from bench to bedside. *J. Hematol. Oncol.* 11:100. doi: 10.1186/s13045-018-0644-y

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A systematic study on drug-response associated genes using baseline gene expressions of the cancer cell line encyclopedia. *Sci. Rep.* 6:22811. doi: 10.1038/srep22811

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B* 39, 539–550. doi: 10.1109/TSMCB.2008.2007853

Lu, K. V., Zhu, S., Cvrljevic, A., Huang, T. T., Sarkaria, S., Ahkavan, D., et al. (2009). Fyn and SRC are effectors of oncogenic epidermal growth factor receptor signaling in glioblastoma patients. *Cancer Res.* 69, 6889–6898. doi: 10.1158/0008-5472.CAN-09-0347.

Matthews, S. P., McMillan, S. J., Colbert, J. D., Lawrence, R. A., and Watts, C. (2016). Cystatin F ensures eosinophil survival by regulating granule biogenesis. *Immunity* 44, 795–806. doi: 10.1016/j.immuni.2016.03.003

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., et al. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE* 8:e61318. doi: 10.1371/journal.pone.0061318

Michaelis, M., Rothweiler, F., Nerreter, T., Van Rikxoort, M., Sharifi, M., Wiese, M., et al. (2014). Differential effects of the oncogenic BRAF inhibitor PLX4032 (vemurafenib) and its progenitor PLX4720 on ABCB1 function. *J. Pharm. Pharm. Sci.* 17, 154–168. doi: 10.18433/J3TW24

Nehs, M. A., Nagarkatti, S., Nucera, C., Hodin, R. A., and Parangi, S. (2010). Thyroidectomy with neoadjuvant PLX4720 extends survival and decreases tumor burden in an orthotopic mouse model of anaplastic thyroid cancer. *Surgery* 148, 1154–1162. doi: 10.1016/j.surg.2010.09.001

Neto, E. C., Jang, I. S., Friend, S. H., and Margolin, A. A. (2014). "The stream algorithm: computationally efficient ridge-regression via bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity," in *Biocomputing 2014*, R. B. Altman, A. K. Dunker, L. Hunter, T. A. Murray, T. E. Klein, and M. D. Ritchie (Hawaii, HI: World Scientific), 27–38.

Nilsson, R., Peña, J. M., Björkegren, J., and Tegnér, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *J. Mach. Learn. Res.* 8, 589–612.

Ospina, J. D., Zhu, J., Chira, C., Bossi, A., Delobel, J. B., Beckendorf, V., et al. (2014). Random forests to predict rectal toxicity following prostate cancer radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* 89, 1024–1031. doi: 10.1016/j.ijrobp.2014.04.02

Riddick, G., Song, H., Ahn, S., Walling, J., Borges-Rivera, D., Zhang, W., et al. (2010). Predicting *in vitro* drug sensitivity using random forests. *Bioinformatics* 27, 220–224. doi: 10.1093/bioinformatics/btq628

Rinehart, J., Adjei, A. A., LoRusso, P. M., Waterhouse, D., Hecht, J. R., Natale, R. B., et al. (2004). Multicenter phase II study of the oral MEK inhibitor, CI-1040, in patients with advanced non-small-cell lung, breast, colon, and pancreatic cancer. *J. Clin. Oncol.* 22, 4456–4462. doi: 10.1200/JCO.2004.01.185

Rose, A. A., Annis, M. G., Frederick, D. T., Biondini, M., Dong, Z., Kwong, L., et al. (2016). MAPK pathway inhibitors sensitize BRAF-mutant melanoma to an antibody-drug conjugate targeting GPNMB. *Clin. Cancer Res.* 22, 6088-6098. doi: 10.1158/1078-0432.CCR-16-1192

Shen, Y., Pan, X., and Zhao, H. (2014). The histone demethylase PHF8 is an oncogenic protein in human non-small cell lung cancer. *Biochem. Biophys. Res. Commun.* 451, 119–125. doi: 10.1016/j.bbrc.2014.07.076

Shin, S., Cha, H. J., Lee, E.-M., Lee, S.-J., Seo, S.-K., Jin, H.-O., et al. (2009). Alteration of miRNA profiles by ionizing radiation in A549 human non-small cell lung cancer cells. *Int. J. Oncol.* 35, 81–86. doi: 10.3892/ijo_00000315

Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., et al. (2015). Pharmacogx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* 32, 1244–1246. doi: 10.1093/bioinformatics/btv723

Soufan, O., Kleftogiannis, D., Kalnis, P., and Bajic, V. B. (2015). DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. *PLoS ONE* 10:e0117988. doi: 10.1371/journal.pone.0117988

Suphavilai, C., Bertrand, D., and Nagarajan, N. (2018). Predicting cancer drug response using a recommender system. *Bioinformatics* 34, 3907–3914. doi: 10.1093/bioinformatics/bty452

Varmeh, S., Borre, P. V., Gunda, V., Brauner, E., Holm, T., Wang, Y., et al. (2016). Genome-wide analysis of differentially expressed miRNA in PLX4720-resistant and parental human thyroid cancer cell lines. *Surgery* 159, 152–162. doi: 10.1016/j.surg.2015.06.046

Wan, Q., and Pal, R. (2014). An ensemble based top performing approach for NCI-dream drug sensitivity prediction challenge. *PLoS ONE* 9:e101183. doi: 10.1371/journal.pone.0101183

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wang, C.-Y., Chen, C.-L., Tseng, Y.-L., Fang, Y.-T., Lin, Y.-S., Su, W.-C., et al. (2012). Annexin A2 silencing induces G2 arrest of non-small cell lung cancer cells through p53-dependent and-independent mechanisms. *J. Biol. Chem.* 287, 32512–32524. doi: 10.1074/jbc.M112.351957

Wang, L., Li, X., Zhang, L., and Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 17:513. doi: 10.1186/s12885-017-3500-5

Wang, L.-P., Cao, J., Zhang, J., Wang, B.-Y., Hu, X.-C., Shao, Z.-M., et al. (2015). The human chemokine receptor CCRl2 suppresses chemotaxis and invasion by

blocking CCL2-induced phosphorylation of p38 MAPK in human breast cancer cells. *Med. Oncol.* 32:254. doi: 10.1007/s12032-015-0696-6

Wilkinson, G. R. (2005). Drug metabolism and variability among patients in drug response. *N. Engl. J. Med.* 352, 2211–2221. doi: 10.1056/NEJMra032424

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2012). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi: 10.1093/nar/gks1111

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput. Biol.* 11:e1004498. doi: 10.1371/journal.pcbi.1004498

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

frontiers in Genetics

METHODS
published: 28 March 2019
doi: 10.3389/fgene.2019.00236

Simultaneous Interrogation of Cancer Omics to Identify Subtypes With Significant Clinical Differences

*Aodan Xu, Jiazhou Chen, Hong Peng, GuoQiang Han and Hongmin Cai**

School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

Recent advances in high-throughput sequencing have accelerated the accumulation of omics data on the same tumor tissue from multiple sources. Intensive study of multi-omics integration on tumor samples can stimulate progress in precision medicine and is promising in detecting potential biomarkers. However, current methods are restricted owing to highly unbalanced dimensions of omics data or difficulty in assigning weights between different data sources. Therefore, the appropriate approximation and constraints of integrated targets remain a major challenge. In this paper, we proposed an omics data integration method, named high-order path elucidated similarity (HOPES). HOPES fuses the similarities derived from various omics data sources to solve the dimensional discrepancy, and progressively elucidate the similarities from each type of omics data into an integrated similarity with various high-order connected paths. Through a series of incremental constraints for commonality, HOPES can take both specificity of single data and consistency between different data types into consideration. The fused similarity matrix gives global insight into patients' correlation and efficiently distinguishes subgroups. We tested the performance of HOPES on both a simulated dataset and several empirical tumor datasets. The test datasets contain three omics types including gene expression, DNA methylation, and microRNA data for five different TCGA cancer projects. Our method was shown to achieve superior accuracy and high robustness compared with several benchmark methods on simulated data. Further experiments on five cancer datasets demonstrated that HOPES achieved superior performances in cancer classification. The stratified subgroups were shown to have statistically significant differences in survival. We further located and identified the key genes, methylation sites, and microRNAs within each subgroup. They were shown to achieve high potential prognostic value and were enriched in many cancer-related biological processes or pathways.

Keywords: similarity integration, omics data, survival analysis, DNA methylation, gene expression, miRNA

OPEN ACCESS

Edited by:
Binhua Tang,
Hohai University, China

Reviewed by:
Xiaofeng Dai,
Jiangnan University, China
Pu-Feng Du,
Tianjin University, China

****Correspondence:***
Hongmin Cai
hmcai@scut.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 November 2018*
Accepted: *04 March 2019*
Published: *28 March 2019*

Citation:
Xu A, Chen J, Peng H, Han G and Cai H (2019) Simultaneous Interrogation of Cancer Omics to Identify Subtypes With Significant Clinical Differences. Front. Genet. 10:236. doi: 10.3389/fgene.2019.00236

1. INTRODUCTION

In current clinical practice, cancer is typically categorized based on its tissue source and pathological histology. However, cancer is also known as a well-characterized pathological system among the molecular level. Most cancers emerge along with complex molecular alterations at the germ and/or somatic level (Kristensen et al., 2014). Molecule-level cancer re-classification and

subtyping based on genome-scale data sets can act as a sally port for precision oncology (Wu et al., 2017), such as for evaluating the metastatic potential of patients and selecting the most promising treatment (Forbes et al., 2010). Although enormous quantities of molecular data have been accumulated from various cancer profiling projects, for example, the Catalog of Somatic Mutations in Cancer (COSMIC) database (Forbes et al., 2008), the International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium et al., 2010), and The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), interpreting such data is difficult. In recent years, many sophisticated statistical and mathematical models have been proposed to analyze biological data, most of which are based on a single data type (e.g., gene expression, methylation). However, all biological mechanisms consist of multiple molecular phenomena and genomes exhibit variation owing to gene mutations, epigenetic changes, individual differences and environmental influences. It is difficult for conventional analysis based on data from a single genome to capture the heterogeneity of all biological processes and clearly differentiate phenotypes. Thus, the focus has now been shifted to how to integrate multi-omics to achieve more promising and stable cancer classification results.

To perform such simultaneous interrogation, there are two major challenges. First, distinct omics data are heterogeneous in scale, dimension, and quality, and such heterogeneity requires subtle processing. Second, there are internal relationships between single data layers (e.g., the promoter DNA methylation may suppress expression). As such, information on these regulatory patterns can improve our integrated analysis. Existing methods can be roughly divided into three categories based on their methodology: latent variable representation methods, probabilistic modeling methods, and network-based methods (Huang et al., 2017; Rappoport and Shamir, 2018). Latent variable representation are mainly committed to mapping diverse features from different data types into a shared low-dimension common space under the assumption that a set of latent variables is shared across multi-omics data. For example, iCluster+ employs an expectation-maximization (EM) algorithm to build regularized regression in modeling latent variables and observed data (Mo et al., 2013). A joint non-negative matrix factorization (jNMF) method is used to detect the shared characteristic space (Zhang et al., 2012). A moCluster algorithm can define a joint latent variable using the modified consensus PCA (CPCA) (Meng et al., 2015). The major drawback of these methods is that, when dimensions and variances of different omics datasets differ greatly, the basic assumption may be unexplainable. The unobserved latent variables possess little biological meaning and have far fewer dimensions than original spaces. Probabilistic models always presume different prior distributions of multi-omics data, constructing a mixture model, and then estimate the parameters and mixture ratios. For instance, a Beta-Gaussian mixture model can integrate gene expression data and protein-DNA binding probabilities into a single probabilistic modeling framework (Dai et al., 2009). Except for modeling original data, we can also model the probability of clusters distribution on the local and global level using the hierarchical Dirichlet mixture model (Gabasova et al., 2017). However, the accuracy relies heavily on the inherent distribution of data and overfitting may occur when sample size far less than features. Instead of searching common latent variables in measurement space, network-based methods begin with each single data layer and propagate information through interactions between samples to construct a global graph structure. A previous work named similarity network fusion (SNF) (Wang et al., 2014) follows this route using the message-passing theory to fuse similarities of each available data type into one network by iteratively updates every network as the similarity matrix product of a single layer and the average of the rest layers. Network structure can effectively handle differences in dimension and scale. However, the main difficulty lies in how to determine the contributions of each local pattern and how to interpret the clustering result in terms of the original features. Hence, there are still-strong demands for efficient and precise multi-omics data integration methods that can overcome the dimension variance and heterogeneous scale.

In this paper, we proposed a method to interrogate omics data simultaneously to achieve multi-scale cancer subtyping. The proposed high-order path elucidated similarity (HOPES) integrates the similarities for each type of omics data into a unified and stable one, thus achieving a simplified link of the underlying mechanism of various types of expression. We modeled integrated similarity as the approximation to various high-order paths across each local dataset, the progressively increased high-order path can represent different consistency requirements. We especially emphasized interaction within each pair of local layers rather than updates using a single layer and average of the rest layers. HOPES models such similarity integration as a minimization problem consist of three subobjective functions, for which an efficient numerical algorithm was designed to obtain the solution. Through the optimization procedure, we strengthened the strong correlation between patients and removed the weak ties mainly caused by noise. Thereby, we successfully subtype cancers with significant clinical differences. Real experiments on five cancer projects of TCGA and a normal control set for cancer diagnosis and prognosis tasks demonstrated the excellent performance of HOPES in subtyping and identifying key oncogenesis pathway. The subsequent biological analysis of the resulted key pathway was shown to possess potential prognostic value and biological significance.

2. MATERIALS AND METHODS

2.1. Tumor Datasets With Comprehensive Omics Measurements

We tested the proposed HOPES on five distinct tumor datasets, downloaded from TCGA. The tested samples consisted of five tumor types: glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), and a cervical cancer dataset (CESC). Each tumor was measured by DNA methylation, gene expression, and miRNA expression. The overall survival information corresponding to each sample was also considered.

The first four projects were the same as the experimental data obtained in a previous study (Wang et al., 2014). The gene expression data for GBM and LUSC were collected using the Broad Institute HT-HG-U133A platform, while COAD was collected by the UNC-Agilent-G4502A-07 platform, and KIRC by the UNC-Illumina-Hiseq-RNASeq platform. The miRNA expression data for GBM were collected by the UNC-miRNA-8X15K platform, while those for LUSC, KIRC, and COAD were collected by the BCGSC-Illumina-GA-miRNAseq. The methylation for GBM was analyzed by the JHU-USC-Illumina-DNA-Methylation platform, while for the others the JHU-USC-Human-Methylation-27 platform was used. The fifth CESC dataset contains data on clinical and pathological features, genomic alterations, DNA methylation profiles, and RNA and proteomic signatures, and is available from TCGA (Cancer Genome Atlas Research Network et al., 2017). We collected gene expression profiles, DNA methylation expression, miRNA expression, and clinical data from the Broad Institute TCGA Genome Data Analysis Center (Broad Institute TCGA Genome Data Analysis Center, 2016). A total of 284 samples with these four types of data were included in the study. For each data type, we removed signatures with a missing rate among all of the samples higher than 20%. For the remaining missing-value data, a K-nearest neighbor (KNN) imputation (Troyanskaya et al., 2001) scheme was used to complement it by filling the empty area with the mean value of non-empty neighbors. Finally, we normalized each dataset across samples and obtained a gene expression dataset of 20,118 genes, a methylation dataset of 396,065 CpG sites, and a miRNA dataset of 885 miRNAs. To reduce computational cost, for analysis involving methylation data, the 1,000 most variable CPG sites based on the standard deviation of beta values were selected.

2.2. Comparative Healthy Dataset as a Control

Besides the tumor samples, we also prepared normal samples as a control set to evaluate the capacity for using HOPES in diagnosis. A few healthy cases with data on gene expression, methylation, and miRNA expression are also included in TCGA. Finally, we merged 35 samples derived from several normal tissues adjacent to cancerous tissue among the six TCGA disease projects(BRCA, GBM, KIRC, COAD, LUSC, and CESC). Preprocessing as mentioned above was also performed on the 35 normal controls. Although we simply integrate healthy samples from different tissues as a control set, the normalization step can remove differences between different tissues, and ensure the separability between cancer samples and healthy controls.

2.3. Methods

2.3.1. SNF

Similarity network fusion(SNF) is a novel algorithm which integrates different omics data through computing and fusing patient similarity networks. SNF conduct the similarity fusing by iteratively updating every similarity network, making it more similar to the others with every iteration as follows:

$$P^{(v)} = S^{(v)} \times \left(\frac{\sum_{k \neq v} P^{(k)}}{m-1} \right) \times (S^{(v)})^T, v = 1, 2, ..., m$$

where P represent the similarity matrix derived from each datasets, S represent the local affinity which only contains the nearest neighbors' information, and m is the number of different data types. Actually the iteration process means updating the similarity between node i and node j in $P^{(v)}$ as the weighted sum of similarities between the K nearest neighbors of node i and those of node j. While neighbors' similarities are derived from the other $m-1$ datasets.

The main contribution of SNF is it can solve the discrepancy of dimensions and variances in different omics datasets which may be the biggest challenge for omics data integration. And it has been widely used in many practical biological tasks. However, it still exists some limitations in this algorithm. (1) This procedure treats each network as the same without weights constraints. (2) There is only one connection path between different datasets that across two intermediate nodes which is insufficient for depicting complex network interaction. (3) The information exchange only exists in one dataset and the average of the others. There are no direct mutual adjustments between different datasets which may cover some interconnection between specific data types. The incomplete network connection model makes it difficult to recover the most precise global similarity pattern or resist high-level noise in biological data.

2.3.2. Similarity Fusion Through High-Order Path

To have a consistent and highly representative global similarity, HOPES simulate three different network connection models with different path length and try to find the fused pattern which retains the maximal commonality. As it was depicted in **Figure 1**, (1) Path-0 similarity preserves the characteristics of each local affinity obtained using K nearest neighbor, (2) Path-1 similarity import one intermediate node to enhance the effect of each local affinity, (3) Path-2 similarity import two intermediate nodes to integrate interaction between different local affinity to enhance the commonality. The detailed numerical expression and constraint of the different order paths are as follows.

Suppose we have C different omics datasets, and their local affinities $S_i (i \in 1, ..., C)$ were evaluated by a scaled exponential similarity kernel (Wang et al., 2014) see details in **Supplementary Methods**. First, for the path-0 similarity, the fused similarity is required to be close to each underlying affinity which can be simply characterized by minimizing average losses as follows:

$$\min_W \sum_{i=1}^{C} \| W \cdot \Omega_i - S_i \|_F^2 \quad (1)$$

where W is a $n \times n$ fused similarity matrix, S_i is local affinity extracted from i-th omics data, and Ω_i is a $n \times n$ matrix whose entries denote whether corresponding entries in S_i are equal to 0. There are C types of omics data.

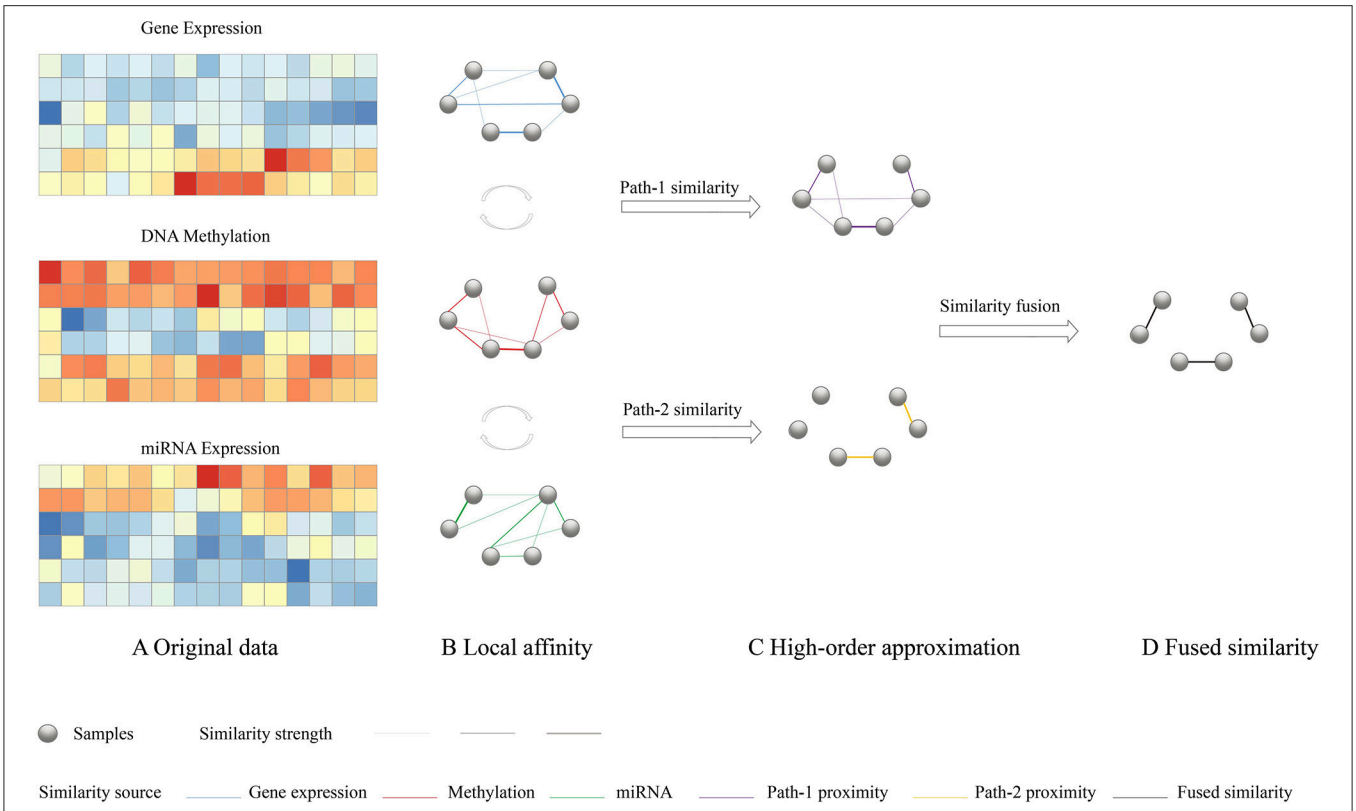

FIGURE 1 | Overview of HOPES. **(A)** Illustrative example of input multi-omics data, including gene expression, DNA methylation, and miRNA expression data sets for the same sample cohort. **(B)** Local affinity of each data type was defined as part of the global similarity matrix that only contains edges among *K* nearest neighbors (*K* = 2). **(C)** Simplified illustration of path-1 elucidated similarity. We used matrix multiplication to transform the directed distance between samples to weighted one-hop distance. The purple edges represent correlations shared by two data types. **(D)** A path-2 elucidated similarity, in which only those edges with large similarity are preserved. Such edges are highlighted in yellow.

Different from the path-0 similarity, we further propose path-1 similarity to retain the maximal commonality when filtering through each underlying affinity. Thus we assume the fused global similarity to be close to every one step transformed similarity by multiple each local affinity.

$$\min_{W} \sum_{i=1}^{C} \| W - S_i W \|_F^2 \tag{2}$$

It can be noted that $(S_i W)_{(m,n)} = \sum S_i(m,k) W(k,n)$ can be interpreted as the weighted sum of distance between the K nearest neighbors of node m and node n while neighbors' information are from dataset i, which represents W filtered by S_i. Therefore, the aim of Equation (2) is to ensure proximity between the global affinity and the transformed affinity after it has been weighted by each local affinity. One can impose a stricter requirement that the fused global similarity is closed to the transformed similarity which has been filtered by each underlying local affinity through higher-order paths. For example, with path-2 proximity,

$$\min_{W} \sum_{i=1}^{C} \sum_{j=1}^{C} \| W - S_i W S_j^T \|_F^2 \tag{3}$$

Where $(S_i W S_j)_{(m,n)} = \sum S_i(m,k) W(k,l) S_j(l,n)$, It also represents the weighted sum of the distance between the K nearest neighbors of node m and those of node n, while neighbors' information of two vertexes is from two different datasets. This interactivity between different local affinity sharply strengthens the commonality requirement. The filtration process is supposed to weaken the original edges in W unless the correlation between node i and j is simultaneously supported by each pair of data types.

Finally, combining the aforementioned constraints for modeling proximities of various path orders, we propose the determination of the global affinity by minimizing the following energy function:

$$\min_{W} \sum_{i=1}^{C} (\| W \cdot \Omega_i - S_i \|_F^2 + \alpha \| W - S_i W \|_F^2 + \beta \sum_{j=1}^{C} \| W - S_i W S_j^T \|_F^2) \tag{4}$$

where α and β are hyperparameters that adjust the weight of different order constraints and can be empirically set. Details on parameter tuning was attached in the **Supplementary Methods**. The optimization problem can be solved through a consensus

alternating direction minimization method (ADMM)(see **Supplementary Methods** for detailed solution procedure).

In conclusion, the three different order paths represent an incremental relationship from specificity to commonality and from weak constraint to strong constraint. They can simulate much more complex network connection models and set increasing consistency requirements on the global similarity. Therefore, we can take all the specialty of every single dataset, the interconnection between datasets, and global consistency into consideration and construct a more comprehensive and robust global similarity network. Moreover, the weights can be adjusted manually based on the real world condition which makes HOPES more flexible.

2.3.3. Downstream Applications

Once we have the fused global similarity matrix, it can be the fundamental structure for much downstream analysis. The most directly is applying the spectral clustering to cluster the samples into different subgroups which can be used for cancer diagnosis or molecular subtyping. In this paper, to eliminate the variations due to clustering initialization, the consensus clustering (Monti et al., 2003) was used to enhance the reliability performance. It records the consensus across multiple clustering repeated trials based on one certain global similarity matrix to assess the stability of the clustering results.

Except for clustering, we also tried to project the global structure into specific characteristics in every single dataset. Since these features are the most relevant to the fused results, they can not only be prognostic valuable but also may indicate some interconnection between different omics layers. We located these features using MCFS, an unsupervised feature selection algorithm for multi-cluster data (Cai et al., 2010). After providing our fused similarity matrix W and the original omics data as input, the feature selection task can be modeled as a $L1-regularized$ regression problem that exports the sparse coefficient vectors of features. In this case, we can easily select

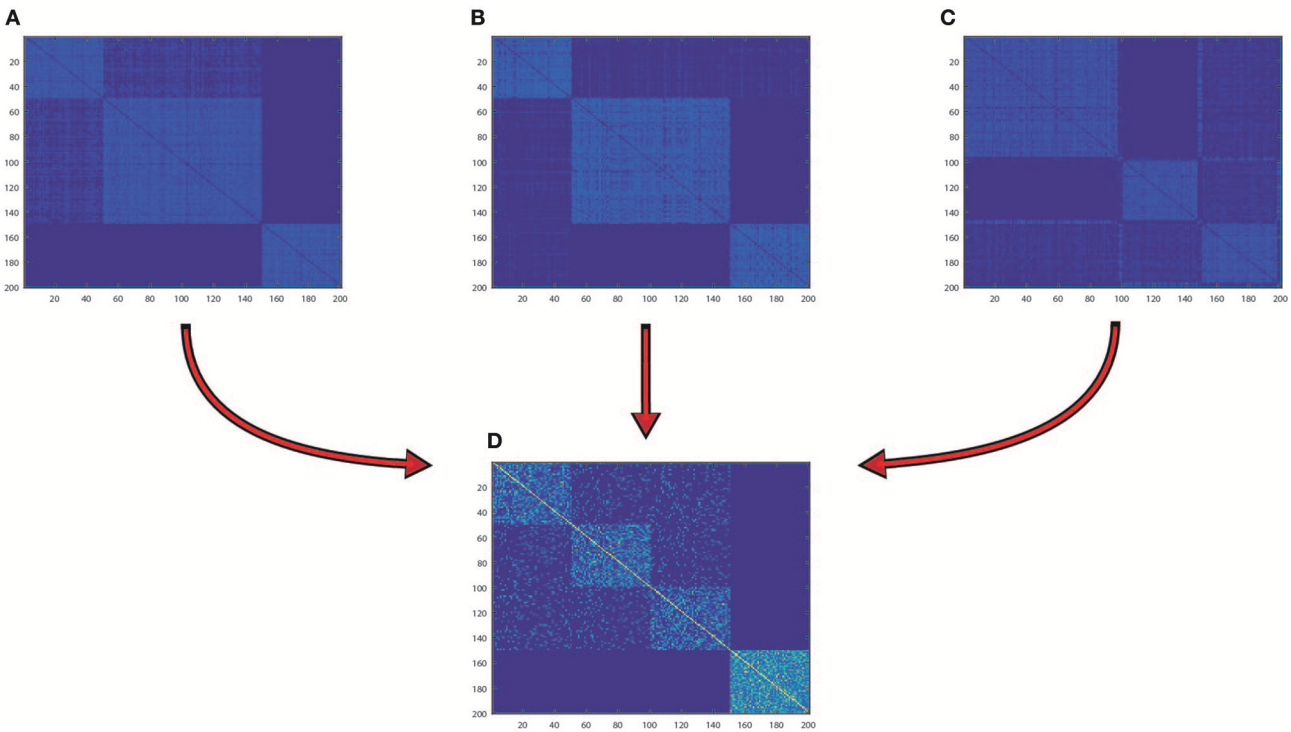

FIGURE 2 | Synthetic data integration. Heatmaps of similarity matrix distinctly derived from RNA **(A)**, methylation **(B)**, miRNA data **(C)**, and fused data **(D)**, for which samples were arranged in the order of spectral clustering results.

TABLE 1 | Performance measured by NMI on simulated datasets.

	SimData1			SimData2		
	Low noise	**Moderate noise**	**High noise**	**Low noise**	**Moderate noise**	**High noise**
HOPES	**0.972 ± 0.025**	**0.921 ± 0.044**	**0.858 ± 0.060**	**0.889 ± 0.056**	**0.838 ± 0.072**	**0.799 ± 0.071**
SNF	0.954 ± 0.061	0.811 ± 0.088	0.750 ± 0.075	0.822 ± 0.109	0.668 ± 0.095	0.619 ± 0.054
moCluster	0.864 ± 0.113	0.778 ± 0.088	0.748 ± 0.104	0.815 ± 0.015	0.786 ± 0.076	0.731 ± 0.108
iCluster+	0.710 ± 0.008	0.707 ± 0.008	0.693 ± 0.016	0.659 ± 0.026	0.617 ± 0.028	0.595 ± 0.036

a series of most relevant features(corresponding to the non-zero coefficients).

3. RESULTS

We designed a series of experiments to demonstrate the progress of HOPES by comparing it with four representative methods belong to three kinds of popular integration framework: network fusion-based SNF (Wang et al., 2014), joint latent variables-based iCluster+ (Mo et al., 2013), moCluster (Meng et al., 2015), and probabilistic model-based Clusternomics (Gabasova et al., 2017). Simulations and real data experiments were performed to evaluate the performance on global cluster structure detection and usability in clinical practice, respectively.

3.1. Experiments to Demonstrate the Accuracy and Robustness of HOPES With Simulated Data

To demonstrate the performance of HOPES in fusing multi-omics data, we first tested it on simulated datasets and

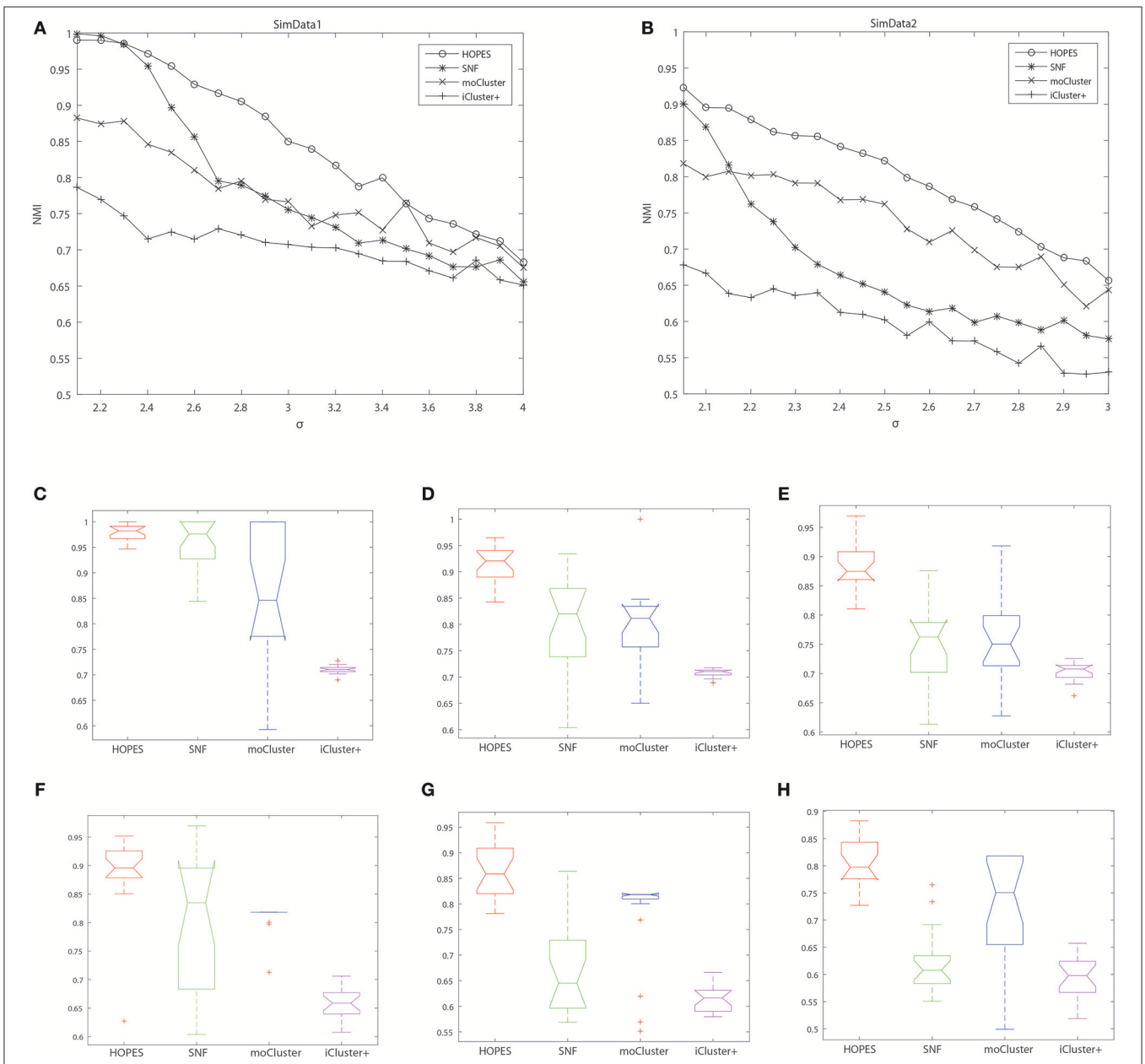

FIGURE 3 | Cluster accuracy comparison between different methods on different simulation datasets. The upper panel represents the NMI of HOPES, SNF, and moCluster in SimData1 **(A)** and SimData2 **(B)** under incremental standard deviation of Gaussian noise. The lower panel shows the NMI boxplots in SimData1 **(C–E)** and SimData2 **(F–H)** among three methods under different noise levels (from left to right in the order of low, intermediate, and high), measuring their accuracy and stability on recovering the integrated pattern through partial layers.

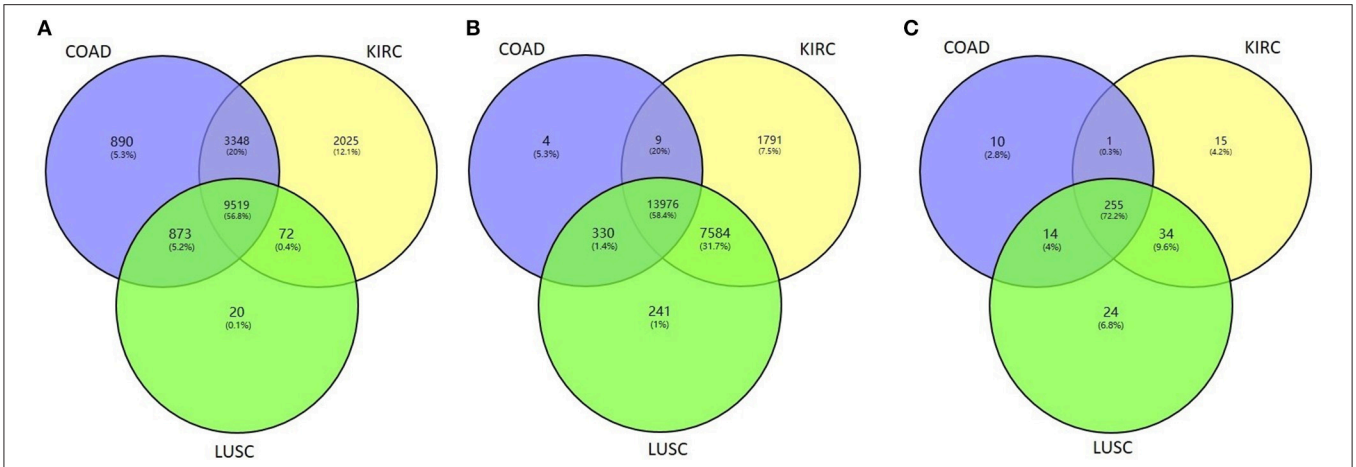

FIGURE 4 | Venn diagram of the overlapping measurements among LUSC, COAD, and KIRC projects for the biological features of **(A)** gene expression, **(B)** methylation, and **(C)** miRNA expression.

TABLE 2 | The accuracy for cancer diagnosis of different methods.

	COAD	KIRC	LUSC
Gene expression	0.8740	0.5159	0.8865
Methylation	0.4882	0.6433	0.6667
miRNA expression	0.8504	0.8471	0.8652
HOPES(fused)	**0.8976**	**0.9236**	**0.9286**
SNF(fused)	0.8976	0.9172	0.9078
iCluster+(fused)	0.6299	0.5923	0.6383
moCluster(fused)	0.7559	0.707	0.7801
Clusternomics(fused)	0.5276	0.6433	0.8865

compared it with SNF and moCluster. The simulated dataset was generated similarly to the one reported elsewhere (Shi et al., 2017). The simulated dataset was created to recapitulate the features of actual genomic data by combining biological variation levels from real data and a pre-defined cluster structure. The actual genomic profiles were downloaded from GEO (Barrett et al., 2013) with the following GEO codes: GSE51557, GSE73002 and GSE106453. These three were focused on DNA methylation (Conway et al., 2015), RNA expression (Nakagawa et al., 2008) and miRNA expression (Shimomura et al., 2016), respectively. Based on these actual genomic data we used the singular value decomposition (SVD) to fuse them with pre-defined cluster structure, and constructed two synthetic data sets (SimData1 and SimData2). SimData1 has a clear boundary between each cluster while SimData2 possesses fuzzy boundaries(see **Supplementary Methods** for more details).

We tested HOPES and the other methods on both simulation datasets under different levels of noise intensity to assess the information integration capability and robustness. We used the normalized mutual information (NMI) as a criterion for performance, and for each noise condition we ran repeated trials 20 times to eliminate accidental error. Collectively, all simulation results suggested that HOPES can always successfully recover the four pre-defined clusters from incomplete layers (**Figure 2**). As we demonstrate in data construction, the three single layers each contained an indivisible part. To dig out the real cluster information, an effective integration method was required. The proposed HOPES used the high order path distance among different data types to approximate the global similarity. The correlation information of nodes i and j will be weakened if it exists in only a single data layer, which ensures the separation of mixed groups in a single data source. Moreover, the progressive proximity model not only sets constraint on the high-order path distance, but also reconcile the extremely specific characteristics in each single data layer. Thus, it is promising for detection of the hidden cluster structure shaped by multi-source data.

The numerical results are shown in **Table 1** and **Figure 3**, which suggest that HOPES outperformed the compared methods irrespective of the set signal and noise conditions, highlighted in bold in **Table 1**. It should be noted that Clusternomics show little tolerance on noise, because the lack of modeling for noise. For the rest three methods we can add the variance of Guassian noise to 3, while Clusternomics can only resist noise with variance lower than 1 (see **Supplementary Figures** for more details). In this section, we mainly discuss the performance on the rest four methods. It can be demonstrated that SNF achieved high precision when the noise level remained low; however, its robustness upon exposure to noise was insufficient. The low stability may be ascribed to SNF updating a fused network through a single local affinity and the other average similarity at every iteration. The update rule raises concern about the enhancement of erroneous information derived from one data layer, especially when edge points exist. However, HOPES provided path-2 elucidated similarity determined by each pair of data types which effectively solve it. In contrast, the latent variables-based methods such as iCluster+ and moCluster showed fairly good stability but poor accuracy for both of the synthetic datasets as noise increased. The iCluster+ modeled continuous variable as the linear combination of specific intercept term, common latent variables, and residual variance which all follow normal distribution. This assumption can fits

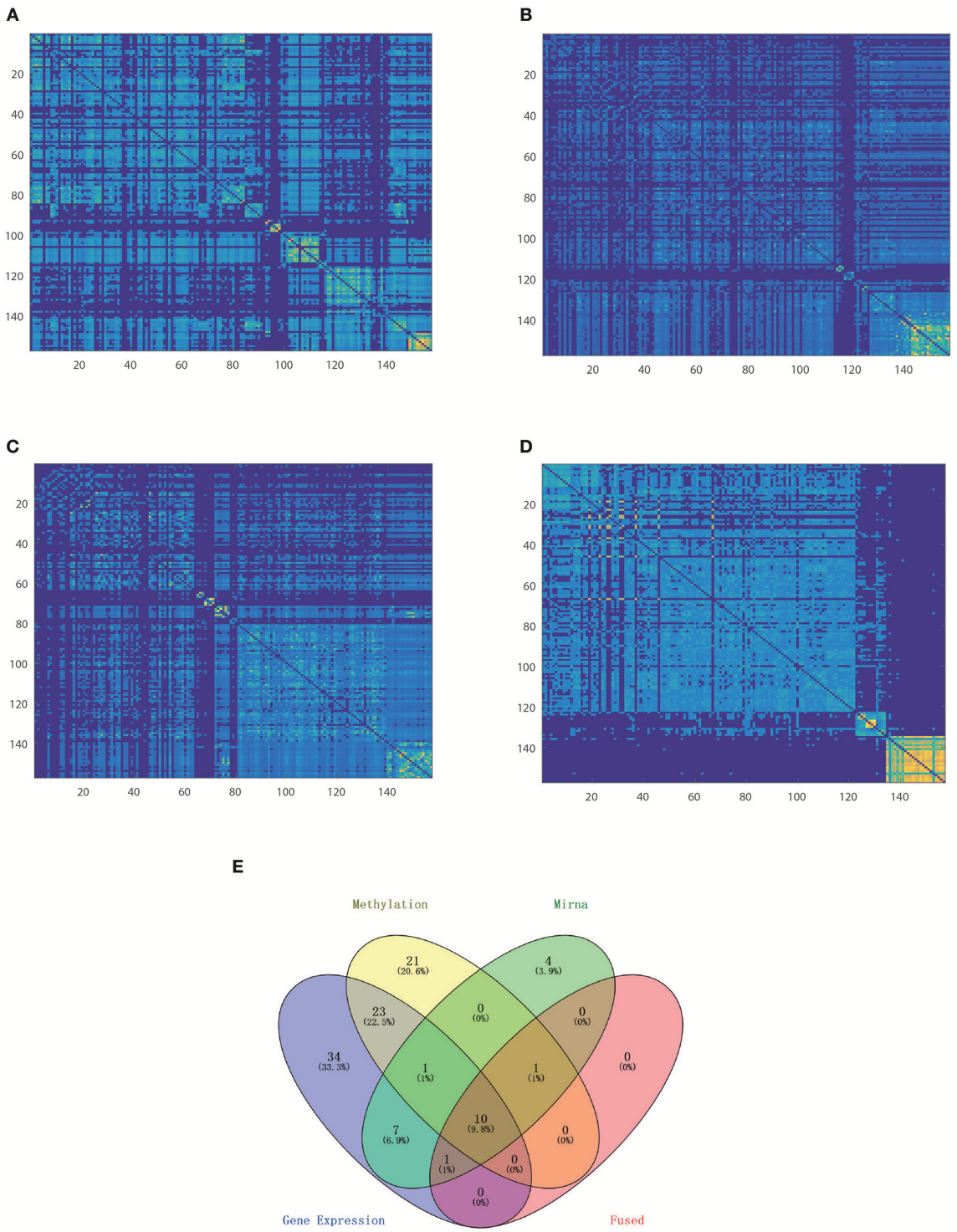

FIGURE 5 | Comparison of classification performance based on single and fused data. Heatmap of similarity matrix derived from **(A)** gene expression data, **(B)** methylation data, **(C)** miRNA expression data, and **(D)** fused data where samples were gathered by classification results on the corresponding dataset. **(E)** Venn plot shows the distribution of mis-assigned specimens in all of the four data sets.

our noise and original data setting, however, it can not accurately model the distribution of latent variables as a discrete sequence. So iCluster+ show good performance on dealing noises but unable to capture the global structures. The moCluster is based on a joint latent variable derived by consensus PCA, so it strongly relies on the selection of principal components. Moreover, the

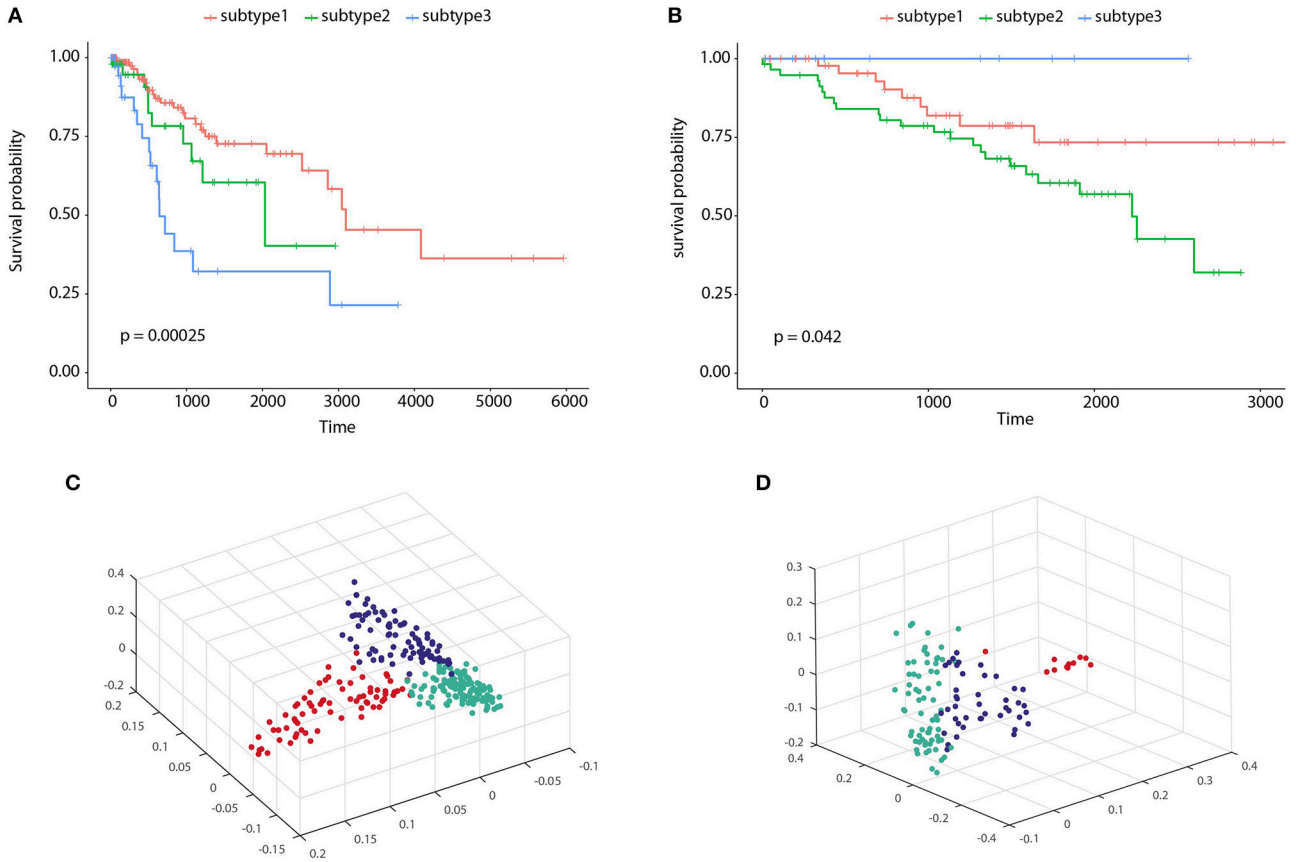

FIGURE 6 | Clustering results of CESC and KIRC. The Kaplan-Meier survival curves *P*-values are recorded in **Table 3** for **(A)** CESC and **(B)** KIRC, and 3D scatter plots for **(C)** CESC and **(D)** KIRC. Vertexes of scatter plots represent samples colored by their cluster label; the *x*-, *y*-, and *z*-axis represent the first three principal components of the fused data matrix.

large gap between feature magnitude of distinct data types also affects the accuracy. More specifically, the boxplots indicate the degree of dispersion and skewness in the data, and show outliers during 20 repeated trials under low, medium, and high noise levels. As depicted in **Figures 3C–H**, HOPES achieved higher accuracy and more stable results within all three methods in SimData1. However, the results of moCluster were highly dispersed during repeated trials which makes the results less credible. After we imported edge points in SimData2, the discreteness of every method slightly increased, but HOPES still performed best, in accordance with the previous results. Interestingly, moCluster appears to be very stable when the noise level is low, but with moderate noise, almost half of the trials were quantified as outliers, which suggests this method exhibits large fluctuations.

3.2. Experiments for Cancer Diagnosis on Actual Cancer Datasets

We then tested whether the proposed method HOPES can distinguish tumor samples from normal controls based on their omics measurements. We applied the HOPES and other comparative methods to combinations of COAD (92 samples), KIRC (122 samples), LUSC (106 samples) and 35 normal controls. The gene expression, methylation, and miRNA expression data for these case/control sets and the overlap among them are shown in **Figure 4**. It can be noted that the amounts and proportion of common variables vary between different data types. The normal samples tested in this work were selected to have the matching characteristics. It can be noted that the amounts of variables vary from the expression of 280 miRNAs to 23,360 methylation sites, and miRNA measurements are shown to have the largest proportion of overlap among all of cancer types.

TABLE 3 | Survival analysis by Log-rank test on five tumor datasets.

	CESC	COAD	GBM	KIRC	LUSC
HOPES	**0.000248**	0.00918	**0.000224**	**0.0417**	**0.00132**
SNF	0.000626	0.038	0.000621	0.124	0.00551
iCluster+	0.63	**0.00316**	0.751	0.206	0.0082
moCluster	0.0567	0.139	0.0207	0.0667	0.00193
Clusternomics	0.162	–	0.048	0.129	0.00504

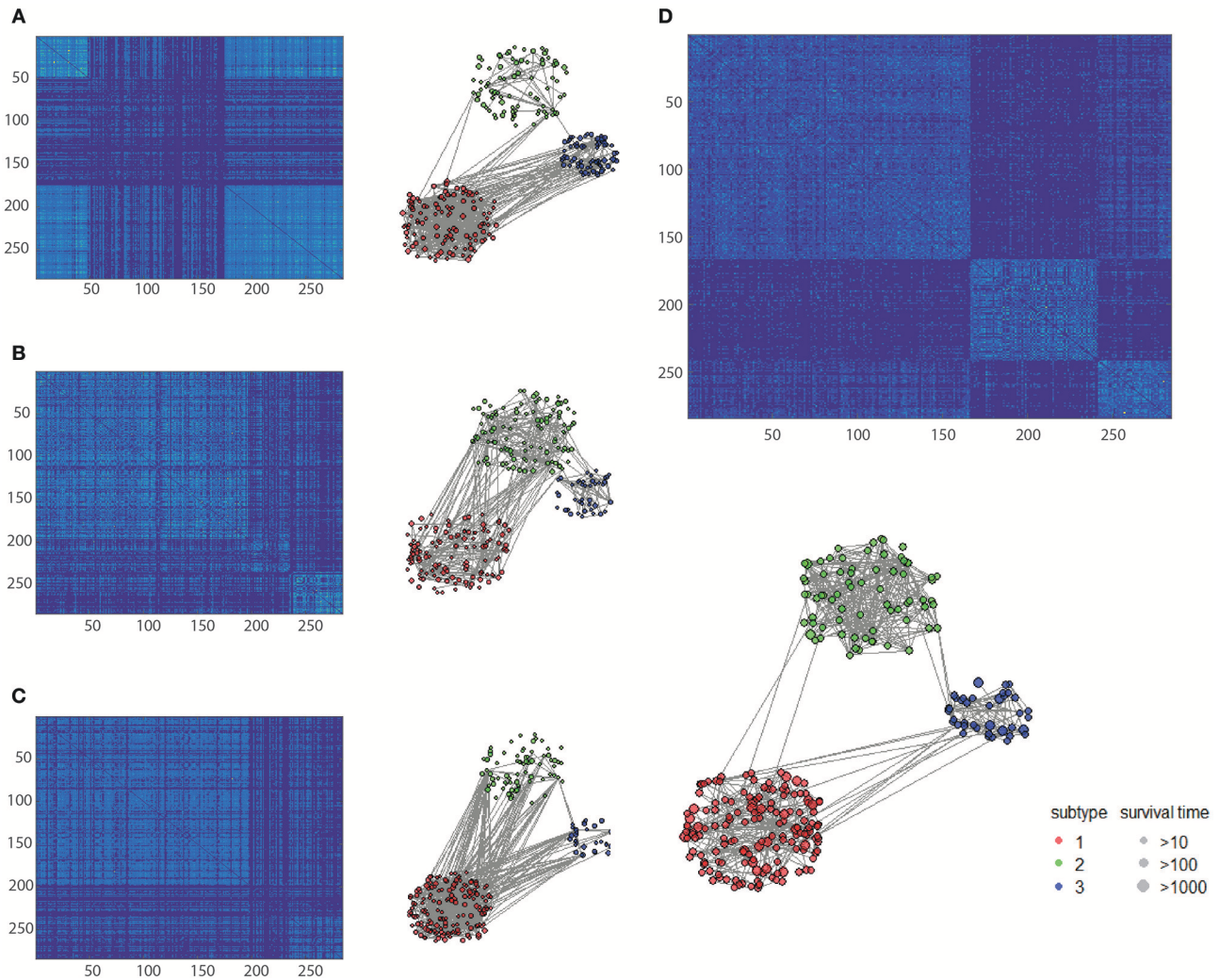

FIGURE 7 | Patient similarity networks for 284 CESC samples of each single data type independently compared to the fused similarity network. Heatmaps on similarity for **(A)** gene expression, **(B)** methylation, **(C)** miRNA expression, and **(D)** integrated matrix are shown.

We calculated the classification accuracy on the collected tumor vs. normal samples. **Table 2** shows the classification performance either by one single set of data or by the fused methods, in which the most highest accuracy were highlighted in bold. The results reflect that, at the single data level, miRNA with the smallest number of measurements showed the best performance regarding sample classification while methylation showed the worst performance. On average, the performance on fused data derived by HOPES and SNF is uniformly better than that for a single source. The good performance of data fusion is attributed to its capability of resisting erroneous correlations or even negative effects, which not only enhances accuracy but also generates more stable results.

Nevertheless, integration methods such as iCluster+ which splices all of the features, strongly rely on a priori gene selection; therefore, if the number of variables is imbalanced, it will be difficult to retain positive information. Thus, the classification accuracy falls in between the worst and best of single level analyse, so as for moCluster. The Clusternomics extract the global assignment based on the mixture of local partitions, so if clustering results were obscure in single data layer the global performance can not be satisfied. The sample size also influence the performance of Clusternomics a lot. We take an example of KIRC dataset for further analysis. One can see that the fused data clustered tightly and uniformly, as shown by the heatmap of the similarity matrix (**Figure 5**). One can see that the clustering result by the proposed HOPES achieved superior performance (**Figure 5D**) to that by each single source (**Figures 5A–C**). In **Figure 5D** shows distinct boundaries between different clusters and uniform structure within each cluster. The fused similarity between healthy samples is far greater than cancer samples, which demonstrates the heterogeneity of cancer. We also created a Venn diagram to examine the sample assignment by each single source or by the fused one. We found that the fused data by HOPES

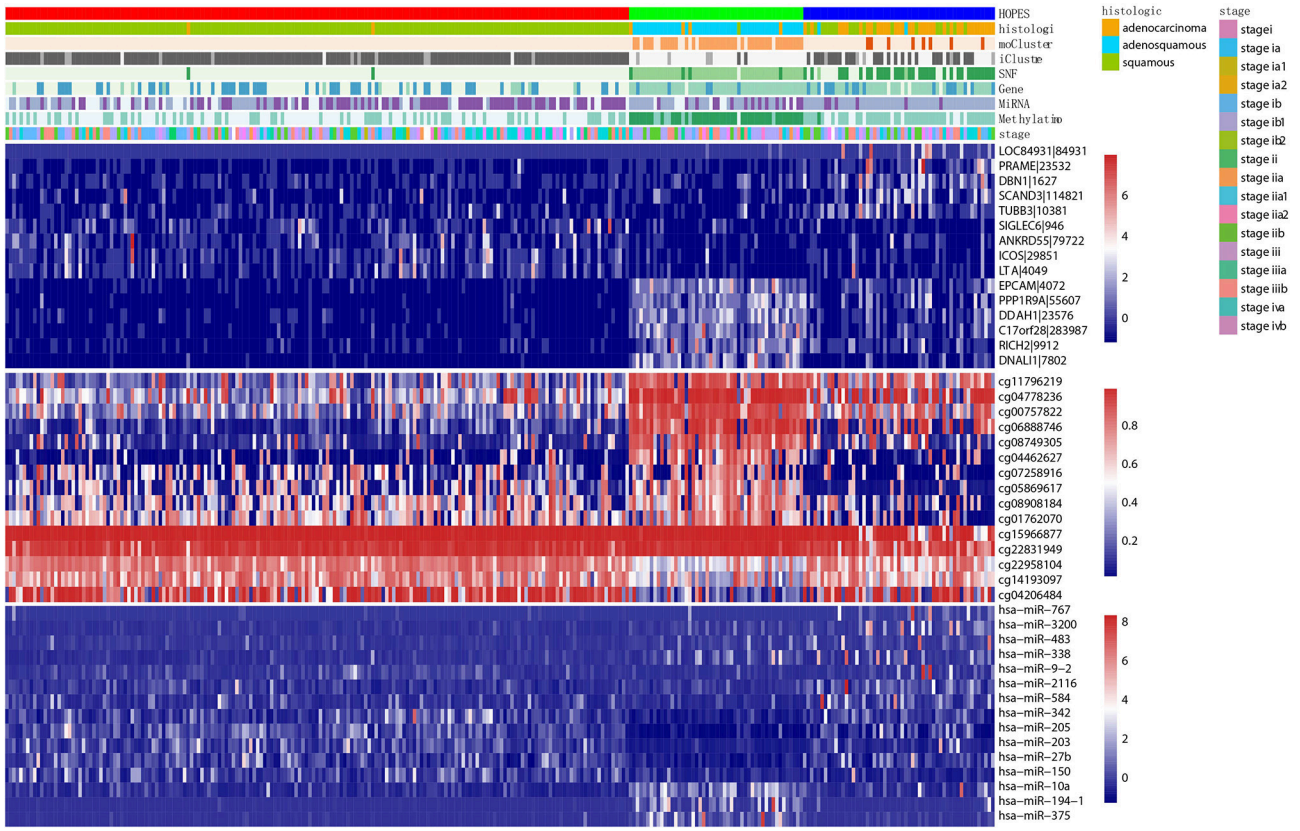

FIGURE 8 | Multiplatform integrative clustering of cervical cancer. Clustering of 284 cervical cancer samples was performed based on different datasets (i.e., gene expression, methylation, miRNA, and fused data) using different methods(HOPES, SNF, iCluster+, moCluster). Moreover, the histological type and clinical stage of patients are also indicated in the legends. The heatmaps show the selected top 15 mRNAs, CpG sites, and miRNAs that are either significantly associated with HOPES groups or have been identified as markers in previous analyses; different datasets correspond to distinct color scale bars.

are robust to mistakes in each single source. More precisely, for 65% (102 of 157) of samples, there were incorrect assignments in at least one single data type analysis, while for 33% (53 of 157) of cases, the classification results were wrong in at least two single data types. However, only 7.6% (12 of 157) of cases were mis-assigned by our method (**Figure 5E**).

3.3. Prognostic Performance on Actual Cancer Datasets

To illustrate the prognostic ability of the elucidated similarity, we applied HOPES to five tumor omics datasets, namely CESC, GBM, COAD, KIRC, and LUSC. The similarities obtained by SNF and HOPES were used to cluster each tumor sample into three subtypes. Their corresponding survival curves were drawn and quantified by the log-rank test. The statistical significance of differences between them was denoted by the *P*-value. To facilitate visual comparisons, the results on both the survival curves and the first three principal components are shown in **Figure 6** and **Supplementary Figure 3**. The survival curves resulting from HOPES can be observed to achieve the smallest *P*-value, highlighted in bold in **Table 3**. Consistent with the results in synetic experiments, HOPES show the most clinical significant and reliable performance in all datasets. Since COAD only contains 92 samples with more than 20,000 gene features, the Clusternomics can not fit a mixture model for COAD.

To clarify the beneficial characteristics of the similarity elucidated by HOPES, we took another example of CESC for further analysis. We compared the clustering results on each single type of omics data alone with those for the elucidated one. The results are plotted in a heatmap as shown in **Figure 7**. Notably, it is difficult to cluster each single type of omics data into sub-clusters. There are no legible block structures in **Figure 7A**, or only tiny sub-clusters in **Figures 7B,C**. Between different clusters, the cross section shows small differences in color, implying that the differences were negligible. In comparison, the clustering results after HOPES were shown to feature three distinct sub-clusters. The last sub-cluster in the bottom-right corner exhibits a fairly homogeneous color within the clusters. Moreover, we can deduce that there are two clusters, upon clustering by gene expression, as shown in **Figure 7A**. There are no obvious sub-clusters either by methylation level (**Figure 7B**) or by miRNA expression (**Figure 7C**). In comparison, the clustering results after HOPES were shown to feature three distinct sub-clusters. The last sub-cluster in the bottom-right corner exhibits a fairly homogeneous color within the clusters. The elucidated similarity makes it markedly easy to find

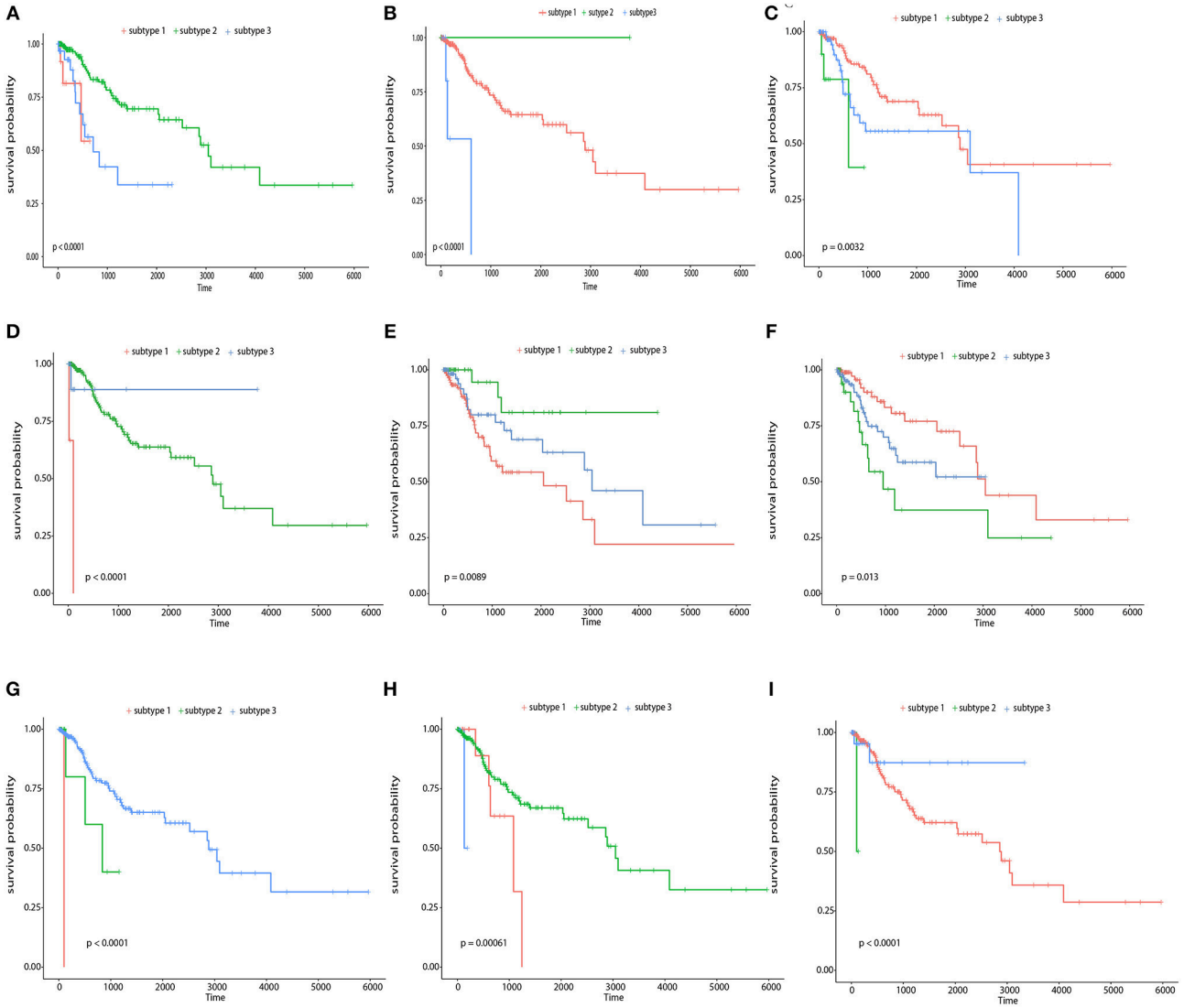

FIGURE 9 | Survival curves of top 3 most significant features in genes, CpG sites, and miRNAs respectively The Kaplan-Meier survival curves for the top 3 most significant genes, namely, LOC84931, TUBB3, and DBN1 **(A–C)**, the top 3 most significant CpG sites, namely, cg1596687, cg07258916, and cg11796219 **(D–F)**, and the top 3 most significant miRNAs, namely, hsa-miR-767, hsa-miR-483, and hsa-miR-9-2 **(G–I)**.

sub-clusters that were concealed in the analyses for each type of omics data alone.

We also found that the elucidated similarity highlights the molecular heterogeneity in cervical carcinomas. The subtyping by HOPES differed depending on the histological classification, showing a discrepancy between phenotype and gene-level types. For instance, the sub-clusters by HOPES largely corresponded to those by methylation level. The CESC project classified samples into six subgroups by histology. To determine the correspondence between the histological classification and HOPES, we merged four different types of adenocarcinoma into one type, as used in studying cervical cancer (Cancer Genome Atlas Research Network et al., 2017). The clusters produced by HOPES strongly correlated with the histological types, but were not the same; our cluster 3 contained all of the adenosquamous cases, while cluster 2 mainly consisted of cervical squamous cell carcinoma samples. We used the χ^2 test to determine whether the two clustering results are significantly associated, and our cluster results showed a strong correlation with each single genomic data cluster, with small P-values (gene expression $P = 1.28 \times 10^{-6}$; methylation $P = 7.94 \times 10^{-9}$; miRNA expression $P = 2.2 \times 10^{-16}$).

3.4. Functional Annotation of Relevant Features Among Cervical Cancer Subtypes

To demonstrate the biological significance of subtype derived by HOPES, we extracted the subset of the most relevant features among the original features and conducted a series of functional analyse on it. We chose the 15 most relevant features in gene expression, methylation, and miRNA data for further analysis.

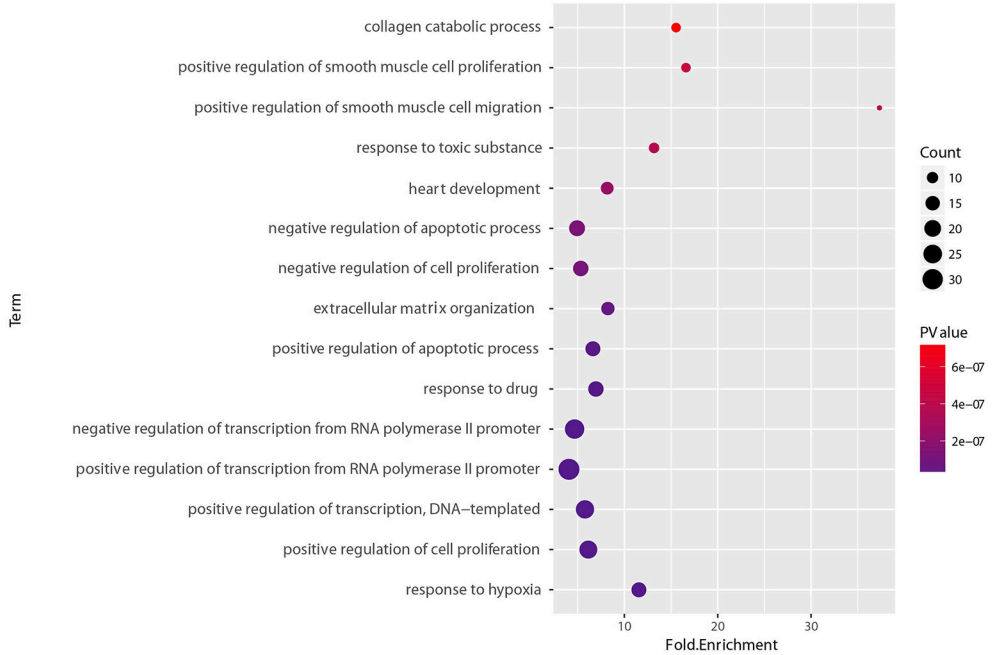

FIGURE 10 | GO biological process enrichment analyses of set of 173 core genes gene count, *p*-value, and fold enrichment of GO biological process terms with *P*-value$< 10^{-6}$.

First, we constructed a corresponding heatmap with different clustering labels, In **Figure 8**, selected signatures of all three data types are merged, showing a clear block form corresponding to the HOPES subgroup. As long as these selected features are differentially expressed following our clustering result, their biological annotation can help us to confirm that the separation created by HOPES is not only clinical meaningful but also biologically significant. In terms of the gene expression pattern, subtype 1 (red), corresponding to lower expression in EPCAM, PPP1R9A, DDAH1, C17orf28, RICH2, and DNALI1, showed a longer survival time, while subtype 2 exhibited completely the opposite performance in the same gene set. The subgroup with the poorest prognosis (blue) significantly corresponded to LOC84931, PRAME, DBN1, SCAND3, and TUBB3 over-expression. The methylation data specifically highlight subgroup 1 in the first five CpG sites(cg11796219, cg04778236, cg00757822, cg06888746, cg08749305); subgroup 2 shows down-regulation in the last three CpG sites (cg22958104, cg14193097, cg04206484); while subgroup 3 is relatively down-regulated in cg07258916, cg05869617, cg15966877, and cg22831949. The heatmap of miRNA shows increased expression of hsa-miR-767, hsa-miR3200, and hsa-miR-483, which correlates with decreased survival probability and clearly up-regulated expression of hsa-miR-10a, hsa-miR-194-1, and hsa-miR-375 in subgroup 2.

Second, we performed survival analysis on each single feature using the kmeans as a general clustering method, and found that more than 1/3 relevant features showed good partition ability with a Log-rank test *p*-value$<$ 0.05 including five genes (LOC84931, DBN1, SCAND3, TUBB3, ICOS), six CpG sites (cg11796219, cg08749305, cg07258916, cg05869617, cg01762070, cg15966877), and six miRNAs (hsa-miR-767, hsa-miR-3200, hsa-miR-483, hsa-miR-9-2, hsa-miR-584, hsa-miR-342). **Figure 9** shows the Kaplan-Meier survival curves of the top 3 most significant features in genes, CpG sites, and miRNAs. Among these genes, DBN1 was detected as a useful oncofetal biomarker (Iyama et al., 2016). It is involved in migration and invasion of glioma, colon, bladder and lung cancer (Mitra et al., 2011; Terakawa et al., 2013; Lin et al., 2014; Zwiener et al., 2014; Xu et al., 2015); TUBB3 was assessed as one of the predictive and prognostic factors in cervical cancer patients under different neoadjuvant regimens (Zwenger et al., 2015). It was also defined to be a useful prognostic biomarker in patients with advanced NSCLS (Li Z. et al., 2014). Moreover, ICOS was also included in one of the genotype combinations (CD28/IFNG/ICOS) that is associated with cervical cancer (Guzman et al., 2008). In analyzing each single CpG site, an R package, "IlluminaHumanMethylation450kanno.ilmn12.hg19" was applied to match each CpG site with reference gene region. The most significant features, included cg22831949, falls in PTPRN2 , which was found to inhibit apoptosis and promote cancer formation in breast cancer (Sorokin et al., 2015); cg07258916 corresponding to PLXNA4 which belongs to the plexin family, and was previously indicated to inhibit tumor cell migration (Balakrishnan et al., 2009); cg11796219 matched with C3orf21, while C3orf21 ablation was proved to promote cell proliferation, inhibite apoptosis and accelerate cell migration in lung cancer. Selected miR-767 contributes to the decrease of TET activity, which is a hallmark of cancer (Loriot et al.,

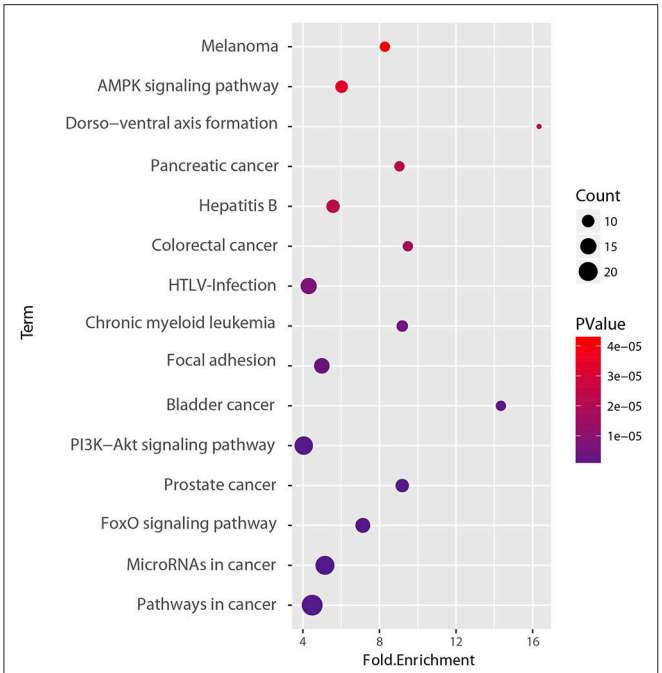

FIGURE 11 | KEGG pathway enrichment analyses of set of 173 core genes gene count, *p*-value, and fold enrichment of KEGG pathways with *P*-value$< 10^{-4}$.

2014). It also known as risky miRNA that significantly correlates with clinical outcomes in GBM (Li R. et al., 2014). Moreover, miR-483 can play the role of an antiapoptotic oncogene in many human cancers, such as Wilms' tumors, colon, liver, and breast cancers (Veronese et al., 2010). It was also identified as predictors of poor prognosis in adrenocortical Cancer (Soon et al., 2009). miR-9 was proved to be correlated with MYCN amplification, tumor grade, and metastatic status (Ma et al., 2010), more specifically, it was found to be associated with clear cell renal cell carcinoma, breast cancer, gastric carcinoma, and brain tumors (Lehmann et al., 2008; Luo et al., 2009; Nass et al., 2009; Hildebrandt et al., 2010).

To determine the functional relevance of the selected features, the identified genes, target genes of CpG sites and miRNAs were merged as a core set. We then performed the GO enrichment analysis (Ashburner et al., 2000) and KEGG pathway analysis (Kanehisa et al., 2011) on it using DAVID tools (Huang et al., 2008, 2009). The genes targeted by miRNAs were predicted by miRTarBase, an experimentally validated miRNA-target interaction database (Chou et al., 2017). We only used the interactions supported by strong experimental evidence (reporter assay or western blot). Finally, the core gene set included 173 genes consisting of 15 original genes, 15 methylation related genes, and 143 miRNA targets. We found that the whole core gene set was enriched in 56 GO biological process terms, with Benjamini-corrected *p*-value $<$ 0.05. **Figure 10** depicts GO terms with *p*-value $<$ 10^{-6}, notably, these significant terms strongly correlate with cancer. An example of this is the most significant term, namely respond to hypoxia. Numerous research has confirmed that pathological hypoxia plays a pivotal role in cancer progression and migration (Muz et al., 2015). In addition, the Hypoxia-inducible factor 1α, which regulates genes involved in response to hypoxia was proved as a strong prognostic marker in early stage cervical cancer (Birner et al., 2000). The regulation of cell proliferation, regulation of transcription from RNA polymerase II promoter, and regulation of apoptotic process participate in the full life-cycle of tumors (Takeshima et al., 2009; Vander Heiden et al., 2009; Wong, 2011). For KEGG analysis, a total of 46 pathways (Benjamini-corrected *p*-value $<$ 0.05) were identified, **Figure 11** shows pathways with *p*-value $< 10^{-4}$. Among these pathways, cancer was the most common subclass such as pathways in cancer, microRNAs in cancer, Bladder cancer, colorectal cancer and pancreatic cancer. Besides direct cancer pathways, the PI3K-AKT-FoxO signaling cascade was identified, which has been previously identified to be involved in cancer and aging (Zhang et al., 2011). The PI3K/Akt signaling pathway leads to the inhibition of the downstream targets FoXO transcription factors, while FoXO is associated with cell cycle progression (Medema et al., 2000), apoptosis (Urbich et al., 2005), and angiogenesis (Tang and Lasky, 2003). There is another research revealed that the activation of AMPK impedes cervical cancer cell growth through this PI3K-AKT-FoxO axis (Yung et al., 2013).

In conclusion, we performed survival analysis, GO enrichment analysis, and KEGG pathway analysis on a subset of the most relevant features of gene expression, methylation and miRNAs corresponding to our HOPES subgroups. We found that these selected features were of great significance in cancer clinical outcomes and biological function such as cancer cell proliferation, apoptosis, and angiogenesis. These findings not only demonstrate the biological meaning of our integrated clustering results, but also indicate that HOPES can act as the anterior work for prognostic biomarker detection.

4. DISCUSSION

The integrated analysis of multi-omics data can facilitate the study of molecular events at different periods of cancer progression and development, and complementary information can remove the effect of noise, leading to precise and useful classification results. Our proposed HOPES method integrates the similarity of different data layers to overcome the dimension and scale heterogeneity that hinders latent variable-based methods. The progressive fusion model based on high-order path similarity can evaluate the strength of single data level specificity and global level consistency together for a consistent and highly representative global similarity. The derived global similarity can filter erroneous or single level specific ties. This procedure can solve the issue of inducing too much noise or distortions by partial structures in a single data set, when we integrate all of the similarity information from each data type. Downstream consensus spectral clustering contributes to the obtainment of reliable clustering results.

In practice, our method shows superior capabilities in distinguishing global patterns through multiple source data. In addition, HOPES show great robustness compared to the other methods which are constrained by sample size or priori feature selection. Since HOPES only used the sample similarity information, its performance is independent of the data source, so it is promising for general usage. The fused similarity matrix shows the higher accuracy of tumor classification than any single data type or other integration methods. Moreover, the clustering results of cancer patients feature significant separation regarding a prognostic indicator (survival time), which can contribute to cancer subtyping at the molecular level and further clinical treatment. The obtained subgroups are also shown to be promising for the identification of potential biomarkers by revealing the key components that drive the differences between subgroups. The enrichment analysis on the key components confirmed the power of HOPES in discriminating the biomarkers.

DATA AVAILABILITY

The CESC dataset generated during and analyzed during the current study are available in the Broad Institute TCGA Genome Data Analysis Center with identifier "https://doi.org/10.7908/C11G0KM9" (Broad Institute TCGA Genome Data Analysis Center, 2016). The BRCA, LUSC, COAD, and GBM datasets that support the findings of this study are provided by Wang et al. (2014). The Code used in this publication is freely available at github.com/scutbioinformatics/HOPES.

AUTHOR CONTRIBUTIONS

AX and HC conceived, designed, and supervised all phases of the project. AX performed experiments and wrote the manuscript. AX and JC performed the bioinformatics analysis. JC, HP, and GH contributed to discussions, and editing of the paper. All authors read and approved the final manuscript.

FUNDING

This work is partially supported by the National Natural Science Foundation of China (61472145, 61372141, 61771007), Science and Technology Planning Project of Guangdong Province (2016A010101013, 2017B020226004), Applied Science and Technology Research and Development Project of Guangdong Province (2016B010127003), Guangdong Natural Science Foundation (2017A030312008), the Fundamental Research Fund for the Central Universities (2017ZD051) and Health Medical Collaborative Innovation Project of Guangzhou City (201803010021).

ACKNOWLEDGMENTS

We thank Liwen Bianji, Edanz Group China, for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00236/full#supplementary-material

REFERENCES

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

Balakrishnan, A., Penachioni, J. Y., Lamba, S., Bleeker, F. E., Zanon, C., Rodolfo, M., et al. (2009). Molecular profiling of the "plexinome" in melanoma and pancreatic cancer. *Hum. Mutat.* 30, 1167–1174. doi: 10.1002/humu.21017

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Birner, P., Schindl, M., Obermair, A., Plank, C., Breitenecker, G., and Oberhuber, G. (2000). Overexpression of hypoxia-inducible factor 1α is a marker for an unfavorable prognosis in early-stage invasive cervical cancer. *Cancer Res.* 60, 4693–4696.

Broad Institute TCGA Genome Data Analysis Center (2016). *Firehose Stddata__2016_01_28 run*. Cambridge, MA: Broad Institute of MIT and Harvard. doi: 10.7908/C11G0KM9

Cai, D., Zhang, C., and He, X. (2010). "Unsupervised feature selection for multicluster data," in *Proceedings of the 16th ACMSIGKDD International Conference on knowledge Discovery and Data Mining (ACM)* (Washington, DC), 333–342.

Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature* 543, 378–384. doi: 10.1038/nature21386

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucl. Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Conway, K., Edmiston, S. N., Tse, C. K., Bryant, C., Kuan, P. F., Hair, B. Y., et al. (2015). Racial variation in breast tumor promoter methylation in the Carolina Breast Cancer Study. *Cancer Epidemiol. Prevent. Biomark.* 24, 921–930. doi: 10.1158/1055-9965.EPI-14-1228

Dai, X., Erkkilä, T., Yli-Harja, O., and Lähdesmäki, H. (2009). A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data. *BMC Bioinformatics* 10:165. doi: 10.1186/1471-2105-10-165

Forbes, S., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., et al. (2008). The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* Unit–10.11. doi: 10.1002/0471142905.hg1011s57

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., et al. (2010). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res.* 39(Suppl. 1):D945–D950. doi: 10.1093/nar/gkq929

Gabasova, E., Reid, J., and Wernisch, L. (2017). Clusternomics: integrative context-dependent clustering for heterogeneous datasets. *PLoS Comput. Biol.* 13:e1005781. doi: 10.1371/journal.pcbi.1005781

Guzman, V. B., Yambartsev, A., Goncalves-Primo, A., Silva, I. D., Carvalho, C. R., Ribalta, J. C., et al. (2008). New approach reveals CD28 and IFNG gene interaction in the susceptibility to cervical cancer. *Hum. Mol. Genet.* 17, 1838–1844. doi: 10.1093/hmg/ddn077

Hildebrandt, M., Gu, J., Lin, J., Ye, Y., Tan, W., Tamboli, P., et al. (2010). Hsa-miR-9 methylation status is associated with cancer development and metastatic recurrence in patients with clear cell renal cell carcinoma. *Oncogene* 29, 5724–5728. doi: 10.1038/onc.2010.305

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084

International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Iyama, S., Ono, M., Kawai Nakahara, H., Husni, R. E., Dai, T., Shiozawa, T., et al. (2016). Drebrin: a new oncofetal biomarker associated with prognosis of lung adenocarcinoma. *Lung Cancer* 102, 74–81. doi: 10.1016/j.lungcan.2016.10.013

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucl. Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen Dale, A. L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Lehmann, U., Hasemeier, B., Christgen, M., Müller, M., Römermann, D., Länger, F., et al. (2008). Epigenetic inactivation of microRNA gene hsa-mir-9-1 in human breast cancer. *J. Pathol.* 214, 17–24. doi: 10.1002/path.2251

Li, R., Gao, K., Luo, H., Wang, X., Shi, Y., Dong, Q., et al. (2014). Identification of intrinsic subtype-specific prognostic microRNAs in primary glioblastoma. *J. Exp. Clin. Cancer Res.* 33:9. doi: 10.1186/1756-9966-33-9

Li, Z., Qing, Y., Guan, W., Li, M., Peng, Y., Zhang, S., et al. (2014). Predictive value of APE1, BRCA1, ERCC1 and TUBB3 expression in patients with advanced non-small cell lung cancer (NSCLC) receiving first-line platinum–paclitaxel chemotherapy. *Cancer Chemother. Pharmacol.* 74, 777–786. doi: 10.1007/s00280-014-2562-1

Lin, Q., Tan, H. T., Lim, T. K., Khoo, A., Lim, K. H., and Chung, M. C. (2014). iTRAQ analysis of colorectal cancer cell lines suggests drebrin (DBN1) is overexpressed during liver metastasis. *Proteomics* 14, 1434–1443. doi: 10.1002/pmic.201300462

Loriot, A., Van Tongelen, A., Blanco, J., Klaessens, S., Cannuyer, J., van Baren, N., et al. (2014). A novel cancer-germline transcript carrying pro-metastatic miR-105 and tet-targeting miR-767 induced by dna hypomethylation in tumors. *Epigenetics* 9, 1163–1171. doi: 10.4161/epi.29628

Luo, H., Zhang, H., Zhang, Z., Zhang, X., Ning, B., Guo, J., et al. (2009). Down-regulated miR-9 and miR-433 in human gastric carcinoma. *J. Exp. Clin. Cancer Res.* 28:82. doi: 10.1186/1756-9966-28-82

Ma, L., Young, J., Prabhala, H., Pan, E., Mestdagh, P., Muth, D., et al. (2010). miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat. Cell Biol.* 12, 247–256. doi: 10.1038/ncb2024

Medema, R. H., Kops, G. J., Bos, J. L., and Burgering, B. M. (2000). AFX-like forkhead transcription factors mediate cell-cycle regulation by Ras and PKB through p27 kip1. *Nature* 404, 782–787. doi: 10.1038/35008115

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2015). moCluster: identifying joint patterns across multiple omics data sets. *J. Proteome Res.* 15, 755–765. doi: 10.1021/acs.jproteome.5b00824

Mitra, R., Lee, J., Jo, J., Milani, M., McClintick, J. N., Edenberg, H. J., et al. (2011). Prediction of postoperative recurrence-free survival in non-small cell lung cancer by using an internationally validated gene expression model. *Clin. Cancer Res.* 17, 2934–2946. doi: 10.1158/1078-0432.CCR-10-1803

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4245–4250. doi: 10.1073/pnas.1208949110

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* 52, 91–118. doi: 10.1023/A:1023949509487

Muz, B., de la Puente, P., Azab, F., and Azab, A. K. (2015). The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* 3, 83–92. doi: 10.2147/HP.S93413

Nakagawa, T., Kollmeyer, T. M., Morlan, B. W., Anderson, S. K., Bergstralh, E. J., Davis, B. J., et al. (2008). A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PLoS ONE* 3:e2318. doi: 10.1371/journal.pone.0002318

Nass, D., Rosenwald, S., Meiri, E., Gilad, S., Tabibian-Keissar, H., Schlosberg, A., et al. (2009). MiR-92b and miR-9/9* are specifically expressed in brain primary tumors and can be used to differentiate primary from metastatic brain tumors. *Brain Pathol.* 19, 375–383. doi: 10.1111/j.1750-3639.2008.00184.x

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucl. Acids Res.* 46, 10546–10562. doi: 10.1093/nar/gky889

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., and Chen, L. (2017). Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* 33, 2706–2714. doi: 10.1093/bioinformatics/btx176

Shimomura, A., Shiino, S., Kawauchi, J., Takizawa, S., Sakamoto, H., Matsuzaki, J., et al. (2016). Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci.* 107, 326–334. doi: 10.1111/cas.12880

Soon, P. S. H., Tacon, L. J., Gill, A. J., Bambach, C. P., Sywak, M. S., Campbell, P. R., et al. (2009). miR-195 and miR-483-5p identified as predictors of poor prognosis in adrenocortical cancer. *Clin. Cancer Res.* 15, 7684–7692. doi: 10.1158/1078-0432.CCR-09-1587

Sorokin, A. V., Nair, B. C., Wei, Y., Aziz, K. E., Evdokimova, V., Hung, M.-C., et al. (2015). Aberrant expression of proPTPRN2 in cancer cells confers resistance to apoptosis. *Cancer Res.* 75, 1846–1858. doi: 10.1158/0008-5472.CAN-14-2718

Takeshima, H., Yamashita, S., Shimazu, T., Niwa, T., and Ushijima, T. (2009). The presence of RNA polymerase ii, active or stalled, predicts epigenetic fate of promoter CpG islands. *Genome Res.* 19, 1974–1982. doi: 10.1101/gr.093310.109

Tang, T. T. L., and Lasky, L. A. (2003). The forkhead transcription factor FOXO4 induces the down-regulation of hypoxia-inducible factor 1 alpha by a von Hippel-Lindau protein-independent mechanism. *J. Biol. Chem.* 278, 30125–30135. doi: 10.1074/jbc.M302042200

Terakawa, Y., Agnihotri, S., Golbourn, B., Nadi, M., Sabha, N., Smith, C. A., et al. (2013). The role of drebrin in glioma migration and invasion. *Exp. Cell Res.* 319, 517–528. doi: 10.1016/j.yexcr.2012.11.008

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* 17, 520–525. doi: 10.1093/bioinformatics/17.6.520

Urbich, C., Knau, A., Fichtlscherer, S., Walter, D. H., Brühl, T., Potente, M., et al. (2005). FOXO-dependent expression of the proapoptotic protein bim: pivotal role for apoptosis signaling in endothelial progenitor cells. *FASEB J.* 19, 974–976. doi: 10.1096/fj.04-2727fje

Vander Heiden, M. G., Cantley, L. C., and Thompson, C. B. (2009). Understanding the warburg effect: the metabolic requirements of cell proliferation. *Science* 324, 1029–1033. doi: 10.1126/science.1160809

Veronese, A., Lupini, L., Consiglio, J., Visone, R., Ferracin, M., Fornari, F., et al. (2010). Oncogenic role of miR-483-3p at the IGF2/483 locus. *Cancer Res.* 70, 3140–3149. doi: 10.1158/0008-5472.CAN-09-4456

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Wong, R. S. (2011). Apoptosis in cancer: from pathogenesis to treatment. *J. Exp. Clin. Cancer Res.* 30:87. doi: 10.1186/1756-9966-30-87

Wu, P. Y., Cheng, C. W., Kaddi, C. D., Venugopalan, J., Hoffman, R., and Wang, M. D. (2017). –Omic and electronic health record big data analytics for precision medicine. *IEEE Trans. Biomed. Eng.* 64, 263–273. doi: 10.1109/TBME.2016.2573285

Xu, S. Q., Buraschi, S., Morcavallo, A., Genua, M., Shirao, T., Peiper, S. C., et al. (2015). A novel role for drebrin in regulating progranulin bioactivity in bladder cancer. *Oncotarget* 6, 10825–10839. doi: 10.18632/oncotarget.3424

Yung, M. M. H., Chan, D. W., Liu, V. W. S., Yao, K. M., and Ngan, H. Y. S. (2013). Activation of AMPK inhibits cervical cancer cell growth through AKT/FOXO3a/FOXM1 signaling cascade. *BMC Cancer* 13:327. doi: 10.1186/1471-2407-13-327

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucl. Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

Zhang, X., Tang, N., Hadden, T. J., and Rishi, A. K. (2011). Akt, FoxO and regulation of apoptosis. *Biochim. Biophys. Acta Mol. Cell Res.* 1813, 1978–1986. doi: 10.1016/j.bbamcr.2011.03.010

Zwenger, A. O., Grosman, G., Iturbe, J., Leone, J., Vallejo, C. T., Leone, J. P., et al. (2015). Expression of ERCC1 and TUBB3 in locally advanced cervical squamous cell cancer and its correlation with different therapeutic regimens. *Int. J. Biol. Mark.* 30, 301–314. doi: 10.5301/jbm.5000161

Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming rna-seq data to improve the performance of prognostic gene signatures. *PLoS ONE* 9:e85150. doi: 10.1371/journal.pone.0085150

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

frontiers in Genetics

TECHNOLOGY REPORT
published: 02 April 2019
doi: 10.3389/fgene.2019.00282

BayesPI-BAR2: A New Python Package for Predicting Functional Non-coding Mutations in Cancer Patient Cohorts

*Kirill Batmanov[1], Jan Delabie[2] and Junbai Wang[1]**

[1] Department of Pathology, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway, [2] Department of Pathology, University Health Network, Toronto, ON, Canada

Most of somatic mutations in cancer occur outside of gene coding regions. These mutations may disrupt the gene regulation by affecting protein-DNA interaction. A study of these disruptions is important in understanding tumorigenesis. However, current computational tools process DNA sequence variants individually, when predicting the effect on protein-DNA binding. Thus, it is a daunting task to identify functional regulatory disturbances among thousands of mutations in a patient. Previously, we have reported and validated a pipeline for identifying functional non-coding somatic mutations in cancer patient cohorts, by integrating diverse information such as gene expression, spatial distribution of the mutations, and a biophysical model for estimating protein binding affinity. Here, we present a new user-friendly Python package BayesPI-BAR2 based on the proposed pipeline for integrative whole-genome sequence analysis. This may be the first prediction package that considers information from both multiple mutations and multiple patients. It is evaluated in follicular lymphoma and skin cancer patients, by focusing on sequence variants in gene promoter regions. BayesPI-BAR2 is a useful tool for predicting functional non-coding mutations in whole genome sequencing data: it allows identification of novel transcription factors (TFs) whose binding is altered by non-coding mutations in cancer. BayesPI-BAR2 program can analyze multiple datasets of genome-wide mutations at once and generate concise, easily interpretable reports for potentially affected gene regulatory sites. The package is freely available at http://folk.uio.no/junbaiw/BayesPI-BAR2/.

Keywords: gene regulation, transcription factors, cancer, bioinformatics, non-coding mutations

OPEN ACCESS

Edited by:
Marko Djordjevic, University of Belgrade, Serbia

Reviewed by:
Dusanka Savic Pavicevic, University of Belgrade, Serbia
Martin Taylor, The University of Edinburgh, United Kingdom
Philipp Bucher, École Polytechnique Fédérale de Lausanne, Switzerland

****Correspondence:***
Junbai Wang
junbai.wang@rr-research.no

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *15 October 2018*
Accepted: *15 March 2019*
Published: *02 April 2019*

Citation:
Batmanov K, Delabie J and Wang J (2019) BayesPI-BAR2: A New Python Package for Predicting Functional Non-coding Mutations in Cancer Patient Cohorts. Front. Genet. 10:282. doi: 10.3389/fgene.2019.00282

INTRODUCTION

Somatic mutations are the primary cause of cancer. Although most studies of cancer genomes to date have focused on mutations occurring within exons, recent efforts have made whole genome sequences of paired tumor and normal samples widely available, facilitating the analysis of non-coding variants in cancer. In many cases, such variants have been shown to affect gene expression

Abbreviations: BayesPI-BAR, Bayesian modeling of Protein-DNA Interaction and Binding Affinity Ranking; FL, follicular lymphoma; PWM, position weight matrix; SNV, single nucleotide variant; TF, transcription factor.

and to promote tumorigenesis (Khurana et al., 2016). One mechanism by which non-coding variants can affect gene expression is the alteration of TF binding to mutated DNA sequences. For example, a mutation may disrupt a TF binding site, preventing the TF from recognizing its target sequence, or a new binding site may be created by a mutation. Several computational tools are available to predict such effects, e.g., GERV (Zeng et al., 2016), atSNP (Zuo et al., 2015), BayesPI-BAR (Wang and Batmanov, 2015), among others. All these tools have the same mode of operation: given a mutation, typically a SNV, and a set of TF-DNA binding models, they produce a list of TFs whose binding is possibly affected by the SNV, ordered by the effect size and/or certainty. However, the predicted list may contain dozens of TFs for every SNV. Adding to the complexity of issue, each cancer sample may have thousands of SNVs, which makes it difficult to interpret the results. Importantly, there is no software package available today to perform such analysis for a patient cohort based on genome-wide sequencing data, considering recurring effects of mutations among several patients.

The BayesPI-BAR2 package presented here aims to solve these problems. It ranks TFs affected by SNV through a new BayesPI-BAR algorithm (Batmanov et al., 2017), augmented with a set of tools to find mutation hotspots among patients and mutations linked to differentially expressed genes. The pipeline collects information about SNVs of all patients in the mutation hotspot regions, and then evaluates the significance of predicted effects against randomly generated background mutation models. The methodology behind BayesPI-BAR2 package and the robustness of predictions were validated in a previous study (Batmanov et al., 2017). Now, a user-friendly Python package is developed based on the proposed pipeline. The package is evaluated in both FL and skin cancer patients, by using mutations called from the whole genome sequencing experiments. BayesPI-BAR2 may reveal novel regulatory sites that are disrupted by mutations in cancer or other diseases, by using genome-wide sequencing data, which is similar to the findings in Weinhold et al. (2014). Additionally, it can identify novel TFs whose binding is altered by non-coding mutations in the genome (Batmanov et al., 2017). It is useful not only for regulatory mutation study in cancer, but also for similar research in other diseases.

MATERIALS AND METHODS

Overview of BayesPI-BAR2 Python Package

The operation of the BayesPI-BAR2 pipeline is illustrated in **Figure 1**. It is motivated by works in Batmanov et al. (2017) where novel mutations affecting gene regulation were discovered in FL patients, by considering diverse genome information. The original analysis pipeline comprised of various scripts that were implemented in different programming languages. Here, a completely new Python package was built with enhanced functionality and user-friendly command line options. Particularly, the old BayesPI-BAR (Wang and Batmanov, 2015) program (a combination of R and Perl programs) was reimplemented in Python with a more efficient algorithm and flexible parallelization. This computationally demanding task can be automatically parallelized now either on a single multi-core machine, or on a cluster supporting the SLURM job queue manager.

BayesPI-BAR2 Python package first finds DNA regions with high mutation density and close to differentially expressed genes, then predicts TF affinity changes in these regions using the new BayesPI-BAR, and finally tests the significance of these predicted changes against a background model. All analysis is carried out by a set of command line tools written in Python 2. The package also includes binary files of the new BayesPI program (Wang and Morigen, 2009) which can infer new TF binding affinity models PWMs such as dinucleotide interdependence (Wang, 2014), DNA shape-restricted dinucleotide models (Batmanov and Wang, 2017), and compute TF-DNA differential binding affinity (dbA) scores (Wang et al., 2015). There is also a demo script in the package that shows a full pipeline execution. BayesPI-BAR2 Python package is a useful tool for identifying functional regulatory mutations in cancers or diseases, based on whole genome sequencing experiments. For a more detailed description of the package, please refer to following sections and (Batmanov et al., 2017).

Identification of Mutation Hot Regions and Patient-Specific Mutation Blocks

In the first step of the BayesPI-BAR2 pipeline, highly mutated DNA sequence (mutation hotspot) regions are identified by a method described in Batmanov et al. (2017), which considers mutations from several patients to define a set of regions. In default setting, BayesPI-BAR2 searches for putative mutation hotspot regions near the transcription start sites (TSS) of differentially expressed genes, because important regulatory sequences (e.g., functional regulatory mutations) are often located in the promoters. To have a robust mutation calling (Alioto et al., 2015) in the promoter region, a minimum sequencing depth of 30 is recommended at this point. The significance of the differential expressions is tested by two-sample Kolmogorov–Smirnov test, where *reads per kilobase of exon model per million mapped reads* (RPKM) values of RNA-seq data of patients are compared to that of the normal samples (e.g., $P < 0.05$). Since RPKM-based differential expression tests may be affected by experimental biases (Bullard et al., 2010) and result in imprecise prediction, a multiple testing correction of *P*-values is not recommended. Nevertheless, by changing the threshold value of the pipeline, it is easy to apply the Bonferroni correction on the *P*-values. Alternatively, user can apply external software to perform the differential gene expression analysis, and directly input the gene list into BayesPI-BAR2 package.

Subsequently, MuSSD (Mutation filtering based on the Space and Sample Distribution) algorithm (Batmanov et al., 2017) is applied on the promoter regions of differentially expressed genes. Based on the identified mutation hotspot regions from MuSSD, patient specific mutation blocks are built: the reference sequence is taken from the reference genome assembly according to the region covered by the mutation hotspot (possibly including

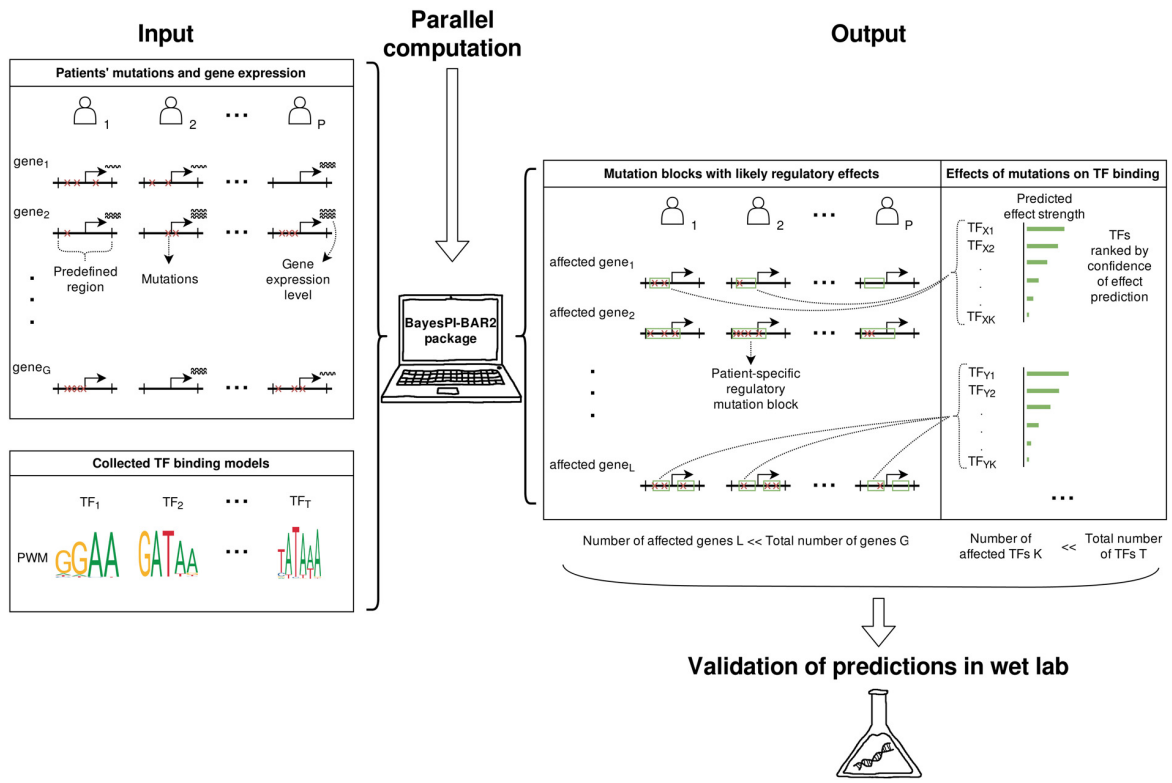

FIGURE 1 | Overview of the BayesPI-BAR2 operation. Using patients' mutation and the corresponding gene expression data, BayesPI-BAR2 finds mutation hotspots and patient-specific mutation blocks. Then, using a given set of PWMs for TFs of interest, it produces a list of significant TFs whose binding is potentially affected at mutation block. The predictions consider all available mutation data among patient samples, to which strict statistical tests are applied to determine the significance of predicted effects. As a result, the number of predicted effects is small enough for performing follow-up wet lab validation.

patient germline variants), and the alternate sequence contains all mutations from the same patient in the region. In BayesPI-BAR2 package, the computational predictions of both the mutation hotspot regions and the patient-specific mutational blocks are implemented in Python, with a more efficient algorithm than the original MATLAB script (Batmanov et al., 2017).

BayesPI TF-DNA Binding Affinity Model

The basic biophysical model for computing TF-DNA binding affinity, named BayesPI, was first reported in Wang and Morigen (2009). The TF-DNA binding probability is derived from the statistical mechanical theory of TF-DNA interactions (Djordjevic et al., 2003; Foat et al., 2006), which can be shown as

$$P(S, w, \mu) = \sum_{i=0}^{N-M} \frac{1}{1 + e^{E_{indep}(S_{i:i+M}, w) - \mu}}$$

where $S_{i,a} = 1$ if the DNA sequence has nucleotide a (one of A, C, G, T) at position i and $S_{i,a} = 1$ otherwise, N is the sequence length, M is the length of the binding motif, μ is the chemical potential of the TF or its concentration in the nucleus. The selection of μ (e.g., $\mu = 0, -10, -13, -15, -18, -20$) is based on a previous study (Wang and Batmanov, 2015) of the effect of DNA sequence variants on TF binding affinity changes, where verified regulatory mutations in human genome were used to infer the dynamical range of chemical potentials.

$$E_{indep}(S, w) = \sum_{j=0}^{M-1} \sum_{a=1}^{4} w_{j,a} S_{j,a}$$

$E_{indep}(S, w)$ is the TF binding energy to a short DNA fragment with length M bp. This model assumes that nucleotides at each binding position contribute to the binding energy independently. The matrix $w \in R^{(M \times 4)}$, called position-specific affinity matrix (PSAM), where $w_{j,a}$ is the binding energy of nucleotide a at position j of the DNA fragment. In BayesPI-BAR2 Python package, a collection of PSAMs derived from a previous published work (Kheradpour and Kellis, 2014) is included, and several new BayesPI features are also added [e.g., PSAM with dinucleotide interdependence (Wang, 2014), and DNA shape-restricted dinucleotide models (Batmanov and Wang, 2017)].

BayesPI-BAR Approach

Bayesian modeling of Protein-DNA Interaction and Binding Affinity Ranking (Wang and Batmanov, 2015) method is used to evaluate the significance of TF binding affinity changes caused by DNA sequence variants. It is based on an idea for distinguishing direct versus indirect TF binding in Wang et al. (2015). A new

quantity, *dbA*, is introduced to measure the binding strength above background level. BayesPI-BAR Python code computes the *shifted differential binding affinity* (δ*dbA*), for each sequence variant and TF:

$$\delta dbA\,(S_{\text{ref}}, S_{\text{alt}}) = dbA\,(S_{\text{alt}}) - dbA\,(S_{\text{ref}})$$

S_{ref}, S_{alt} represent the reference and alternate sequences, respectively. δ*dbA* is the measure of the affinity change used by BayesPI-BAR. More details about the BayesPI-BAR approach are available in the supplementary and (Batmanov et al., 2017).

Significance Testing for TF Binding Affinity Changes

To test the significance of disruption of TF-DNA binding by patient SNVs, patient-specific δdbA values of a given regulatory mutation block are compared to that of the randomly generated background mutation blocks, using the two-sided Rank-sum test. BayesPI-BAR2 has three alternative mutation models to generate the background: a tumor-derived mutation model, a k-mer *mutation signature* such as those available from COSMIC (Tate et al., 2018), and a uniform mutation model. A list of TF binding effects which are significantly stronger than estimated by the background model is exported by BayesPI-BAR2.

Since patient mutation blocks are pre-filtered by MuSSD algorithm based on the space and sample distribution of mutations, there are several constraints on the background mutation blocks: (a) both the size and the mutation counts of the background mutation blocks are kept same as that of patient ones. (b) DNA sequence is selected randomly from the same regions as the patient mutation block. (c) distributions of the mutation positions and the nucleotide changes are based on specific mutation signature such as tumor-derived mutations. To evaluate the relationship between the number of background blocks and the precision of background δ*dbA* model, a few simulations are displayed in **Figure 2**. It shows the fraction of significant TFs reaches a plateau when there are more than 1000 blocks used. The significance test for TF-DNA binding affinity changes proceeds in following three steps:

(1) Background mutation blocks are extracted randomly from regions of interest, with the same sequence length as patient block. Reference sequence of a background mutation block is taken from the reference genome. The alternate sequence is generated by random alteration of nucleotides in reference sequence, using either the tumor-derived mutations or the given k-mer mutation probability distribution (the mutation signature).
(2) For each given TF, BayesPI-BAR computes δ*dbA* of a patient regulatory mutation block. Then, it computes δ*dbA* values for about 2000 background blocks that represent the background distribution of δ*dbA* scores.
(3) Wilcoxon rank-sum test is used to compare the distribution of δ*dbA* values between the patients' and the background mutation blocks. Bonferroni correction of *P*-values is applied.

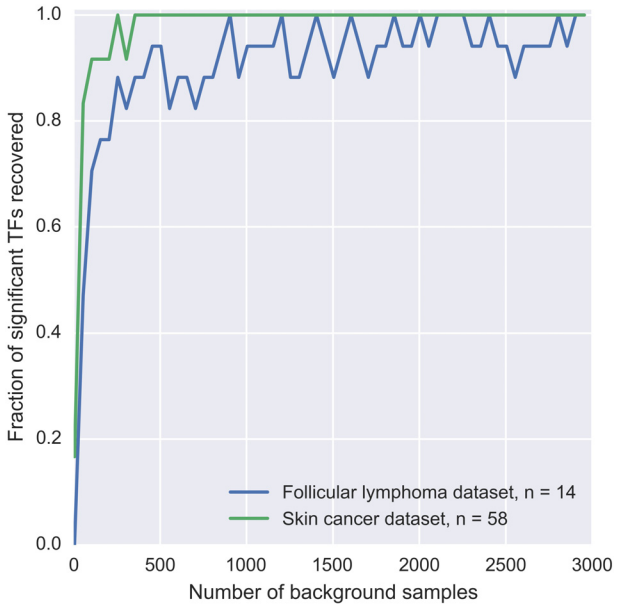

FIGURE 2 | Estimation of sufficient background samples for BayesPI-BAR2 package. The plot displays the dependency of significant TF discovery on the number of background samples used. Significant TFs in the mutation blocks from two different datasets are considered: (1) two *BCL2* blocks from FL dataset with 14 patients affected, blue line; (2) and the *TERT* block from skin cancer dataset with 58 patients affected, green line. On the *X*-axis, we plot the number of background mutation blocks taken. On the *Y*-axis, we plot the number of significant TFs found when using X background mutation blocks, which are *also* significant when using the full set of 10000 background blocks. Y is normalized by the number of significant TFs discovered using the full background set. Therefore, $Y = 1$ corresponds to the same result as using the full background set.

The significance testing considers both the strength of TF binding affinity change and the recurrence of δ*dbA* values across samples, using the Bonferroni correction for the number of TFs tested. A stronger *P*-value correction procedure may not be suitable here. For example, Benjamini–Hochberg (BH) false discovery rate requires the *P*-values to be independent (or have limited dependencies) (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001), but there are strong dependencies among *P*-values of the significance testing for TF binding affinity changes. Often, *P*-values of very similar PWMs are close to each other, which may result in unreliable correction by the BH procedure. Bonferroni correction has no assumptions about the process used to generate the *P*-values which is suited in the current study. At least 10 samples are needed to perform proper statistical test in BayesPI-BAR2. If the sample size is too small, there will be a problem in achieving the statistical significance by Rank-sum test, even if the effects are large (Wild and Seber, 2011).

Algorithm Efficiency and Parallel Computation

Computation of scores is the most time-consuming task that is needed for both the patient and the background mutation blocks.

The old R program (Wang and Batmanov, 2015) was designed to evaluate TF binding affinity changes in a single mutation and was unable to process multiple mutations simultaneously. In the new Python package, a parallel computation paradigm is developed by using more efficient data processing library. Additionally, the efficiency of BayesPI code was improved by applying a new sub-expression for TF binding probability (please refer to BayesPI TF-DNA binding affinity model section):

$$e^{\sum_{j=0}^{M-1}\sum_{a=1}^{4} w_{j,a}S_{j,a}-\mu} = e^{-\mu}\prod_{j=0}^{M-1}\prod_{a=1}^{4}\left(e^{w_{j,a}}\right)^{S_{j,a}}$$

Where the terms $e^{w_{j,a}}$ and $e^{-\mu}$ in the right side of the formula are precomputed and stored in order to avoid computing the exponent term in every sliding window. The new implementation reduces the computational time by about 90%. In addition, in BayesPI-BAR2 Python package, all calculations are parallelized across either multiple local CPUs or multiple nodes on a cluster using the SLURM workload manager. For instance, it takes about 5 h to process all mutation blocks in the skin cancer dataset (263 patients; ∼100000 mutations), by using 8 nodes of 8 CPUs in each. The overall waiting time can be further reduced if more parallel processes are used or few mutation blocks are selected for testing. User guide and package architecture of BayesPI-BAR2 are available in the **Supplementary Section**.

RESULTS

Validating New Python Code in Verified Regulatory Mutations

The precision of the new BayesPI-BAR Python program, which is the basis of BayesPI-BAR2 package, was first assessed by a benchmark dataset of 67 SNVs with experimentally verified effects of TF binding. The results match the previous study (Wang and Batmanov, 2015).

Evaluating the New BayesPI-BAR2 Package in Follicular Lymphoma

A previous analysis of regulatory mutations in FL cancer patients was performed by running various scripts manually. The new BayesPI-BAR2 Python package is applied on the same FL patients, by considering only the gene promoter regions (e.g., TSS ± 1000 bp with 795 called SNVs) as were investigated before (Batmanov et al., 2017). Putative mutation hot blocks near *BCL6*, *BCL2*, and *HIST1H2BM* genes are detected automatically, where containing 34, 40, and 2 SNVs, respectively. The results match with the earlier report (Batmanov et al., 2017). Also, the mutation effects on TF binding at the promoter of two important FL genes (*BCL6* and *BCL2*) (Pasqualucci et al., 2014) were recovered: for example, regulatory activities of two TFs (*FOXD2* and *FOXD3*) on *BCL6* and *BCL2* were confirmed previously by knockdown experiments in SUDHL4 lymphoma cell (Batmanov et al., 2017). The new BayesPI-BAR2 Python package can reproduce the previous results (Batmanov et al., 2017) and is robust in predicting functional regulatory mutations.

Applying BayesPI-BAR2 on Genome-Wide Sequencing Data of Skin Cancer

The somatic mutations and RNA-Seq counts for the skin cancer evaluation were downloaded from the public DCC data release 23 at the International Cancer Genome Consortium (ICGC) data portal, from the MELA-AU, SKCA-BR, and SKCM-US projects. The dataset contains 23 million mutations called from whole genome sequence analysis of 263 patients. Melanoma or skin cancer has the highest prevalence of somatic mutations across human cancer types, which is more than ten times higher than that in Lymphoma cancer (Alexandrov et al., 2013). There are frequent driver coding mutations in melanoma cancer (Hodis et al., 2012; Roberts and Gordenin, 2014). Therefore, DNA regions from 2 Kbp upstream to 100 bp downstream of TSS of protein-coding genes [e.g., GENCODE (Harrow et al., 2012)] were selected, and genes differentially expressed between the patient RNA-Seq data and the normal melanocyte RNA-Seq (Haltaufderhyde and Oancea, 2014) were used in this study (10015 genes with ∼99173 mutations).

After applying BayesPI-BAR2 Python package, 166 putative regulatory mutation blocks were detected (containing 2746 mutations). A list of the 15 most highly mutated blocks is shown in a **Supplementary Table 1**, where blocks matched to previous findings are marked and the corresponding publications are cited. A mutation block near *TERT* gene has the most patients affected, 58 in number, closely followed by blocks near several housekeeping genes (*RPL**, *RPS**, and others). This is in agreement with the previous studies (Weinhold et al., 2014; Poulos et al., 2015). It has been suggested that these mutations are due to vulnerability of some DNA positions to ultraviolet light damage (Fredriksson et al., 2017). In the *TERT* mutation block, significantly affected TFs were also predicted by BayesPI-BAR2 automatically (e.g., Wilcoxon rank-sum test $P < 0.001$ with Bonferroni correction; **Figure 3**), which split into two groups: positive change (creation of binding sites) at the top, in orange; and negative change (destruction of existing binding sites) on the bottom, in blue. The heatmap of **Figure 3** shows the variation of affinity changes among 58 patients, who harbor at least one mutation in the *TERT* block. Nine out of seventeen positively affected TFs belong to the ETS protein family, which are the most significantly affected ones. This is also in agreement with the well-known pathomechanisms of melanoma (Huang et al., 2013). When testing significance of affinity changes against the skin cancer specific mutation signature model and a uniform model, the same significantly affected TFs were found in the *TERT* block, with small differences in the ranking (**Supplementary Figures 1**, **2**).

Additionally, BayesPI-BAR2 discovers novel regulatory mutations which affect gene expression in skin cancer. For instance, binding of TFs from Sp/KLF family and ETS family

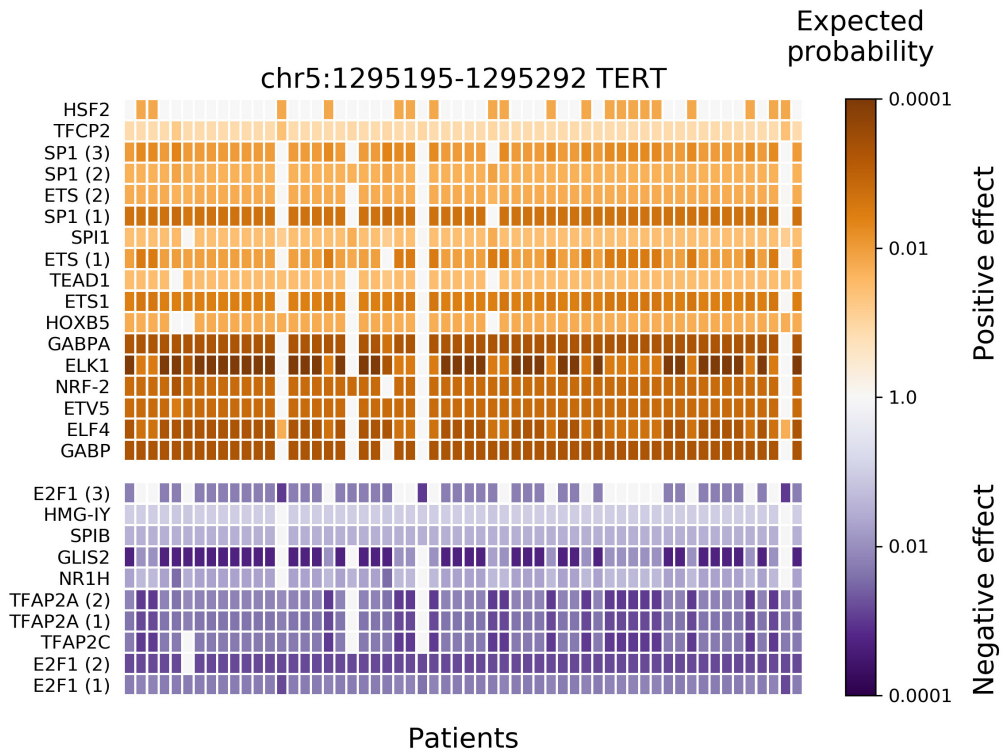

FIGURE 3 | Results of new BayesPI-BAR2 package – TFs significantly affected by skin cancer somatic mutations in the *TERT* promoter mutation block. The heatmap displays the distribution of predicted TF binding effects of *TERT* promoter somatic SNVs across 58 skin cancer patients. The columns represent patients, the rows represent predicted significantly affected TFs, and the color represents the binding effect size. Reduced binding is shown in blue and increased binding in orange. The color shade represents the $\log_{10}$-scaled fraction of background δdbA values which are more extreme than the observed δdbA of a patient, which is an indication of effect size. The darker cells the larger effect. Some TFs are represented by multiple PWM models, their instances are indicated by a number in parentheses. Only significantly affected TFs (Wilcoxon rank-sum test *P*-value < 0.001 after Bonferroni correction) are shown. The following TFs belong to the ETS family (Gutierrez-Hartmann et al., 2007): GABPα, ELF4, ETV5, ELK1, ETS1, SPI1, and SPIB. Here, background mutation model in BayesPI-BAR2 is based on tumor samples.

were found to be disrupted (e.g., about 47 patients with mutations; **Supplementary Table 1**) in a mutation block near RALY. RALY is differentially expressed between the skin cancer patients and the normal control samples. It is an RNA-binding protein that may play a role in pre-mRNA splicing. Based on human phenotype association evidence for RALY from the GWAS Catalog (MacArthur et al., 2017), we found mutations of this gene associated with melanoma, skin pigmentation, and skin sensitivity to sun. The next most frequent mutation block was predicted near RPS27 (e.g., 46 patients with mutations), where binding of TBP, ETS, and IRF TF families are interrupted. RPS27 mutation and its elevated expression have been detected in many melanoma patients and in various human cancers (Dutton-Regester et al., 2014). The two newly discovered regulatory mutation blocks may contribute to the dysregulation of RALY and RPS27 and are worthy for further investigation because both genes are known to be significantly associated with melanoma. Thus, BayesPI-BAR2 not only can automatically recover known gene regulatory disturbance, but also can discover the novel ones which can be tested in wet-lab. BayesPI-BAR2 Python package comes with the code to perform the complete analysis of this melanoma dataset.

DISCUSSION AND CONCLUSION

The new BayesPI-BAR2 Python package has been evaluated in both small (e.g., 14 FL patients) and large (e.g., 263 skin cancer patients) cancer patient cohorts, based on whole genome sequencing experiments. It achieves good prediction accuracy and automatically reproduces the published results. The new package can be used to investigate previously unknown regulatory effects, even if the sample size is small and the recurrent mutation frequency is low. Nevertheless, the robustness of significance test in BayesPI-BAR2 is dependent on the sample size (Biau et al., 2008), a small sample size may pose difficulty in achieving the significance difference. For example, there are 3 mutation blocks from 14 FL patients that pass the test of significant TF binding affinity changes (*P*-values <0.05), but there are 15 mutation blocks from 263 skin cancer samples that pass a more stringent criteria (*P*-values <0.001). Therefore, a large sample size is preferred when using BayesPI-BAR2 to predict putative functional non-coding mutations.

BayesPI-BAR2 approach is more general than a previous mutation recurrence analysis (Weinhold et al., 2014), because it takes into account the recurrence of both the mutation

among multiple patients and the effect on TF binding. In other words, different mutations may contribute to the creation or disruption of the same regulatory link in different patients. For example, there are two canonical highly recurrent mutations in the *TERT* promoter mutations: C > T at chr5:1,295,228 and chr5:1,295,250. Both of these mutations create ETS binging sites. Though six of fifty-eight patients did not have these two mutations, some ETS factors are positively affected in five of them (**Figure 3**). It indicates that other non-canonical mutations at TERT promoter may also create ETS binding sites.

Although BayesPI-BAR2 needs heavy computation to achieve the goal, the waiting time can be significantly reduced by distributing more jobs in a high performance computing system. In the study of 263 skin cancer patients, the total waiting time was reduced to 1 h and 30 min while using 10 nodes of 10 CPUs of ABEL computer cluster at University of Oslo. On average, approximately 6 min are used for completing the calculation of one mutation block. Efficiency of BayesPI-BAR2 can be further improved by applying advanced sampling method and parallel algorithm, or by implementing it in Graphical Processing unit (GPU) (Zou et al., 2018). Alternatively, if more prior information regarding mutation blocks (e.g., differential methylation, nucleosome occupancy, active enhancer/promoter histone markers, and predicted long distance gene regulations) (Wang et al., 2013; Cao et al., 2017; Dhingra et al., 2017) is available, then fewer mutation blocks will be selected for testing against the background models. Thus additional information can also reduce the total computation time significantly. The new features will be implemented in the future.

The new BayesPI-BAR2 Python package allows analysis of non-coding mutations in cancer patient cohorts, discovering mutation hotspots, and predicting effects of these mutations on TF-DNA binding. Unlike previously available tools, it considers the frequency of mutations, their recurrence across patients, and integrates this information with the predicted affinity changes employing a simple and statistically sound approach. Although in principle, it is applicable to any mutation dataset, BayesPI-BAR2 is designed for the typical cancer use case, with the goal to find few non-random effects among many somatic mutations. The package can be a useful tool for in-depth analysis of non-coding mutations detected in whole genome sequencing experiments, as well as for predicting their effects on genome regulation in cancer. All in all, it provides a reasonable number of predictions for further experimental validation.

DATA AVAILABILITY

The package source code, binaries for Linux and OS X, and demo datasets are available at http://folk.uio.no/junbaiw/BayesPI-BAR2/; Project name: BayesPI-BAR2 Package; Operating system(s): Linux and OS X; Programming language: Python; License: General Public License (GNU GPLv3); Any restrictions to use by non-academics: None; The datasets analyzed during the current study are available in the public DCC data release 23 at the ICGC data portal: https://dcc.icgc.org/releases/release_23/Projects.

AUTHOR CONTRIBUTIONS

KB implemented the BayesPI-BAR2 pipeline in Python. JD validated study. JW conceived project, designed BayesPI-BAR2 pipeline, and contributed in developing package. KB and JW drafted manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Norwegian Cancer Society (DNK 2192630-2012-33376, DNK 2192630-2013-33463, and DNK 2192630-2014-33518), South-Eastern Norway Regional Health Authority (HSØ 2017061 and HSØ 2018107), and the Norwegian Research Council NOTUR project (nn4605k).

ACKNOWLEDGMENTS

The authors thank Prof. Magnar Bjørås for proofreading the article and Ms. Amna Farooq for manuscript editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00282/full#supplementary-material

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* 6:10001. doi: 10.1038/ncomms10001

Batmanov, K., and Wang, J. (2017). Predicting variation of DNA shape preferences in protein-DNA interaction in cancer cells with a new biophysical model. *Genes* 8:233. doi: 10.3390/genes8090233

Batmanov, K., Wang, W., Bjoras, M., Delabie, J., and Wang, J. (2017). Integrative whole-genome sequence analysis reveals roles of regulatory mutations in BCL6 and BCL2 in follicular lymphoma. *Sci. Rep.* 7:7040. doi: 10.1038/s41598-017-07226-4

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1186/1471-2105-9-114

Biau, D. J., Kerneis, S., and Porcher, R. (2008). Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin. Orthopaedics Relat. Res.* 466, 2282–2288. doi: 10.1007/s11999-008-0346-9

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Cao, Q., Anyansi, C., Hu, X. H., Xu, L. L., Xiong, L., Tang, W. S., et al. (2017). Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genetics* 49, 1428–1436. doi: 10.1038/ng.3950

Dhingra, P., Martinez-Fundichely, A., Berger, A., Huang, F. W., Forbes, A. N., Liu, E. M., et al. (2017). Identification of novel prostate cancer drivers using RegNetDriver: a framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network. *Genome Biol.* 18:141. doi: 10.1186/s13059-017-1266-3

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res.* 13, 2381–2390. doi: 10.1101/gr.1271603

Dutton-Regester, K., Gartner, J. J., Emmanuel, R., Qutob, N., Davies, M. A., Gershenwald, J. E., et al. (2014). A highly recurrent RPS27 5′ UTR mutation in melanoma. *Oncotarget* 5, 2912–2917. doi: 10.18632/oncotarget.2048

Foat, B. C., Morozov, A. V., and Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22, e141–e149. doi: 10.1093/bioinformatics/btl223

Fredriksson, N. J., Elliott, K., Filges, S., Van Den Eynden, J., Stahlberg, A., and Larsson, E. (2017). Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLoS Genet.* 13:e1006773. doi: 10.1371/journal.pgen.1006773

Gutierrez-Hartmann, A., Duval, D. L., and Bradford, A. P. (2007). ETS transcription factors in endocrine systems. *Trends Endocrinol. Metab.* 18, 150–158. doi: 10.1016/j.tem.2007.03.002

Haltaufderhyde, K. D., and Oancea, E. (2014). Data set for the genome-wide transcriptome analysis of human epidermal melanocytes. *Data Brief* 1, 70–72. doi: 10.1016/j.dib.2014.09.002

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., Theurillat, J. P., et al. (2012). A landscape of driver mutations in melanoma. *Cell* 150, 251–263. doi: 10.1016/j.cell.2012.06.024

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., and Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science* 339, 957–959. doi: 10.1126/science.1229259

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* 42, 2976–2987. doi: 10.1093/nar/gkt1249

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* 17, 93–108. doi: 10.1038/nrg.2015.17

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Pasqualucci, L., Khiabanian, H., Fangazio, M., Vasishtha, M., Messina, M., Holmes, A. B., et al. (2014). Genetics of follicular lymphoma transformation. *Cell Rep.* 6, 130–140. doi: 10.1016/j.celrep.2013.12.027

Poulos, R. C., Thoms, J. A., Shah, A., Beck, D., Pimanda, J. E., and Wong, J. W. (2015). Systematic screening of promoter regions pinpoints functional cis-regulatory mutations in a cutaneous melanoma genome. *Mol. Cancer Res.* 13, 1218–1226. doi: 10.1158/1541-7786.MCR-15-0146

Roberts, S. A., and Gordenin, D. A. (2014). Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* 14, 786–800. doi: 10.1038/nrc3816

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2018). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015

Wang, J. (2014). Quality versus accuracy: result of a reanalysis of protein-binding microarrays from the DREAM5 challenge by using BayesPI2 including dinucleotide interdependence. *BMC Bioinformatics* 15:289. doi: 10.1186/1471-2105-15-289

Wang, J., and Batmanov, K. (2015). BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations. *Nucleic Acids Res.* 43:e147. doi: 10.1093/nar/gkv733

Wang, J., Malecka, A., Trøenand, G., and Delabie, J. (2015). Comprehensive genome-wide transcription factor analysis reveals that a combination of high affinity and low affinity DNA binding is needed for human gene regulation. *BMC Genomics* 16(Suppl. 7):S12. doi: 10.1186/1471-2164-16-S7-S12

Wang, J., and Morigen. (2009). BayesPI - a new model to study protein-DNA interactions: a case study of condition-specific protein binding parameters for Yeast transcription factors. *BMC Bioinformatics* 10:345. doi: 10.1186/1471-2105-10-345

Wang, J. B., Lan, X., Hsu, P. Y., Hsu, H. K., Huang, K., Parvin, J., et al. (2013). Genome-wide analysis uncovers high frequency, strong differential chromosomal interactions and their associated epigenetic patterns in E2-mediated gene regulation. *BMC Genomics* 14:70. doi: 10.1186/1471-2164-14-70

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* 46, 1160–1165. doi: 10.1038/ng.3101

Wild, C., and Seber, G. (2011). "The Wilcoxon rank-sum test," in *Chance Encounters: A First Course in Data Analysis and Inference*, ed. G. Seber (New York, NY: Wiley&Sons).

Zeng, H., Hashimoto, T., Kang, D. D., and Gifford, D. K. (2016). GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* 32, 490–496. doi: 10.1093/bioinformatics/btv565

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2018). A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18. doi: 10.1038/s41588-018-0295-5

Zuo, C., Shin, S., and Keles, S. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* 31, 3353–3355. doi: 10.1093/bioinformatics/btv328

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 May 2019
doi: 10.3389/fgene.2019.00420

Multi-Omic Data Interpretation to Repurpose Subtype Specific Drug Candidates for Breast Cancer

Beste Turanli[1,2,3†], Kubra Karagoz[4†], Gholamreza Bidkhori[2], Raghu Sinha[5], Michael L. Gatza[4], Mathias Uhlen[2], Adil Mardinoglu[2,5,6] and Kazim Yalcin Arga[1*]*

[1] Department of Bioengineering, Marmara University, Istanbul, Turkey, [2] Science for Life Laboratory, KTH – Royal Institute of Technology, Stockholm, Sweden, [3] Department of Bioengineering, Istanbul Medeniyet University, Istanbul, Turkey, [4] Department of Radiation Oncology, Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, United States, [5] Department of Biochemistry and Molecular Biology, Penn State College of Medicine, Hershey, PA, United States, [6] Faculty of Dentistry, Oral and Craniofacial Sciences, Centre for Host-Microbiome Interactions, King's College London, London, United Kingdom, [7] Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

OPEN ACCESS

Edited by:
Junbai Wang,
Oslo University Hospital, Norway

Reviewed by:
Woonyoung Choi,
The Johns Hopkins Hospital, United States
Diego Bonatto,
Federal University of Rio Grande do Sul, Brazil

***Correspondence:**
Adil Mardinoglu
adilm@scilifelab.se
Kazim Yalcin Arga
kazim.arga@marmara.edu.tr

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *23 November 2018*
Accepted: *17 April 2019*
Published: *07 May 2019*

Citation:
Turanli B, Karagoz K, Bidkhori G, Sinha R, Gatza ML, Uhlen M, Mardinoglu A and Arga KY (2019) Multi-Omic Data Interpretation to Repurpose Subtype Specific Drug Candidates for Breast Cancer. Front. Genet. 10:420. doi: 10.3389/fgene.2019.00420

Triple-negative breast cancer (TNBC), which is largely synonymous with the basal-like molecular subtype, is the 5th leading cause of cancer deaths for women in the United States. The overall prognosis for TNBC patients remains poor given that few treatment options exist; including targeted therapies (not FDA approved), and multi-agent chemotherapy as standard-of-care treatment. TNBC like other complex diseases is governed by the perturbations of the complex interaction networks thereby elucidating the underlying molecular mechanisms of this disease in the context of network principles, which have the potential to identify targets for drug development. Here, we present an integrated "omics" approach based on the use of transcriptome and interactome data to identify dynamic/active protein-protein interaction networks (PPINs) in TNBC patients. We have identified three highly connected modules, EED, DHX9, and AURKA, which are extremely activated in TNBC tumors compared to both normal tissues and other breast cancer subtypes. Based on the functional analyses, we propose that these modules are potential drivers of proliferation and, as such, should be considered candidate molecular targets for drug development or drug repositioning in TNBC. Consistent with this argument, we repurposed steroids, anti-inflammatory agents, anti-infective agents, cardiovascular agents for patients with basal-like breast cancer. Finally, we have performed essential metabolite analysis on personalized genome-scale metabolic models and found that metabolites such as sphingosine-1-phosphate and cholesterol-sulfate have utmost importance in TNBC tumor growth.

Keywords: breast cancer, drug repositioning, non-cancer therapeutics, repurposing, basal subtype, personalized metabolic models

INTRODUCTION

Breast cancer is the most commonly diagnoses and second leading cause of cancer-related deaths in women in the United States with an estimated 268,600 new cases and 41,760 deaths in 2019 (Siegel et al., 2019). Although overall survival has significantly improved over the past several decades owing in part to advances in early diagnostic techniques and an increasing understanding of the underlying biological basis of the disease, which has led to improved treatment strategies. On a molecular level, breast cancer can be defined as five predominant molecular subtypes including the luminal A (LumA), luminal B (LumB), and Normal-like (NL) subtypes which are predominantly estrogen receptor (ER) and progesterone receptor (PR) positive; the HER2 Enriched subtype (HER2E) subtype; and basal-like tumors which are largely synonymous with Triple Negative Breast cancer (TNBC) and are ER/PR/HER2 negative. The considerable differences among these molecular subtypes are a consequence of dramatically altered genomic and proteomic profiles which manifest as changes in activated signaling networks (Gatza et al., 2014) and manifest as differences in risk factors, incidence, age, prognosis and response to treatment. Therefore, there is a clear need to develop reliable biomarkers and to identify potential drug targets in each molecular and clinical subtype (Perou et al., 2000; Curtis et al., 2012; Weigman et al., 2012; Gatza et al., 2014; Ciriello et al., 2015; Mertins et al., 2016).

Basal-like breast cancers disproportionally affect younger women and women of African American decent. This subtype, which is highly concordant with TNBC, accounts for ∼15–20% of diagnosed breast tumors but more than 1-in-4 breast cancer related deaths each year. This is, due in part, to the lack of effective therapeutic options for TNBC patients aside from multi-agent chemotherapy, which remains the standard-of-care treatment despite a limited and varied response among patients and the related toxic side-effects (Solzak et al., 2017). In this context, we and others, have proposed that systems level analyses can assist in revealing the underlying molecular mechanism of the diseases, discovery of biomarkers for specific subtypes, identification of subtype specific drug targets and reposition of drugs that can be used in effective treatment of patients (Mardinoglu and Nielsen, 2015; Mardinoglu et al., 2018; Turanli et al., 2018).

Publicly available "omics" datasets including The Cancer Genome Atlas (TCGA) (Ciriello et al., 2015), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012), and the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (NCI-CPTAC) (Mertins et al., 2016) enhance our understanding of the subtype specific molecular mechanisms of breast cancer. Moreover, integrative and comparative analysis of "omics" data together with network modeling provided a comprehensive platform for the drug repositioning and multi-target drug design (Kibble et al., 2015; Vitali et al., 2016; Turanli et al., 2017). A number of studies also combined genomic, transcriptomic, proteomic data with protein-protein interaction networks (PPINs) and identified putative druggable candidates in breast cancer by analyzing topological features of the reconstructed networks (Karagoz et al., 2015; Liu et al., 2017; Li et al., 2018; Nuncia-Cantarero et al., 2018). These bioinformatics pipelines have their own power through decreasing the number of candidate therapeutic targets/drugs and proposing potential treatment strategies for subsets of breast cancer patients.

The overall prognosis for patients with basal-like breast cancer remains poor and there is an urgent need to identify molecular targets to develop effective therapeutic strategies. To take advantage of the extensive publicly available "omics" data, we integrated transcriptome with interactome data and calculated network entropy for each protein-protein interactions (PPIs) to identify the dynamic states in basal-like breast cancer. Our analyses identified modules as systems biomarkers at gene expression level and these networks were confirmed at the proteomic level. Importantly, functional annotation and analysis of module activity scores demonstrated that these modules were subtype specific. Using these models essential metabolites and drug candidates were identified within the context of basal-like specific modules. Collectively, these analyses suggest that the proposed strategy incorporating multi-omics analyses of human breast tumors has the capacity to define novel signaling networks and link these features to existing therapeutic opportunities.

MATERIALS AND METHODS

Data Collection

Throughout the study, we integrated multi-omics data including genomics, transcriptomics, and proteomics using network analysis (**Table 1**). TCGA data were obtained from https://gdac.broadinstitute.org/, METABRIC and CPTAC data were collected from **Supplementary Files** of these studies. At transcriptomic level, gene expressions were obtained from two major initiatives presenting RNA-Seq data from the TCGA study and microarray data from the METABRIC study. Normalized gene expression values for 179 basal and 852 non-basal like breast cancer samples (n = 1031) from TCGA, and 331 basal and 1665 non-basal samples from the METABRIC project (n = 1992) were used in integrative analysis. At the protein level, two different sources were used, (i) expression data of 160 basal and 777 non-basal like samples (n = 937) in TCGA, using Reverse Phase Protein Array (RPPA)- based analysis of 226 proteins, and (ii) expression data of 19 basal and 58 non-basal like samples (n = 77) from CPTAC which performed comprehensive mass-spectrometry methods including around 10,000 proteins (Mertins et al., 2016).

RNA sequencing data from TCGA (n = 1031) were used as a discovery set whereas, microarray data from METABRIC and proteomic data from TCGA and/or CPTAC were used as independent validation data sets in the study (**Table 1**).

Differential Interactome

To obtain a differential view of human interactome between two different phenotypes, and to identify PPIs that are up- or down-regulated in each phenotype relative to the other one, we used the gene expression profiles of interacting protein pairs and recruited the differential interactome analysis as previously described (Ayyildiz et al., 2017). For this purpose, normalized gene expression profiles from TCGA (179 basal-like

and 852 non-basal like samples) were categorized into three levels: high (1), moderate (0), and low (-1) expression levels according to comparison of each gene expression with the average expression within each sample. The probability distributions for any possible co-expression profile of gene pairs (encoding proteins interacting with each other) were estimated, and the uncertainty of determining whether or not a PPI in encountered in a phenotype was estimated through an entropy formulation. In order to define possible PPIs, we used the high confidence human PPIs (Karagoz et al., 2016), comprising 147,923 interactions among 13,213 proteins. Karagoz and coworkers assembled and integrated physical PPIs of Homo sapiens from six publicly available databases including BioGRID (Chatr-Aryamontri et al., 2015), DIP (Salwinski, 2004), IntAct (Orchard et al., 2014), HIPPIE (Schaefer et al., 2012), HomoMINT (Persico et al., 2005), and HPRD (Prasad et al., 2009). Then, PPIs analyzed the differential view of human interactome between the basal and non-basal subtypes of breast cancer; $P < 0.05$ was considered statistically significant for these analyses.

Differentially Expressed Genes and Proteins

Both differentially expressed genes (DEGs) between 179 basal and 852 non-basal samples in TCGA cohort, and differentially expressed proteins (DEPs) between 19 basal and 61 non-basal samples in CPTAC cohort were identified by using the Significance Analysis of Microarrays (SAM) method implemented in R software (Tusher et al., 2001; Hu et al., 2016; Gámez-Pozo et al., 2017). False Discovery Rate (FDR), adjusted p-value was set at $p < 0.05$, and fold changes > 1 between basal-like and non-basal samples were considered as up-regulated DEGs and proteins in basal tumors.

Module Extraction From Basal Specific Networks

Basal subtype specific PPI networks were constructed by using the differential interactome from basal-like tumors. The interactions associated with proteins corresponding to DEGs that are up-regulated in basal-like tumors were identified and used to construct up-regulated PPI networks specific to basal-like breast cancer. The networks were visualized by using Cytoscape software (version 3.4.0) (Lopes et al., 2011). The topological analysis of the networks was performed via CytoNCA plugin of Cytoscape (version 2.1) (Tang et al., 2015). Two different topological metrics, degree, which is defined by the number of adjacent nodes of a node in the network, and betweenness centrality, which characterizes nodes by how often they occur on the shortest path between two other nodes in the network, were simultaneously employed to define hub nodes. Hub nodes with higher degree or betweenness values were reported to have significant roles in cellular signal trafficking and could be potential candidate biomarkers or drug targets Modules were identified as highly connected subnetworks within up-regulated networks. Gene expression data from METABRIC were used for validation of the gene expression modules in basal-like breast cancer.

Functional Annotation

Functional enrichment analysis associated with the three protein-protein interaction modules were analyzed using QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City)[1].

Module Activity

In order to convert the identified EED, AURKA, and DHX9 modules to gene expression signatures that can be used to quantify pathway activity in a given sample from independent datasets, the module was converted to a gene list and the mean expression of unweighted gene list was used to calculate a pathway score. For these studies, a score was calculated for each sample in the TCGA (discovery) and METABRIC cohort (validation). Analysis of variance (ANOVA) tests were used to quantify differences between the EED-module, DHX9-module and AURKA-module activity scores between breast cancer subtypes in each dataset. A Student's t-test was used to evaluate levels of EED, DHX9e and AURKA signature scores between adjacent normal breast tissue and basal-like tumors. To infer the functional roles of these modules, a panel of 270 experimentally derived gene expression signatures that predict activation of various oncogenic signaling pathways, was performed by integrating gene expression data as described previously (Gatza et al., 2014). To identify the association of the modules with oncogenic pathways, a Spearman's rank correlation was used between oncogenic pathway activity scores and EED, DHX9 and AURKA activity scores.

Module Specific Drug Repositioning

To identify small molecules that can potentially reverse gene expression of basal-like tumors, we utilized the Library of Integrated Network-based Cellular Signatures (LINCS) – L1000 data which includes gene expression data from ∼50 human cell line in response to ∼20,000 compounds (Campillos et al., 2008). We queried basal-like specific module genes which are all up-regulated and down-regulated DEGs (Fold Change < 0.2)

[1]www.qiagen.com/ingenuity

TABLE 1 | Validation and discovery sets used in this study.

Data type	Data portal	"Omic" level	Number of basal samples	Number of non-basal samples	Set type
Gene expression levels	TCGA	Transcriptomic	179	852	Discovery
Gene expression levels	METABRIC	Transcriptomic	331	1655	Validation
Protein expression levels	CPTAC	Proteomic	19	58	Validation
Protein expression levels	TCGA	Proteomic	160	777	Validation

signatures as input. We used the L1000CDS2 (Duan et al., 2016) search engine, which contains 30,000 significant signatures that were processed from the LINCS L1000 data, to identify small molecule signatures associated with each module. The identified drugs were ranked based on their scores and the top 50 were acquired for each query. Drugs were checked through literature review and publicly available datasets such as CTD (Davis et al., 2017) and KEGG DRUG (Kanehisa et al., 2012) to identify those that were previously investigated within the context of breast cancer.

Subtype Specific Essential Metabolites

We next acquired 917 personalized genome scale metabolic models (GEMs) of breast cancer patients (Uhlen et al., 2017). We analyzed each patient GEM to identify essential metabolites for tumor growth by removing the reactions in which the metabolite functions as substrate regardless of compartmentalization (Bidkhori et al., 2018). Next, we categorized personalized models based on clinical information to create subtype-specific patient metabolic models and found the percentage of subtype representation of each metabolite. A Fisher exact test was applied to identify statistically significant difference between basal-like and non-basal-like (i.e., all other tumors) for each metabolite. Significant difference between subtypes was determined based on a $P < 0.05$.

RESULTS

Basal-Like Subtype Specific PPI Elucidation via Differential Interactome

Cancer cells are characterized by increase in network entropy comprising high uncertainty, pathway redundancy and promiscuous signaling resulting from intra-sample heterogeneity. Recently, a differential interactome network analysis were presented to show the uncertainties of PPIs in ovarian cancer (Ayyildiz et al., 2017). In this study, we employed differential interactome algorithm utilizing the entropy concept using a comprehensive gene expression data and human PPI network to reveal the heterogeneity among the breast cancer subtypes (i.e., basal-like vs. non-basal-like). To do so, we categorized the expression of each gene and for each patient using 179 basal and 852 non-basal-like samples from TCGA into three classes as -1, 0, 1, These classes were then integrated with a high confident PPI network (Karagoz et al., 2016) and the frequency of PPIs estimated for both basal-like and non-basal-like tumors. Using a 95% confidence interval ($p < 0.05$), significant values <0.2 and >0.8 as well as corresponding $H < 0.7$ were calculated for each class. As a result, 3,002 interactions among 1,652 proteins were considered significant across the entire dataset. These analyses identified 2,291 interactions among 1,391 proteins as being significantly activated in basal-like tumors whereas 712 interactions among 612 proteins were identified as significant in non-basal-like samples; 351 proteins were common across both subgroups of tumors (**Supplementary Table S1**).

Since low entropy presents low uncertainty, low redundancy and deterministic signaling resulting with homogeneity in the population, we next focused on the basal-like subtype to identify low entropy interactions ($H < 0.1$). These analyses identified the EED protein network which is defined by 82 interactions within the group of 98 proteins. Importantly, the lowest entropy profile of the EED centroid network only identified an interaction with one protein (CTCF) in non-basal-like tumors. We further identified a sub-set of proteins, excluding 351 common signatures evident in both basal-like and non-basal-like tumors to identify a basal-like subtype specific network (**Supplementary Table S2**). All differential interactome networks and basal-like subtype specific networks were delimited regarding up-regulated genes in the basal-like tumors through 2-class SAM analysis (Tusher et al., 2001; **Supplementary Table S3**). Through the integration of SAM analysis and the above detailed differential interactome framework, we identified three significant modules: EED centroid module, covering relatively low entropy PPIs (**Figure 1A**); the DHX9 centroid module, covering mixed of low and high entropy PPIs (**Figure 1B**); and the AURKA centroid module, covering relatively high entropy PPIs (**Figure 1C**).

Further analyses of the EED, DHX9, and AURKA modules determined that genes included in EED-module play roles in cyclins and cell cycle regulation (p = 6.1e-19), cell cycle: G1/S checkpoint regulation (p = 3.5e-18), regulation of cellular mechanics by calpain protease (p = 1.6e-11), aryl hydrocarbon receptor signaling (p = 4.3e-11), apoptosis signaling (p = 7.0e-10), TWEAK signaling (p = 1.8e-09), and GADD45 signaling (p = 4.3e-9), In contrast, the genes in DHX9-module contribute to mTOR signaling (p = 4.1e-06), regulation of eIF4 and p70S6K signaling (p = 7.9e-06), EIF2 signaling (p = 7.2e-05), Inflammasome pathway (p = 1.4e-04), assembly of RNA Polymerase I Complex (p = 1.1e-03), DNA double strand break repair (p = 1.8e-03) and cell cycle (p = 3.5e-03) while the genes associated with the AURKA-module are involved in DNA damaged-induced 14-3-3A signaling (p = 1.8e-10), mitotic roles of Polo like kinase (p = 2.1e-09), role of CHK proteins in cell cycle checkpoint control (p = 6.0e-08), ATM signaling (p = 9.3e-07) and mismatch repair (p = 3.1e-06), role of BRCA1 in DNA damage response (p = 1.3e-05), and cell cycle (p = 9.8e-05). These data suggest that each module represent a unique aspect of basal-like breast cancer signaling. Some of these pathways such as TWEAK signaling, apoptosis signaling, mTOR signaling, ATM signaling showed that the chemotherapy targeted pathways are also activated in basal-like tumors in which chemotherapy is the front-line treatment option, nowadays (**Supplementary Figure S1**).

Proteomic Analysis of Basal Specific Modules

We next reconstructed PPI networks using transcriptome data and validated our findings at proteomic level by leveraging orthogonal genomic and proteomic data from the TCGA and CPTAC projects. Transcriptome data from 937 sample was compared to RPPA analysis of the same samples to assess the relationship between each network at the 226 proteins and phosphoproteins from TCGA. Likewise the gene expression data from a subset of 77 of these samples was used to

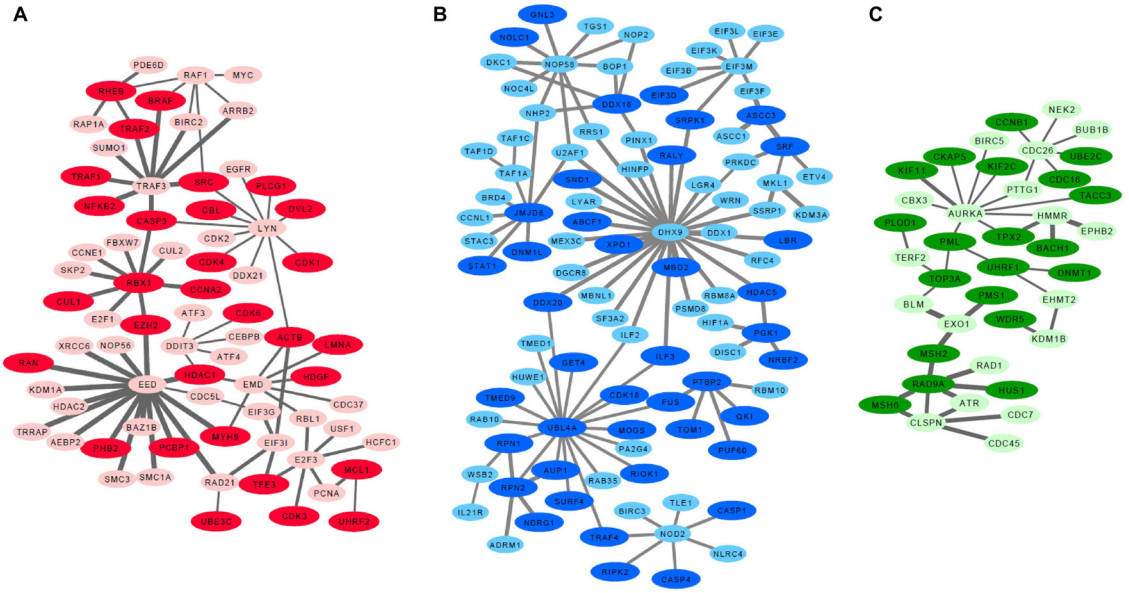

FIGURE 1 | Basal like breast cancer specific highly connected protein-protein interaction modules. **(A)** EED module, **(B)** DHX9 module, **(C)** AURKA module. Darker nodes indicate the statistically significant positive correlations between mRNA and protein pairs. Thicker edges indicate lowest entropy levels between interacting pairs.

examine the relationship between each module and 10,062 proteins and phosphoproteins using mass spectrometry-derived proteomic data from the CPTAC project. First, we used CPTAC proteome data to compare each gene to its corresponding protein across all basal-like tumors and assessed correlation for those pairs. Overall, 52.6–64.5% of the mRNA-protein pairs showed statistically significant positive Spearman correlations ($P < 0.05$) when changes in mRNA abundance were compared to changes in relative protein abundance. These proteins in basal-like samples are shown in darker colors in **Figures 1A–C**. Then, we identified DEPs between basal-like and non-basal-like samples by using both RPPA and CPTAC data. Although RPPA data has limited number of proteins, we identified several up-regulated proteins including CCNE1, RAF1, SRC, CDK1, EGFR, MYC, MYH9, PCNA associated with the EED-module. Similarly, NDGR1 and CCNB1 were associated with the DHX9 and AURKA modules, respectively. We also analyzed DEPs between basal-like and non-basal-like tumors by using CPTAC data which is more comprehensive than RPPA data and it covered 69.4–56.4% of the module genes and 29.4–36.4% of these proteins were identified as being up-regulated in basal-like tumors (**Supplementary Table S3**).

Modules as Basal Specific Signatures

In order to quantitatively assess the activity of each modular in each patient sample, we next generated a gene expression signature on the basis of median expression of each gene in the module. This strategy was used to calculate a module score for each sample in the TCGA (discovery set) and METABRIC (validation set) datasets. We then quantitatively evaluated the differences in the module activities across breast cancer subtypes by an ANOVA test. These analyses demonstrated that EED (P = 1.13e-244), DHX9 (P = 2.4e-236), and AURKA (P = 2.05e-175) activity was highest in basal-like tumors in the TCGA cohort (**Figures 2A–C**); these findings were validated by analysis of module activity in the METABRIC cohort (**Figures 2D–F**). Finally, we determined that the EED (P = 1.06-e96), DHX9 (P = 2.44e-85), and AURKA modules (P = 6.61e-127) were expressed at significantly higher levels in in basal-like tumors compared to adjacent normal tissue (**Figures 3A–C**).

Functionality of Basal Specific Modules

We examined the functional roles of these modules by exploring the correlations with a series of previously published gene expression signatures which are capable of measuring oncogene or tumor suppressor pathway activity, aspects of the tumor microenvironment and other tumor characteristics. We identified pathway activities, which were positively (or negatively) correlated with module activities using a Spearman rank correlation to assess the relationship between pathway activity and the EED, DHX9, or AURKA module activity scores. As expected, our data recapitulated known characteristics of basal-like tumors including low hormone receptor signaling and high expression of proliferation pathway activity and demonstrated the relationship between these characteristics and the expression of each module (i.e., EED, DHX9, and AURKA). Moreover, these modules were associated with multiple indicators of proliferation including, RB_LOSS, RB_LOH, and bMYB highly correlated with these module activities as well as RAS, PIK3CA, β-catenin, MYC and HER1_Cluster 1, HER1_Cluster 2, and HER1_Cluster 3 signatures (**Figure 4A**). Consistent results were obtained using the METABRIC data

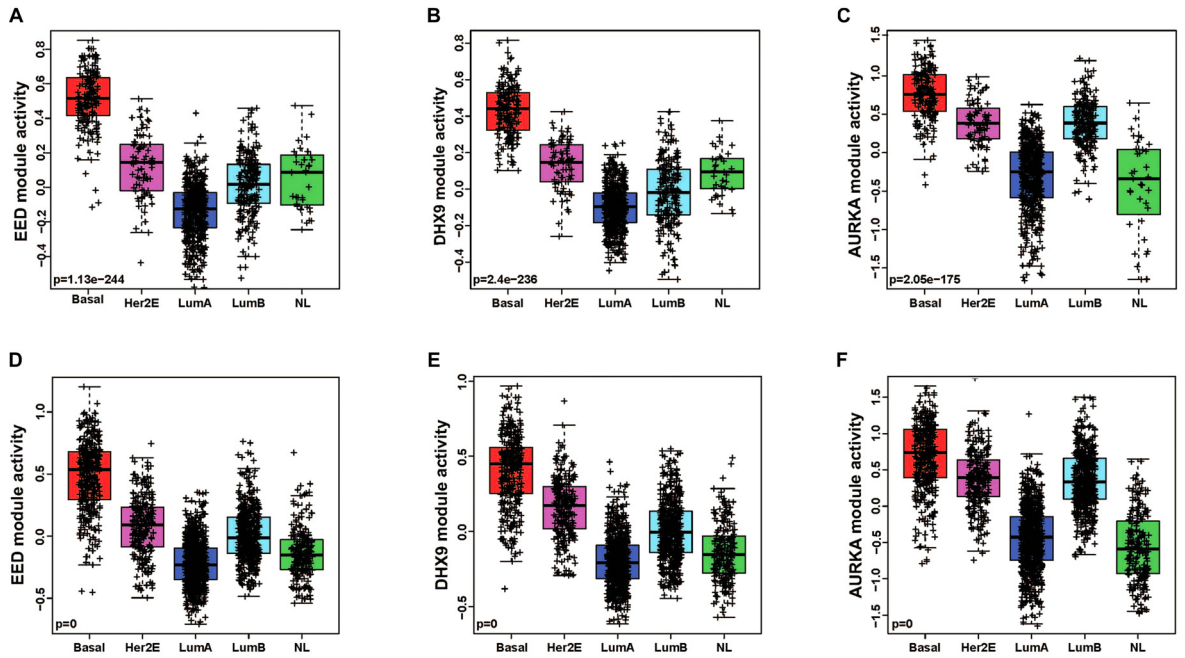

FIGURE 2 | The pattern of basal like breast cancer specific modules activity across breast cancer subtypes. **(A–C)** EED, DHX9, and AURKA modules are highly activated in basal like tumors by using TCGA cohort-discovery set. **(D–F)** EED, DHX9, and AURKA modules are highly activated in basal like tumors by using METABRIC cohort-validation set.

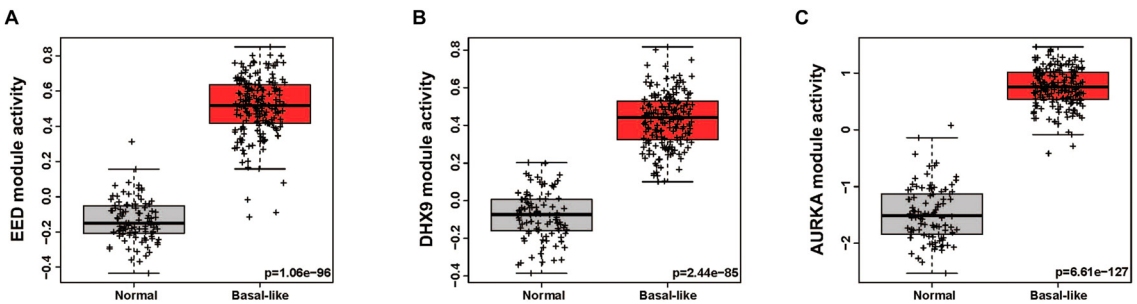

FIGURE 3 | The activity levels of basal like breast cancer specific modules in normal and basal like tumors. **(A)** EED module, **(B)** DHX9 module, **(C)** AURKA module.

(**Figure 4B**). Importantly, we also confirmed the ability of the transcriptomic module signatures to assess the functional roles of EED, DHX9, and AURKA modules by exploring relationships between the module signature scores and protein expression. Analysis of RPPA data from basal-like samples confirmed that these tumors with high module scores have significantly higher levels of CHK1, CHK2, CDK1, Cyclin B1, Cyclin E1, FOXM1, and PCNA protein expression consistent with their role in cell cycle regulation and proliferation (**Figure 4C**).

Drug Repositioning Based on Basal Subtype Specific Modules

As discussed above, the EED, DHX9, and AURKA modules were converted to gene expression signatures on the basis of up-regulated genes specific to each module; as would be expected down-regulated genes (Fold Change < 0.2) were common for all modules. We asked the question of whether each module/signature identified potential therapeutic opportunities. To do so, we queried each gene signatures separately against the LINCS database L1000CDS2 (Duan et al., 2016) in order to identify concordant and discordant patterns of gene expression between each module and gene expression profiles associates with drug-induced and/or disease expression. Drugs that resulted in a gene expression profile that was negatively correlated with each module were identified and selected as potential candidate compounds that had the potential to reverse the activity of each module network that was associated with basal-like tumors (**Supplementary Figure S2**). Since we have demonstrated specificity of the modules to basal-like tumors, we may also propose that our candidate drugs are specifically targeting basal-like tumors.

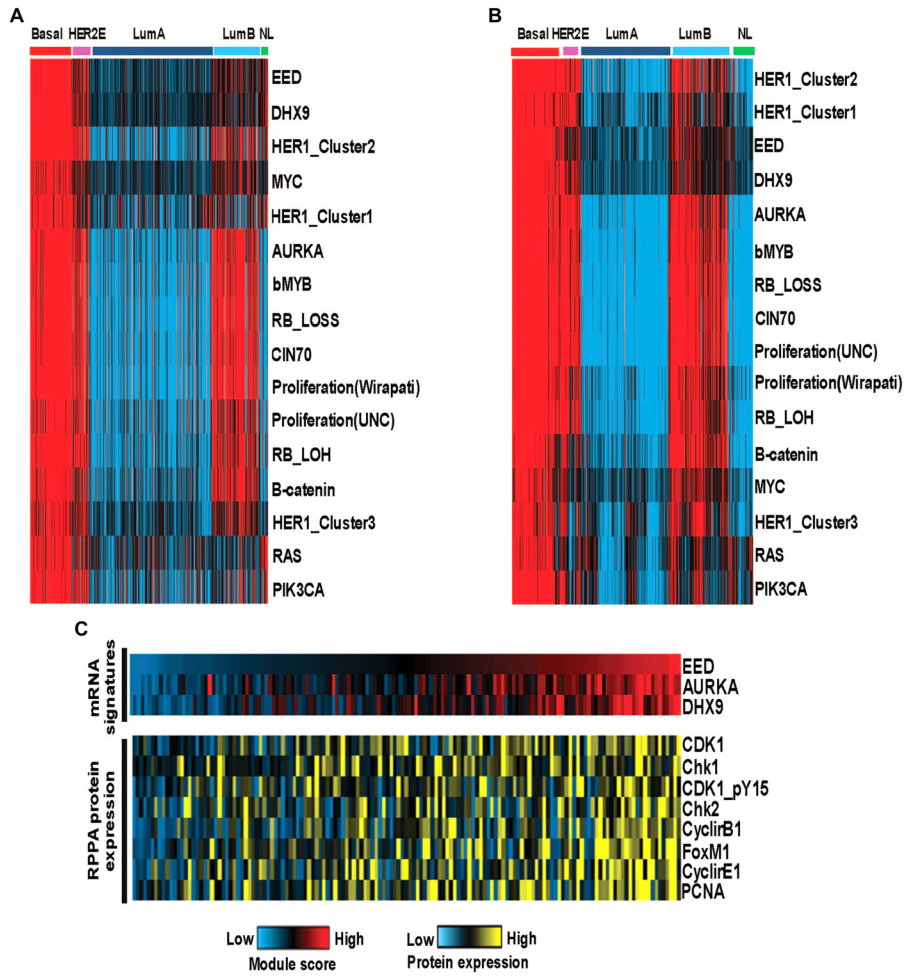

FIGURE 4 | The functional analysis of basal like breast cancer specific modules **(A)** The activity of oncogenic pathways correlated with module activities in TCGA cohort-discovery set. **(B)** The activity of oncogenic pathways correlated with module activities in METABRIC cohort-validation set. **(C)** High module activities characterized by high expression of cell cycle proteins.

After removing the duplicated drugs from query results, we found that EED and AURKA modules were associated with 41 candidate compounds while DHX9 was associated with 31 candidate small molecules. Networks comprising drug candidates and modules were found to have 114 interaction between three modules and 80 drugs (**Figure 5A**). The 80 identified drugs were categorized as molecular inhibitors (23%), anti-neoplastic agents (15%), heterocyclic compounds (10%), anti-infective agents (6%), or steroids (6%). Moreover, a number of the drugs specific to each module (as well as some common candidates) were also identified in each drug category (**Figure 5B**). There are at least 19 approved, 24 investigational, and 6 experimental drugs listed in DrugBank (version 5.1.1), however there are perturbagens used in L1000 platform without detailed information (**Supplementary Table S4**).

Nine of the drugs including selumetinib, trametinib, and several other investigational drugs were common to each of the three modules. Consistent with our results, selumetinib as MEK inhibitor was reported to suppresses cell proliferation, migration, and trigger apoptosis, following G1 arrest in TNBC cells (Zhou et al., 2016). Furthermore, the MEK inhibitor, trametinib is also a therapy of significant interest for the treatment of TNBC since TNBC cell lines have been shown to be especially sensitive to this drug (Jing et al., 2012; Davis et al., 2014). Finally, we noted some overlap between drugs associated with each module. For instance, the three common drugs (i.e., wortmannin, mestanolone, NVP-TAE684) are associated with both the EED and AURKA modules while 12 drugs (i.e., radicicol, lapatinib, alvocidib, zileuton, geldanamycin, exemestane) are consistent between the EED and DHX9 module (**Supplementary Table S4**). Intriguingly, 10 of our candidate drugs were previously associated with the breast cancer based on at least one of the sources including CTD, KEGG Drug, Clinical Trials, and scientific literature (**Table 2**).

Since EED module has the lowest entropy level between PPIs, we focused on 17 drug candidates which are only related to EED module in addition to common drugs. Three of these drugs are anti-neoplastic agents and five of them are unknown, however, others belonged to steroids (BRD-A94793051,

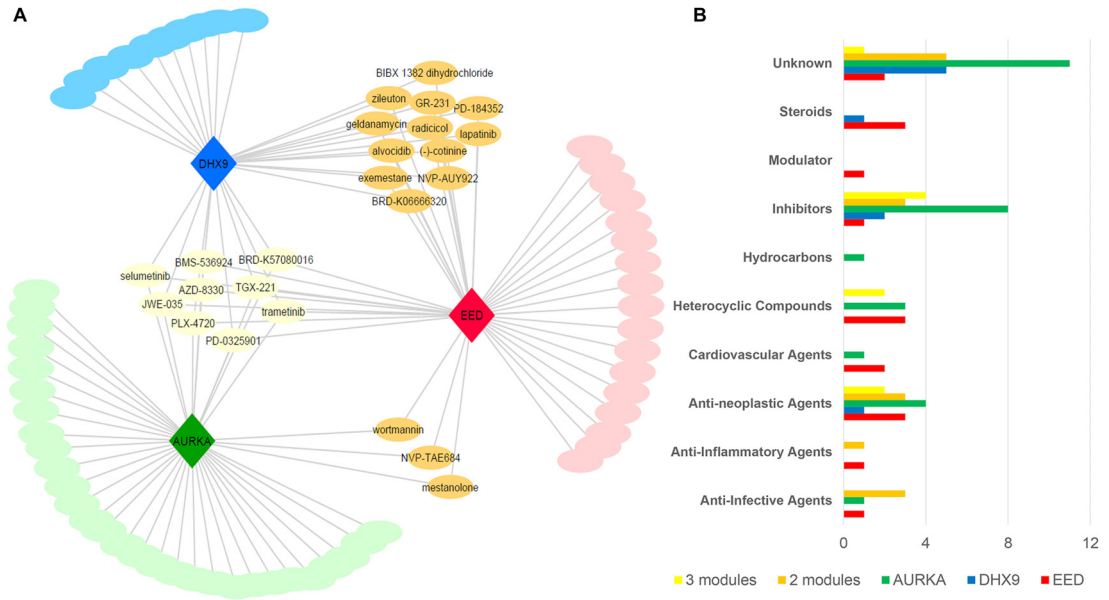

FIGURE 5 | Drug repositioning for basal like breast cancer specific modules **(A)** Module-drug networks **(B)** Drug categories of module specific drugs and common drugs among modules.

Oxymetholone, Testosterone propionate), PLK inhibitor (BI-2536), heterocyclic compounds (BRD-K17953061, GDC-0980, TG101348), cardiovascular agents (BRD-K52080565, S-2500), anti-inflammatory (oxaprozin), and anti-infective agents (5-fluorocytosine).

Essential Metabolites and Anti-metabolites as Drug Candidates

GEMs reconstructed for different cancer tissues have been used for characterization of metabolic modifications; disease stratification and determination of drug targets using essential genes or metabolites (Folger et al., 2011; Agren et al., 2012; Bidkhori et al., 2018). To address this question, we first identified a panel of 917 personalized GEMs derived from breast cancer patients (Uhlen et al., 2017). We then categorized each GEMs based on clinical information to create subtype-specific patient metabolic models. These models were then used to identify subtype-specific metabolites essential for tumor growth. After categorization of BCS, percentage of abundance for each essential metabolite was calculated. Significant alteration between the abundance of basal-like and non-basal BCS were determined based on FDR adjusted P-value threshold (P-adj < 0.05) (**Supplementary Table S5**). These analyses identified 27 essential metabolites (**Supplementary Table S6**); 11 were significantly enriched in basal-like tumors while the remaining 16 were enriched in non-basal-like samples. Further analyses determined that the essential metabolites that are expressed at higher levels in basal-like tumors were associated with steroid metabolism, biotin metabolism, nucleotide metabolism, sphingolipid metabolism and transport. Conversely, the identified metabolites down-regulated in basal-like samples were involved in beta-alanine metabolism, arginine and proline metabolism, cysteine and methionine metabolism, and carnitine shuttle (**Figure 6**).

TABLE 2 | Various drug candidates that already associated with breast cancer via different sources.

Drug name	Literature evidence	CTD	KEGG drug	Clinical trials
Epirubicin	Warm et al., 2010	✓	✓	NCT00176488
Erlotinib	Catania et al., 2006			NCT01650506
Lapatinib	Giampaglia et al., 2010	✓		NCT00694252
Exemestane	Goss et al., 2013	✓	✓	NCT00810797
Wortmannin	Li et al., 2012	✓		
Alvocidib	Murphy and Dickler, 2015	✓		NCT00039455
Tyrphostin ag 1478	Zhang et al., 2008	✓		
Canertinib	Gschwantler-Kaulich et al., 2016	✓		NCT00051051
Danazol	Coombes et al., 1983	✓		
Palbociclib	Finn et al., 2016	✓	✓	NCT02513394

DISCUSSION

The dynamics of cells are regulated by PPIs and properties of networks such as entropy provide information about the current state of the network. Given that cancer cells are reported to have an increase in network entropy, several previous studies have integrated gene expression data with PPI network information

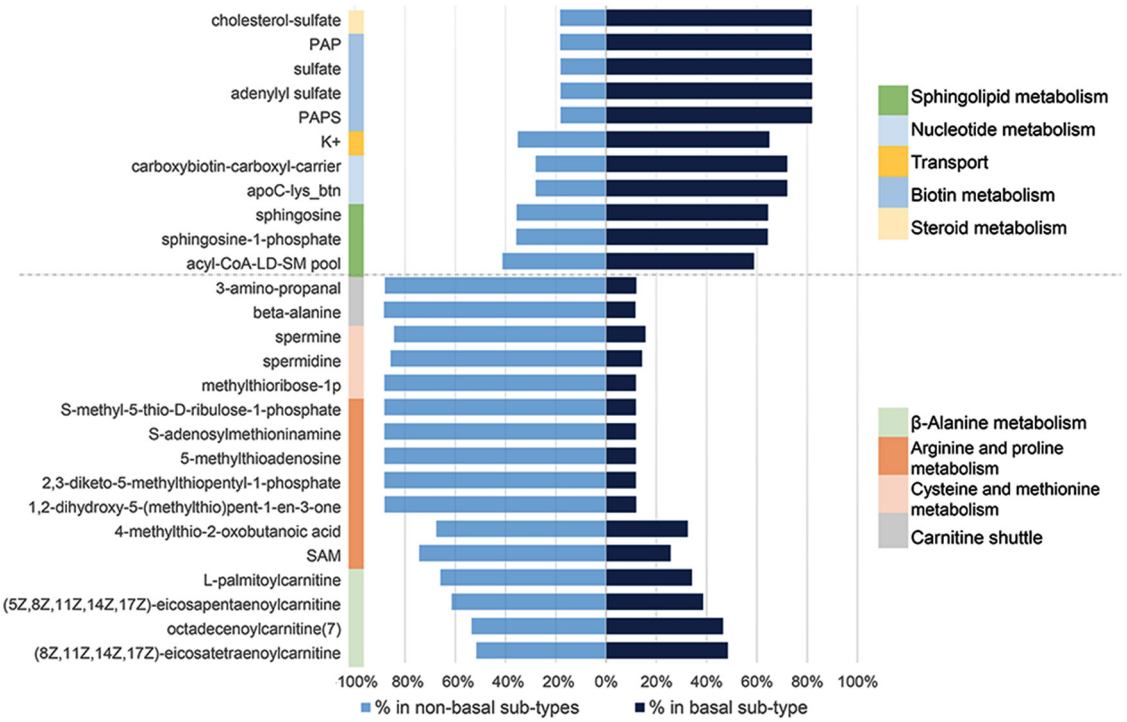

FIGURE 6 | Significant essential metabolite differences between non-basal and basal like breast cancer specific personalized metabolic models and their associated pathways.

to compute the energetic state of cancer cells by calculating entropy (West et al., 2012; Teschendorff et al., 2015; Rietman et al., 2016). Likewise, a number of studies have used a network-based entropy approach to identify disease specific PPIs as biomarker candidates, proliferative and prognostic markers in lung and breast cancer, as well as to demonstrate the association between network entropy and tumor initiation, progression, and anticancer drug responses (Varadan and Anastassiou, 2006; Xiong et al., 2010; Banerji et al., 2013; Lecca and Re, 2015; Cheng et al., 2016; Ayyildiz et al., 2017).

The current study employed a novel multi-omics-based approach to integrate genomic, proteomic and metabolomic tumor data. Our analyses of mRNA expression data identified three highly connected modules which are centered on the activation of the EED, DHX9, and AURKA signaling networks. These data demonstrated that each module is highly activated in basal-like tumors compared to non-basal-like tumors as well as adjacent normal tissues. Importantly, by analyzing proteome data, our results confirmed the correlation between the expression of genes and proteins that comprise each identified module. By analyzing the association between module expression and oncogenic signaling using a panel of more than 250 gene expression signatures, we were able to assess the functional relationship of these modules with known oncogenic and signaling features. Our results demonstrated the correlation between EED, DHX9, and AURKA module activity and proliferative oncogenic pathways including RAS, PI3K, and Rb/E2F signaling in basal-like tumors. Consistent with these results, CHK1, CHK2, CDK1, Cyclin B1, Cyclin E1, and PCNA protein expression levels were identified higher in tumors with high module scores. Through integrated analyses, we identified candidate drugs to target three modules by drug repositioning. Utilizing multiple omics data including genome, transcriptome, and interactome, we repurposed 519 agents for breast cancer by incorporating data from the LINCS project (Duan et al., 2016) into our analyses. In another drug repositioning study, five of the identified repurposed candidate agents showed superior therapeutic indices compared to doxorubicin in *in vitro* assays in basal sub-type cell line (SUM149) in addition to luminal cell line (MCF7) (Chen et al., 2016). Moreover, Lee et al. (2016) developed an integrative approach for drug repositioning using the expression signature, chemical structure, target signatures and LINCS data. They applied this strategy to identify candidate anti-cancer drugs for breast cancer (Lee et al., 2016). Although there are previous computational drug-repositioning efforts that utilized LINCS as mentioned, the methodologies are focused on breast cancer regardless of disease heterogeneity and subtype information.

In addition, our analyses identified subtype-specific metabolites, including several specific to basal-like tumors, which may provide opportunity to design anti-metabolite drugs for breast cancer. Results in essential metabolite analysis emphasized sphingolipids and steroid metabolism for basal-like breast cancer. Sphingolipid levels in breast cancer tissue are generally higher than normal breast tissue and bioactive sphingolipids, such as sphingosine-1-phosphate (S1P) has many cellular functions like

cell proliferation, migration, survival, immune cell trafficking, and angiogenesis which are related to cancer progression and metastasis (Nagahashi et al., 2016). However, sphingosine and S1P were recently highlighted as important for signaling mechanisms in metastatic TNBC and its targeted therapy (Maiti et al., 2017). A recent lipidomics profiling of TNBC tumors also supported sphingolipids as potential prognostic markers and associated enzymes as candidate therapeutic targets (Purwaha et al., 2018) in parallel to our results.

TNBC was associated with expression pattern of 2-pore domain potassium (K2p) channels which enable background leak of potassium (K^+). Differential expression on K2p-channels may be suggested as a novel molecular marker related to potassium levels in basal like BCS (Dookeran et al., 2017). In another study, expression of calcium-activated potassium (SK4) channels were also associated with TNBC and cellular functions such as proliferation, migration, apoptosis, and EMT processes (Zhang et al., 2016).

Breast cancer is known as one of the malignancies in which steroid hormones drive cellular proliferation (Capper et al., 2017). As steroid metabolism associated metabolite, cholesterol sulfate, is quantitatively the most important known sterol sulfate in human plasma and may play a role in cell adhesion, differentiation and signal transduction (Strott and Higashi, 2003). Given that current standard-of-care therapy for TNBC is largely limited to multi-agent cytotoxic chemotherapy, the potential of incorporating identified repurposed drugs and/or targeting identified modules and/or metabolites represents a potential therapeutic opportunity for a subset of patents with limited treatment options.

Given these data, we would propose that the strategy outlined here can be used to repurposed drugs in order to identify novel candidate compounds or drugs to be utilized in not only monotherapy but also in combination therapy for the treatment of TNBC. Consistent with this argument, a number of the candidate drugs identified by our analyses have been incorporated in ongoing clinical trials. For instance, TNBC patients who received pre-operative sequential epirubicin and cyclophosphamide followed by docetaxel were found to have a significant increase in pathological complete response (PCR) (Warm et al., 2010). Although a great number of pre-clinical trials will be necessary to support the *in silico* modeling detailed in the current study prior to initiation of clinical trials, a large number of identified candidates have significant *in vitro* and *in vivo* support to indicate that these represent potential therapeutic opportunities. For instance, drugs inhibiting cyclin-dependent kinases (CDKs), including the CDK9 inhibitor alvocidib have been reported to be effective against TNBC (Ocana and Pandiella, 2015).

Erlotinib also showed anti-tumor effect on TNBC in a xenograft model (Ueno and Zhang, 2011). Likewise, targeting the MET and EGFR receptors, which regulate RAS/ERK and PI3K/AKT signaling, resulted in improved treatment compared to monotherapy (Linklater et al., 2016).

The current study has defined a novel approach to identify breast cancer subtype-specific network modules via a network entropy-based approach. This strategy can be used for both the identification of potentially novel signaling networks but also to identify subtype-specific therapeutic opportunities through drug repositioning. Importantly, we demonstrate that this approach can be used to link signaling networks with and subtype-specific essential metabolites which represents additional therapeutic opportunities. As such, the current studies have the potential enhancing the impact of existing therapeutics or multi-agent therapeutic strategies by identifying novel drug/target networks in the context of breast cancer and in breast cancer subtypes. On a broader scale, this strategy is largely applicable to all cancer and disease type/subtypes where multi-platform genomic, proteomic, and metabolomic data exists and thus represents a potential strategy to define novel signaling networks unique to each disease and identify disease/subtype-specific therapeutic strategies.

AUTHOR CONTRIBUTIONS

BT and KK designed the study, performed the all other analyses, and wrote the manuscript. GB performed essential metabolite analysis. MG, RS, AM, MU, and KA supervised the work and contributed to the manuscript during the progress of the work. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported by Knut and Alice Wallenberg Foundation and Marmara University Scientific Projects Committee (BAPKO) in the context of the project FEN-C-DRP-250816-0417. R00CA166228 from the National Cancer Institute of the National Institutes of Health and V2016-013 from the V Foundation for Cancer Research to MG and DHFS-18PPC-024 from the New Jersey Commission for Cancer Research to KK.

ACKNOWLEDGMENTS

We thank the TUBITAK BIDEB 2211 National Doctoral Fellowship Program provided to BT.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00420/full#supplementary-material

FIGURE S1 | Functional enrichment results of the genes involved in each basal-like module using Ingenuity Pathway Analysis (IPA).

FIGURE S2 | The gene signatures of three modules separately on L1000CDS2 for elucidating the differences and similarities between drug-induced expression profiles and disease expression. Drugs were ranked for each module and we elected drugs that showed negatively correlated action mechanisms with the module gene signatures to reverse disease gene expression.

TABLE S1 | Non-basal and basal-like subtype specific PPI elucidation via differential interactome.

TABLE S2 | Three modules for only in basal-like subtype specific networks.

TABLE S3 | Statistical values of differential expressed genes and proteins.

TABLE S4 | Information of repurposed module specific and common drug signatures.

TABLE S5 | Essential metabolite and personalized model matrix and breast cancer categorization of personalized GEMs.

TABLE S6 | Significant essential metabolites between non-basal and basal-like breast cancer.

REFERENCES

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.* 8:e1002518. doi: 10.1371/journal.pcbi.1002518

Ayyildiz, D., Gov, E., Sinha, R., and Arga, K. Y. (2017). Ovarian cancer differential interactome and network entropy analysis reveal new candidate biomarkers. *Omi. A J. Integr. Biol.* 21, 285–294. doi: 10.1089/omi.2017.0010

Banerji, C. R. S., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J. X., et al. (2013). Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Sci. Rep.* 3:3039. doi: 10.1038/srep03039

Bidkhori, G., Benfeitas, R., Elmas, E., Kararoudi, M. N., Arif, M., Uhlen, M., et al. (2018). Metabolic network-based identification and prioritization of anticancer targets based on expression data in hepatocellular carcinoma. *Front. Physiol.* 9:916. doi: 10.3389/fphys.2018.00916

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321, 263–266. doi: 10.1126/science.1158140

Capper, C. P., Rae, J. M., and Auchus, R. J. (2017). The metabolism, analysis, and targeting of steroid hormones in breast and prostate cancer. *Horm Cancer* 7, 149–164. doi: 10.1007/s12672-016-0259-0

Catania, C., De Pas, T. M., Pelosi, G., Manzotti, M., Adamoli, L., Nolè, F., et al. (2006). Erlotinib-induced breast cancer regression. *Ann. Pharmacother.* 40, 2043–2047. doi: 10.1345/aph.1H252

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 43, D470–D478. doi: 10.1093/nar/gku1204

Chen, H.-R., Sherr, D. H., Hu, Z., DeLisi, C., Jin, G., Fu, C., et al. (2016). A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer. *J. Natl. Compr. Canc. Netw.* 8, 1–21. doi: 10.1186/s12920-016-0212-7

Cheng, F., Liu, C., Shen, B., and Zhao, Z. (2016). Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. *BMC Syst. Biol.* 10:65. doi: 10.1186/s12918-016-0309-9

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163, 506–519. doi: 10.1016/j.cell.2015.09.033

Coombes, R. C., Perez, D., Gazet, J.-C., Ford, H. T., and Powles, T. J. (1983). Danazol treatment for advanced breast cancer. *Cancer Chemother. Pharmacol.* 10, 194–195. doi: 10.1007/BF00255761

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., et al. (2017). The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.* 45, D972–D978. doi: 10.1093/nar/gkw838

Davis, S. L., Eckhardt, S. G., Tentler, J. J., and Diamond, J. R. (2014). Triple-negative breast cancer: bridging the gap from cancer genomics to predictive biomarkers. *Ther. Adv. Med. Oncol.* 6, 88–100. doi: 10.1177/1758834013519843

Dookeran, K. A., Zhang, W., Stayner, L., and Argos, M. (2017). Associations of two-pore domain potassium channels and triple negative breast cancer subtype in the cancer genome atlas: systematic evaluation of gene expression and methylation. *BMC Res. Notes* 10:475. doi: 10.1186/s13104-017-2777-2774

Duan, Q., Reid, S. P., Clark, N. R., Wang, Z., Fernandez, N. F., Rouillard, A. D., et al. (2016). L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst. Biol. Appl.* 2:16015. doi: 10.1038/npjsba.2016.15

Finn, R. S., Martin, M., Rugo, H. S., Jones, S., Im, S.-A., Gelmon, K., et al. (2016). Palbociclib and letrozole in advanced breast cancer. *N. Engl. J. Med.* 375, 1925–1936. doi: 10.1056/NEJMoa1607303

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* 7, 1–10. doi: 10.1038/msb.2011.35

Gámez-Pozo, A., Trilla-Fuertes, L., Berges-Soria, J., Selevsek, N., López-Vacas, R., Díaz-Almirón, M., et al. (2017). Functional proteomics outlines the complexity of breast cancer molecular subtypes. *Sci. Rep.* 7:10100. doi: 10.1038/s41598-017-10493-w

Gatza, M. L., Silva, G. O., Parker, J. S., Fan, C., and Perou, C. M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat. Genet.* 46, 1051–1059. doi: 10.1038/ng.3073

Giampaglia, M., Chiuri, V. E., Tinelli, A., De Laurentiis, M., Silvestris, N., and Lorusso, V. (2010). Lapatinib in breast cancer: clinical experiences and future perspectives. *Cancer Treat. Rev.* 36(Suppl. 3), S72–S79. doi: 10.1016/S0305-7372(10)70024-4

Goss, P. E., Ingle, J. N., Pritchard, K. I., Ellis, M. J., Sledge, G. W., Budd, G. T., et al. (2013). Exemestane versus anastrozole in postmenopausal women with early breast cancer: NCIC CTG MA.*27* - A randomized controlled phase III trial. *J. Clin. Oncol.* 31, 1398–1404. doi: 10.1200/JCO.2012.44.7805

Gschwantler-Kaulich, D., Grunt, T. W., Muhr, D., Wagner, R., Kölbl, H., and Singer, C. F. (2016). HER specific TKIs exert their antineoplastic effects on breast cancer cell lines through the involvement of STAT5 and JNK. *PLoS One* 11:e0146311. doi: 10.1371/journal.pone.0146311

Hu, J., Ye, F., Cui, M., Lee, P., Wei, C., Hao, Y., et al. (2016). Protein profiling of bladder urothelial cell carcinoma. *PLoS One* 11:e0161922. doi: 10.1371/journal.pone.0161922

Jing, J., Greshock, J., Holbrook, J. D., Gilmartin, A., Zhang, X., McNeil, E., et al. (2012). Comprehensive predictive biomarker analysis for MEK inhibitor GSK1120212. *Mol. Cancer Ther.* 11, 720–729. doi: 10.1158/1535-7163.MCT-11-0505

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, 109–114. doi: 10.1093/nar/gkr988

Karagoz, K., Sevimoglu, T., and Arga, K. Y. (2016). Integration of multiple biological features yields high confidence human protein interactome. *J. Theor. Biol.* 403, 85–96. doi: 10.1016/j.jtbi.2016.05.020

Karagoz, K., Sinha, R., and Arga, K. Y. (2015). Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *OMICS* 19, 115–130. doi: 10.1089/omi.2014.0135

Kibble, M., Saarinen, N., Tang, J., Wennerberg, K., Mäkelä, S., and Aittokallio, T. (2015). Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products. *Nat. Prod. Rep.* 32, 1249–1266. doi: 10.1039/c5np00005j

Lecca, P., and Re, A. (2015). Detecting modules in biological networks by edge weight clustering and entropy significance. *Front. Genet.* 6:265. doi: 10.3389/fgene.2015.00265

Lee, H., Kang, S., and Kim, W. (2016). Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. *PLoS One* 11:e0150460. doi: 10.1371/journal.pone.0150460

Li, C., Luo, L., Wei, S., and Wang, X. (2018). Identification of the potential crucial genes in invasive ductal carcinoma using bioinformatics analysis. *Oncotarget* 9, 6800–6813. doi: 10.18632/oncotarget.23239

Li, J., Li, F., Wang, H., Wang, X., Jiang, Y., and Li, D. (2012). Wortmannin reduces metastasis and angiogenesis of human breast cancer cells via nuclear factor-κB-dependent matrix metalloproteinase-9 and interleukin-8 pathways. *J. Int. Med. Res.* 40, 867–876. doi: 10.1177/147323001204000305

Linklater, E. S., Tovar, E. A., Essenburg, C. J., Turner, L., Madaj, Z., Winn, M. E., et al. (2016). Targeting MET and EGFR crosstalk signaling in triple-negative breast cancers. *Oncotarget* 7, 69903–69915. doi: 10.18632/oncotarget.12065

Liu, Y., Yin, X., Zhong, J., Guan, N., Luo, Z., Min, L., et al. (2017). Systematic identification and assessment of therapeutic targets for breast cancer based

on genome-wide RNA interference transcriptomes. *Genes* 8:E86. doi: 10.3390/genes8030086

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., Bader, G. D., et al. (2011). Cytoscape web: an interactive web-based network browser. *Bioinformatics* 26, 2347–2348. doi: 10.1093/bioinformatics/btq430

Maiti, A., Takabe, K., and Hait, N. C. (2017). Metastatic triple-negative breast cancer is dependent on SphKs/S1P signaling for growth and survival. *Cell. Signal.* 32, 85–92. doi: 10.1016/j.cellsig.2017.01.021

Mardinoglu, A., Boren, J., Smith, U., Uhlen, M., and Nielsen, J. (2018). Systems biology in hepatology: approaches and applications. *Nat. Rev. Gastroenterol. Hepatol.* 15, 365–377. doi: 10.1038/s41575-018-0007-8

Mardinoglu, A., and Nielsen, J. (2015). New paradigms for metabolic modeling of human cells. *Curr. Opin. Biotechnol.* 34, 91–97. doi: 10.1016/j.copbio.2014.12.013

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55–62. doi: 10.1038/nature18003

Murphy, C. G., and Dickler, M. N. (2015). The role of CDK4/6 inhibition in breast cancer. *Oncologist* 20, 483–490. doi: 10.1634/theoncologist.2014-0443

Nagahashi, M., Tsuchida, J., Moro, K., Hasegawa, M., Tatsuda, K., Woelfel, I. A., et al. (2016). High levels of sphingolipids in human breast cancer. *J. Surg. Res.* 204, 435–444. doi: 10.1016/j.jss.2016.05.022

Nuncia-Cantarero, M., Martinez-Canales, S., Andrés-Pretel, F., Santpere, G., Ocaña, A., and Galan-Moya, E. M. (2018). Functional transcriptomic annotation and protein–protein interaction network analysis identify NEK2, BIRC5, and TOP2A as potential targets in obese patients with luminal a breast cancer. *Breast Cancer Res. Treat.* 168, 613–623. doi: 10.1007/s10549-017-4652-3

Ocana, A., and Pandiella, A. (2015). Targeting oncogenic vulnerabilities in triple negative breast cancer: biological bases and ongoing clinical studies. *Oncotarget* 8, 22218–22234. doi: 10.18632/oncotarget.14731

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project - intAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Persico, M., Ceol, A., Gavrila, C., Hoffman, R., Florio, A., and Cesareni, G. (2005). HomoMINT: An inferred human network based on orthodology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics.* 6(Suppl. 1):S21. doi: 10.1186/1471-2105-6-S1-S21

Prasad, T. S. K., Kandasamy, K., and Pandey, A. (2009). Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* 577, 67–79. doi: 10.1007/978-1-60761-232-2_6

Purwaha, P., Gu, F., Piyarathna, D. W. B., Rajendiran, T., Ravindran, A., Omilian, A. R., et al. (2018). Unbiased lipidomic profiling of triple-negative breast cancer tissues reveals the association of sphingomyelin levels with patient disease-free survival. *Metabolites* 8, 1–14. doi: 10.3390/metabo8030041

Rietman, E. A., Platig, J., Tuszynski, J. A., and Lakka Klement, G. (2016). Thermodynamic measures of cancer: gibbs free energy and entropy of protein–protein interactions. *J. Biol. Phys.* 42, 339–350. doi: 10.1007/s10867-016-9410-y

Salwinski, L. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 31, 248–250. doi: 10.1093/nar/gkh086

Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). Hippie: integrating protein interaction networks with experiment based quality scores. *PLoS One* 7:e31826. doi: 10.1371/journal.pone.0031826

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA. Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Solzak, J. P., Atale, R. V., Hancock, B. A., Sinn, A. L., Pollok, K. E., Jones, D. R., et al. (2017). Dual PI3K and Wnt pathway inhibition is a synergistic combination against triple negative breast cancer. *NPJ Breast Cancer* 3:17. doi: 10.1038/s41523-017-0016-8

Strott, C. A., and Higashi, Y. (2003). Cholesterol sulfate in human physiology. *J. Lipid Res.* 44, 1268–1278. doi: 10.1194/jlr.R300005-JLR200

Tang, Y., Li, M., Wang, J., Pan, Y., and Wu, F. X. (2015). CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *BioSystems* 127, 67–72. doi: 10.1016/j.biosystems.2014.11.005

Teschendorff, A. E., Banerji, C. R. S., Severini, S., Kuehn, R., and Sollich, P. (2015). Increased signaling entropy in cancer requires the scale-free property of protein interaction networks. *Sci. Rep.* 5:9646. doi: 10.1038/srep09646

Turanli, B., Grøtli, M., Boren, J., Nielsen, J., Uhlen, M., Arga, K. Y., et al. (2018). Drug repositioning for effective prostate cancer treatment. *Front. Physiol.* 9:500. doi: 10.3389/fphys.2018.00500

Turanli, B., Gulfidan, G., and Arga, K. Y. (2017). Transcriptomic-guided drug repositioning supported by a new bioinformatics search tool: geneXpharma. *Omi. A J. Integr. Biol.* 21, 584–591. doi: 10.1089/omi.2017.0127

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10869–10874. doi: 10.1073/pnas.091062498

Ueno, N. T., and Zhang, D. (2011). Targeting EGFR in triple negative breast cancer. *J. Cancer* 2, 324–328. doi: 10.7150/jca.2.324

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 80:357. doi: 10.1126/science.aan2507

Varadan, V., and Anastassiou, D. (2006). Inference of disease-related molecular logic from systems-based microarray analysis. *PLoS Comput. Biol.* 2:e68. doi: 10.1371/journal.pcbi.0020068

Vitali, F., Cohen, L. D., Demartini, A., Amato, A., Eterno, V., Zambelli, A., et al. (2016). A network-based data integration approach to support drug repurposing and multi-Target therapies in triple negative breast cancer. *PLoS One* 11:e0162407. doi: 10.1371/journal.pone.0162407

Warm, M., Kates, R., Große-Onnebrink, E. M., Stoff-Khalili, M., Hoopmann, M., Mallmann, P., et al. (2010). Impact of tumor biology, particularly triple-negative status, on response to pre-operative sequential, dose-dense epirubicin, cyclophosphamide followed by docetaxel in breast cancer. *Anticancer Res.* 30, 4251–4259.

Weigman, V. J., Chao, H.-H., Shabalin, A. A., He, X., Parker, J. S., Nordgard, S. H., et al. (2012). Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res. Treat.* 133, 865–880. doi: 10.1007/s10549-011-1846-y

West, J., Bianconi, G., Severini, S., and Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Sci. Rep.* 2:802. doi: 10.1038/srep00802

Xiong, J., Liu, J., Rayner, S., Li, Y., and Chen, S. (2010). Protein-protein interaction reveals synergistic discrimination of cancer phenotype. *Cancer Inform* 9, 61–66.

Zhang, P., Yang, X., Yin, Q., Yi, J., Shen, W., Zhao, L., et al. (2016). Inhibition of SK4 potassium channels suppresses cell proliferation, migration and the epithelial-mesenchymal transition in triple-negative breast cancer cells. *PLoS One* 11:e0154471. doi: 10.1371/journal.pone.0154471

Zhang, Y. G., Du, Q., Fang, W. G., Jin, M. L., and Tian, X. X. (2008). Tyrphostin AG1478 supresses proliferation and invasion of human breast cancer cells. *Int. J. Oncol.* 33, 595–602. doi: 10.3892/ijo_00000045

Zhou, Y., Lin, S., Tseng, K. F., Han, K., Wang, Y., Gan, Z. H., et al. (2016). Selumetinib suppresses cell proliferation, migration and trigger apoptosis, G1 arrest in triple-negative breast cancer cells. *BMC Cancer* 16:818. doi: 10.1186/s12885-016-2773-4

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 May 2019
doi: 10.3389/fgene.2019.00468

Gene Co-expression Network and Copy Number Variation Analyses Identify Transcription Factors Associated With Multiple Myeloma Progression

*Christina Y. Yu[1,2], Shunian Xiang[3,4], Zhi Huang[2,5], Travis S. Johnson[1,2], Xiaohui Zhan[2,4], Zhi Han[2,6], Mohammad Abu Zaid[2] and Kun Huang[2,6]**

[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, United States, [2] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, United States, [3] Department of Medical and Molecular Genetics, Indiana University, Indianapolis, IN, United States, [4] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China, [5] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States, [6] Regenstrief Institute, Indianapolis, IN, United States

OPEN ACCESS

Edited by:
Victor Jin,
The University of Texas Health Science Center at San Antonio, United States

Reviewed by:
Zhengqing Ouyang,
The Jackson Laboratory for Genomic Medicine, United States
Vishal Acharya,
Institute of Himalayan Bioresource Technology (CSIR), India

***Correspondence:**
Kun Huang
kunhuang@iu.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *01 December 2018*
Accepted: *01 May 2019*
Published: *17 May 2019*

Citation:
Yu CY, Xiang S, Huang Z, Johnson TS, Zhan X, Han Z, Abu Zaid M and Huang K (2019) Gene Co-expression Network and Copy Number Variation Analyses Identify Transcription Factors Associated With Multiple Myeloma Progression. Front. Genet. 10:468. doi: 10.3389/fgene.2019.00468

Multiple myeloma (MM) has two clinical precursor stages of disease: monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM). However, the mechanism of progression is not well understood. Because gene co-expression network analysis is a well-known method for discovering new gene functions and regulatory relationships, we utilized this framework to conduct differential co-expression analysis to identify interesting transcription factors (TFs) in two publicly available datasets. We then used copy number variation (CNV) data from a third public dataset to validate these TFs. First, we identified co-expressed gene modules in two publicly available datasets each containing three conditions: normal, MGUS, and SMM. These modules were assessed for condition-specific gene expression, and then enrichment analysis was conducted on condition-specific modules to identify their biological function and upstream TFs. TFs were assessed for differential gene expression between normal and MM precursors, then validated with CNV analysis to identify candidate genes. Functional enrichment analysis reaffirmed known functional categories in MM pathology, the main one relating to immune function. Enrichment analysis revealed a handful of differentially expressed TFs between normal and either MGUS or SMM in gene expression and/or CNV. Overall, we identified four genes of interest (*MAX*, *TCF4*, *ZNF148*, and *ZNF281*) that aid in our understanding of MM initiation and progression.

Keywords: multiple myeloma, MGUS, SMM, gene co-expression, copy number variation

INTRODUCTION

Multiple myeloma (MM) is a B-cell malignancy caused by the proliferation of aberrant clonal plasma cells that secrete monoclonal immunoglobulin protein, also known as M protein. MM is consistently preceded by a premalignant phase called monoclonal gammopathy of undetermined significance (MGUS) and clinically defined by thresholds in serum M protein and clonal bone

marrow plasma cell content with the absence of hypercalcemia, renal insufficiency, anemia, and bone lesions (known as CRAB features) or amyloidosis relating to the plasma cell proliferative disorder (Landgren et al., 2009; Rajkumar et al., 2014). The risk of developing MGUS is low, thought to be around 3.2% of individuals aged 50 or older and increases to 5.3% for those aged 70 or older (Kyle et al., 2006). An individual with MGUS lives with an increased risk of developing MM at a rate of 1% per year (Kyle et al., 2002). Additionally, there is an intermediate precursor between MGUS and MM known as smoldering multiple myeloma (SMM). This phase is clinically defined by a higher threshold in M-protein or clonal bone marrow plasma cell content with the continued absence of CRAB features (Rajkumar et al., 2014). The risk of progression for SMM increases at a variable rate, as 10% per year for the first 5 years, 3% per year for the next 5 years, and 1% per year in the following 10 years (Kyle et al., 2007). Understanding the biological basis of MM progression from these precursors is still unclear.

Gene expression profiling studies have been applied to MM to identify subgroups and biomarkers in order to better understand the molecular basis of disease, improve prognostic models, and characterize features associated with a high risk of disease progression (Davies et al., 2003; Zhan et al., 2006; Chng et al., 2007a; Shaughnessy et al., 2007; Broyl et al., 2010; Dhodapkar et al., 2014; López-Corral et al., 2014; Shao et al., 2018). A few studies have analyzed the disease precursors using hierarchical clustering and differential expression analysis to identify gene signatures (Davies et al., 2003; Zhan et al., 2007; López-Corral et al., 2014). We approached gene expression profiling analysis from the transcription factor (TF) perspective, using gene co-expression networks (GCNs).

Gene co-expression networks have been widely used in discovery of new gene functions and regulatory relationships (Langfelder and Horvath, 2008; Zhang et al., 2010, 2012; Kais et al., 2011; Yin et al., 2015; Zhang and Huang, 2016; Miao et al., 2018). GCNs have been implemented in a few MM studies albeit these studies focused on differential gene expression and not co-expression (Dong et al., 2015; Wang et al., 2016; Liu et al., 2017). We applied GCN analysis on two publicly available MM datasets to identify regulatory genes specifically associated with or disrupted in MM precursors.

The GCN algorithm we employed is local maximal Quasi-Clique Merger (lmQCM) (Zhang and Huang, 2016), previously developed to mine densely correlated gene modules in weighted GCNs (Zhang et al., 2010; Zhang and Huang, 2016, 2017; Xiang et al., 2018). The advantages that lmQCM has over a similar method such as WGCNA (Langfelder and Horvath, 2008) is the ability to allow genes to belong to more than one module and the ability to produce smaller sized modules many of which are related to copy number variations (CNVs) in cancers (Han et al., 2016; Zhang and Huang, 2016; Xiang et al., 2018).

We further supported and validated our gene expression findings with CNVs from microarray technology based on single-nucleotide polymorphism (SNP) arrays. SNP arrays can be used in numerous ways to identify genomic imbalances (She et al., 2008; López-Corral et al., 2012; Johnson et al., 2016; Mitchell et al., 2016; Mikulasova et al., 2017). We surmised that some gene expression changes from normal to MM precursors can be explained by CNVs in order to better understand the genomic changes of myeloma progression.

MATERIALS AND METHODS

Gene Expression Profiling Datasets: Processing and GCN

We applied an integrative network-based approach to identify modules of co-expressed genes associated with MM precursors. MM microarray datasets GSE5900 and GSE6477 from the Gene Expression Omnibus (GEO) were obtained, annotated, and filtered using the TSUNAMI web-tool[1]. The web-tool retrieved the gene expression matrices via the R package GEOquery. We converted probe IDs to corresponding HGNC symbols according to GEO Platform accession number. In the case of duplicate gene symbols, we retained the one with the largest mean expression value. Probes without gene symbols were removed. We further filtered the data by removing the lowest 20% of genes quantified by absolute average value. The lowest 50% of genes quantified by variance in GSE5900 were removed, while filtering GSE6477 was accomplished by removing the lowest 10% of genes quantified by absolute average value and lowest 10% of genes quantified by variance. We applied a stricter cutoff on GSE5900 because the microarray platform had a much larger probeset than the platform in GSE6477 (54,675 vs. 22,283 probes). This was conducted in order to obtain expression sets with similar numbers of genes. The resulting datasets had 15,388 and 12,530 genes for GSE5900 and GSE6477, respectively. Normalization of the datasets was confirmed by inspecting the boxplots of the samples for consistent median values.

SNP Array Dataset: Processing and CNV Analysis

We obtained raw CEL files from GEO study GSE31339, sequenced on Affymetrix Genome-Wide Human SNP Array 6.0. The CEL files were analyzed by the R package Rawcopy (Mayrhofer et al., 2016) and then aggregated by the following conditions: normal (n = 10), MGUS (n = 20), and SMM (n = 19). SMM sample GSM777173 was removed from our analysis after the sample identity distogram suggested some cell or DNA contamination with other samples (**Supplementary Figure S1**). CNVs were detected in genomic segments using PSCBS, an enhanced method of circular binary segmentation (Bengtsson et al., 2010; Olshen et al., 2011). We used the reference data included in Rawcopy for calculating logarithm (base 2) ratios ($\log_2$ ratios) of genome segmentation. Rawcopy defined the thresholds for copy number gain as segment median $\log_2$ ratio > 0.2 and copy number losses as segment median $\log_2$ ratio < -0.3 (Mayrhofer et al., 2016). The package also annotated probes with their corresponding genes.

[1]https://apps.medgen.iupui.edu/rsc/tsunami/

Gene Co-expression Network Mining

We separated GSE5900 into three datasets: normal (n = 22), MGUS (n = 44), and SMM (n = 12). The GSE6477 dataset was separated in the same fashion into three datasets: normal (n = 15), MGUS (n = 22), and SMM (n = 22). GCN mining was conducted using the R package lmQCM. The lmQCM algorithm has an option for normalizing the edge weights of the weighted co-expression network by setting the sums of both rows and columns of the weight matrix to be all ones similar to the weight normalization in spectral clustering (Ng et al., 2001). Another important parameter for lmQCM is gamma that controls the initiation of new gene modules in the iterative mining process. Here, we applied the edge weight normalization and also tested varying gamma values; the rest of the parameters were kept as the default. The normalization process suppresses high weights between nodes and boosts edges with relatively lower weights, which overcomes the issue of unbalanced edge weights in dense module mining algorithms (Zhang and Huang, 2016). The gamma variable ranges from 0 to 1 and controls for the number of generated modules and the maximum module size. For normalized weights, the suggested range of gamma is 0.3–0.75. A higher gamma results in more total modules with fewer genes in the largest module. A lower gamma results in less total modules with more genes in the largest module. We selected gamma values that struck a balance between these two outcomes and elected to keep the largest module under 500 genes. Different values for gamma were selected to obtain a similar number of modules between the same conditions (i.e., normal, MGUS, or SMM) in GSE5900 and GSE6477. This allows the identified modules to be more comparable between datasets of the same condition. We chose the following gamma values for GSE5900: 0.60 for normal, 0.40 for MGUS, and 0.75 for SMM. The following gamma values were chosen for GSE6477– normal: 0.65, MGUS: 0.60, and SMM: 0.55.

For comparison, we also applied the widely used weighted GCN mining algorithm WGCNA (Langfelder and Horvath, 2008) on the same datasets specifying a minimum module size of 10 and using power 5 or 6 as appropriate, leaving the rest of the settings as default. We then selected the most similar modules from lmQCM and WGCNA and calculated gene-wise Spearman correlations to quantify the co-expression density of each module. The most similar modules were determined using the Jaccard index between lmQCM and WGCNA modules in the same condition, where the Jaccard index is simply defined as the size of the intersection between two gene modules divided by the size of the union of the same two modules.

Identification of Condition-Specific Modules

Condition-specific modules are those in which the expression profile of the genes in one module is more correlated in one condition compared to others (e.g., normal, MGUS, or SMM). We utilized a previously developed metric called Centralized Concordance Index (CCI) that evaluates the co-expression of genes within modules identified from GCN analysis (Han et al., 2016). The CCI describes how strongly genes co-express and is calculated from a subset of gene expression data containing the genes from a module and samples from a single condition. CCI values range from 0 to 1, with a higher number indicating more densely correlated genes. For each gene module identified from lmQCM, we calculated the corresponding CCI in normal, MGUS, and SMM. The CCIs for each module were then compared across the three conditions, and a difference of ∼0.2 in CCI values between MM precursors (MGUS or SMM) and normal were identified as potentially interesting.

Module Similarity Between Datasets

We further reduced our modules of interest by identifying modules with similar genes between GSE5900 and GSE6477. The Jaccard index, described above in Section "Gene Co-expression Network Mining," was used to calculate the similarity of modules in the same conditions between GSE5900 and GSE6477. This calculation was conducted between every pair of modules in each condition: normal, MGUS, and SMM. Each resulting matrix was then transformed into a z-score where the top one percentile of similar module pairs from each condition were kept to filter the list of potentially interesting modules for enrichment analysis.

Functional Enrichment Analysis and Identification of Upstream Regulators

We used the R package enrichR (Kuleshov et al., 2016) to conduct enrichment analysis of the genes in each module of interest. We specified the "GO Biological Process 2017b" and "KEGG 2016" databases for functional and pathway enrichment analyses. For determining the significance of GO and KEGG pathway terms, we used Bonferroni significance cutoffs of 0.05/nMods where nMods is the number of modules corresponding to the specific dataset. For instance, the p-value cutoff for GSE5900 normal-specific data is 0.05/31 = 0.00161. We took GO terms with significant p-values and summarized them using the web-tool REVIGO (Supek et al., 2011).

Using enrichR, we specified the "TRANSFAC and JASPAR PWMs" database to identify TFs that regulate the genes in our modules of interest, using a less stringent Bonferroni cutoff of 0.1/nMods. We then narrowed down the list of TFs by identifying those that were differentially expressed among the three conditions by either gene expression data or CNV segment median data by conducting Mann–Whitney tests between normal and MGUS and between normal and SMM samples.

Network Analysis of TF Targets

We used Ingenuity Pathways Analysis (IPA, Qiagen) for network analysis of TFs and their targets determined from enrichR to explore possible signaling pathways. We conducted *core analyses* (which is a function of IPA) for each TF and its targets, using experimentally observed knowledge in the Ingenuity Knowledge Base and specifying direct and indirect gene relationships in human tissue and cell lines.

RESULTS

lmQCM Produces Smaller-Sized Modules Than WGCNA With Stronger Gene Correlations

Our workflow is shown in **Figure 1**. After applying the lmQCM algorithm using the specified gamma values to the GSE5900 datasets, we obtained 78, 60, and 95 modules for normal, MGUS, and SMM, respectively; module sizes ranged from 10 to 400 genes. In GSE6477, using the specified gamma values, we obtained 79, 85, and 70 modules for the normal, MGUS, and SMM samples, respectively; module sizes ranged from 10 to 352 genes. Applying WGCNA to GSE5900, we obtained 40, 41, and 98 modules for normal, MGUS, and SMM, respectively; module sizes ranged from 11 to 4694 genes. In applying WGCNA to GSE6477, we obtained 34, 99, and 74 modules for normal, MGUS, and SMM, respectively; module sizes ranged from 11 to 4324 genes. Detailed breakdowns by sample type are shown in **Table 1**.

The most similar gene modules were identified from two SMM modules in lmQCM and WGCNA. The lmQCM module contained 224 genes and the WGCNA module contained 393 genes. The Jaccard index was 0.396, with an overlap of 175 genes. Within each respective module, we calculated the Spearman correlation in a gene-wise manner and conducted a two-sided Mann–Whitney test between the absolute value of the correlation coefficients in each population. The correlation coefficients were significantly higher in the lmQCM module (median: 0.399) compared to the WGCNA module (median: 0.322) with a p-value of 2.2E-16 (**Supplementary Figure S2**).

Module Reduction Using CCI and Jaccard Similarity

Normal-, MGUS-, and SMM-specific modules were identified by calculating the CCI difference between normal and MGUS samples and normal and SMM samples and setting a cutoff of around 0.2 CCI difference. This resulted in 68 and 79 normal-specific modules, 45 and 72 MGUS-specific modules, 95 and 63 SMM-specific modules across GSE5900 and GSE6477 datasets, respectively. An example of a normal-specific gene module is visualized using Spearman correlation heatmaps in **Supplementary Figure S3**.

To further reduce modules of interest, we used Jaccard similarity. After module similarity comparison using the Jaccard index, we reduced the interesting modules to more manageable numbers than solely using CCI and were left with 31 and 39 normal-specific modules, 22 and 31 MGUS-specific modules, and 47 and 30 SMM-specific modules across GSE5900 and GSE6477 datasets, respectively. The module sizes ranged from 10 to 400 genes.

Frequency of CNVs Increase From MGUS to SMM

Chromosomes 2, 4, 10, 11, 12, and 21 were mostly unchanged and showed 10% or less allelic imbalance in all conditions. Chromosomes 1q, 3, 5, 6, 7, 9, 15, 18, and 19 were slightly amplified in MGUS and more amplified in SMM, with chromosomes 1q, 5, 9, and 19 showing the highest frequencies of change in SMM of around 40%. For instance, 1q had about 10% of MGUS samples amplified and around 40% of SMM samples amplified. We observed an increased frequency of deletions in chromosomes 1p, 6, 7, 8p, 10, 12p, 13, 14q, 16q, 18, 20, and 22q; the highest deletion frequency was around 25% and was observed in 8p, 13, 16q, and 22q of SMM patients. The CNV landscape across conditions is shown in **Figure 2**.

EnrichR GO Results Are Highly Enriched in Immune-Related Terms

The top GO BP terms from all condition-specific modules are shown in **Supplementary Table S1**. In the normal-specific data, there were 95 significant GO BP terms that appeared in both GSE5900 and GSE6477, the top few being neutrophil degranulation, antigen processing and presentation of exogenous peptide antigen via MHC class II, and antifungal humoral response. These GO terms are mostly related to immune system response.

The MGUS-specific data had 40 significant GO BP terms in common from GSE5900 and GSE6477, with many immune function terms such as positive regulation of B cell activation, response to interferon-alpha, and B cell receptor signaling pathway.

The SMM-specific data shared 125 GO BP terms between GSE5900 and GSE6477 data, the most significant ones relating to the process of transcription and translation. There were also terms related to immune function such as B cell receptor signaling pathway.

Condition-Specific Modules From Four Identified TFs Describe Different Aspects of Myeloma

We identified these TFs as interesting: *MAX*, *TCF4*, *ZNF148*, and *ZNF281*. *MAX* was identified from a normal-specific module, *TCF4* and *ZNF148* were identified from MGUS-specific modules, and *ZNF281* was identified from a SMM-specific module. Three TFs (*MAX*, *TCF4*, and *ZNF148*) were differentially expressed between normal and a MM precursor (MGUS or SMM) in the gene expression datasets and/or the CNV dataset (**Table 2**). While *ZNF281* was not differentially expressed, it showed an interesting increase in copy number gain from normal to MGUS and to SMM.

Module Descriptions

The gene co-expression module containing *MAX* was functionally enriched in bleb assembly and activation of MAPKKK activity involved in innate immune response.

The gene co-expression module containing *ZNF148* was functionally enriched in antigen processing and presentation of exogenous peptide antigen via MHC class II and negative regulation of peptide hormone processing.

In the gene co-expression module containing *TCF4*, multiple assembly complexes containing the genes *GEMIN5*, *PPARGC1A*, and *TEAD1* were significantly enriched. They

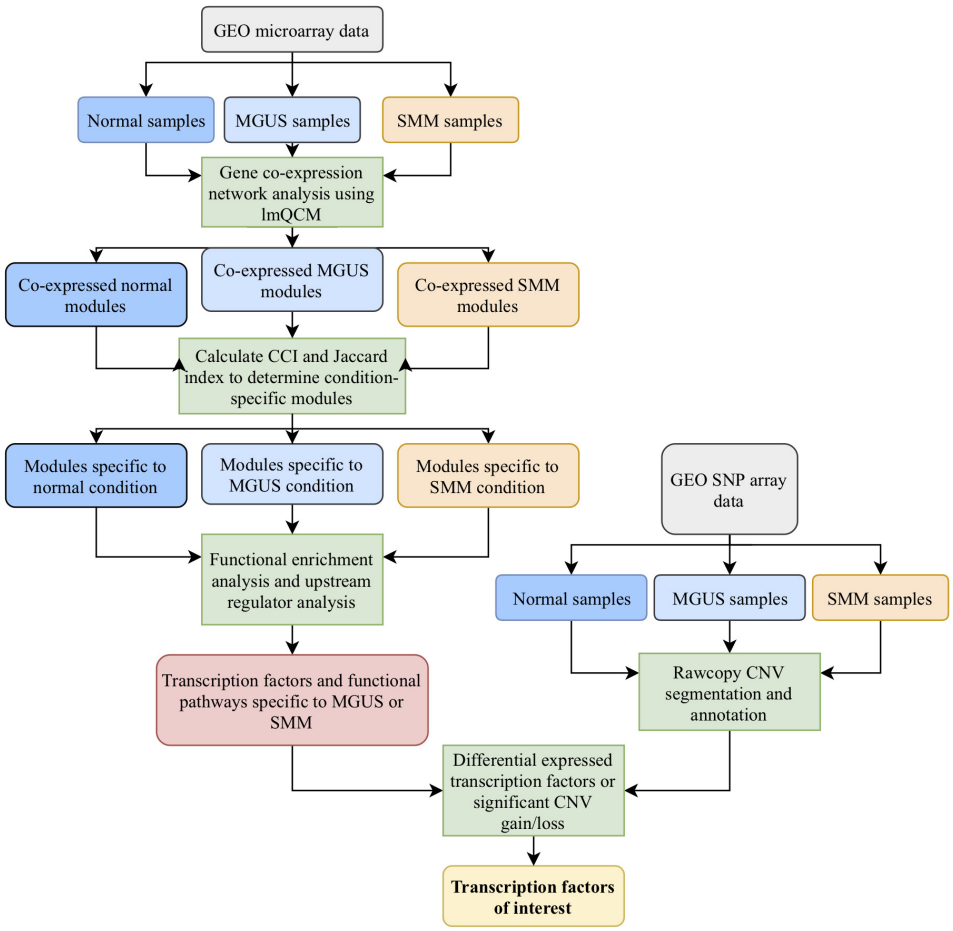

FIGURE 1 | Workflow of the methods employed in this study.

TABLE 1 | GCN results from algorithms lmQCM and WGCNA.

Dataset	Sample type	Sample size	lmQCM total modules	lmQCM module sizes	WGCNA total modules	WGCNA module sizes
GSE5900	Normal	22	78	10–400	40	12–1943
GSE5900	MGUS	12	60	10–332	41	12–4694
GSE5900	SMM	44	95	10–236	98	11–2732
GSE6477	Normal	15	79	10–119	34	11–4324
GSE6477	MGUS	22	85	10–352	99	13–1494
GSE6477	SMM	24	70	10–248	74	11–1652

The total number of resulting modules and size range are detailed by dataset and sample type.

include apoptosome assembly, mitotic checkpoint complex assembly, and Wnt signalosome assembly.

The gene co-expression module containing *ZNF281* is functionally enriched in genes involved in transcription. These include transcription, DNA-templated, transcription from RNA polymerase II promoter, telomeric repeat-containing RNA transcription, and mRNA transcription.

The details of GO BP enrichment results (top enriched terms and *p*-values) for these modules with their corresponding *p*-values are listed in **Supplementary Table S1**.

TFs Exhibit Consistent CNV and Gene Expression Trends During the Course of Myeloma Progression

MAX did not show differential gene expression; however, its copy number significantly decreased in MGUS and SMM compared to normal (*p*-val = 1.17E-05 and 6.10E-04, respectively, **Figures 3A,B**). The CNV pattern showed deletions in MGUS and amplification and deletions in SMM (**Figure 3B**).

ZNF148 was the only TF that showed significantly different CNV aberrations and gene expression, with gene expression and

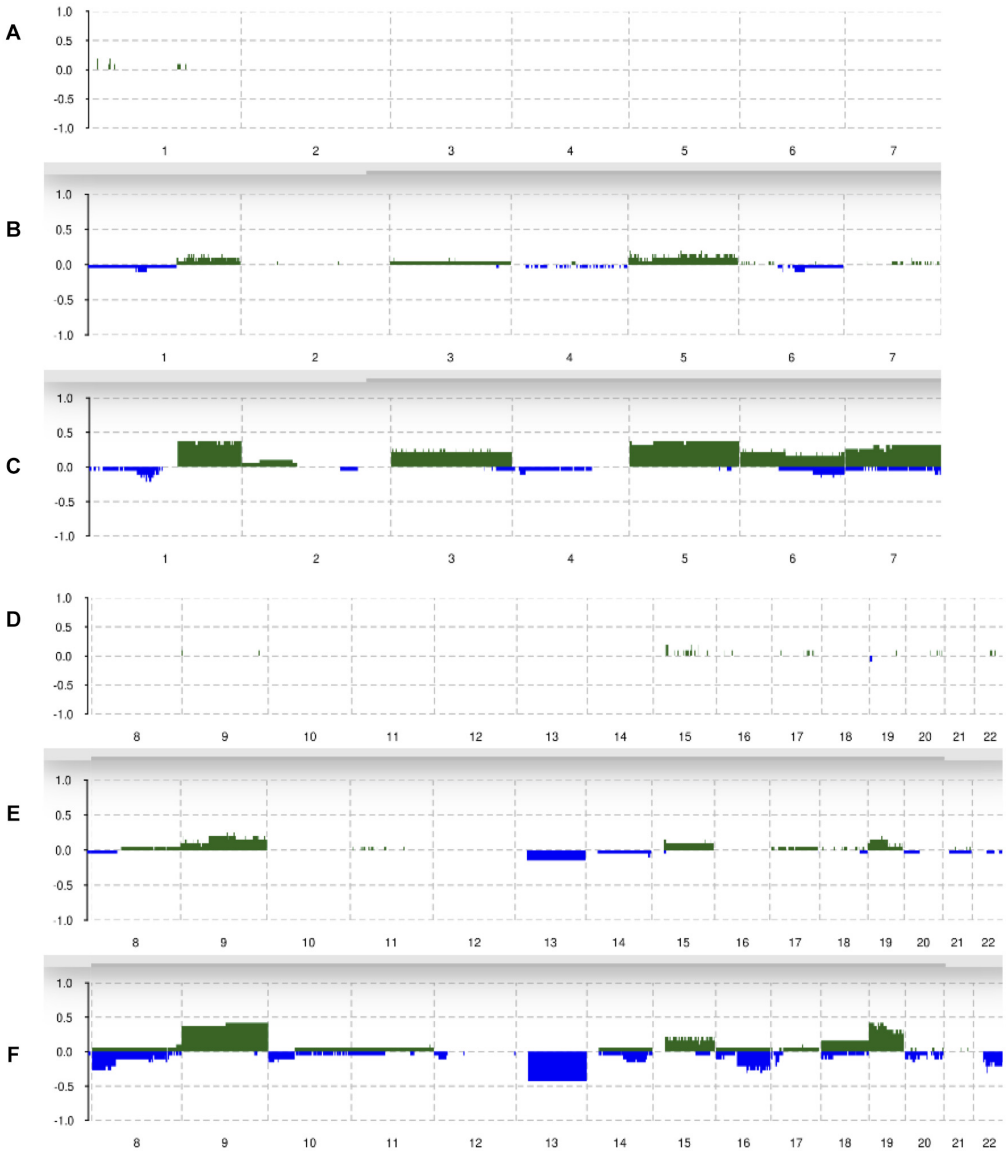

FIGURE 2 | Summary of CNVs across the genome from chromosomes 1–7 in **(A)** normal, **(B)** MGUS, and **(C)** SMM samples. Summary of CNVs across the genome from chromosomes 8–22 in **(D)** normal, **(E)** MGUS, and **(F)** SMM samples. The *y*-axis indicates the frequency of the chromosomal aberration. Green indicates amplification; blue indicates deletion.

copy number amplification both increasing in MGUS and SMM (*p*-val range: 1.75E-02–3.11E-04, **Figures 4A,B**).

TCF4 was differentially expressed between normal/MGUS (*p*-val = 3.65E-03) and normal/SMM (*p*-val = 1.49E-02), with gene expression progressively increasing from MGUS to SMM (**Figure 5A**). In regard to CNVs, *TCF4* exhibited amplifications in MGUS and amplifications and deletions in SMM (**Figure 5B**).

ZNF281 did not show differential gene expression (**Figure 6A**). *ZNF281* showed increasing CNV amplifications from MGUS to SMM, but it was not considered significant by Mann–Whitney tests (**Figure 6B**).

TF Signaling Networks Are Related to Cancer Progression

IPA network analysis showed *MAX* and its targets interact with other TFs *CCNT1*, *KLF10*, and *MYC*. *MAX* is further predicted to target *CCNG2* and *TXNIP*. *BRD4* is shown to regulate expression of *BHLHE40* and *SLC7A2* (**Figure 3C**).

ZNF148 and its targets were shown to interact with TFs *TP53*, *FOXO1*, *SP1*, *TCF3*, *HSF1*, *SMARCA4*, and *E2F1*. Additionally, *CDKN1A* was shown to be a common target of the TFs listed above (**Figure 4C**).

TABLE 2 | Transcription factors of interest, identified from condition-specific modules in normal, MGUS, and SMM samples.

Transcription factor	Chromosomal region	TF targets
MAX	14q12-q24	*NLGN4X*, *VEGFB*, *STMN3*, *CTSW*, *OVOL1*, *SGSH*, *PDP1*, *LYL1*, *DRAM1*, *SH3BP1*, *ZMIZ1*, *NFIC*, *RGL3*, *PTPRCAP*, *FGF13*, *CUEDC1*
ZNF148	3q13-q22	*NLGN4X*, *VEGFB*, *STMN3*, *CTSW*, *OVOL1*, *SGSH*, *PDP1*, *LYL1*, *DRAM1*, *SH3BP1*, *ZMIZ1*, *NFIC*, *RGL3*, *PTPRCAP*, *FGF13*, *CUEDC1*
TCF4	18q11-q23	*UEVLD*, *DSP*, *ALS2CR11*, *NT5E*, *RALYL*, *EFEMP1*, *GEMIN5*, *PPARGC1A*
ZNF281	1q32-q44	*HRK*, *SLC26A1*, *TNXB*, *CRABP2*, *IBA57*, *LOC728392*, *ESPN*, *AGPAT2*, *HS6ST1*, *DLL3*, *IL4I1*, *RGS3*, *FUT7*, *PDLIM2*, *NUP62*, *POLR2F*, *GGT1*, *SLC38A3*, *ZBTB7B*, *POLR2J*, *WNT2*, *MUC6*, *POLR2J3*, *WWTR1*, *PDIA2*, *KLF12*, *ZFHX3*, *ACE*, *POLR2J2*, *SLC2A11*, *GP1BB*, *ABCA3*, *XRCC1*, *FNDC11*, *CTAG2*, *RENBP*, *CLDN5*, *DLG4*, *TRPV4*, *NOX5*, *IGFALS*, *HOXB8*

The chromosomal regions were determined by Rawcopy. The TF targets were identified by enrichR.

TCF4 and its targets were shown to interact with TFs *RUNX2*, *CCND1*, and *HNF4A* in addition to nuclear receptor *PPARG* and junction protein *JUP* (**Figure 5C**).

ZNF281 and its targets were shown to interact with TFs *CREB1*, *CTNNB1*, *RELA*, *NPM1*, and *POU5F1*. *ZNF281* was shown to directly target *GADD45A*. *TP53* was shown to be an intermediate interactor that connected each subnetwork (**Figure 6C**).

DISCUSSION

We conducted GCN analyses on two publicly available MM datasets and identified four TFs by a condition-specific method. This pipeline has previously not been applied to studying MM precursors. Our approach identified TFs expressed in condition-specific gene modules in publicly available MM data. We then validated our TFs with CNV data taken from a third publicly available dataset, looking for genes located on chromosomal segments that showed a consistent trend in aberration from normal to SMM and identified four TFs: *MAX*, *ZNF148*, *TCF4*, and *ZNF281*.

The gene module that *MAX* belongs to was determined to be condition-specific in normal samples. This means that the genes in the module were observed to be co-expressed in normal samples and less so in MGUS and SMM samples. This suggests that *MAX* is dysregulated in MGUS and SMM, which we observed to be true in the CNV data. MAX is known to complex with MYC to regulate transcription (Kato et al., 1992) and MYC is commonly known to be constitutively active in MM. The MAX–MYC relationship has been targeted in previous studies to inhibit c-MYC activity in MM cell lines (Holien et al., 2012). This association appears to conflict with our data, which shows the chromosomal region of *MAX* deleted in some MGUS and SMM samples and decreased gene expression in some SMM samples. An alternate explanation can be found in studies that show MYC can function independently of MAX in pheochromocytoma and small cell lung cancer (Ribon et al., 1994; Romero et al., 2014). MAX-independent expression of MYC in MM and its precursors requires further investigation; a recent abstract identified *MAX* as a tumor suppressor driver gene in MM (Garcia et al., 2017), which is a promising start.

ZNF148 has been implicated in other MM studies (Magrangeas et al., 2003; Dong et al., 2015), but to our knowledge, none have directly associated this gene with MGUS or SMM. The associated chromosomal segment of *ZNF148* was progressively amplified from normal to MGUS and to SMM, corresponding with increased *ZNF148* gene expression. This suggests that this TF is involved as a driver in disease progression earlier than previously thought.

TCF4 was differentially overexpressed in MGUS and SMM compared to normal. *TCF4* was not significantly amplified in MGUS, although this may be due to small sample size. We suggest that copy number amplification may play a part in *TCF4* dysregulation and may be involved in the initiation of MGUS but not SMM. This reasoning is due to the observation that the *TCF4* region is solely amplified in MGUS whereas there is a mix of amplified and deleted regions in SMM. This is consistent with our identification of *TCF4*'s gene module as MGUS-specific. Module enrichment and network analysis suggest Wnt signaling through *TCF4* contributes to *RUNX2* and *CCND1* overexpression. RUNX2 overexpression has been shown to be a driver of MM progression (Li et al., 2014; Trotter et al., 2015). *CCND1* overexpression has typically been observed to occur in MM precursors with chromosomal 11 and 14 translocations (Miura et al., 2003; Zhan et al., 2006). In gastric cancer, CCND1 has been shown to directly interact with TCF4 through the Wnt signaling pathway (Zheng et al., 2018), suggesting that other mechanisms of *CCND1* overexpression may also occur in MM.

ZNF281 was increasingly amplified from MGUS to SMM patients. However, this is not considered statistically significant, possibly due to small sample size. Module enrichment results suggest transcriptional genes are more active in SMM, consistent with the fact that cancer cells require continued transcription in order to grow and proliferate. Increased transcription increases the chances of mutations in the DNA, which would activate tumor suppressor p53 and lead to cell cycle arrest or apoptosis in normal functioning cells. Cancer cells commonly have mutated *TP53* to avoid transcriptional control and apoptosis. However, *TP53* mutations are relatively rare in newly diagnosed MM patients (Chng et al., 2007b; Abdi et al., 2017). Our IPA network analysis suggests that TP53 may be regulated by CTNNB1. A previous study showed CTNNB1 suppressed TP53 in smooth muscle cells during artery

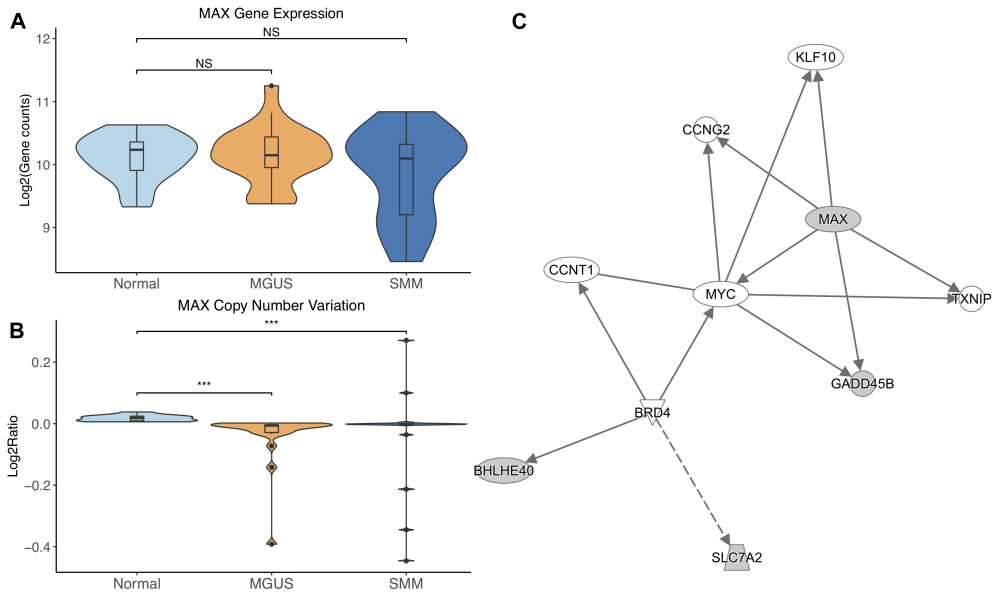

FIGURE 3 | (A) *MAX* expression across sample groups. Mann–Whitney tests between groups showed no significant difference. **(B)** Observations of *MAX* copy number. Mann–Whitney tests showed significant copy number variation between Normal and MGUS (p = 1.17E-05) and between Normal and SMM (p = 6.10E-04). **(C)** A predicted interaction network of *MAX* and its downstream targets. The gray nodes indicate genes from our module and the white nodes are gene interactions defined in IPA. Solid lines between nodes indicate a direct interaction supported by the Ingenuity Knowledge Base while the dashed line indicates an indirect interaction. Significance levels: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

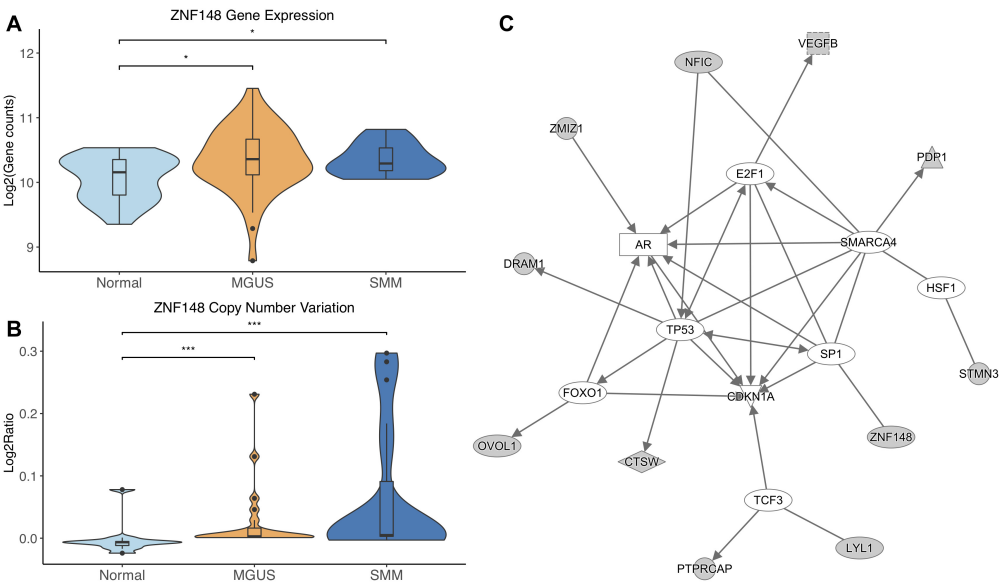

FIGURE 4 | (A) *ZNF148* expression across sample groups. Mann–Whitney tests showed significant differential expression between Normal and MGUS (p = 1.75E-02) and between Normal and SMM (p = 4.05E-02). **(B)** Observations of *ZNF148* copy number. Mann–Whitney tests showed significant copy number variation between Normal and MGUS (p = 4.76E-04) and between Normal and SMM (p = 3.11E-04). **(C)** A predicted interaction network of *ZNF148* and its downstream targets. The gray nodes indicate genes from our module and the white nodes are gene interactions defined in IPA. Solid lines between nodes indicate a direct interaction supported by the Ingenuity Knowledge Base. Significance levels: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

formation (Riascos-Bernal et al., 2016). Something similar may be occurring in MM.

As previously observed by the original authors (López-Corral et al., 2012), the incidence of CNVs progressively increased from normal to MGUS and to SMM. Our analysis with Rawcopy identified similar regions of amplification and deletion from normal to MGUS and from MGUS to SMM. While not all the chromosomal regions were

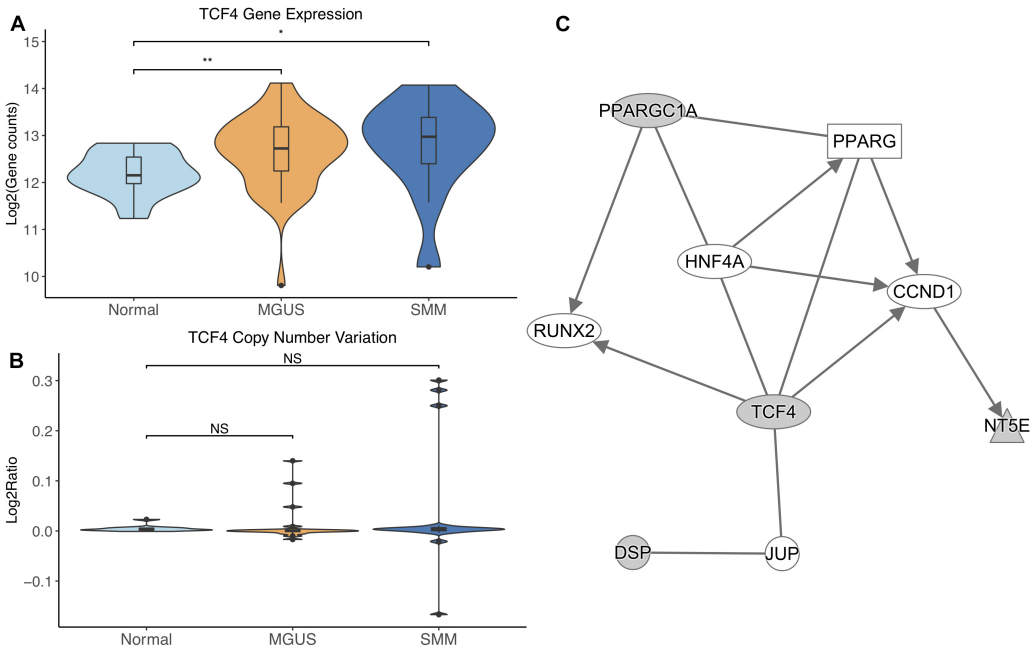

FIGURE 5 | (A) *TCF4* expression across sample groups. Mann–Whitney tests showed significant differential expression between Normal and MGUS (p = 3.65E-03) and between Normal and SMM (p = 1.49E-02). **(B)** Observations of *TCF4* copy number. Mann–Whitney tests showed no significant differences between any groups. **(C)** A predicted interaction network of *TCF4* and its downstream targets. The gray nodes indicate genes from our module and the white nodes are gene interactions defined in IPA. Solid lines between nodes indicate a direct interaction supported by the Ingenuity Knowledge Base. Significance levels: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

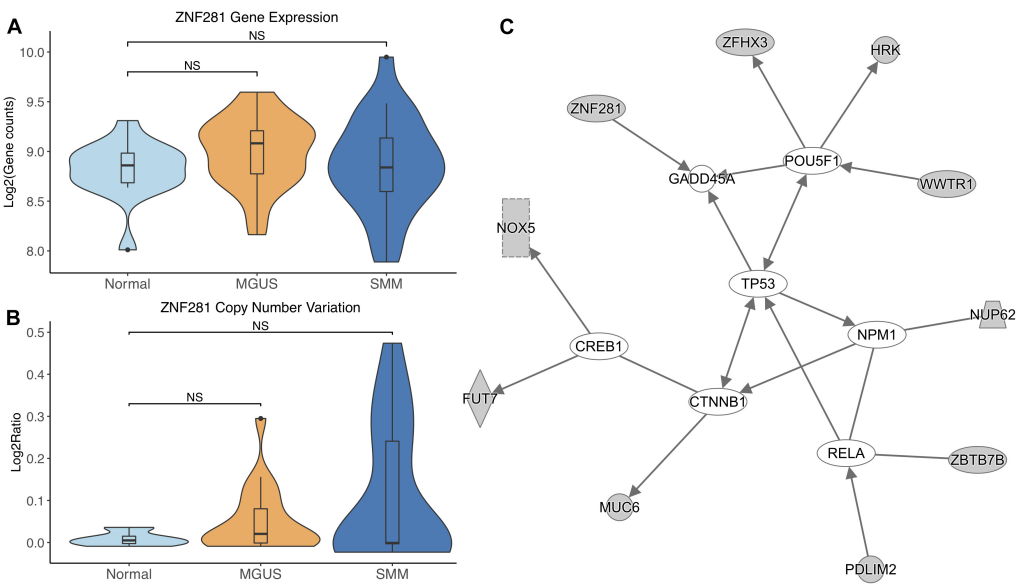

FIGURE 6 | (A) *ZNF281* expression across sample groups. Mann–Whitney tests between groups showed no significant difference. **(B)** Observations of *ZNF281* copy number. Mann–Whitney tests showed no significant differences between any groups. **(C)** A predicted interaction network of *ZNF281* and its downstream targets. The gray nodes indicate genes from our module and the white nodes are gene interactions defined in IPA. Solid lines between nodes indicate a direct interaction supported by the Ingenuity Knowledge Base. Significance levels: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

considered statistically different in the original study, it is visually striking how the frequency of chromosomal aberrations increase in patients from MGUS to SMM. The chromosomal regions of our identified TFs exhibited copy number changes. We suggest that these copy number alterations affect gene expression to an extent. The limitation is that we

cannot offer direct evidence for this, therefore we suggest further exploration of this relationship in the laboratory.

There are other limitations to our study we should acknowledge. We filtered our gene lists down to 12,000–15,000 genes out of ∼22,000 and ∼54,000 microarray probes and identified TFs that showed consistent trends across groups. We may have removed or overlooked genes that could also play a part in myelomagenesis or progression. Although we inferred potential biological mechanisms of the four TFs from literature, the clinical significance of these genes remains to be investigated. Further research can be conducted to assess the pertinence of our TFs in addition to integrating other data modalities into more analyses. Despite these drawbacks, the biological details for these genes appear to have a relevant role in MM initiation and progression.

CONCLUSION

In conclusion, we interrogated the role that TFs have in MM progression using a pipeline of GCN analysis, condition-specific gene module selection, TF enrichment analysis, and CNV analysis. We identified the TFs *MAX*, *ZNF148*, *TCF4*, and *ZNF281* from gene expression data and validated that their CNVs change from normal to MGUS and SMM. We examined the biological relevance of these TFs in MM and suggest further study of these genes in the laboratory.

AUTHOR CONTRIBUTIONS

KH and CY conceptualized the study. CY analyzed and interpreted the multiple myeloma data and was the major contributor in writing the manuscript. TJ, SX, and ZHu contributed to the design of the experiments and data interpretation. MA and XZ critically reviewed the manuscript. ZHa and KH gave research direction.

FUNDING

The funding for this study was partially supported by the NLM grant (4 T15 LM 11270-5), Indiana University Precision Health Initiative, and NCI ITCR U01 CA188547.

ACKNOWLEDGMENTS

The authors thank Dr. Jie Zhang for her suggestions on module analyses and Ms. Megan Metzger for proofreading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00468/full#supplementary-material

FIGURE S1 | Sample identity distograms of SMM samples produced by Rawcopy. **(A)** Distogram including GSM777173 that suggests this sample has some relatedness to other samples. **(B)** Distogram after removing GSM777173.

FIGURE S2 | Gene-wise correlation heatmap of the two most highly similar modules in **(A)** ImQCM (n = 224) and **(B)** WGCNA (n = 393). The correlation coefficients are the absolute value of the Spearman correlation. The median correlation coefficient is higher in ImQCM (0.403) compared to WGCNA (0.344). SCC, Spearman correlation coefficient.

FIGURE S3 | Gene-wise correlation heatmap of a normal-specific gene module. The genes in the module were identified by ImQCM in the normal samples. Gene-wise correlation coefficients are calculated from gene expression in each respective condition: **(A)** Normal, **(B)** MGUS, and **(C)** SMM. The correlation coefficients are the absolute value of the Spearman correlation. The genes are more correlated in normal samples and decrease in correlation in MGUS and SMM samples. The CCI values are 0.697, 0.226, and 0.252, respectively. SCC, Spearman correlation coefficient.

TABLE S1 | GO BP enrichment results identified by enrichR. The most relevant enrichment terms are included along with the enrichment size and p-value associated with the corresponding dataset.

REFERENCES

Abdi, J., Rastgoo, N., Li, L., Chen, W., and Chang, H. (2017). Role of tumor suppressor p53 and micro-RNA interplay in multiple myeloma pathogenesis. *J. Hematol. Oncol.* 10:169. doi: 10.1186/s13045-017-0538-4

Bengtsson, H., Neuvial, P., and Speed, T. P. (2010). TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinform.* 11:245. doi: 10.1186/1471-2105-11-245

Broyl, A., Hose, D., Lokhorst, H., Knegt, Y., de Peeters, J., Jauch, A., et al. (2010). Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood* 116, 2543–2553. doi: 10.1182/blood-2009-12-261032

Chng, W. J., Kumar, S., VanWier, S., Ahmann, G., Price-Troska, T., Henderson, K., et al. (2007a). Molecular dissection of hyperdiploid multiple myeloma by gene expression profiling. *Cancer Res.* 67, 2982–2989. doi: 10.1158/0008-5472.CAN-06-4046

Chng, W. J., Price-Troska, T., Gonzalez-Paz, N., Wier, S. V., Jacobus, S., Blood, E., et al. (2007b). Clinical significance of *TP53* mutation in myeloma. *Leukemia* 21, 582–584. doi: 10.1038/sj.leu.2404524

Davies, F. E., Dring, A. M., Li, C., Rawstron, A. C., Shammas, M. A., O'Connor, S. M., et al. (2003). Insights into the multistep transformation of MGUS to myeloma using microarray expression analysis. *Blood* 102, 4504–4511. doi: 10.1182/blood-2003-01-0016

Dhodapkar, M. V., Sexton, R., Waheed, S., Usmani, S., Papanikolaou, X., Nair, B., et al. (2014). Clinical, genomic, and imaging predictors of myeloma progression from asymptomatic monoclonal gammopathies (SWOG S0120). *Blood* 123, 78–85. doi: 10.1182/blood-2013-07-515239

Dong, L., Chen, C. Y., Ning, B., Xu, D. L., Gao, J. H., Wang, L. L., et al. (2015). Pathway-based network analysis of myeloma tumors: monoclonal gammopathy of unknown significance, smoldering multiple myeloma, and multiple myeloma. *Genet. Mol. Res.* 14, 9571–9584. doi: 10.4238/2015.August.14.20

Garcia, S. B., Ruiz-Heredia, Y., Via', M. D., Gallardo, M., Garitano-Trojaola, A., Zovko, J., et al. (2017). Role of MAX as a tumor suppressor driver gene in multiple myeloma. *Blood* 130:4347.

Han, Z., Zhang, J., Sun, G., Liu, G., and Huang, K. (2016). A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules. *BMC Genom.* 17:519. doi: 10.1186/s12864-016-2912-y

Holien, T., Våtsveen, T. K., Hella, H., Waage, A., and Sundan, A. (2012). Addiction to c-MYC in multiple myeloma. *Blood* 120, 2450–2453. doi: 10.1182/blood-2011-08-371567

Johnson, D. C., Weinhold, N., Mitchell, J. S., Chen, B., Kaiser, M., Begum, D. B., et al. (2016). Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma. *Nat. Commun.* 7:10290. doi: 10.1038/ncomms10290

Kais, Z., Barsky, S. H., Mathsyaraja, H., Zha, A., Ransburgh, D. J. R., He, G., et al. (2011). KIAA0101 interacts with BRCA1 and regulates centrosome number. *Mol. Cancer Res.* 9, 1091–1099. doi: 10.1158/1541-7786.MCR-10-0503

Kato, G. J., Lee, W. M., Chen, L. L., and Dang, C. V. (1992). Max: functional domains and interaction with c-Myc. *Genes Dev.* 6, 81–92. doi: 10.1101/gad.6.1.81

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Kyle, R. A., Remstein, E. D., Therneau, T. M., Dispenzieri, A., Kurtin, P. J., Hodnefield, J. M., et al. (2007). Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma. *N. Engl. J. Med.* 356, 2582–2590. doi: 10.1056/NEJMoa070389

Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Larson, D. R., Plevak, M. F., Offord, J. R., et al. (2006). Prevalence of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* 354, 1362–1369. doi: 10.1056/NEJMoa054494

Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Offord, J. R., Larson, D. R., Plevak, M. F., et al. (2002). A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* 346, 564–569. doi: 10.1056/NEJMoa01133202

Landgren, O., Kyle, R. A., Pfeiffer, R. M., Katzmann, J. A., Caporaso, N. E., Hayes, R. B., et al. (2009). Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* 113, 5412–5417. doi: 10.1182/blood-2008-12-194241

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9:559. doi: 10.1186/1471-2105-9-559

Li, M., Trotter, T. N., Peker, D., Rowan, P. D., Qianying, P., Larry, S. J., et al. (2014). Myeloma cell-derived Runx2 promotes myeloma progression and bone-homing. *Blood* 124, 724–724. doi: 10.1182/blood-2014-12-613968

Liu, Z., Huang, J., Zhong, Q., She, Y., Ou, R., Li, C., et al. (2017). Network-based analysis of the molecular mechanisms of multiple myeloma and monoclonal gammopathy of undetermined significance. *Oncol. Lett.* 14, 4167–4175. doi: 10.3892/ol.2017.6723

López-Corral, L., Corchete, L. A., Sarasquete, M. E., Mateos, M. V., García-Sanz, R., Fermiñán, E., et al. (2014). Transcriptome analysis reveals molecular profiles associated with evolving steps of monoclonal gammopathies. *Haematologica* 99, 1365–1372. doi: 10.3324/haematol.2013.087809

López-Corral, L., Sarasquete, M. E., Beà, S., García-Sanz, R., Mateos, M. V., Corchete, L. A., et al. (2012). SNP-based mapping arrays reveal high genomic complexity in monoclonal gammopathies, from MGUS to myeloma status. *Leukemia* 26, 2521–2529. doi: 10.1038/leu.2012.128

Magrangeas, F., Nasser, V., Avet-Loiseau, H., Loriod, B., Decaux, O., Granjeaud, S., et al. (2003). Gene expression profiling of multiple myeloma reveals molecular portraits in relation to the pathogenesis of the disease. *Blood* 101, 4998–5006. doi: 10.1182/blood-2002-11-3385

Mayrhofer, M., Viklund, B., and Isaksson, A. (2016). Rawcopy: improved copy number analysis with Affymetrix arrays. *Sci. Rep.* 6:36158. doi: 10.1038/srep36158

Miao, L., Yin, R.-X., Pan, S.-L., Yang, S., Yang, D.-Z., and Lin, W.-X. (2018). Weighted gene co-expression network analysis identifies specific modules and hub genes related to hyperlipidemia. *Cell. Physiol. Biochem.* 48, 1151–1163. doi: 10.1159/000491982

Mikulasova, A., Wardell, C. P., Murison, A., Boyle, E. M., Jackson, G. H., Smetana, J., et al. (2017). The spectrum of somatic mutations in monoclonal gammopathy of undetermined significance indicates a less complex genomic landscape than that in multiple myeloma. *Haematologica* 102, 1617–1625. doi: 10.3324/haematol.2017.163766

Mitchell, J. S., Li, N., Weinhold, N., Försti, A., Ali, M., van Duin, M., et al. (2016). Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nat. Commun.* 7:12050. doi: 10.1038/ncomms12050

Miura, K., Iida, S., Hanamura, I., Kato, M., Banno, S., Ishida, T., et al. (2003). Frequent occurrence of CCND1 deregulation in patients with early stages of plasma cell dyscrasia. *Cancer Sci.* 94, 350–354. doi: 10.1111/j.1349-7006.2003.tb01445.x

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). "). On spectral clustering: analysis and an algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic NIPS'01*, (Cambridge, MA: MIT Press), 849–856.

Olshen, A. B., Bengtsson, H., Neuvial, P., Spellman, P. T., Olshen, R. A., and Seshan, V. E. (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* 27, 2038–2046. doi: 10.1093/bioinformatics/btr329

Rajkumar, S. V., Dimopoulos, M. A., Palumbo, A., Blade, J., Merlini, G., Mateos, M.-V., et al. (2014). International myeloma working group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* 15, e538–e548. doi: 10.1016/S1470-2045(14)70442-5

Riascos-Bernal, D. F., Chinnasamy, P., Cao, L. L., Dunaway, C. M., Valenta, T., Basler, K., et al. (2016). β-Catenin C-terminal signals suppress p53 and are essential for artery formation. *Nat. Commun.* 7:12389. doi: 10.1038/ncomms12389

Ribon, V., Leff, T., and Saltiel, A. R. (1994). c-Myc does not require max for transcriptional activity in PC-12 cells. *Mol. Cell. Neurosci.* 5, 277–282. doi: 10.1006/mcne.1994.1032

Romero, O. A., Torres-Diz, M., Pros, E., Savola, S., Gomez, A., Moran, S., et al. (2014). MAX inactivation in small cell lung cancer disrupts MYC–SWI/SNF programs and is synthetic lethal with BRG1. *Cancer Discov.* 4, 292–303. doi: 10.1158/2159-8290.CD-13-0799

Shao, W., Cheng, J., Sun, L., Han, Z., Feng, Q., Zhang, D., et al. (2018). *Ordinal Multi-modal Feature Selection for Survival Analysis of Early-Stage Renal Cancer*. Cham: Springer, 648–656.

Shaughnessy, J. D., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., et al. (2007). A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* 109, 2276–2284. doi: 10.1182/blood-2006-07-038430

She, X., Cheng, Z., Zöllner, S., Church, D. M., and Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nat. Genet.* 40, 909–914. doi: 10.1038/ng.172

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

Trotter, T. N., Li, M., Pan, Q., Peker, D., Rowan, P. D., Li, J., et al. (2015). Myeloma cell–derived Runx2 promotes myeloma progression in bone. *Blood* 125, 3598–3608. doi: 10.1182/blood-2014-12-613968

Wang, X.-G., Peng, Y., Song, X.-L., and Lan, J.-P. (2016). Identification potential biomarkers and therapeutic agents in multiple myeloma based on bioinformatics analysis. *Eur. Rev. Med. Pharmacol. Sci.* 20, 810–817.

Xiang, S., Huang, Z., Wang, T., Han, Z., Yu, C. Y., Ni, D., et al. (2018). Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. *BMC Med. Genom.* 11:115. doi: 10.1186/s12920-018-0431-1

Yin, D.-X., Zhao, H.-M., Sun, D.-J., Yao, J., and Ding, D.-Y. (2015). Identification of candidate target genes for human peripheral arterial disease using weighted gene co-expression network analysis. *Mol. Med. Rep.* 12, 8107–8112. doi: 10.3892/mmr.2015.4450

Zhan, F., Barlogie, B., Arzoumanian, V., Huang, Y., Williams, D. R., Hollmig, K., et al. (2007). Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis. *Blood* 109, 1692–1700. doi: 10.1182/blood-2006-07-037077

Zhan, F., Huang, Y., Colla, S., Stewart, J. P., Hanamura, I., Gupta, S., et al. (2006). The molecular classification of multiple myeloma. *Blood* 108, 2020–2028. doi: 10.1182/blood-2005-11-013458

Zhang, J., and Huang, K. (2016). Normalized lmQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. *Cancer Inform.* 13, 137–146. doi: 10.4137/CIN.S14021

Zhang, J., and Huang, K. (2017). Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *BMC Genom.* 18:1045. doi: 10.1186/s12864-016-3259-0

Zhang, J., Lu, K., Xiang, Y., Islam, M., Kotian, S., Kais, Z., et al. (2012). Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Comput. Biol.* 8:e1002656. doi: 10.1371/journal.pcbi.1002656

Zhang, J., Xiang, Y., Ding, L., Keen-Circle, K., Borlawsky, T. B., Ozer, H. G., et al. (2010). Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. *BMC Bioinform.* 11(Suppl. 9):S5. doi: 10.1186/1471-2105-11-S9-S5

Zheng, L., Liang, X., Li, S., Li, T., Shang, W., Ma, L., et al. (2018). CHAF1A interacts with TCF4 to promote gastric carcinogenesis via upregulation of c-MYC and CCND1 expression. *EBioMedicine* 38, 69–78. doi: 10.1016/j.ebiom.2018.11.009

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 May 2019
doi: 10.3389/fgene.2019.00489

Abundance of HPV L1 Intra-Genotype Variants With Capsid Epitopic Modifications Found Within Low- and High-Grade Pap Smears With Potential Implications for Vaccinology

Jane Shen-Gunther[1], Hong Cai[2,3], Hao Zhang[2] and Yufeng Wang[2,3]**

[1] Gynecologic Oncology and Clinical Investigation, Department of Clinical Investigation, Brooke Army Medical Center, Fort Sam Houston, TX, United States, [2] Department of Biology, University of Texas at San Antonio, San Antonio, TX, United States, [3] South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX, United States

Background: The aim of this study was to explore the Human Papillomavirus (HPV) genotype composition and intra-genotype variants within individual samples of low- and high-grade cervical cytology by deep sequencing. Clinical, cytological, sequencing, and functional/structural data were forged into an integrated variant profiling pipeline for the detection of potentially vaccine-resistant genotypes or variants.

Methods: Low- and high-grade intraepithelial lesion (LSIL and HSIL) cytology samples with +HPV were subjected to amplicon (L1 gene fragment) sequencing by dideoxy (Sanger) and deep methods. Taxonomic, abundance, diversity, and phylogenetic analyses were conducted to determine HPV genotypes/sub-lineages, relative abundance, species diversity and phylogenetic distances within and between samples. Variant detection and functional analysis of translated L1 amino acid sequences determined structural variations of interest.

Results: Pure and mixed HPV infections were common among LSIL (n = 6) and HSIL (n = 6) samples. Taxonomic profiling revealed loss of species richness and gain of dominance by carcinogenic genotypes in HSIL samples. Phylogenetic analysis showed excellent correlation between HPV-type specific genetic distances and carcinogenic potential. For combined LSIL/HSIL samples (n = 12), 11 HPV genotypes and 417 mutations were detected: 375 single-nucleotide variants (SNV), 29 insertion/deletion (indel), 12 multi-nucleotide variants (MNV), and 1 replacement variant. The proportion of nonsynonymous mutations was lower for HSIL (0.38) than for LSIL samples (0.51)

OPEN ACCESS

Edited by:
Junbai Wang,
Oslo University Hospital, Norway

Reviewed by:
Xiangyun Wang,
Novartis, United States
Manoj Kumar,
Institute of Microbial Technology (CSIR), India
Luis Felipe Jave-Suarez,
Centro de Investigación Biomédica de Occidente (CIBO), Mexico

***Correspondence:**
Jane Shen-Gunther
jane.shengunther.mil@mail.mil;
shengunther@livemail.uthscsa.edu
Yufeng Wang
yufeng.wang@utsa.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *04 December 2018*
Accepted: *06 May 2019*
Published: *24 May 2019*

Citation:
Shen-Gunther J, Cai H, Zhang H and Wang Y (2019) Abundance of HPV L1 Intra-Genotype Variants With Capsid Epitopic Modifications Found Within Low- and High-Grade Pap Smears With Potential Implications for Vaccinology. Front. Genet. 10:489. doi: 10.3389/fgene.2019.00489

($p < 0.05$). HPV variant analysis pinpointed nucleotide-level mutations and amino acid-level structural modifications.

Conclusion: HPV L1 intra-host and intra-genotype variants are abundant in LSIL and HSIL samples with potential functional/structural consequences. An integrated multi-omics approach to variant analysis may provide a sensitive and practical means of detecting changes in HPV evolution and dynamics within individuals or populations.

Keywords: human papillomavirus, HPV genotyping, HSIL, late major capsid protein L1, metagenome, next generation sequencing, protein structure prediction, vaccine

INTRODUCTION

In 1932, Richard Shope isolated the first papillomavirus (PV) from crude extracts of "warty" tumors found on the skin of a wild cottontail rabbit (Shope, 1932). Since then, 183 animal and 225 HPV have been discovered and classified in The Papillomavirus Episteme (PaVE) (Van Doorslaer et al., 2017)[1] With the advent of metagenomic sequencing, the rate of HPV discovery has accelerated rapidly (Bzhalava et al., 2014) and the resolution of HPV viromes and variants have sharpened immensely (Shen-Gunther et al., 2017) to allow in-depth analysis of genetic variations and functional consequences (van der Weele et al., 2017; Dube Mandishora et al., 2018).

The PV is believed to have co-evolved with their hosts over 350 million years (Doorbar et al., 2015). Through phylogenetic analysis, Chen et al. (2018a) demonstrated that viral niche-adaptation to host ecosystems (tissue tropism) anteceded viral-host codivergence. The PV-host tissue tropism apparently played a vital role in shaping the molecular evolution of oncogenic HPV from archaic hominins to modern humans. HPV-16, an extraordinary result of evolutionary processes over the last 40 million years (Chen et al., 2018a) has emerged as a highly potent carcinogen with a predilection for human mucosa. HPV-16 is now the leading cause of invasive cervical cancer and other cancers of the oropharyngeal and anogenital tracts (Bosch et al., 2013).

The HPV genome is a ∼8,000 base pair (bp), double stranded, circular DNA packaged within a protein capsid. The prototypical genome encodes 6 early genes (E1, E2, E4, E5, E6, and E7) and 2 late genes (L1 and L2) (Van Doorslaer et al., 2017). Specifically, the L1 gene encodes the major capsid protein which forms a pentameric capsomer that self-arranges into a 72-subunit icosahedral capsid. The capsid is essential for viral binding and entry into host-specific tissues (Buck et al., 2013). Furthermore, the L1 coding sequences of the immunogenic surface loops are distinctively poorly conserved due to selective pressures for mutagenesis and immune evasion (Buck et al., 2013).

Recently, whole-genome Sanger and deep sequencing studies have shown a surprisingly high level of intra-host diversity of HPV-16, −18, −52, and −58 (van der Weele et al., 2017, 2018; Hirose et al., 2018). Extensive intra-host HPV L1 sequence variability in 35 HPV genotypes was also discovered in samples from Zimbabwean women by deep sequencing (Dube Mandishora et al., 2018). Such intra-host viral sequence variability is believed to be caused by error-prone host replication machinery used for viral replication and HPV-induced APOBEC deaminase activity with ensuing selective shaping by host tissues and immune responses (Dube Mandishora et al., 2018; Hirose et al., 2018). These remarkable findings of L1 genetic variability are clinically important due to potential structural changes on the epitopes of virions arising from nonsynonymous mutations. The result may be ineffectual binding by host neutralizing antibodies induced by either natural infections or prophylactic vaccines (Bissett et al., 2016; El-Aliani et al., 2017).

Using a multi-omics approach, we aimed to explore the HPV genotype composition and intra-genotype variants within individual samples of low and high-grade cervical cytology. We also focused on the genetic and translated amino acid sequence variations of L1 informed by next-generation sequencing (NGS) for mapping onto the structure of HPV antigenic loops as a means of variant profiling and visualization.

MATERIALS AND METHODS

Subjects and Samples

Residual liquid-based cervical cytology samples were consecutively procured from the Department of Pathology after completion of cytological diagnosis. Demographic and cytohistological data were abstracted from the electronic health record (AHLTA) of the Department of Defense (DoD) and linked to each sample. In our previous study, three categories of samples, i.e., negative for intraepithelial lesion or malignancy (NILM), low-grade squamous intraepithelial lesion (LSIL) and high-grade squamous intraepithelial lesion (HSIL) were collected for HPV genotyping and DNA methylation analysis (Shen-Gunther et al., 2016). For this pilot study, we randomly selected a subset of HPV-positive LSIL ($n = 6$) and HSIL ($n = 6$) for characterization and comparison of viral diversity and variant analysis.

Abbreviations: BLAST, Basic Local Alignment Search Tool; CIN, cervical intraepithelial neoplasia; HPV, Human Papillomavirus; HSIL, high-grade squamous intraepithelial lesion; IARC, International Agency for Research on Cancer; Indel, insertion/deletion; LSIL, low-grade squamous intraepithelial lesion; MCL, Maximum Composite Likelihood; ML, Maximum Likelihood; MNV, multi-nucleotide variant; NGS, next-generation sequencing; NJ, Neighbor-Joining; ORF, open reading frame; Pap, Papanicolaou smear; PaVE, papillomavirus genome database; PCoA, principal coordinate analysis; PDB, Protein Data Bank; QC, Quality Control (QC); SNV, single nucleotide variant; WHO, World Health Organization.

[1]https://pave.niaid.nih.gov

HPV L1 DNA Amplification and Deep Sequencing

DNA extraction from residual liquid-based cervical cytology for HPV DNA amplification and deep sequencing was performed as described previously (Shen-Gunther et al., 2017). Briefly, HPV DNA was amplified using the consensus primer set: MY09/11 to target a 450 bp region (corresponding to flanking nucleotide positions 6584/7035 on HPV-16) of the L1 gene for genotype identification (Shen-Gunther and Yu, 2011). The PCR products were then purified for construction of DNA libraries using the Nextera XT kit (Illumina). Each DNA sample (1 ng) with a standardized concentration of 0.1–0.2 ng/μL was "tagmented" (fragmented and tagged with sequencing adapters) and barcoded with dual index adaptors. The DNA libraries were normalized quantitatively for equal representation from each sample prior to pooling and sequencing. Paired-end bi-directional sequencing (2 × 300 bp) was performed on the MiSeq (Illumina) instrument using the MiSeq Reagent Kit v3 (600-cycle) for bridge amplification. Quality sequences were subjected to nucleotide BLAST (Altschul et al., 1997) against the HPV sequences in the papillomavirus genome database (PaVE) (Van Doorslaer et al., 2017)[2], to determine the HPV genotype(s) (Shen-Gunther and Yu, 2011).

The PCR products were concurrently subjected to dideoxy (Sanger) sequencing for validation of deep-sequenced results. Briefly, amplicons (∼200 ng DNA/sample) were sequenced using primer MY11 at Eurofins Operon (United States). The resulting quality sequences were BLAST aligned for HPV genotyping as described above.

Next-Generation Sequencing (NGS) Data Analysis, Genotyping, and Taxonomic Profiling

The pre-configured, automated Quality Control (QC) workflow implemented in Illumina MiSeq output a series of QC metrics including the summary statistics of the reads, and the Phred quality scores Q which correspond to the base-calling error probabilities (Ewing and Green, 1998; Ewing et al., 1998). The reads were processed using the CLC Genomics Workbench 11.0.1 (QIAGEN). The Core NGS workflow was implemented, including: (1) Preprocessing reads with quality trimming based on quality scores with a limit cutoff 0.05, and the ambiguity number $\leq$2, and adapter trimming. (2) Merging overlapping pairs to improve the read quality. The parameter setting was mismatch cost 2, gap cost 3, and minimum score 8. (3) Mapping to the nonredundant HPV reference genome database, which was constructed based on the collection and annotation of the PaVE database (Van Doorslaer et al., 2017). Mapping parameters included read alignment match score 1, mismatch penalty 2, linear gap cost for insertion or deletion of 3. (4) Taxonomic profiling. The Microbial Genomics Module was implemented to perform qualification by assigning the read to a HPV genotype if a match is found and quantification of the abundance of each qualified HPV genotype to generate an abundance table for each sample. Reads matching to the host genome were filtered.

Diversity Analysis of HPV Communities in LSIL and HSIL Samples

The diversity of the HPV genotypes was analyzed for each sample using the Microbial Genomics Module of the CLC Genomics Workbench 11.01.1 (QIAGEN). α-diversity of the HPV communities was computed to measure within-sample variation by (1) the Simpson's index (Simpson, 1949): $SI = 1 - \sum_{i=1}^{n} p_i^2$, and (2) Shannon entropy (Shannon, 1948): $H = \sum_{1}^{n} p_i \log_2 p_i$, where n was the number of HPV genotypes found in the sample, and p_i was the proportion of reads that were identified as the i^{th} HPV genotype. β-diversity analysis was performed with the principle coordinate analysis (PCoA) of Bray-Curtis distances (Bray and Curtis, 1957): $B = \frac{\sum_{i=1}^{n} |x_i^A - x_i^B|}{\sum_{i=1}^{n} (x_i^A - x_i^B)}$, where n is the number of operational taxonomic unit (OTU) i and x_i^A and x_i^B are the respective abundances of OTU i in samples A and B, to measure the dissimilarity or "distance" of HPV genotype composition between samples. Principal component analysis (PCA) was used to determine the correlative relationship between variables (HPV genotypes) in the LSIL or HSIL group. PCA was performed on the covariance matrix of natural log-transformed abundance data [ln (n +1)] of HPV genotypes within each sample (Rencher and Christensen, 2012). Log transformation was applied to reduce the influence (skewness) of highly abundant genotypes. PCA was performed using STATA/IC 15.0 (StataCorp).

Phylogenetic Analysis and Tree Construction of HPV Genotypes

Multiple alignment of consensus sequences of each HPV genotype detected in the HSIL and LSIL samples was obtained using the T-coffee program (Notredame et al., 2000). The evolutionary history of the HPV L1 sequences was inferred by using the Maximum Likelihood (ML) method (Felsenstein, 1981) and Tamura-Nei model (Tamura and Nei, 1993). Initial trees for the heuristic search were obtained automatically by applying Neighbor-Joining (NJ) (Saitou and Nei, 1987) and BioNJ (Gascuel, 1997) algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The bootstrap resampling with 1,000 pseudo-replicates was carried out to assess support for each individual branch (Felsenstein, 1985). Bootstrap values of <50% were collapsed and treated as unresolved polytomies. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018).

Detection of HPV L1 Sequence Variants, Amino Acid Alterations, and Structural Modifications

Variants were detected by comparing to reference sequences of each HPV type, using the Low Frequency Variant Detection Module in the CLC Bio Genomics Workbench 11.0.1 (QIAGEN), where an error model was included to exclude variants that were likely due to sequencing errors. Variants were classified

[2]https://pave.niaid.nih.gov

into four categories: SNV, MNV, indel, or replacement of one or more bases.

The functional consequences of detected variants in each sample were inferred based on the predicted changes at the codon level. These changes were classified as nonsynonymous (with amino acid changes), synonymous (silent mutation without alteration in amino acid designation), or indels which can lead to reading frame shift or early stop codon. To map the amino acid changes to protein structure, BLAST searches were conducted to identify the homologous HPV L1 structure(s) collected in the Protein Data Bank (PDB)[3] (Berman et al., 2000). 3D models showing the structure of HPV L1 protein with variant and reference sites was created using the CLC Bio Genomics Workbench 11.0.1 (QIAGEN). Another protein structural feature, i.e., surface probability, useful for identification of antigenic determinants was calculated using the protein module of CLC Bio Genomics Workbench 11.0.1 (QIAGEN). The surface probability (accessibility) of an amino acid is predicted using Emini's formula: $S_n = [\prod_{i=1}^{6} \delta_{n+4-i}] * (0.37)^{-6}$ where S_n is the surface probability of amino acid n equating to the normalized product of fractional surface probabilities (δ_x) of six amino acids flanked by positions $n-2$ and $n+3$ (Emini et al., 1985). The S_n of a random hexapeptide is 1.0 (threshold); a value >1.0 indicates increased surface probability.

HPV Taxonomy and Carcinogenicity Classifications

The genotype classification of PV is based on the DNA sequence of the L1 gene (de Villiers et al., 2004; Bernard et al., 2010). The definitions for taxonomic ranks (PaVE) are as follows: (1) Genera: members of the same genus share >60% nucleotide sequence identity in the L1 open reading frame (ORF), (2) Species: PV types within a species share between 71 and 89% nucleotide identity within the complete L1 ORF, (3) Genotypes: PV of the same type share ≥90% nucleotide sequence identity, (4) Variants: <2% sequence difference from a known type, (5) Variant lineage: PV genomes with approximately 1.0% nucleotide sequence difference (proposed nomenclature), and (6) Sub-lineage: PV genomes with 0.5–1.0% nucleotide sequence difference (proposed nomenclature).

The World Health Organization (WHO) International Agency for Research on Cancer (IARC) Working Group assessed carcinogenic potential of HPV types and classified them into three categories (International Agency for Research on Cancer, 2012) (1) carcinogenic: HPV types 16, 31, 33, 35, 52, and 58 in α-9, HPV types 18, 39, 45, 59, and 68 in α−7, HPV type 51 in α−5, HPV type 56 in α−6, (2) possibly carcinogenic: HPV types 26, 69, and 82 in α-5, HPV types 30, 53, 66 in α-6, HPV types 70, 85, and 97 in α−7, HPV types 67 in α−9, and HPV types 34 and 73 in α−11, and (3) not classifiable/not carcinogenic: The viruses in this group are from α−1, −2, −3, −4, −8, −10, −13, −14/15. HPV types 6 and 11 were not classifiable, and all others were probably not carcinogenic (Schiffman et al., 2009; Bernard et al., 2010).

[3]https://www.rcsb.org/

RESULTS

Deep Sequencing Resolved Viromes and Genotypes of Mixed HPV Infections for Differentiation Between LSIL and HSIL Samples

This study included 12 cytology samples, classified as LSIL (n = 6) and HSIL (n = 6) (**Table 1**). The median age of the cohort was 28 years (range, 21–40). For the LSIL group, the median age [34 years (range, 22–40)] was slightly greater than that of the HSIL group [27 years (range, 21–29)]. Histological results from cervical biopsies or excisions were available for 9 of 12 (75%) samples. Histological validation of the cytology samples showed overall good agreement (78%) (**Table 1**).

Both traditional Sanger and NGS platforms were used to detect HPV genotypes and sub-lineages within each sample. Sanger sequencing resolved the single dominant HPV genotype within each sample. Compared to Sanger sequencing, NGS achieved a better resolution in detection of mixed genotypes (up to four in this cohort) and low-abundance genotypes (**Table 2**). Comparing the dominant genotypes and sub-lineages derived from both sequencing methods, the inter-assay agreement was 100%. Tabulated summary of NGS reads is shown in **Table 2**. The median of reads that passed quality check for 12 samples was 328,197. The proportion of merged reads that were mapped to reference HPV genotype (s) ranged from 94.9 to 99.8%.

TABLE 1 | Cytohistological correlation.

	Cytohistological correlation		
Histology	**Total**	**LSIL**	**HSIL**
Samples, n	12	6	6
Histology (biopsy or excision)[a]			
Documented, n (%)	9 (75)	4 (67)	5 (83)
Not documented, n (%)	3 (25)	2 (33)	1 (17)
Histological grade[a]			
CIN 0, n (%)	0	0	0
CIN I, n (%)	4 (44)	3 (75)	1 (20)
CIN II/III, n (%)	5 (56)	1 (25)	4 (80)
Cytohistological agreement[b]			
Agreement, %	78		
Expected agreement, %	51		
Kappa	0.55		
Std. Error	0.33		
p-value	0.05		

CIN, cervical intraepithelial neoplasia; HSIL, high-grade squamous intraepithelial lesion; HPV, human papillomavirus; LSIL, low-grade squamous intraepithelial lesion. [a]Cervical histopathology is based on the highest grade documented on cervical biopsy or therapeutic excisional biopsy, i.e., cold knife conization (CKC) and loop excisional procedure (LEEP). Absence or presence of pathology reports in the DoD electronic health records was categorized as "Documented" or "Not documented," respectively. [b]Cytohistological agreement was calculated using samples with documented histopathology. LSIL cytology corresponds to CIN I histology; HSIL cytology corresponds to CIN II/III histology.

TABLE 2 | HPV L1 sequencing results.

Sample Info and PCR[a]			Dideoxy seq[b]					Deep sequencing[c]							
								HPV genotypes and sub-lineages							
ID	**PAP**	**PCR band (n)**	**HPV type**	**IARC carc**	**Total merged reads (n)**	**Total mapped reads (n)**	**Total HPV types (n)**	**HPV #1**	**HPV#1 mapped reads (n)**	**HPV #2**	**HPV#2 mapped reads (n)**	**HPV #3**	**HPV#3 mapped reads (n)**	**HPV #4**	**HPV#4 mapped reads (n)**
179	HSIL	1	**16** D3	CARC	226981	224923	3	**16** D3	218169	**70** B1	4258	**39** A2	2150		
305	HSIL	1	**53** D1	POSC	447187	434307	4	**53** D1	404329	**18** A5	18828	**66** B1	5095	**70** B1	3197
313	HSIL	1	**16** A4	CARC	462487	456343	1	**16** A4	451363						
319	HSIL	1	**16** A4	CARC	521199	493925	2	**16** A4	466086	**70** B1	24150				
386	HSIL	1	**83**	NC	283025	282650	2	**83**	278890	**16** A4	2015				
399	HSIL	1	**16** A4	CARC	240743	239279	1	**16** A4	238954						
81	LSIL	1	**58** C1	CARC	120075	119929	2	**58** C1	91062	**61**	28293				
137	LSIL	1	**66** A1	CARC	459454	458465	1	**66** A1	457259						
138	LSIL	1	**81**	NC	323374	323167	2	**81**	295600	**83**	24988				
140	LSIL	3	**61**	NC	112792	112662	3	**61**	102636	**39** A1	8525	**62**	1075		
141	LSIL	1	**53** A1	POSC	255851	254268	2	**53** A1	247587	**61**	4781				
160	LSIL	1	**16** A4	CARC	429808	428796	2	**16** A4	417946	**18** A5	10551				

CARC, carcinogenic HPV; HSIL, high-grade squamous intraepithelial lesion; HPV, human papillomavirus; ID, sample identification; IARC Carc, International Agency for Research on Cancer – classification of carcinogenicity; L1, HPV L1 gene amplified by PCR; LSIL, low-grade squamous intraepithelial lesion; PCR, polymerase chain reaction; POSC, possibly carcinogenic; NC, not classifiable/probably not carcinogenic; Sample ID No., sequentially numbered samples grouped as LSIL or HSIL; PAP, Pap smear; Seq, sequencing. [a]Cytologically derived DNA samples amplified by PCR using consensus primers to target the HPV L1 loci. The number of PCR amplicon bands was determined by high-resolution capillary gel electrophoresis. [b]HPV genotype determined by BLAST alignment after amplicon sequencing (dideoxy method). HPV genotype number in bold and sub-lineage in alphanumeric fonts. [c]HPV genotype determined by BLAST alignment after amplicon sequencing (deep method). HPV genotype number in bold and sub-lineage in alphanumeric fonts.

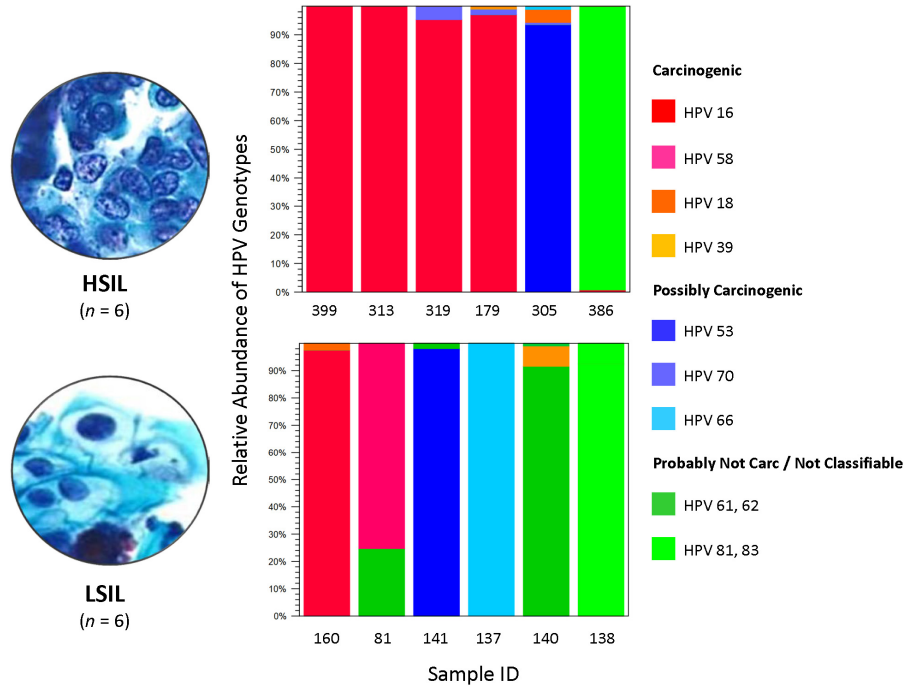

FIGURE 1 | HPV genotype composition found in LSIL and HSIL samples. Deep sequencing of HPV L1 amplicons derived from each LSIL or HSIL sample identified one to four HPV genotypes and quantitated their composition (%) based on number of mapped reads to total mapped reads. Each sample contained a dominant genotype. For HSIL, HPV-16 prevailed in 4 of 6 (67%) samples. In contrast, a mixture of carcinogenic, possibly carcinogenic and probably not carcinogenic HPV genotypes (33% for each category) were identified in LSIL samples. The HPV carcinogenicity is based on IARC's classification of human carcinogens (International Agency for Research on Cancer, 2012). Carc, carcinogenic; HSIL, high-grade squamous intraepithelial lesion; IARC, International Agency for Research on Cancer; ID, identification; LSIL, low-grade squamous intraepithelial lesion.

HPV Communities Were Dissimilar Between LSIL and HSIL With Loss of Species Richness and Gain of HPV-16 Dominance in HSIL Samples

The composition of HPV genotypes in each sample is illustrated in **Figure 1**. For six LSIL L1 samples, the number of genotype(s) per sample was distributed as: 1 (16.7%), 2 (66.6%), and 3 (16.7%). For six HSIL samples, the number of genotype(s) per sample was distributed as: 1 (33.3%), 2 (33.3%), 3 (16.7%), and 4 (16.7%). Notably, all HSIL samples contained at least one carcinogenic HPV genotype, whereas only half of the LSIL samples were found to have a carcinogenic genotype (**Table 2**).

We analyzed the HPV diversity, dominance and community structure between LSIL and HSIL samples. A total of 10 different genotypes were found in single and mixed-infected LSIL samples, whereas seven different genotypes were identified in HSIL samples. The respective Shannon Entropy Indices for LSIL and HSIL samples were 0.32 and 0.16, suggesting reduced diversity in HSIL samples (**Figure 2**). The dominant (most abundant) genotype in LSIL samples was HPV-61 versus HPV-16 for HSIL. HPV-16, one of the most important carcinogens responsible for almost half of the cervical cancer incidences (Taylor et al., 2016; Mirabello et al., 2017), was found in 5 of 6 (83.3%) HSIL samples. Two additional carcinogenic genotypes HPV-18 and HPV-39 were also discovered in HSIL samples. By contrast, HPV-61, which was considered noncarcinogenic, had 50% occurrence in LSIL samples, indicative of low risk for cervical cancer (Schiffman et al., 2009). It is worthy to note that two LSIL samples contained carcinogenic genotypes (HPV-58 in Sample 81, and HPV-16 in Sample 160), suggesting a finer resolution by HPV molecular profiling than cytological grading for carcinogenic potential.

We further examined the diversity of each sample estimated through read counts and Simpson-Index (**Figure 2**). The reduced diversity in high grade cytology samples is supported by the mean Simpson's indices (0.12 versus 0.05 for LSIL and HSIL, respectively). Sample 81 showed a relatively high diversity among LSIL samples, likely due to the presence of two abundant genotypes HPV58 and HPV 61. Sample 305 had the highest diversity in HSIL samples with mixed infection of four genotypes (carcinogenic HPV18, and possibly carcinogenic HPV53, HPV 66, and HPV 70). Samples with pure HPV genotypes, 137 and 399, exhibited low diversity.

Dissimilarity of HPV communities across HSIL and LSIL samples was visualized by principle coordinate analysis (PCoA) of Bray-Curtis distances (Bray and Curtis, 1957; **Figure 3**). PCoA showed HPV-16 (PCo 1, 60%) as being the most influential genotype in HSIL. In contrast, LSIL was influenced about equally (PCo 1–3, 21–26%) by carcinogenic, possibly carcinogenic, and probably not carcinogenic/not classifiable genotypes. As for the PCA results, the component loadings plot for LSIL and HSIL showed the correlative relationship between HPV

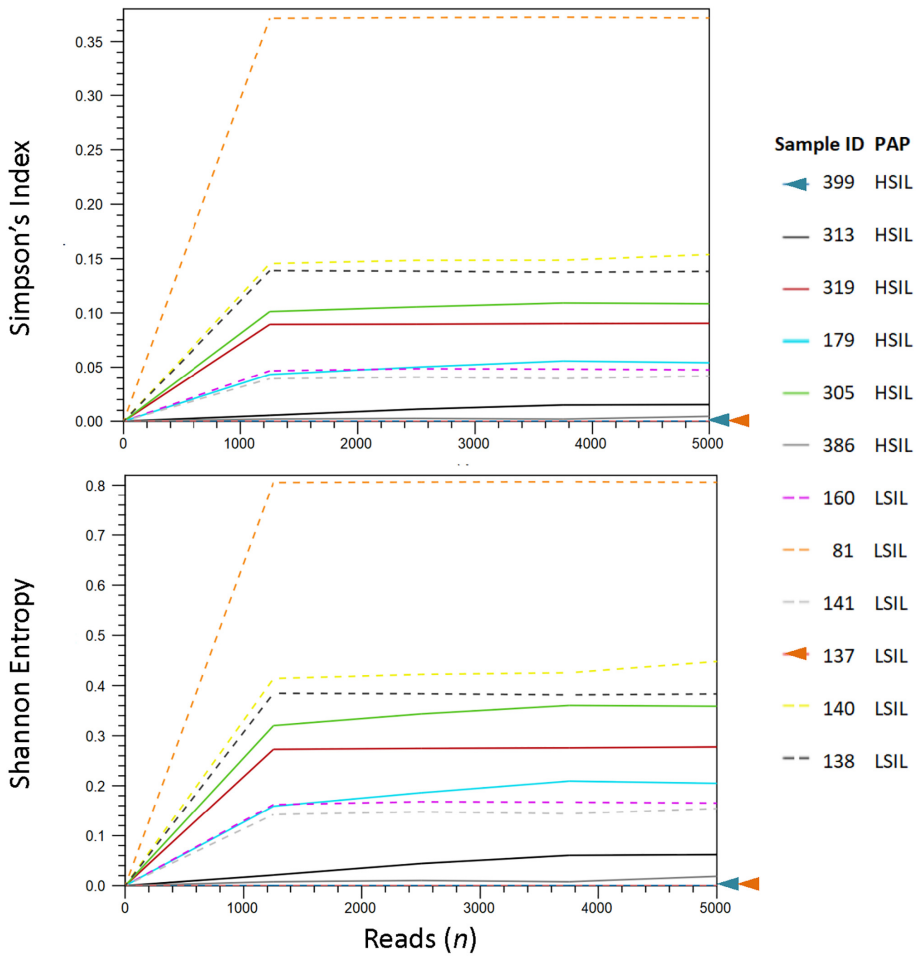

FIGURE 2 | HPV diversity analysis for LSIL and HSIL based on L1 deep sequencing. A total of 10 genotypes out of 6 samples were found in LSIL versus 7 genotypes out of 6 samples for HSIL. The respective Shannon Entropy Indices for LSIL (dashed line) and HSIL samples (solid line) were 0.32 and 0.16, suggesting reduced diversity in HSIL samples. Similarly, species richness measured by Simpson's index showed a reduction in high-grade cytology (0.12 vs. 0.05 for LSIL and HSIL, respectively). Two samples (137 and 399) contained pure species or zero diversity are indicated by arrowheads. The dominant (most abundant) genotype in LSIL samples was HPV-61 versus HPV-16 for HSIL.

genotypes along the first two principal components axes (PC1 and PC2) (**Supplementary Figure 3**). The sum of PC1 and PC2 explained 51.6 and 96.2% of the total variance for LSIL and HSIL, respectively. Comparing LSIL and HSIL, HPV-16 emerged from all other genotypes as the dominant component in HSIL. The score variables plots displayed each sample's contribution to the principal components. HSIL compared to LSIL had a preponderance of samples containing a high composition of HPV-16.

Molecular Taxonomy of HPV Genotypes Based on NGS Is Highly Discriminatory and Correlated With IARC-Defined Carcinogenicity

Prototypical HPV genome based on the genetic information of HPV-16 (GenBank ID: K02718) is created using the CLC Bio Genomics Workbench 11.0.1 (QIAGEN) and shown in **Figure 4**. The L1 (450 bp) gene fragment of each sample was the target used for sequencing, genotyping, and phylogenetic analysis. A maximum likelihood tree was inferred from the L1 sequences derived from single and multi-infected samples (**Figure 5**). The tree topology is consistent with the HPV species trees (Schiffman et al., 2009; Bernard et al., 2010; International Agency for Research on Cancer, 2012). These L1 sequences were clustered into four clades with strong bootstrap support: (1) α-9 clade included HPV-16 from four HSIL and two LSIL samples, and HPV-58 from an LSIL sample 81. Both HPV-16 and HPV-58 are carcinogenic. (2) α-7 clade included carcinogenic HPV-18 and HPV-39, and a possibly carcinogenic HPV-70, which were shown in three mixed-infected HSIL samples. (3) α-6 clade included possibly carcinogenic HPV-53 and HPV-66. (4) α-3 clade included all the probably not carcinogenic genotypes found in HSIL and LSIL samples (Chen et al., 2018b). Clearly, the broad categorical grade designation based on precancerous cervical lesions (HSIL versus LSIL) was imprecise at predicting

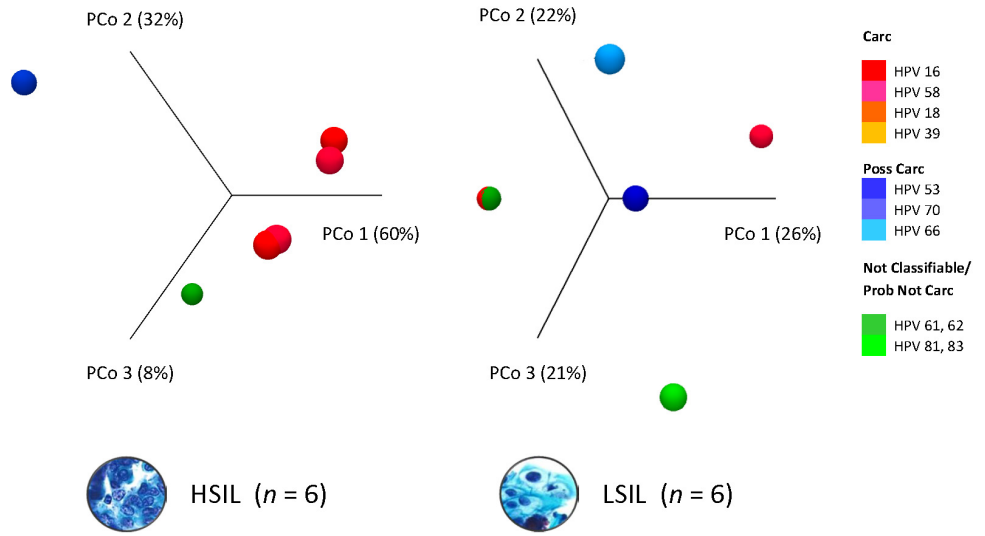

FIGURE 3 | HPV dominance and community structure between LSIL and HSIL. HPV L1 3D-Principal Coordinate Analysis (PCoA) plots of HSIL and LSIL samples showing dissimilarity between the two HPV communities with HPV-16 (PCoA 1) being the most influential genotype in HSIL versus HPV-61 for LSIL (PCoA 1). β-diversity was measured by Bray-Curtis index. Carc, carcinogenic; HSIL, high-grade squamous intraepithelial lesion; ID, identification; LSIL, low-grade squamous intraepithelial lesion.

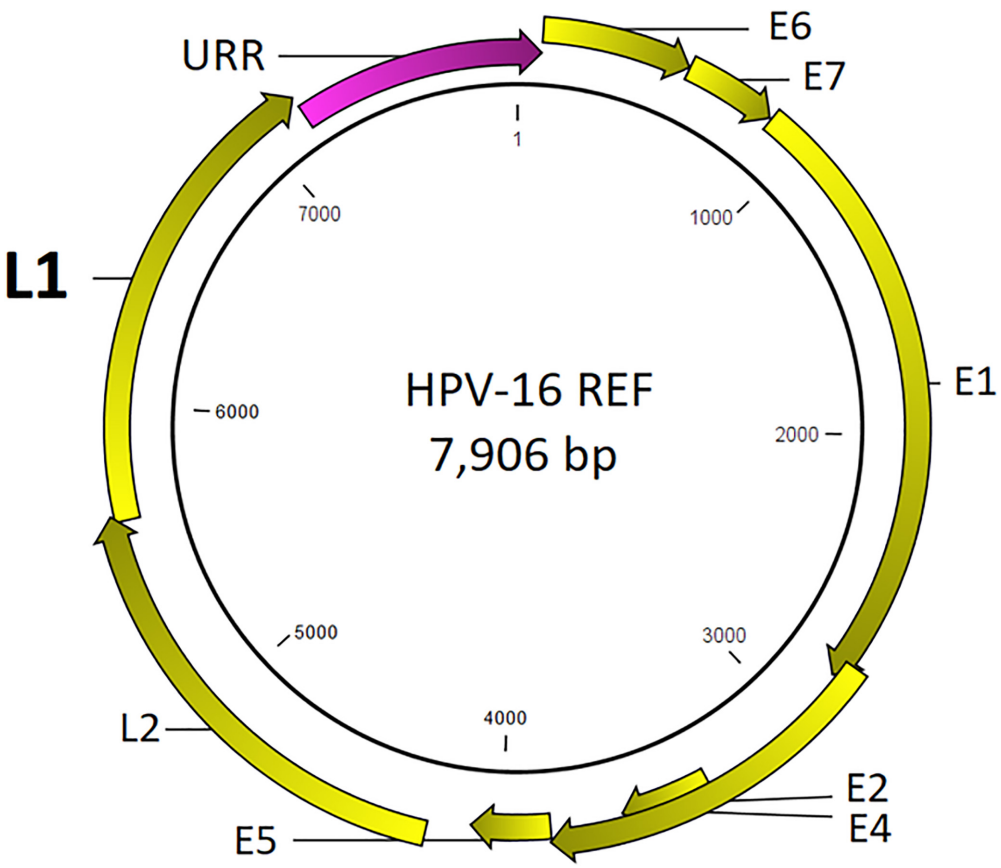

FIGURE 4 | Prototypical HPV genome based on the genetic information of HPV-16 (GenBank ID: K02718). The L1 (450 bp) gene segment of each sample was the target used for sequencing, genotyping, and phylogenetic analysis.

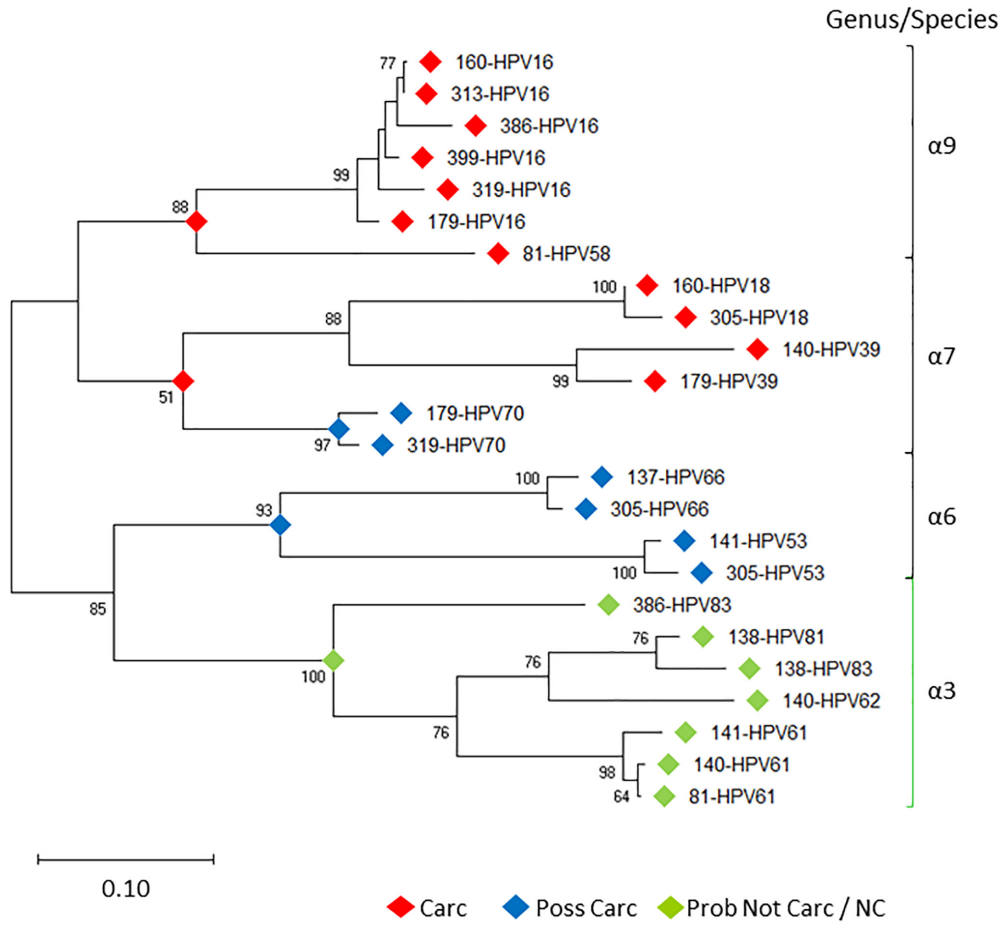

FIGURE 5 | Evolutionary relationships of HPV L1 sequences derived from LSIL and HSIL samples. Phylogenetic tree of L1 nucleotide sequences revealed clades of species (−6, 7, 9) [black bracket] and −3 [green bracket] within the -genus, correlating with the level of IARC-defined carcinogenicity. Between individual samples (sample ID-HPV*n*), HPV intra-type genetic differences are prevalent as distinguished by nonoverlapping branches. The evolutionary history of the HPV L1 sequences was inferred by using the Maximum Likelihood method (Felsenstein, 1981) and Tamura-Nei model (Tamura and Nei, 1993). The tree with the highest log likelihood (−7492.54) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining (Saitou and Nei, 1987) and BioNJ (Gascuel, 1997) algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. Bootstrap resampling with 1,000 pseudo-replicates was carried out to assess support for each individual branch (Felsenstein, 1985). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018). Carc, carcinogenic; NC, not classifiable; Poss Carc, possibly carcinogenic; Prob Not Carc, probably not carcinogenic; REF, reference genome; URR, upstream regulatory region.

carcinogenicity. Conversely, the molecular taxonomy based on NGS is highly discriminatory and correlated well with IARC-defined carcinogenicity.

Sequence and Structural Variations Identified at HPV Antigenic Sites May Alter Viral Recognition by Innate or Vaccine-Induced Host Defense

We hypothesized that variation in HPV L1 within and among the clinical samples can reveal critical details about the genetic basis for evolution of HPV immune evasion and host-pathogen interactions, because L1 encodes the major capsid protein that plays an important role in virion attachment and entry to the host (Knappe et al., 2007; Dasgupta et al., 2011; Surviladze et al., 2015; Chabeda et al., 2018). Being a natural antigen, the capsid surface is the target of HPV prophylactic vaccines (Harper, 2009; Harper and Williams, 2010; Yang et al., 2016). **Supplementary Table 1** lists the position, predicted mutation type and change at the coding region for HPV variants, compared to the respective reference HPV types. For the combined LSIL/HSIL samples (n = 12), a total of 417 mutations were detected, including 375 SNVs, 29 indels, 12 MNVs, and one replacement variant. The distribution of these variants for the 12 samples by Pap grade and HPV genotype is shown in **Figures 6A,B**, respectively. The proportion of nonsynonymous mutations was lower for HSIL (0.38) than for LSIL samples (0.51) (p = 0.017, Fisher's exact test) (**Figure 7**). On the other hand, probably or probably not carcinogenic HPV types in LSIL samples appeared to be under relaxed functional constraint to accumulate mutations.

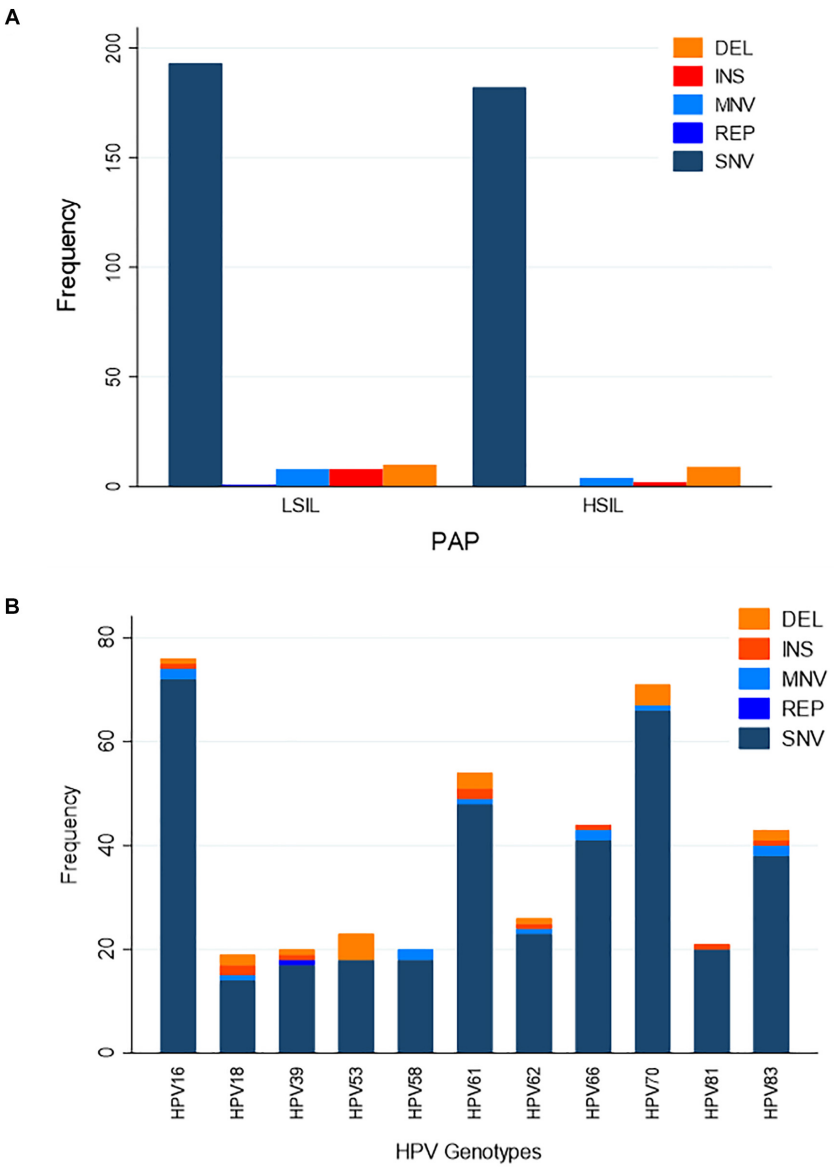

FIGURE 6 | Distribution of variants in HPV L1 sequences. **(A)** Distribution of variants by Pap grade. The predominant type of variants identified in LSIL (*n* = 6) and HSIL (*n* = 6) samples was single nucleotide variant (SNV). **(B)** Distribution of variants by HPV genotype. Eleven HPV genotypes were identified in the deep sequenced LSIL/HSIL (*n* = 12) samples. The top three genotypes with the highest total number of variants were HPV-16, -61, and -70. DEL, deletion; INS, insertion; MNV, multi-nucleotide variant (two or more SNVs in succession); REP, replacement; SNV, single nucleotide variant.

The distribution of variants in HPV L1 by amino acid positions according to HPV genotype is shown in **Supplementary Figure 1**.

It is important to identify mutations that is potentially driven by vaccine- or natural infection-induced host immune response. To visualize mutations in 3D, first the structural model of HPV-16 L1 (PDB ID: 2R5H) (Bishop et al., 2007) was reconstructed with demarcated hypervariable surface loops: BC, DE, EF, FG, and HI (**Figure 8**). Additionally, the HPV-16 L1 protein sequence with surface probability plot for prediction of antigenic determinants on surface proteins (Emini et al., 1985) is provided in **Supplementary Figure 2**. In the case of HSIL Sample 179, we identified seven nonsynonymous mutations in HPV-16. **Figure 9** shows 3D conformational changes visualized by overlying the mutated amino acid residues (cyan) to those (purple) in the reference HPV-16 structure (PDB ID: 1DZL) (Chen et al., 2000). It is particularly noticeable that the mutation at position 353 corresponded to a threonine to proline change (T353P) located at the HI-Loop. The T353P change also increased the surface probability from 3.40 to 3.63 (range, 0–6.47; threshold = 1.0) (**Supplementary Figure 2**). HI Loop is one of the loops in L1 protein that extends to the outer surface of the capsid complex (Chen et al., 2000; Bishop et al., 2007). This hypervariable HI loop (AA 339–365) contains an HPV-16 immunodominant epitope (Christensen et al., 2001). As seen in human Influenza virus,

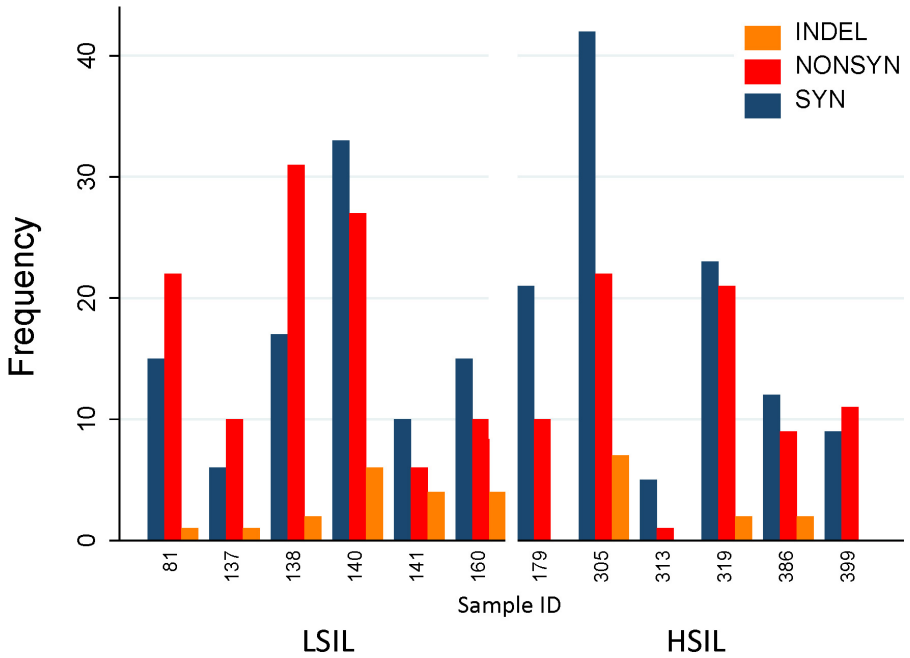

FIGURE 7 | Distribution of variants in HPV L1 sequences found in HSIL and LSIL samples. Variants are categorized as nonsynonymous, synonymous, and insertion/deletion and displayed as frequency counts. The proportion of nonsynonymous mutations was lower for HSIL (0.38) than for LSIL samples (0.51) (p = 0.017, Fisher's exact test). This finding suggests that the inherent competitive advantage of carcinogenic HPV genotypes, e.g., HPV-16 further shaped by intra-host selection may contribute to viral carcinogenesis. One replacement variant of HPV-39 discovered in Sample 140 is not shown in the figure. The annotated variant table including predicted amino acid changes is presented in **Supplementary Table 1**. HSIL, high-grade squamous intraepithelial lesion; ID, identification; INDEL, insertion/deletion; LSIL, low-grade squamous intraepithelial lesion; NONSYN, nonsynonymous; SYN, synonymous.

antigen drift, where mutations are accumulated in antigenic sites, is a potent force driving the evolution of immune evasion and reduced vaccine efficacy (Fitch et al., 1997; Bush et al., 1999; Smith et al., 2004). Similarly, codon changes like T353P at the antigenic regions may confer selective advantage by increasing the likelihood of immune evasion. In addition to T353P, other mutations in this sample may lead to changes in the secondary structure, including W325C at G2 β-sheet, G367P at β-I sheet, T389S at α-2 helix, L441I at a β-turn, Q461P and F462Y near α-5 helix (Bishop et al., 2007).

DISCUSSION

This study revealed the complex genetic diversity of HPV viromes within low- and high-grade Pap samples. Both pure and mixed infections were common as shown by deep amplicon sequencing. Taxonomic profiling revealed the difference between LSIL and HSIL viral communities with loss of species richness and gain of dominance by carcinogenic genotypes, particularly HPV-16, in HSIL samples. Deep sequencing allowed the detection of carcinogenic HPVs constituting a minor component of a virome which was undisclosed by Sanger sequencing or cytological grading. Phylogenetic inference of the patient-derived L1 sequences showed excellent correlation between HPV type-specific distances and IARC-defined carcinogenic potential. Together with taxonomic profiling, this "Taxo-Phylo" approach holds promise as a molecular taxonomy-based classifier of cervical cytology.

HPV variant detection and analysis pinpointed the nucleotide-level mutations and potential functional, as well as, structural consequences. Localizing mutations to primary sequences and structures can help understand the functional consequence of mutations and identify causal or adaptive mutations. Furthermore, *in silico* modeling of mutations may direct laboratory testing and confirmation of its significance through antigen-antibody binding assays. For example, hepatitis B virus (HBV) genotypes are known to vary by ethno-geography. Mutations in the major hydrophilic regions (MHR) of the hepatitis B surface antigen (HBSAg) have resulted in stable, vaccine-escape mutant virions that are infectious and pathogenic (Carman et al., 1990; Gencay et al., 2018). Recently, investigators have used ultra-deep sequencing and clinical immunoassays (monoclonal antibodies) to detect single-nucleotide, vaccine-escape mutations and associated changes in the HBSAg amino acid residues in clinical samples (Gencay et al., 2018). Similarly, liquid-based cervical cytology samples may be interrogated by deep sequencing and multiplexed immunoassays, e.g., Luminex xMAP® (Peters et al., 2013) to survey HPV L1 mutant virions that may escape from innate or vaccine-induced immunity.

Longitudinal HPV metagenomic surveillance may also provide a sensitive means of detecting changes in HPV evolution and dynamics within individuals or populations. This is clinically important because virulent genotype(s) of low abundance may

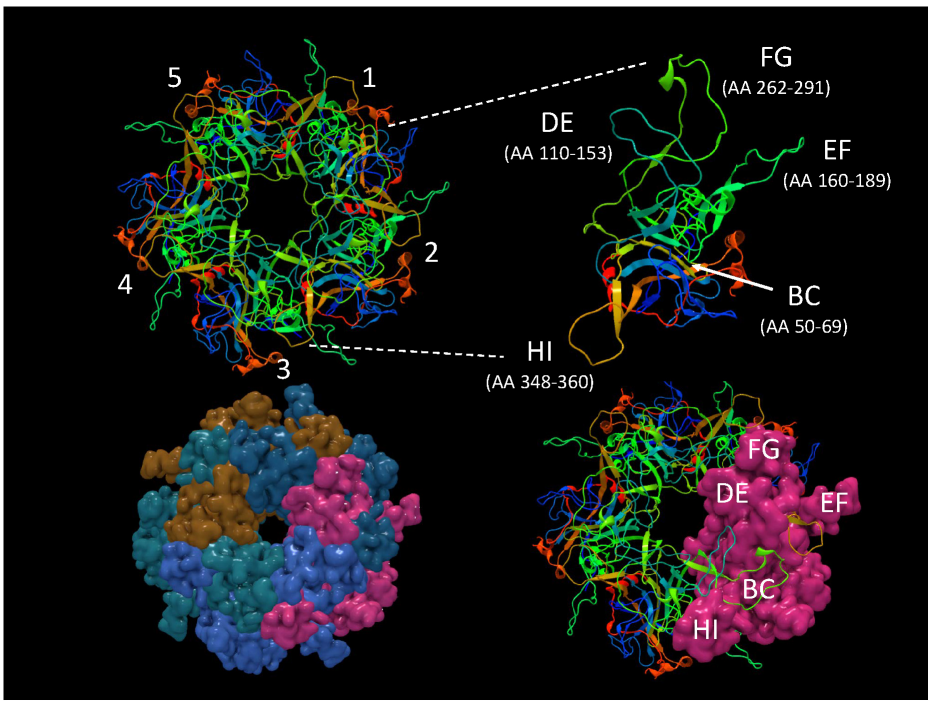

FIGURE 8 | Structure of HPV-16 L1 capsomer. Structural model of HPV-16 L1 capsomer reconstructed from the coordinates and crystal structure filed in Protein Data Bank (PDB ID: 2R5H) (Bishop et al., 2007). The capsomer composed of five L1 subunits (numbered 1–5) are displayed in backbone and surface views to highlight the hypervariable surface loops: BC, DE, EF, FG, and HI and amino acid (AA) positions. These loops are antigenic regions of interest in vaccinology.

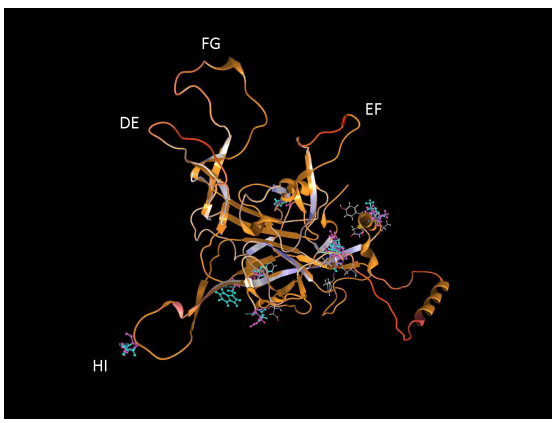

FIGURE 9 | Structural location of L1 variants. Visualization of L1 variants from a HSIL sample (Sample 179) linked to a 3D protein structure. The reference structure is a HPV 16 L1 monomer with accession number 1DZL (Chen et al., 2000) shown in backbone representation. Variant consequences in 3D are identified by the variant in cyan collocated on top of the reference amino acid in purple with attention toward the surface loops. AA, amino acid position.

later dominate the virome if it is inherently more carcinogenic or confers a selective advantage with ensuing clonal expansion. Current published literature on HPV L1 variant analysis is scarce. As noted previously, a high intra-type L1 sequence variability was discovered in 35 HPV genotypes by deep sequencing. These investigators also found unique genotypes and its variants associating with distinct anatomical sites supporting the notion of viral niche-adaptation as shapers of viral evolution (Chen et al., 2018a; Dube Mandishora et al., 2018). However, functional consequences of these mutations were not studied. Another investigation found multiple mutations within the L1 fragment of HPV-16 (MY09/11-primed amplicons) of 35 invasive cervical cancer samples from Morocco (El-Aliani et al., 2017). A distinct mutation in the HI loop (T389P) found in 51.4% of cases could potentially interact with vaccine-induced neutralizing antibodies (El-Aliani et al., 2017). In view of this information, our results are highly consistent with the findings of high intra-host and intra-type L1 sequence variability that could potentially impact vaccine efficacy.

The strength of this study lies in the multi-omics approach developed herein. Integration of clinical metadata, genomic data, and functional/structural information to reveal patient-specific metagenomic profiles and variant structures in 3D is novel and practical. Such individualized virome profiling may provide guidance to clinicians on the risk of cervical cancer and potentially deleterious viral variants/mutations. We acknowledge that our study has limitations in that the sample size was small and a fragment of L1 was studied so overreaching generalizable conclusions cannot be drawn. However, an integrated, holistic approach was established from this dataset to further HPV metagenomics research. Our future direction will be to conduct a large scale, whole-genome or full-sequence L1 variant analysis to survey type-specific variant patterns by cytological grades.

CONCLUSION

In this pilot study, NGS provided a cost-effective platform for an unbiased discovery of HPV communities in clinical samples. The HPV genotype composition was shown to be correlated with clinical severity and the carcinogenic risk for cervical cancer. Multi-omics analyses afforded an unprecedented opportunity to better characterize the L1 complexity in clinical samples. Ultimately, this approach will lead to greater understanding of the dynamic interplay between virus and host in HPV pathogenesis.

AUTHOR'S NOTE

This paper has undergone PAO review at Brooke Army Medical Center and was cleared for publication. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or reflecting the views of the U.S. Department of the Army, U.S. Department of Defense, or the U.S. government.

ETHICS STATEMENT

This study was approved by the institutional review board of Brooke Army Medical Center, Fort Sam Houston, Texas.

AUTHOR CONTRIBUTIONS

JS-G and YW conceived and designed the study and participated in the acquisition of data. JS-G, YW, HC, and HZ analyzed and interpreted the data. JS-G, YW, and HC wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

Laboratory materials for this work were supported in part by the Department of Clinical Investigation Intramural Funding Program at Brooke Army Medical Center, Fort Sam Houston, Texas.

ACKNOWLEDGMENTS

We thank the staff at the Greehey Children's Cancer Research Institute and Bioanalytics and Single-Cell Core of the University of Texas Health Science at San Antonio for their invaluable service in supporting the next-generation sequencing experiments; and the staff, Ms. Roxanne Toscano and Ms. Rosalyn Miller, at the Cytopathology Laboratory of Brooke Army Medical Center for their invaluable service for collecting the clinical samples in support of the HPV Research Program in the Department of Clinical Investigation at Brooke Army Medical Center.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00489/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Berman, H. M., Westbrook, J., Feng, J., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Bernard, H. U., Burk, R. D., Chen, Z., van Doorslaer, K., zur Hausen, H., and de Villiers, E. M. (2010). Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 401, 70–79. doi: 10.1016/j.virol.2010.02.002

Bishop, B., Dasgupta, J., Klein, M., Garcea, R. L., Christensen, N. D., Zhao, R., et al. (2007). Crystal structures of four types of human papillomavirus L1 capsid proteins: understanding the specificity of neutralizing monoclonal antibodies. *J. Biol. Chem.* 282, 31803–31811. doi: 10.1074/jbc.m706380200

Bissett, S. L., Godi, A., and Beddows, S. (2016). The DE and FG loops of the HPV major capsid protein contribute to the epitopes of vaccine-induced cross-neutralising antibodies. *Sci. Rep.* 22:39730. doi: 10.1038/srep39730

Bosch, F. X., Broker, T. R., Forman, D., Moscicki, A. B., Gillison, M. L., Doorbar, J., et al. (2013). Comprehensive control of human papillomavirus infections and related diseases. *Vaccine* 31(Suppl. 5), F1–F31. doi: 10.1016/j.vaccine.2013.10.001

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.* 27, 326–349.

Buck, C. B., Day, P. M., and Trus, B. L. (2013). The papillomavirus major capsid protein L1. *Virology* 445, 169–174. doi: 10.1016/j.virol.2013.05.038

Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* 16, 1457–1465. doi: 10.1093/oxfordjournals.molbev.a026057

Bzhalava, D., Mühr, L. S., Lagheden, C., Ekström, J., Forslund, O., Dillner, J., et al. (2014). Deep sequencing extends the diversity of human papillomaviruses in human skin. *Sci. Rep.* 24:5807. doi: 10.1038/srep05807

Carman, W. F., Zanetti, A. R., Karayiannis, P., Waters, J., Manzillo, G., Tanzi, E., et al. (1990). Vaccine-induced escape mutant of hepatitis B virus. *Lancet* 11, 325–329. doi: 10.1016/0140-6736(90)91874-a

Chabeda, A., Yanez, R. J. R., Lamprecht, R., Meyers, A. E., Rybicki, E. P., and Hitzeroth, I. I. (2018). Therapeutic vaccines for high-risk HPV-associated diseases. *Papillomavirus Res.* 5, 46–58. doi: 10.1016/j.pvr.2017.12.006

Chen, X. S., Garcea, R. L., Goldberg, I., Casini, G., and Harrison, S. C. (2000). Structure of small virus-like particles assembled from the L1 protein of human papillomavirus 16. *Mol. Cell.* 5, 557–567. doi: 10.1016/s1097-2765(00)80449-9

Chen, Z., DeSalle, R., Schiffman, M., Herrero, R., Wood, C. E., Ruiz, J. C., et al. (2018a). Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog.* 14:e1007352. doi: 10.1371/journal.ppat.1007352

Chen, Z., Schiffman, M., Herrero, R., DeSalle, R., Anastos, K., Segondy, M., et al. (2018b). Classification and evolution of human papillomavirus genome variants: alpha-5 (HPV26, 51, 69, 82), Alpha-6 (HPV30, 53, 56, 66), Alpha-11 (HPV34, 73), Alpha-13 (HPV54) and Alpha-3 (HPV61). *Virology* 516, 86–101. doi: 10.1016/j.virol.2018.01.002

Christensen, N. D., Cladel, N. M., Reed, C. A., Budgeon, L. R., Embers, M. E., Skulsky, D. M., et al. (2001). Hybrid papillomavirus L1 molecules assemble into virus-like particles that reconstitute conformational epitopes and

induce neutralizing antibodies to distinct HPV types. *Virology* 291, 324–334. doi: 10.1006/viro.2001.1220

Dasgupta, J., Bienkowska-Haba, M., Ortega, M. E., Patel, H. D., Bodevin, S., Spillmann, D., et al. (2011). Structural basis of oligosaccharide receptor recognition by human papillomavirus. *J. Biol. Chem.* 286, 2617–2624. doi: 10.1074/jbc.M110.160184

de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U., and zur Hausen, H. (2004). Classification of papillomaviruses. *Virology* 324, 17–27. doi: 10.1016/j.virol.2004.03.033

Doorbar, J., Egawa, N., Griffin, H., Kranjec, C., and Murakami, I. (2015). Human papillomavirus molecular biology and disease association. *Rev. Med. Virol.* 25(Suppl.1), 2–23. doi: 10.1002/rmv.1822

Dube Mandishora, R. S., Gjøtterud, K. S., Lagström, S., Stray-Pedersen, B., Duri, K., Chin'ombe, N., et al. (2018). Intra-host sequence variability in human papillomavirus. *Papillomavirus Res.* 5, 180–191. doi: 10.1016/j.pvr.2018.04.006

El-Aliani, A., Alaoui, M. A. E., Chaoui, I., Ennaji, M. M., Attaleb, M., and Mzibri, M. E. (2017). Naturally occurring capsid protein variants L1 of human papillomavirus genotype 16 in Morocco. *Bioinformation* 13, 241–248. doi: 10.6026/97320630013241

Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* 55, 836–839.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194. doi: 10.1101/gr.8.3.186

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185. doi: 10.1101/gr.8.3.175

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi: 10.1007/bf01734359

Felsenstein, J. (1985). Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x

Fitch, W. M., Bush, R. M., Bender, C. A., and Cox, N. J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7712–7718. doi: 10.1073/pnas.94.15.7712

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14, 685–695. doi: 10.1093/oxfordjournals.molbev.a025808

Gencay, M., Vermeulen, M., Neofytos, D., Westergaard, G., Pabinger, S., Kriegner, A., et al. (2018). Substantial variation in the hepatitis B surface antigen (HBsAg) in hepatitis B virus (HBV)-positive patients from South Africa: reliable detection of HBV by the Elecsys HBsAg II assay. *J. Clin. Virol.* 101, 38–43. doi: 10.1016/j.jcv.2018.01.011

Harper, D. M. (2009). Currently approved prophylactic HPV vaccines. *Expert Rev. Vaccines* 8, 1663–1679. doi: 10.1586/erv.09.123

Harper, D. M., and Williams, K. B. (2010). Prophylactic HPV vaccines: current knowledge of impact on gynecologic premalignancies. *Discov. Med.* 10, 7–17.

Hirose, Y., Onuki, M., Tenjimbayashi, Y., Mori, S., Ishii, Y., Takeuchi, T., et al. (2018). Within-host variations of human papillomavirus reveal APOBEC signature mutagenesis in the viral genome. *J. Virol.* 92, e00017-18. doi: 10.1128/JVI.00017-18

International Agency for Research on Cancer (2012). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans-Human Papillomaviruses*. Geneva: World Health Organization, 255–313.

Knappe, M., Bodevin, S., Selinka, H. C., Spillmann, D., Streeck, R. E., Chen, X. S., et al. (2007). Surface-exposed amino acid residues of HPV16 L1 protein mediating interaction with cell surface heparan sulfate. *J. Biol. Chem.* 282, 27913–27922. doi: 10.1074/jbc.m705127200

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Mirabello, L., Yeager, M., Yu, K., Clifford, G. M., Xiao, Y., Zhu, B., et al. (2017). HPV16 E7 genetic conservation is critical to carcinogenesis. *Cell* 170, 1164–1174.e6. doi: 10.1016/j.cell.2017.08.001

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217. doi: 10.1006/jmbi.2000.4042

Peters, J., Thomas, D., Boers, E., de Rijk, T., Berthiller, F., Haasnoot, W., et al. (2013). Colour-encoded paramagnetic microbead-based direct inhibition triplex flow cytometric immunoassay for ochratoxin A, fumonisins and zearalenone in cereals and cereal-based feed. *Anal. Bioanal. Chem.* 405, 7783–7794. doi: 10.1007/s00216-013-7095-7

Rencher, A. C., and Christensen, W. F. (2012). *Methods of Multivariate Analysis*, 3rd Edn. New Jersey, NJ: John Wiley & Sons, 405–433.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Schiffman, M., Clifford, G., and Buonaguro, F. M. (2009). Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect. Agent Cancer* 4, 8. doi: 10.1186/1750-9378-4-8

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 623–656.

Shen-Gunther, J., Wang, C. M., Poage, G. M., Lin, C. L., Perez, L., Banks, N. A., et al. (2016). Molecular Pap smear: HPV genotype and DNA methylation of ADCY8, CDH8, and ZNF582 as an integrated biomarker for high-grade cervical cytology. *Clin. Epigenet.* 13:96. doi: 10.1186/s13148-016-0263-9

Shen-Gunther, J., Wang, Y., Lai, Z., Poage, G. M., Perez, L., and Huang, T. H. (2017). Deep sequencing of HPV E6/E7 genes reveals loss of genotypic diversity and gain of clonal dominance in high-grade intraepithelial lesions of the cervix. *BMC Genomics* 18:231. doi: 10.1186/s12864-017-3612-y

Shen-Gunther, J., and Yu, X. (2011). HPV molecular assays: defining analytical and clinical performance characteristics for cervical cytology specimens. *Gynecol. Oncol.* 123, 263–271. doi: 10.1016/j.ygyno.2011.07.017

Shope, R. E. (1932). A transmissible tumor-like condition in rabbits. *J. Exp. Med.* 30, 793–802. doi: 10.1084/jem.56.6.793

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163, 688–688.

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., et al. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science* 305, 371–376. doi: 10.1126/science.1097211

Surviladze, Z., Sterkand, R. T., and Ozbun, M. A. (2015). Interaction of human papillomavirus type 16 particles with heparan sulfate and syndecan-1 molecules in the keratinocyte extracellular matrix plays an active role in infection. *J. Gen. Virol.* 96, 2232–2241. doi: 10.1099/vir.0.000147

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526.

Taylor, S., Bunge, E., Bakker, M., and Castellsague, X. (2016). The incidence, clearance and persistence of non-cervical human papillomavirus infections: a systematic review of the literature. *BMC Infect. Dis.* 16:293. doi: 10.1186/s12879-016-1633-9

van der Weele, P., Meijer, C. J. L. M., and King, A. J. (2017). Whole-genome sequencing and variant analysis of human papillomavirus 16 infections. *J. Virol.* 91, e00844-17. doi: 10.1128/JVI.00844-17

van der Weele, P., Meijer, C. J. L. M., and King, A. J. (2018). High whole-genome sequence diversity of human papillomavirus type 18 isolates. *Viruses* 10:E68. doi: 10.3390/v10020068

Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., et al. (2017). The papillomavirus episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res.* 45, D499–D506. doi: 10.1093/nar/gkw879

Yang, A., Farmer, E., Wu, T. C., and Hung, C. F. (2016). Perspectives for therapeutic HPV vaccine development. *J. Biomed. Sci.* 23:75.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 July 2019
doi: 10.3389/fgene.2019.00624

Survival Analysis of Multi-Omics Data Identifies Potential Prognostic Markers of Pancreatic Ductal Adenocarcinoma

Nitish Kumar Mishra[†], *Siddesh Southekal*[†] *and Chittibabu Guda* **

Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, United States

OPEN ACCESS

Edited by:
Junbai Wang,
Oslo University Hospital, Norway

Reviewed by:
Ashok Sharma,
Augusta University, United States
Hui-Chen Wu,
National University of Tainan, Taiwan

***Correspondence:**
Chittibabu Guda
babu.guda@unmc.edu

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

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *01 December 2018*
Accepted: *14 June 2019*
Published: *18 July 2019*

Citation:
Mishra NK, Southekal S and Guda C (2019) Survival Analysis of Multi-Omics Data Identifies Potential Prognostic Markers of Pancreatic Ductal Adenocarcinoma. Front. Genet. 10:624. doi: 10.3389/fgene.2019.00624

Pancreatic ductal adenocarcinoma (PDAC) is the most common and among the deadliest of pancreatic cancers. Its 5-year survival is only ~8%. Pancreatic cancers are a heterogeneous group of diseases, of which PDAC is particularly aggressive. Like many other cancers, PDAC also starts as a pre-invasive precursor lesion (known as pancreatic intraepithelial neoplasia, PanIN), which offers an opportunity for both early detection and early treatment. Even advanced PDAC can benefit from prognostic biomarkers. However, reliable biomarkers for early diagnosis or those for prognosis of therapy remain an unfulfilled goal for PDAC. In this study, we selected 153 PDAC patients from the TCGA database and used their clinical, DNA methylation, gene expression, and micro-RNA (miRNA) and long non-coding RNA (lncRNA) expression data for multi-omics analysis. Differential methylations at about 12,000 CpG sites were observed in PDAC tumor genomes, with about 61% of them hypermethylated, predominantly in the promoter regions and in CpG-islands. We correlated promoter methylation and gene expression for mRNAs and identified 17 genes that were previously recognized as PDAC biomarkers. Similarly, several genes (B3GNT3, DMBT1, DEPDC1B) and lncRNAs (PVT1, and GATA6-AS) are strongly correlated with survival, which have not been reported in PDAC before. Other genes such as EFR3B, whose biological roles are not well known in mammals are also found to strongly associated with survival. We further identified 406 promoter methylation target loci associated with patients survival, including known esophageal squamous cell carcinoma biomarkers, cg03234186 (ZNF154), and cg02587316, cg18630667, and cg05020604 (ZNF382). Overall, this is one of the first studies that identified survival associated genes using multi-omics data from PDAC patients.

Keywords: Dm-CpG: Differentially methylated CpG, DMR: differentially methylated region, DEG: differentially expressed gene, HR: hazard ratio, TCGA: The Cancer Genome Atlas, GDC: The Genomic Data Commons, FDR: false discovery rate

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) originates from the ductal epithelial cells of the pancreas and it is the most common malignancy of the pancreas. Due to lack of early symptoms, PDAC is commonly presented in the metastatic stage, and as a result, fewer than 20% patients can be considered for surgical removal of the tumors (Adamska et al., 2017). Unfortunately, removing frank tumors from the pancreas cannot be expected to cure a metastatic disease, which is reflected in the current statistics of 5-year survival, which remains pegged at a dismal 8% (Chiaravalli et al., 2017). By 2030, PDAC is projected to become the second leading cause of mortality from cancer, only behind lung cancer (Rahib et al., 2014). This is the most alarming situation, and we have an urgent need for developing early detection and effective treatment regimens.

Recent studies regarding molecular profiling and epigenetic regulation in PDAC pathophysiology have provided a valuable roadmap for this effort. We are beginning to gather information about the early-onset and PDAC-specific epigenetic alterations that alter gene expression (Neureiter et al., 2014), especially those that induce metastatic changes such as genome structure reorganization and affect tumor grade, stage, and patient survival (Thompson et al., 2015). Such studies are helping in identifying targets for designing epigenetic inhibitors to treat PDAC. Not surprisingly, these targets belong to growth signaling and tumor suppressor-silencing pathways, and also those that affect cell cycle checkpoints (Paradise et al., 2018).

There is also no doubt that early detection and early beginning of therapy will be key for defeating PDAC. Identification of early-onset DNA methylations in PDAC target genes should provide biomarker candidates for early diagnosis. We also know from earlier studies that certain critical genes are hypomethylated in pancreatic cancer. The mucin 4 (MUC4) gene is one example of promoter hypomethylation in pancreatic cancer (Zhu et al., 2011). However, pancreatic cancer appears to be affected by both hyper- and hypomethylated genes (Mishra and Guda, 2017). In particular, inside the promoters of ~72% of human genes, there are stretches of CpG dinucleotides (known as CpG islands), which are hypermethylated in cancer (Saxonov et al., 2006). Frequently, transcription of tumor suppressor genes is silenced by CpG island hypermethylation, while hypomethylation of promoters appears to cause overexpression of oncogenes and genomic instability (Tan et al., 2009). Abnormal DNA methylation affects many genes of cancer patients. In PDAC, genes involved in axon guidance, cell adhesion, epithelial-mesenchymal transition (EMT), and other pathways of tumor development, as well as genes involved in pancreatic development including the HOX-family genes, show abnormal DNA methylation (Nones et al., 2014; Mishra and Guda, 2017). Some of these genes may be useful for diagnosing PDAC stage and for the prognosis of successful therapy.

The availability of bisulfite-sequencing and array-based DNA methylation data in The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013; Tomczak et al., 2015), and International Cancer Genome Consortium (ICGC) (Zhang et al., 2011) has given our pursuit for identifying candidate biomarkers a great fillip. The study of differentially methylated loci between tumor and normal samples has great scientific merit for cataloging the genomic changes in PDAC. But integrated genomic analysis of differences in DNA methylations, their impact on expression of the genes, and correlating those data with patient survival will bring us closer to the goal of identifying the candidate biomarkers. Until recently, integrative analyses have mostly been done for examining methylation status of promoters and CpG islands (Vincent et al., 2011). For example, Raphael et al. used integrative analysis of TCGA pancreatic ductal cancer data (Raphael et al., 2017), but their focus was somatic alterations and molecular subtyping. Using the TCGA data, a number of DNA methylation pattern analyses have been reported for multiple cancers (Noushmehr et al., 2010; Aine et al., 2015; Yang et al., 2015); but for PDAC, this is still lacking. Unlike in our previous study (Mishra and Guda, 2017), in which we performed integrative analysis of all types of pancreatic cancers (PC) in the TCGA database, the present work is focused exclusively on PDAC, that is, this report does not contain any other subtypes of PC. In this PDAC study, we analyzed differential DNA methylation, gene expression, miRNA and lncRNA expression, and association of promoter DNA methylation with gene expression and lncRNA expression (**Figure S1**). Next, we examined whether those genomic and transcriptional changes corresponded with patient survival in a significant way. Overall, in the current study, we identified several prognostic markers for pancreatic ductal cancer.

MATERIALS AND METHODS

Clinical Data and Samples

We downloaded the current study view clinical data as of August 2018 from cBioPortal (Gao et al., 2013). The TCGA database has a total of 186 pancreatic cancer patients. Based on the described neoplastic and histological information of these patients in the clinical files, we selected 154 patients who had PDAC unambiguously. We excluded the other patients who had endocrine, invasive adenocarcinoma, undifferentiated or mixed pancreatic cancers (**Table S2**). CpGs/genes/miRNAs/lncRNAs with missing values in ≥20% samples, and similarly, samples with missing values of ≥20% of CpGs/genes/miRNAs/lncRNAs were excluded from further analysis.

DNA Methylation, RNAseq, and miRNAseq Data

The Bioconductor tool *TCGAbiolinks* (Colaprico et al., 2016) was used to download the TCGA level-3 data on DNA methylation (Illumina HumanMethylation450 BeadArray), gene expression (IlluminaHiSeq RNASeqV2), and lncRNA and microRNA expression (IlluminaHiSeq miRNAseq). The DNA methylation data also contains β values for 485,577 CpG sites with annotations for transcripts from GENCODE v22, the associated CpG island (CGI), CpG sites' distance from the nearest transcription start site (TSS), and CpG coordinates as per GRCh38 reference genome. The β values are calculated as (M/M+U) which ranges

between 0 and 1, where M is the methylated allele frequency and U is the unmethylated allele frequency. Therefore, a higher β values indicate a higher level of methylation. The gene expression data were obtained for each of the 60,483 GENCODE v22 genes in each sample. The miRNASeq data for each sample have single raw read counts and reads-per-million (RPM) counts for 1,881 miRNAs that are annotated in miRBase v21. As TCGA PDAC samples were processed in batches at different sites of the consortium, the data can be vulnerable to batch effects. Before starting the PDAC data analysis we first checked for possible batch effect in different types of data using Mbatch (Akbani et al., 2010).

Methylation Data Processing

Beta values of CpG probes mapped against X, Y, and mitochondrial chromosomes were excluded from analyses to eliminate gender bias. CpGs with missing β values (approximately 20% of the samples) were also excluded. To estimate the remaining missing values in the data, we used the k-nearest neighbor-based imputation method using the *imputeKNN* module of the R tool (R Core Team, 2019), *impute* (Troyanskaya et al., 2001). We also removed the data from CpG probes which overlapped with repeat masker and SNPs from dbSNP v151 with minor allele frequency (MAF) > 1% (Zhou et al., 2017). Statistical analyses of DNA methylation of 162 samples (153 primary tumors and nine normal samples) were performed at two different levels, i.e., the CpG site level, and the region level.

CpG probes were independently mapped in six different subregions of the genes: TSS200 (the region from TSS to 200 bp upstream of TSS), TSS1500 (200–1,500 bp upstream from TSS), 5'UTR, 1st exon, gene body, and 3'UTR. DNA methylation characteristics in the known UCSC CpG island, shores (regions 0–2 kb from CpG islands), and shelfs (regions 2–4 kb from CpG islands) were also analyzed.

Logistic Regression Analysis

We used logistic regression in R to classify the tumor and normal samples on the basis of their DNA methylation, gene expression, lncRNA expression, and miRNA expression data. Logistic regression was performed by using *lm* function in R. R package, *ROCR* was used to evaluate logistic regression performance, calculate the area under curve (AUC), and generate receiver operating characteristic (ROC) curve plots (Sing et al., 2005).

Differential Methylation Analysis

The β values for CpGs after preprocessing and imputation analyses were further normalized by using the beta mixed integer-quantile normalization (BMIQ) tool to adjust for type I and type II probes in data by using R tool, *BMIQ* (Teschendorff et al., 2013). The R package, *limma* was used for conducting supervised differential methylation analyses. For a CpG site to be considered differentially methylated, the primary tumor and normal samples were to have a mean β value difference of at least 0.2 ($\Delta\beta \geq 0.2$), and the BH adjusted *p*-value less than 0.005. Using the R tool, *gtrellis*, we generated circular plots of 10 Mb sliding windows for each chromosome to examine differentially methylated CpGs that had differential methylation frequencies (Gu et al., 2016). Next, we determined the methylation frequency per megabase pair (Mb) for each chromosome by calculating the total number of dm-CpGs in the chromosome and dividing by the length of the chromosome (Mb) using the GRCh38. Hypermethylation and hypomethylation frequencies were also calculated for each autosomal chromosome in a similar manner. For each chromosome, when the ratio between hypermethylation to hypomethylation frequencies was ≥1.5, we considered that chromosome to be predominantly hypermethylated. On the other hand, if the hypomethylation to hypermethylation frequency ratio is ≥1.5 we considered that chromosome to be predominately hypomethylated.

Differentially Methylated Regions (DMRs) Analysis

Differentially methylated region (DMR) analyses were performed using the Bioconductor tool *DMRcate* (Peters et al., 2015). *DMRcate* first calculates differential methylation at individual CpG sites derived by using moderated *t*-statistic from *limma* (Ritchie et al., 2015). After correcting for false discovery rate (FDR), regions of significant dm-CpGs were agglomerated into groups where the distance between two consecutive probes is within 1 kb. Only those DMRs that have at least two dm-CpGs with adjusted *p*-value < 0.01 within 1-kb distance were considered for DMR analysis. Next, we annotated the overlapping promoter regions (+/−2,000 bp from TSS) and generated a plot of DMRs by using the Bioconductor package *Gviz*.

RNASeq and miRNASeq Data Processing

The TCGA level-3 RNASeq data contain a single raw read count and a normalized expression value for each gene. In contrast, the GDC data portal has different types of level-3 data. From the GDC, we used HT-Seq raw read counts data for differential gene expression and the FPKM-UQ for correlation analysis. These expression values were generated by aligning the reads with the GRCh38 reference genome and then quantifying the mapped reads for the genes. TCGA level-3 miRNASeq data contain raw read count for each miRNA in the miRBase database, which was derived by exact mapping of miRNASeq data (Chu et al., 2016).

Differential Gene Expression Analysis

For differential gene expression analysis, the expected counts data from 146 primary PDAC and three normal samples were used. Before differential expression analysis, we removed all genes with missing expression values (~20% of the samples) and also genes which had CPM (count per million) numbers less than one (about 25% of the samples). After preprocessing, we used the Bioconductor tool, *DESeq2* (Love et al., 2014) for differential gene expression analysis, for which, a cutoff value of 0.01 for both raw *p*-value and Benjamini–Hochberg (BH) (Benjamini and Hochberg, 1995) adjusted *p*-value were applied. For differential miRNA analysis, we used raw read counts in *DESeq2* with a BH adjusted *P*-value of ≤0.01.

Correlation Between DNA Methylation and Gene Expression

For the correlation analysis, primary tumor samples of 146 patients that contained both DNA methylation and gene expression data were used. Correlation between promoter DNA methylation and corresponding gene expression was done by using linear regression function in the R package, *cor.test*. Methylation and expression levels (log2 (FPKM-UQ + 1) of genes were tested for non-zero correlation using Spearman's correlation, after excluding all samples with a correlation value of zero. Any association between DNA methylation and gene expression was considered as significant if the p-value ≤ 0.005 and rho ≥|0.25|.

Pathway Enrichment Analysis

Bioconductor package, *clusterProfiler* (Yu et al., 2012) was used for enrichment analysis of differentially expressed genes (DEG). KEGG canonical pathways were used for pathway enrichment analysis. We used BH adjustment p-values of 0.05 and a minimum of five and maximum of 500 genes as selection criteria for every significant pathway. For the pathway enrichment analysis of dm-CpGs, we used '*gometh*' module of Bioconductor tool *missMethyl* (Phipson et al., 2016). Genes associated with dmCpGs (Δβ ≥ 0.2) in the Illumina Human 450K BeadChip are obtained from the annotation package, *IlluminaHumanMethylation450kanno.ilmn12.hg19*. All GO and KEGG terms were tested using '*gometh*' function, and false discovery rates were calculated using the BH method.

Survival Analysis

To reveal the roles of differentially expressed genes and miRNAs on patient survival, PDAC patients were classified into high and low expression groups, using the median expression of genes as the cut-off value. For the analysis of promoter region DNA methylations, we used β value cutoff of ≥0.5 (high) and ≤0.3 (low) groups. We analyzed only those CpG sites that were differentially methylated (±1,500 bp from TSS) and also negatively correlated with gene expression. We used the R tool, *survival*, for survival analysis, and Kaplan–Meier (KM) survival plots were generated. In addition, we performed Cox-regression analyses. For both analyses, we selected CpGs that had p-value ≤ 0.05. For gene expression, miRNA, and lncRNA expression and patient survival analyses, we used all available genes in the analysis and divided PDAC patients into two classes based on the median expression. PDAC patients that were above the median, were classed as the high expression group, and those below the median were classed as the low expression group.

RESULTS

We downloaded level-3 DNA methylation, gene expression, and miRNA expression data from TCGA using Bioconductor tool, *TCGAbiolinks*, and systematically carried out data cleaning, global unsupervised analyses, and detailed individual and integrative analyses on DNA methylation, mRNA, and miRNA expression datasets. To understand the functional significance and relevance of the differentially-expressed and differentially-methylated genes in PDAC, we also performed downstream analyses using pathway enrichment tools and Cox-regression and Kaplan–Meier survival plots. Complete flow-chart of the data analysis is available in **Figure S1**.

Global DNA Methylation Analysis

We performed the Wilcoxon rank test to analyze the overall difference in DNA methylation levels in six different gene sub-regions (TSS200, TSS1500, 1st exon, 5′UTR, 3′UTR, and gene-body) and five methylated genomic regions (CpG-island, s-shore, n-sore, s-shelf, and n-shelf). For this analysis, we combined the β values of all CpGs in corresponding regions for tumor and normal samples. Our analyses revealed that CpG segments close to TSS and also the islands themselves have, in general, a higher level of DNA methylation in tumor samples (**Figure S2**). Specifically, DNA methylation levels of TSS200, TSS1500, 1st exon, 5′UTR, island, s-shore, and n-shore regions were higher in the tumor. In contrast, DNA methylation levels were low in genomic regions that are away from the TSS and the CpG islands (**Figure S2**).

We observed a total 12,083 differentially methylated CpGs (dm-CpGs) with Δβ ≥ |0.2| between tumor and normal samples; out of these 7,378 were hypermethylated and 4,705 were hypomethylated (**Table S3**, **Figure S3**). At even higher thresholds (Δβ ≥ |0.3|), the number of dm-CpG sites dwindled to 1,741. **Figure 1A** shows all dm-CpG results from each autosomal chromosome at Δβ ≥|0.2| depicted in the outer circle of the circos plot. The two innermost circles show the density of hyper- and hypomethylation in a 10 Mb sliding window across the genome. The distribution of dm-CpGs in twelve different genomic subregions is shown in **Table 1** and **Figure 1C**. A total 4,610 dm-CpGs were observed within the promoter regions of genes i.e., ±1.5 Kb from the TSS of genes. We also observed that the regions close to the CpG islands (island, shore) and the promoters (TSS200, TSS1500, promoter, 1st Exon, 5′UTR), were predominantly hypermethylated (**Figure S2**—1.5kb distribution plot), while regions away from promoter (shelf) and promoter (3′UTR, gene body) are hypomethylated (**Table 1**, **Figure 1C**).

In PDAC tumors, we observed that chromosome 1 and 2 contained the highest numbers of dm-CpGs, while chromosome 14, 15 had the lowest. Such differences are expected given the large sizes of chromosomes 1 and 2. To size-normalize for all chromosomes, we calculated the methylation frequency/Mb for each chromosome to compare the net differential methylation. The size-normalized DNA methylation frequencies indicated that chromosome 20 has the highest differential methylation frequency (14.76 dm-CpGs/Mb) while chromosome 18 has the lowest (0.82 dm-CpG/Mb), as shown in **Table 2** and **Figure 1B**. Except in chromosome 9, hypermethylated CpG sites were more prominent than hypomethylated sites in all the other chromosomes (**Table 2**). We also observed that chromosomes 10 and 18 were extensively hypermethylated to the extent that the hypermethylation frequencies for these two chromosomes were three times higher than the hypomethylation frequencies (**Figure 1B**, **Figure S4**).

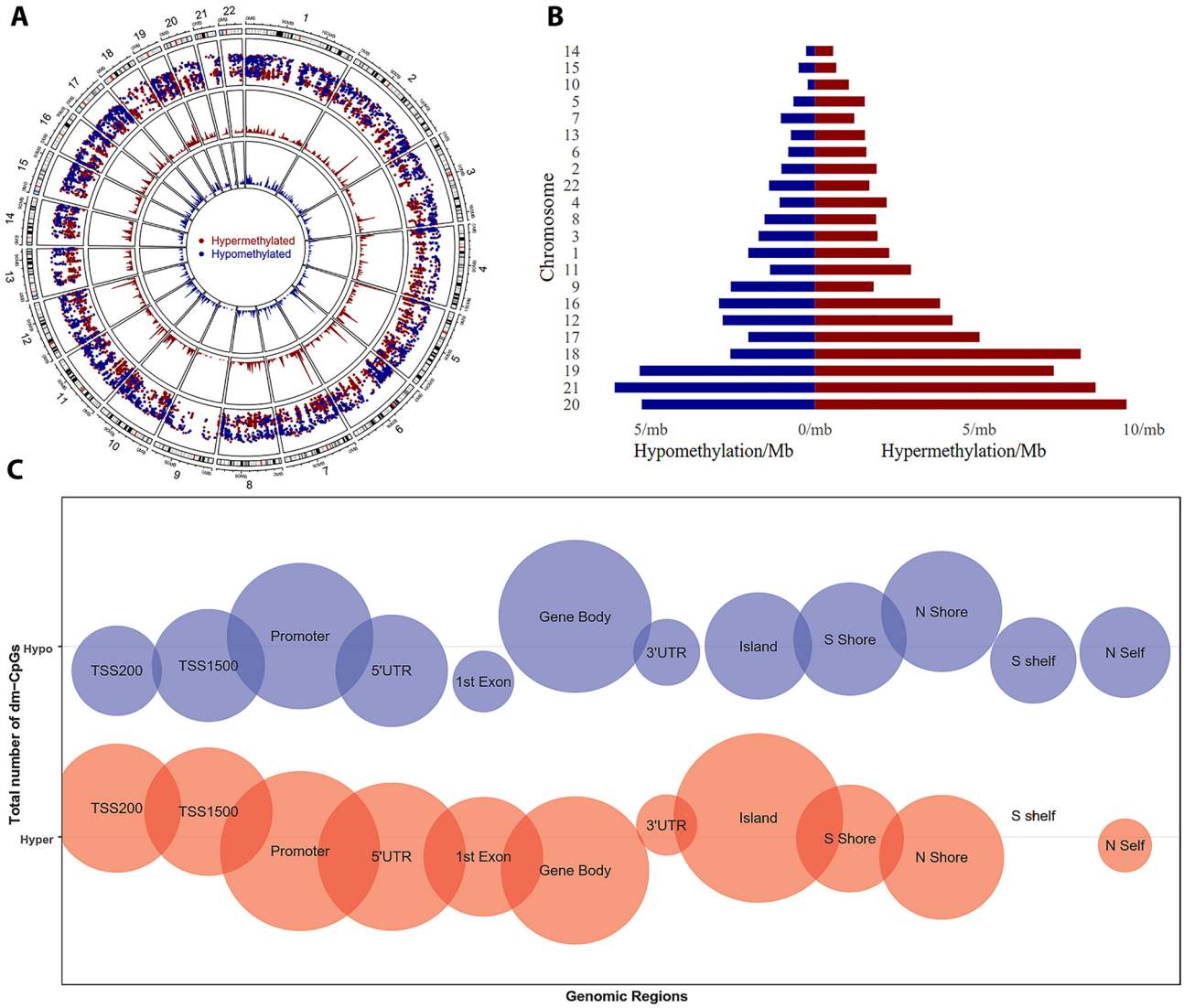

FIGURE 1 | Differential DNA methylation distribution. **(A)** Circular plot of CpGs, chromosomes are shown in a clockwise direction from 1 to 22 in the outermost circle. Chromosomes X, Y, and M were excluded from analysis. The two innermost circles represent the differential hypermethylation and hypomethylation frequencies in a 10 Mb sliding window across the genome. **(B)** Pyramid (stacked) plot for differential hyper and hypomethylation frequencies for each chromosome. Chromosomes are sorted based on total differential methylation in per megabase pair length of the chromosomes. **(C)** Bubble plot of differentially methylated CpGs in genomic regions. Size of bubble represents a total number of dm-CpGs.

To locate genomic regions with high epigenomic perturbations, we calculated dm-CpG frequencies of chromosomal segments in 10 MB sliding windows. Our analysis revealed that chr7:27,000,001–28,000,000 has the highest dm-CpG frequency with the entire region mostly hypermethylated (**Figure 1A**, inner red circle). The region contains several HOX-family genes as HOXA1, HOXA3, HOXA7, HOXA10, HOXA11, and HOXA13.

Genome-Wide Analysis of Differentially Methylated Regions (DMRs)

The normal differential methylation analysis process does statistical testing for individual CpG sites, but regulatory methylation targets are most commonly clustered into short regions. Clusters of hypermethylated CpG sites in the promoter region of a gene are usually associated with epigenetic silencing of the gene (Jones and Baylin, 2002). Differentially methylated regions (DMRs) comprise multiple consecutive methylated CpG sites with at least two dm-CpGs, therefore detecting DMRs is more biologically relevant (Weaver et al., 2004; Bert et al., 2013).

In all, we identified 779 DMRs across the genome in PDAC. Chromosome 7 showed the highest (74) and chromosome 21 showed the lowest (6) DMRs (**Table S3**). The DMRs were of different lengths, ranging from 3bp to ~11kb. There were 116 short (<100 bp) DMRs, 84 long (>2 kb) DMRs. The number of dm-CpGs within DMRs ranges from 2 to 45. These DMRs

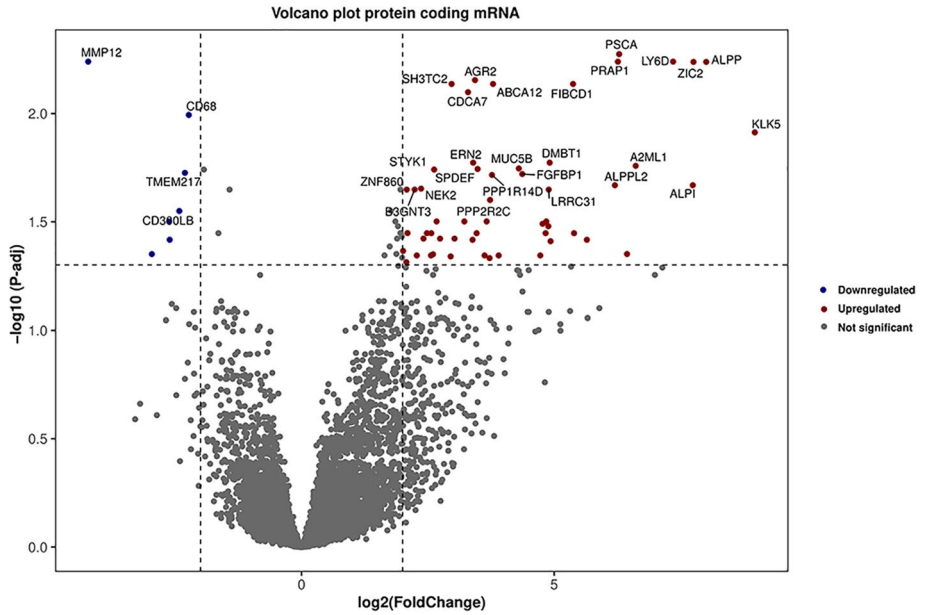

FIGURE 2 | Volcano plot for the differentially expressed genes. Genes which are in red and blue colors are highly upregulated and downregulated, respectively in PDAC. Vertical and horizontal dot line represents a cutoff point for log fold-change *p*-value respectively.

TABLE 1 | Distribution of differentially methylated CpG sites in different genomic and gene regions in pancreatic ductal adenocarcinoma ($\Delta\beta \geq 0.2$).

Genomic region	dm-CpG	Hypermethylated	Hypomethylated
3UTR	310	165	145
5UTR	1,144	1,935	433
1st Exon	815	682	133
Body	3,815	1,935	1,880
TSS200	1,172	935	237
TSS1500	1,536	915	441
Island	5,241	4,870	371
N Shore	1,378	807	571
N Self	388	148	240
S Shore	916	472	444
S shelf	320	105	215
Promoter	4,610	3,174	1,436

TABLE 2 | Differential methylation frequency per mega base-pair (Mb) for each autosomal chromosomes.

	Mb	CpG/Mb	Hyper/Mb	Hypo/Mb	Hyper vs Hypo
chr10	133.8	1.26	1.03	0.22	4.6
chr18	80.37	10.66	8.09	2.58	3.14
chr17	83.26	7.04	5.01	2.03	2.47
chr5	181.54	2.17	1.51	0.66	2.31
chr11	135.09	4.29	2.92	1.37	2.14
chr13	114.36	2.26	1.52	0.73	2.07
chr14	107.04	0.82	0.55	0.27	2.03
chr4	190.21	3.26	2.18	1.08	2.02
chr6	170.81	2.38	1.56	0.81	1.92
chr2	242.19	2.9	1.88	1.02	1.83
chr20	64.44	14.76	9.48	5.28	1.8
chr12	133.28	7	4.19	2.81	1.49
chr21	46.71	14.64	8.54	6.1	1.4
chr19	58.62	12.61	7.27	5.34	1.36
chr16	90.34	6.73	3.81	2.92	1.3
chr15	101.99	1.15	0.65	0.5	1.29
chr8	145.14	3.4	1.86	1.54	1.21
chr22	50.82	3.05	1.65	1.4	1.18
chr7	159.35	2.23	1.19	1.04	1.14
chr1	248.96	4.29	2.26	2.03	1.11
chr3	198.3	3.62	1.9	1.72	1.11
chr9	138.39	4.35	1.78	2.56	0.7

also overlap with the promoters of several HOX-family genes (**Table S3**). Examples of DMRs showing contrasting methylation patterns between normal and tumor samples on chromosome 9 and chromosome 2 are presented in **Figure S5**.

Differential Gene Expression Analysis

HTSeq read-counts for 146 PDAC patient tumors and three normal samples were downloaded from TCGA and differential gene expression analysis was performed on them using DESeq2 package. Trimmed mean of M-values (TMM) normalization was employed to account for library size variations among samples (Robinson and Oshlack, 2010). We identified 90 differentially expressed genes (80 protein-coding, seven lncRNA, two antisenses, and one Ig-V gene) after adjusting to *p*-value < 0.05 (significance corrected using the Benjamini-Hochberg method) (**Figure 2**, **Table S4**). From the 147 tumors and three normal samples, 10 differentially expressed miRNAs were found (**Table S4**).

Promoter DNA Methylation and Gene Expression Correlation Analysis

We used Spearman's test to examine correlations between promoter DNA methylation (within 1.5kb from TSS) and gene

expression using the R function, *cor.test*. Correlations that had rho values of ≥ |0.25| and BH adjusted *p*-values of < 0.005 were taken as significant. We observed correlations of 30,619 promoter CpGs with the expression of 8,932 genes, the majority of which were negatively correlated (25,077 CpGs with 7,518 genes), with only a minority (5,605 CpGs with 2,937 genes) showing positive correlations. At higher rho threshold values (|0.5|) and low FDR (<0.005), we observed correlations of 4,971 CpGs with the expression of 1,744 genes, out of which most (4,568 CpGs with 1,602 genes) were negatively correlated and fewer (407 CpGs with 212 genes) were positively correlated (**Table S5**, **Figure S6**).

Similar Spearman's analyses were performed for finding correlations between CpGs and lncRNAs. We identified 1,216 CpGs that were significantly correlated with 442 lncRNAs, out of these the great majority (1,039 CpGs with 368 lncRNAs) were negatively correlated and fewer (177CpGs with 95 lncRNAs) were positively correlated. At higher thresholds (rho ≥ |0.5| and BH adjusted *p*-value ≤ 0.005), we observed that 199 CpGs were correlated with 84 lncRNAs, out of which 174 CpGs showed negative correlations with 72 lncRNAs, and 25 CpGs were positively correlated with 12 lncRNAs (**Table S5**).

Pathway Enrichment Analysis

Analyses of differentially methylated CpGs using the Bioconductor missMethyl pathway tool indicated the enrichment of several KEGG pathways (**Table 3**). Several critical cancer-related pathways such as MAPK signaling, Rap1 signaling, calcium signaling were shown in the list. We also observed the enrichment of the nicotine addiction pathway as corroborated by the fact that these patients were cigarette smokers (**Table S3**). In case of differential expression, we observed only 80 differentially expressed genes and no significant pathways were enriched from that list of genes.

Survival Analysis

We used an in-house R code to perform survival analysis base on the DNA methylation, gene expression, miRNA, and lncRNA results. This R code uses the R tools, *survival,* and *survMiner* in the background and performs the Cox regression and log-odd tests, and generates KM-plots for CpGs, genes, miRNAs, and lncRNAs—all in the context of significant difference in patient survival in the high and low expression groups. In Cox regression analysis, we used low expression and methylation group of samples as reference. The hazard ratio (HR) > 1 indicates high expression group patients have low survival and <1 suggests high survival.

We conducted survival analysis of PDAC patients with respect to differentially methylated CpGs (*p*-values for both log-odd and Cox regression ≤ 0.05). The results identified 439 CpGs that may have survival roles. Out of these, 80 showed survival relationship at a stringent selection criterion (*p*-value ≤ 0.01). In contrast, survival analysis of the gene expression data indicated 1,954 genes that may influence PDAC patient survival with *p*-value ≤ 0.05 (**Table S5**). When we reduced survival *p*-value cutoff to 0.01, this gene number goes down to 518. Similarly, we observed 236 lncRNAs which correlated with survival at *p*-value ≤ 0.05, whereas this number came down to 74 at *p*-value cutoff of 0.01. For miRNA, these numbers were 25 at *p*-value ≤ 0.05 that were reduced to 7 at *p*-value ≤ 0.01.

Correlative Analysis of Gene Expression and Survival

Genes and genomic regulatory loci that are differentially expressed and correlated with patients' survival could be important for understanding the initiation and progression of PDAC. Integrative analysis of patient survival and differential expression identified 17 genes that passed our tests at BH adjusted *P*-value ≤ 0.05 for both differential expression and patient survival or five genes when the thresholds were decreased to 0.01 for both DEG and survival analysis (**Table 4**). In these tests, we did not observe any differentially expressed lncRNAs that correlated with PDAC patient survival.

Further analysis of genes that have dm-CpGs in the promoter regions (Δβ ≥|0.2|, FDR < 0.005) and showing a negative correlation in corresponding gene expression (rho ≤ -0.5, FDR < 0.005) showed that a total of 93 CpGs have a significant difference (*p*-value ≤ 0.05) in survival between high and low patient groups. This number further goes down to 4 if we use *p*-value ≤ 0.01 in the survival analysis (**Figure 3**).

In the case of lncRNA, we observed that three promoter dm-CpGs showing a negative association with lncRNA expression have a role in overall patient survival (*p*-value ≤ 0.05). This number goes down to two if we further reduce survival *p*-value to 0.01. List of these CpGs with survival details are shown in **Table S7**.

Analysis of Genes of Mucin Family

Our DEG analysis showed that MUC2, MUC5B, and MUC13 were significantly upregulated in PDAC (**Table S8**). MUC1, MUC6, and MUC16 showed overexpression but it was not statistically significant (BH adjusted *P*-value > 0.05). We noted that MUC5B, which was overexpressed in PDAC (BH adjusted *P*-value = 0.018) has also two hypomethylated CpGs (cg20911165 and cg03609102) in its promoter region, which also showed a negative correlation with MUC5B expression (**Figure 4**). We also observed that expression of MUC1, MUC3, MUC4, MUC6, MUC15, MUC17, MUC20, and MUC21 genes was negatively correlated with the promoter methylation (**Table S5**).

TABLE 3 | KEGG pathway analysis for differentially methylated genes. We used *missMethyl* tool for pathway analysis. For each enriched pathways, N is the total gene in given pathways, DN is the number of mapped genes in hg38 against differentially methylated CpGs, P.DM is the *p*-value, and FDR is the BH adjusted *P*-value.

Pathway	N	DM	P.DM	FDR
Neuroactive ligand-receptor interaction	252	90	1.37E-09	4.51E-07
Calcium signaling pathway	173	73	5.36E-07	8.84E-05
Rap1 signaling pathway	203	77	3.40E-05	0.00374
Nicotine addiction	36	20	0.00014	0.01117
MAPK signaling pathway	283	96	0.00031	0.02041
cAMP signaling pathway	191	66	0.00037	0.02043
Salivary secretion	81	31	0.00064	0.03027
Circadian entrainment	95	40	0.00102	0.03834
Morphine addiction	88	38	0.00105	0.03834
Mucin type O-glycan biosynthesis	29	14	0.00132	0.04339

TABLE 4 | List of probable prognostic gene/miRNA biomarkers for pancreatic ductal adenocarcinoma. List of genes and miRNA which have very low *p*-value in survival and *DESeq2* differential gene expression analysis, and high area under curve (AUC).

Gene	log2FC (DESeq)	*P*-value (DESeq)	*P*-adj (DESeq)	*P*-value (log Rank)	*P*-value (Cox)	Beta (Cox)	HR (95% CI)	AUC
ASPM	1.986477	0.000131	0.036978	0.05	0.052	0.43	1.5 (1–2.4)	0.96
B3GNT3	2.237597	4.19E-05	0.022447	0.011	0.012	0.57	1.8 (1.1–2.8)	0.93
BMF	-1.41789	4.63E-05	0.022447	0.03	0.032	-0.47	0.62 (0.4–0.96)	0.89
CD300LB	-2.4129	6.62E-05	0.028185	0.008	0.0091	-0.58	0.56 (0.37–0.87)	0.83
CD68	-2.22506	8.73E-06	0.010149	0.035	0.037	-0.46	0.63 (0.41–0.97)	0.92
CENPF	1.890233	0.000144	0.037802	0.018	0.019	0.52	1.7 (1.1–2.6)	0.96
DEPDC1B	2.074024	0.000251	0.048577	0.005	0.0054	0.63	1.9 (1.2–2.9)	0.95
DMBT1	4.911577	1.72E-05	0.016867	0.023	0.024	-0.51	0.6 (0.39–0.94)	0.89
DTL	1.640861	0.000222	0.045194	0.026	0.028	0.49	1.6 (1.1–2.5)	0.97
ERCC6L	1.957758	4.48E-05	0.022447	<0.001	0.00056	0.79	2.2 (1.4–3.5)	0.97
FAM111B	1.768394	6.58E-05	0.028185	0.022	0.024	0.51	1.7 (1.1–2.6)	0.96
HIST1H2BC	2.740626	0.000145	0.037802	0.003	0.0039	0.66	1.9 (1.2–3)	0.95
HIST1H2BJ	2.949891	0.000228	0.045665	0.032	0.034	0.48	1.6 (1–2.5)	0.93
HIST1H3H	3.383461	0.000154	0.038262	0.016	0.017	0.55	1.7 (1.1–2.7)	0.91
KIF4A	1.855617	8.42E-05	0.031507	0.013	0.014	0.55	1.7 (1.1–2.7)	0.95
NEK2	2.364569	3.95E-05	0.022212	0.001	0.0019	0.71	2.0 (1.3–3.2)	0.95
RASSF4	-1.6378	0.000121	0.035665	0.046	0.048	-0.43	0.65 (0.45–1.0)	0.97
hsa-mir-196b	3.542765	0.000697	0.020469	0.002	0.0022	0.69	2.0 (1.3–3.1)	0.83

Log2FC, log2 fold change; HR, hazard ratio; 95% CI, upper and lower 95% confidence interval values of hazard ratio (HR), and beta is β coefficient of a given variable for the Cox regression analysis.

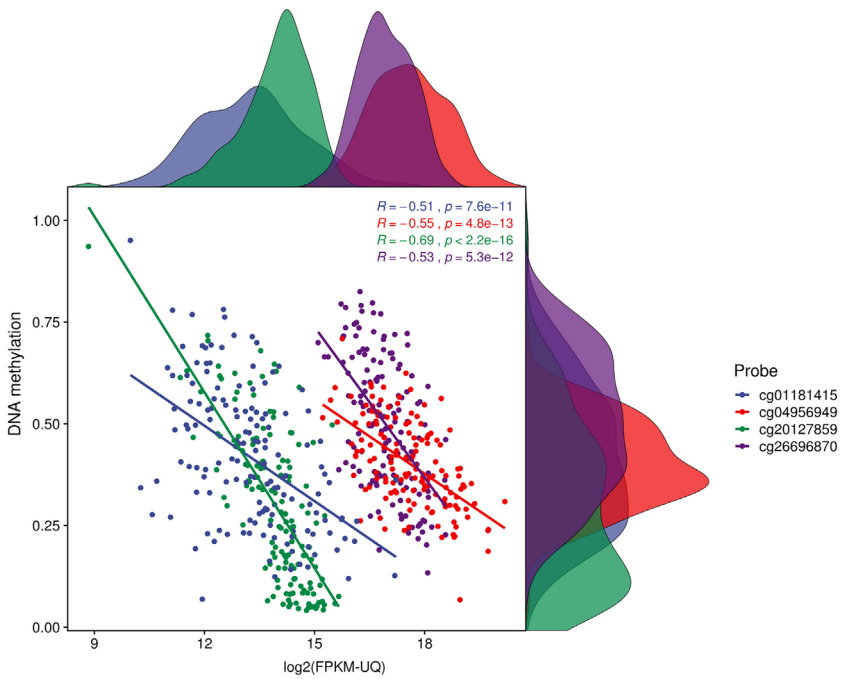

FIGURE 3 | Correlation plot for survival associated CpGs. We used CpGs which have survival p < = 0.01 and Spearman correlation > 0.5 (*p*-value < 0.005). This plot is for four promoter CpGs which are negatively correlated with genes expression and also strongly associated with patients' survival. Distribution of DNA methylation and gene expression in PDAC patients on the right side and top respectively.

DISCUSSION

Alterations in the promoter DNA methylation, as well as miRNA and lncRNA expression, play critical roles in cancer biology by up- or downregulating gene expression (Merlo et al., 1995; Ramachandran et al., 2016). DNA methylation pattern alterations can serve as useful biomarkers for distinguishing tumors from normal samples (Oh et al., 2013). Two previous studies by (Sato et al., 2008) and (Tan et al., 2009) had explored DNA methylation patterns in pancreatic cancer. Sato et al. used methylation-site specific PCR, and Tan et al. used GoldenGate methylation cancer panel array. Both of these technique have limited

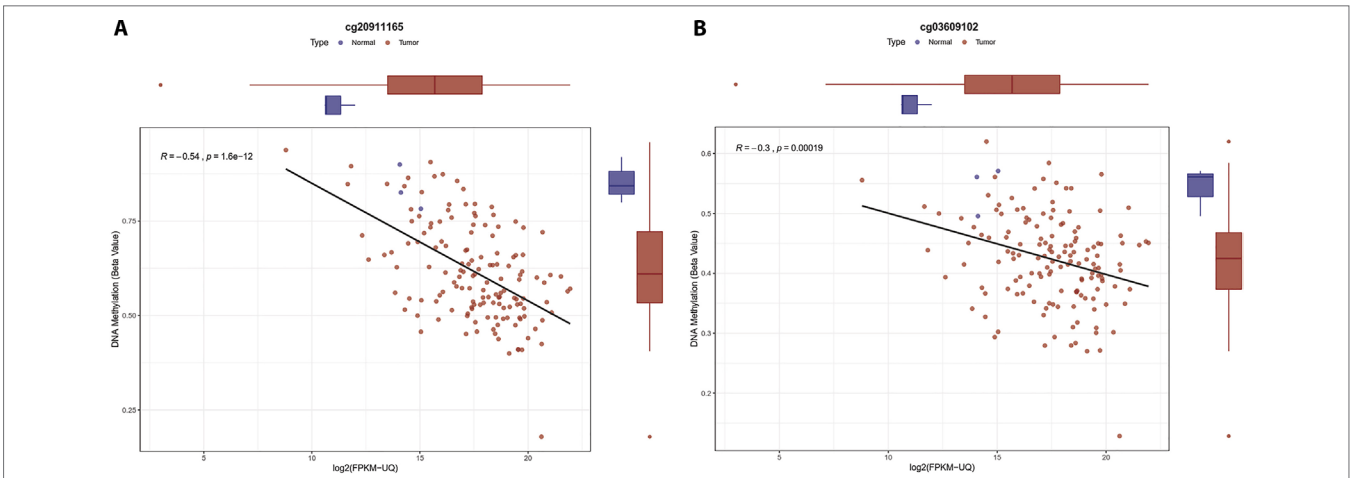

FIGURE 4 | Correlation plot for the MUC5B promoter methylation sites. Boxplot for gene expression and DNA methylation on top and right side respectively, tumor samples are in red and normal samples in blue color. **(A)** Correlation plot and boxplot for cg20911165. **(B)** Correlation plot and boxplot for cg03609102.

genome coverage and sensitivity. In addition, those studies used formalin-fixed paraffin embedded samples, xenografts, and pancreatic cancer cell lines, which might affect the quality of the results. On the other hand, the current study is based on TCGA Illumina HumanMethylation450 chip from fresh tissue samples, which has higher genome coverage with greater consistency and accuracy. Our study is more comprehensive, since we scoped for differential methylation, differential gene expression, differential miRNA, differential lncRNA in a genome-wide manner, and we also correlated these results with patient survival. To avoid gender bias, we excluded all CpG probe and gene expression data from X and Y chromosomes. Our results demonstrated that all chromosomes had dm-CpGs in PDAC (**Figure 1A**, **Table 2**). CpG islands, promoter, and their proximal regions had more hypermethylated CpG sites compared to regions away from islands and promoters (**Figure 1C**, **Table 1**, **Figure S6**). We observed that several chromosomal regions which have a high frequency of dm-CpGs are also a region which is differentially methylated.

In this study, CpG sites in the zinc finger protein 154 (ZNF154) promoter region were hypermethylated and showed a negative correlation with ZNF154 gene expression. We found that promoter of ZNF158 overlap with a region which has the highest differential methylation frequency in chromosome 19. The survival analyses indicated that the cg03234186 high methylation group patients had a low overall survival (HR = 1.7) in PDAC (**Table 5**). ZNF154 hypermethylation is a urine-based prognostic biomarker for bladder cancer, where hypermethylation correlates with recurrence-free survival of the patients (Reinert et al., 2012). ZNF154 hypermethylation may also be a blood-based prognostic biomarker for solid tumors (Sanchez-Vega et al., 2013; Margolin et al., 2016). Recently, Zhang *et al.* located CpG hypermethylations at ZNF154 promoter (cg03234186, cg12506930, cg26465391) by studying the TCGA prostate cancer archive. Hypermethylation downregulates ZNF154 expression and survival analysis suggest that hypermethylation of this site is associated with poor survival of patients (Zhang, Shu et al., 2018).

KRAB zinc-finger tumor suppressor ZNF382 expression is suppressed by promoter methylation in esophageal squamous cell carcinoma (Zhang, Xiang et al., 2018). In PDAC, we identified hypermethylations in five CpG sites in the ZNF382 promoter region, which are negatively correlated with gene expression. Logistic regression-based classification showed an AUC of 1.0 for all these CpGs. Hypermethylation of (cg02587316, cg18630667, and cg05020604) was associated with low survival of PDAC patients (**Table 5**). Above findings suggest that methylation of cg03234186 (ZNF154), and cg02587316, cg18630667, cg05020604

TABLE 5 | List of probable prognostic DNA methylation biomarkers for pancreatic ductal adenocarcinoma.

CpG	log-rank	HR (95% CI)	*P*-value Cox	Correlation	*P*-value	*P*-adj	AUC
cg02587316	0.029	1.8 (1.1–3.2)	0.032	-0.56	<1.0E-21	<1.0E-21	0.95
cg18630667	0.012	2 (1.1–3.4)	0.014	-0.56	<1.0E-21	<1.0E-21	0.96
cg05020604	0.015	1.9 (1.1 3.3)	0.017	-0.55	<1.0E-21	<1.0E-21	0.96
cg03234186	0.043	1.7 (1.0–2.9)	0.046	-0.67	<1.0E-21	<1.0E-21	0.90

AUC, area under curve; HR, hazard ration; P-value Cox, the P-value for cox regression analysis. P-value and P-adj are the raw P-value and BH adjusted P-value respectively for the Spearman rank correlation.

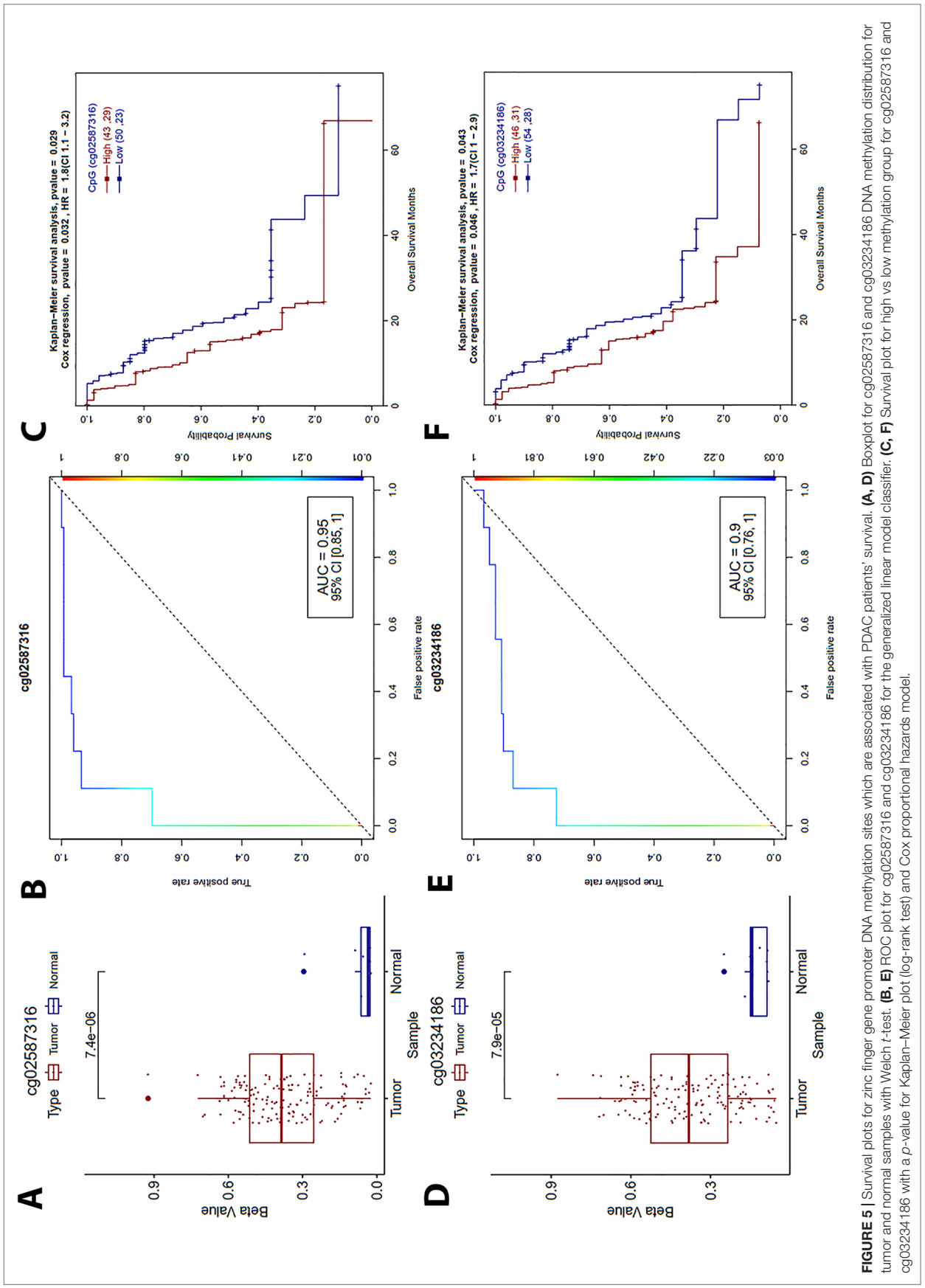

FIGURE 5 | Survival plots for zinc finger gene promoter DNA methylation sites which are associated with PDAC patients' survival. **(A, D)** Boxplot for cg02587316 and cg03234186 DNA methylation distribution for tumor and normal samples with Welch *t*-test. **(B, E)** ROC plot for cg02587316 and cg03234186 for the generalized linear model classifier. **(C, F)** Survival plot for high vs low methylation group for cg02587316 and cg03234186 with a *p*-value for Kaplan–Meier plot (log-rank test) and Cox proportional hazards model.

(ZNF382) have the potential to serve as prognostic biomarkers for PDAC (**Figure 5**).

The differentially expressed miRNAs include hsa-mir-196-a1/2 and hsa-mir-196b, both of which are HOX-cluster embedded miRNA members of the evolutionarily conserved miR-196 gene family (Mansfield and McGlinn, 2012; Fantini et al., 2018). The hsa-mir-196-a1 gene is located in the intergenic region between HOXB9 and HOXB13 on human chromosome 17; the hsa-mir-196a-2 between HOXC9 and HOXC10 on chromosome 12, and the hsa-mir-196b is on chromosome 7. HOX genes such as HOX-B7 (Braig et al., 2010), HOXB8 (Yekta et al., 2004), and HOXA9 (Li et al., 2012) are targets of the miR-196 family. MiR-196b directly targets HOXA9, whose overexpression is associated with bad prognosis in leukemia (Li et al., 2012). The hsa-mir-196a-regulated HOX-B7 expression has a role in melanoma (Braig et al., 2010), it would be worth investigating the role of HOX-cluster gene regulation by miRNA and/or promoter methylations in pancreatic cancers.

Hsa-mir196-b has been reported as a biomarker for digestive tract cancers (Lu et al., 2016) and familial pancreatic cancer (Slater et al., 2014). Multiple studies indicate that hsa-mir196-b overexpression is bad for the cancer patient. For example, hsa-mir196-b overexpression is associated with poor prognosis in gastric cancer (Lim et al., 2013; Ge et al., 2014), and is also associated with accelerated invasiveness in epithelial ovarian cancer (Chong et al., 2017). Kanno et al., (2017) reported that hsa-mir-196b overexpression might be a prognostic biomarker for a bad outcome. In our current study, we also found that PDAC patients with hsa-mir-196b overexpression showed worse survival (**Table 4**, **Figure 6**), which further corroborates the role of hsa-mir-196b as a biomarker for PDAC.

MiR-125a is a tumor suppressor that induces apoptosis, mitochondrial energy disorder, and cellular migration through suppressing mitochondrial fission, and play an important role in pancreatic cancer (Pan et al., 2018). Metastatic colorectal cancer patients treated with bevacizumab in combination with FOLFOX have better progression-free survival (Kiss et al., 2017). In the current study, we observed that hsa-mir-125a is overexpressed but *P*-value was not significant, however, univariate Cox regression analysis suggested that patients with higher expression of mir-125a had a better overall survival (HR = 0.57) (**Table S6**). This finding suggests that hsa-mir-125a might be useful as a prognostic biomarker for PDAC.

Hsa-mir-135a-2 is a precursor of hsa-mir-135a; univariate log-rank test (*P*-value = 0.01) and Cox-regression analysis (HR = 0.55) suggest that higher expression is associated with better overall survival of PDAC patients. Cheng *et al*. reported that mir-135a is a metastasis inhibitor, and they observed similar survival trends in gastric cancer cell line data (Cheng et al., 2017). In our study, we also observed that hsa-mir-3200 expression is associated with good prognosis of PDAC (HR = 0.5) (**Table S6**).

From the survival analyses of protein-coding genes in PDAC, we observed 518 genes that had significant correlations with patient survival both in high and low expression cohorts. The aryl hydrocarbon receptor nuclear translocator like 2 (ARNTL2) gene, which codes for a helix-loop-helix transcription factor, was the most significant among all. Overexpression of this gene was reported to predict poor outcome for lung adenocarcinoma patients (Brady et al., 2016). To our knowledge, the role of ARNTL2 in PDAC was not explored before, and the current study showed that ARNTL2 overexpression had a strong association with poor survival (HR = 2.2) in PDAC patients.

In the contrary, overexpression of certain genes was also found to help extend patient survival. Overexpression of CELF2 and EFR3B were correlated with better PDAC patient survival (**Table S6**). CELF2 is a tumor suppressor (Subramaniam et al., 2011; Ramalingam et al., 2012), and EFR3B contributes to the control of the phosphorylation state and could affect the responsiveness of G-protein-coupled receptors in higher eukaryotes (Bojjireddy et al., 2015). The role of EFR3B is mammalian is still unexplored, nevertheless, our results indicated that its expression is a key indicator of patient survival.

The abnormal expression of many long non-coding RNAs (lncRNAs) has been reported as effectors in the progression of various cancers. Some of these lncRNAs may be useful as diagnostic indicators and anti-cancer targets (Petrovics et al., 2004; Gutschner et al., 2013). We explored whether lncRNAs were involved in PDAC and whether we can find any indication for their utility for the diagnosis and treatment of PDAC. However, none of their expression patterns were correlated with patient survival. It is possible that we needed more than the three tumor-adjacent normal samples for examining lncRNAs. Unfortunately, the present TCGA database has expression values for only three lncRNAs. However, we did find a few lncRNA expression and survival correlations at low *P*-value thresholds (P-value ≤ 0.05) that could be further tested for their role in patient survival (**Table S6**).

LINC00941 is an epigenetically-silenced lncRNA found in pan-cancer TCGA data analysis (Wang et al., 2018). In our study, we found that LINC00941 is overexpressed (*P*-value = 0.02) and that high expression correlated with poor prognosis (HR = 1.8). PVT1 is another lncRNA, which is upregulated in lung cancer and plays a crucial role in lung cancer progression (Li et al., 2018). In our study, PVT1 also turned up overexpressed (*P*-value = 0.009) and correlated with poor PDAC patient survival (HR = 1.60), logistic regression classification AUC is 0.88 (**Figure 7**). Therefore, PVT1 may prove useful as a potential biomarker for PDAC therapy. RP11-54H7.4 is another overexpressed lncRNA in the TCGA database that was reported as a candidate biomarker for lung squamous cell carcinoma prognosis (Tang et al., 2017). We also observed elevated expression of RP11-54H7.4 (not significant), and high expression group PDAC patients had worse survival (HR = 1.6) (**Figure 7**).

A few other lncRNAs had contributory roles in PDAC patient survival, but they did not differentially express. The cancer susceptibility candidate 11 (CASC11) lncRNA is among them. Based on a knockdown study, CASC11 is thought to have a promoting role in colorectal cancer growth and metastasis (Zhang et al., 2016). Our current study showed that CASA11 overexpression associated with low survival. The antisense lncRNA of GATA6 (GATA6-AS) interacts with an epigenetic regulator LOXL2 to regulate endothelial gene expression *via* changes in histone methylation (Neumann et al., 2018). Our

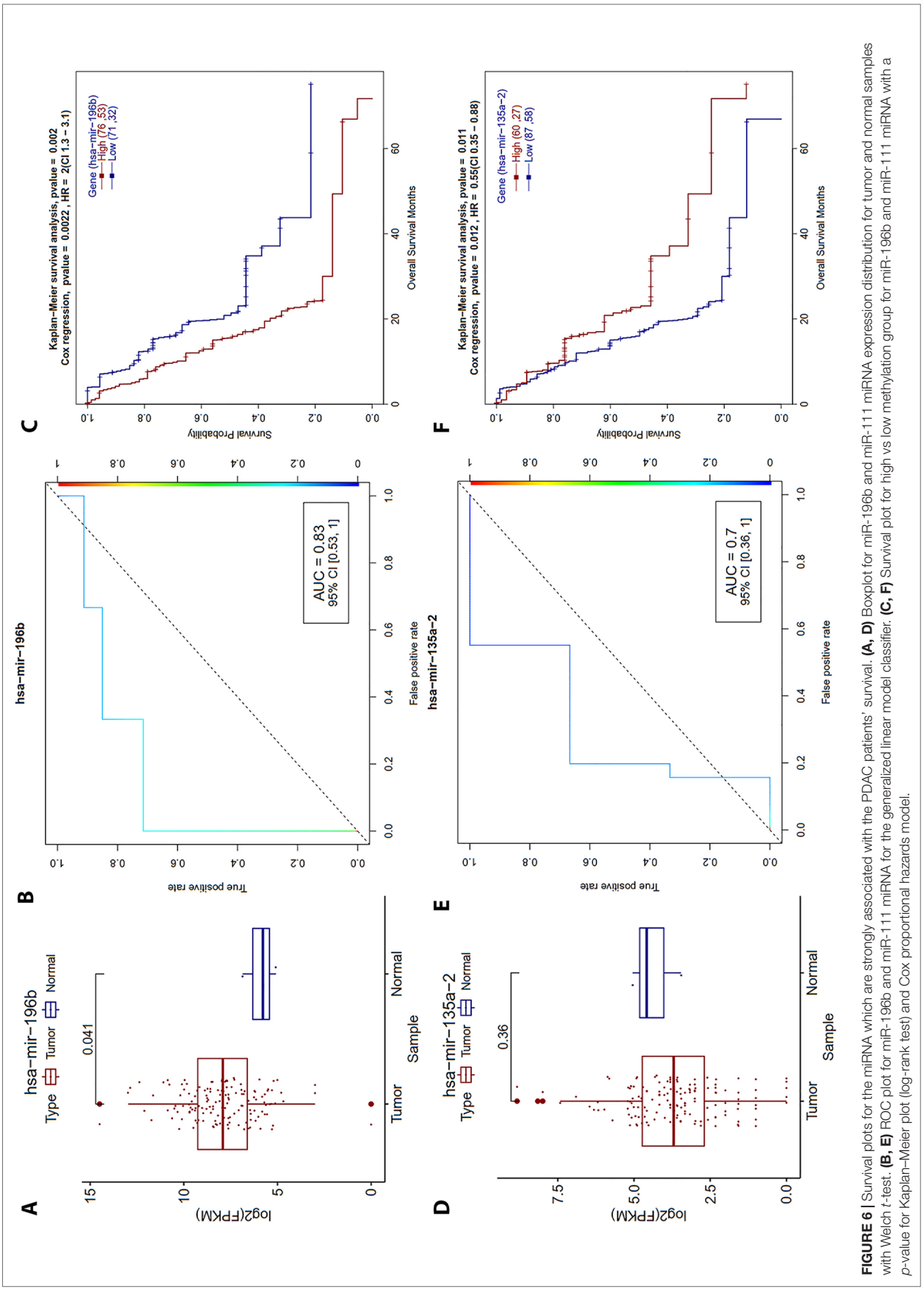

FIGURE 6 | Survival plots for the miRNA which are strongly associated with the PDAC patients' survival. **(A, D)** Boxplot for miR-196b and miR-111 miRNA expression distribution for tumor and normal samples with Welch *t*-test. **(B, E)** ROC plot for miR-196b and miR-111 miRNA for the generalized linear model classifier. **(C, F)** Survival plot for high vs low methylation group for miR-196b and miR-111 miRNA with a *p*-value for Kaplan–Meier plot (log-rank test) and Cox proportional hazards model.

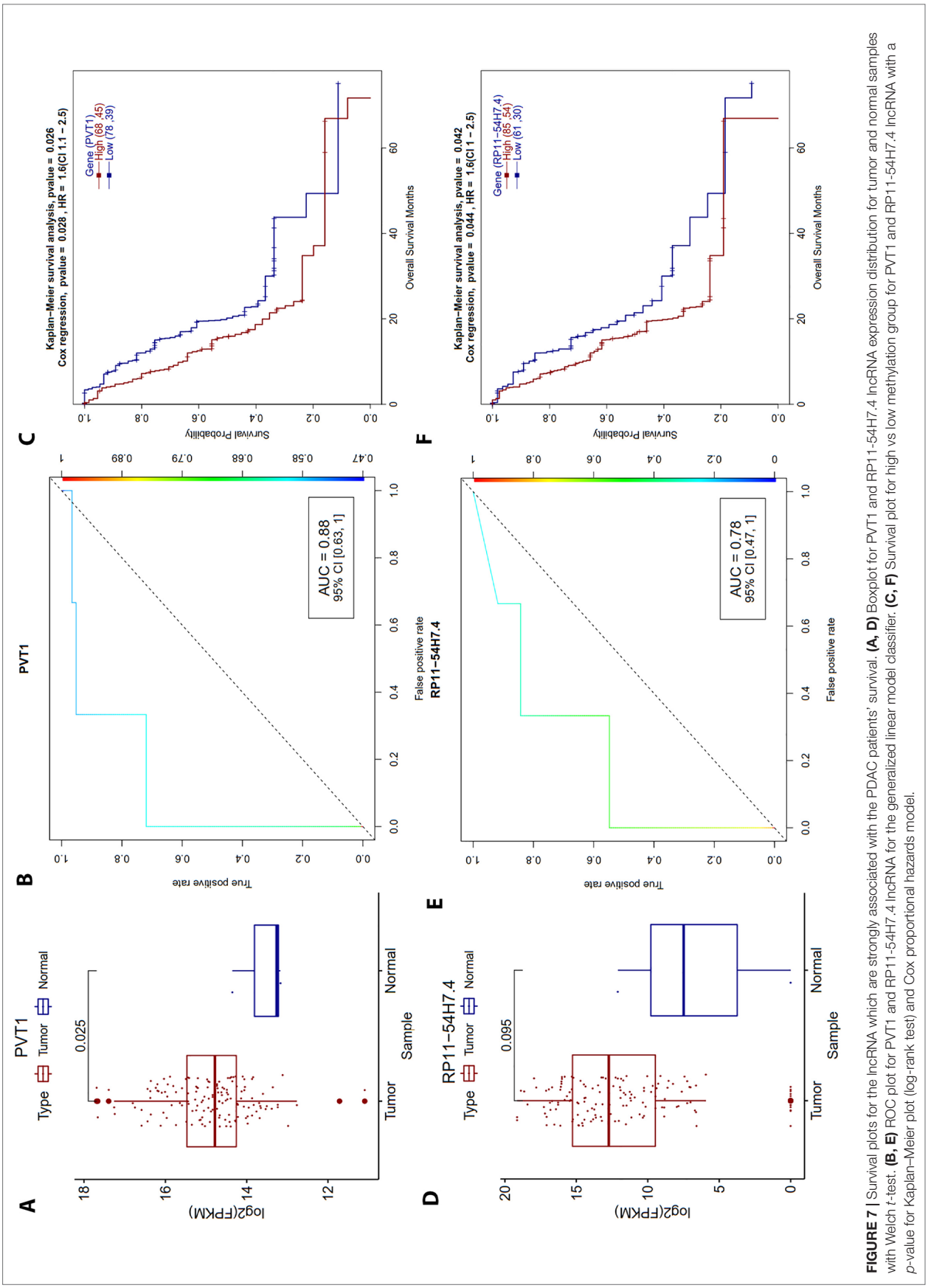

FIGURE 7 | Survival plots for the lncRNA which are strongly associated with the PDAC patients' survival. **(A, D)** Boxplot for PVT1 and RP11-54H7.4 lncRNA expression distribution for tumor and normal samples with Welch *t*-test. **(B, E)** ROC plot for PVT1 and RP11-54H7.4 lncRNA for the generalized linear model classifier. **(C, F)** Survival plot for high vs low methylation group for PVT1 and RP11-54H7.4 lncRNA with a *p*-value for Kaplan–Meier plot (log-rank test) and Cox proportional hazards model.

study showed that GATA6-AS overexpression correlated with poor prognosis of PDAC patients. A second similar lncRNA (GATA6-AS1) also was overexpressed and correlated with poor survival of PDAC patients (HR = 0.5) (**Table S6**).

Regarding protein-coding genes (**Table 4**), our study found 17 differentially expressed genes but five of them were identified at a stringent *P*-value of ≤ 0.01 that also correlated with PDAC patient survival. Expression of ASPM, Nek2, B3GNT3, DMBT1, and DEPDC1 is associated with better survival of PDAC patients in this study. ASPM (abnormal spindle-like microcephaly associated) is an oncogene that promotes tumor aggression in PDAC, and overexpression is associated with poor prognosis (Wang et al., 2013). We also observed that the ASPM overexpressing patient group showed low survival. NIMA-related kinase 2 (Nek2) is a serine/threonine kinase that plays a critical role in mitosis. Nek2 was reported as a prognostic biomarker for lung cancer (Shi et al., 2017), and knockdown of Nek2 gene with siRNA in xenograft mice decreased tumor size and increased survival for liver metastasized pancreatic cancer (Kokuryo et al., 2016). This gene was also reported as a prognostic biomarker for PDAC, as patients with high Nek2 expression showed shorter survival (Ning et al., 2014). In the current study, we observed a similar trend, our logistic regression model analysis also suggests that Nek2 expression may be a distinctive trait in PDAC vs. normal samples (AUC = 0.95). Our finding further reconfirms that Nek2 is a potential prognostic biomarker of PDAC.

We observed that overexpression of B3GNT3 (beta-1,3-N-acetylglucosaminyltransferase-3) is associated with shorter survival in PDAC (**Figure 8**). High AUC for the logistic regression model (AUC = 0.93) and low *P*-value with the high hazard ratio in Cox regression analysis suggests that this can be a potential prognostic biomarker for PDAC. Previous reports also confirmed that B3GNT3 overexpression was associated with shorter survival of patients in the cervical (Zhang et al., 2015) and non-small lung cell (Gao et al., 2018) cancers. Similarly, overexpression of the DEP domain containing 1 (DEPDC1) is associated with shorter overall survival of PDAC patients. Overexpression of DEPD1B is already reported in several types of human cancers (Su et al., 2014; Huang et al., 2017), we also observed overexpression in PDAC. High classification AUC (0.95) and Cox regression HR (1.9) suggest that it's a good candidate for prognostic biomarker in PDAC (**Table 4**). These findings suggest that our proposed methodology is working well for detecting known biomarkers, so it can as well detect novel prognostic biomarkers.

On the other hand, overexpression of DMBT1 and Bcl2-modifying factor (Bmf) is shown to improve survival in our study. DMBT1 (deleted in malignant brain tumors 1) expression cohorts have better survival (HR = 0.6) and high logistic regression classification AUC (0.95) suggests its role as a potential biomarker (**Figure 8**). DMBT1 is a tumor suppressor and involved in immune defense and epithelial differentiation in cancer (Mollenhauer et al., 2000). Expression of DMBT1 goes down in breast cancer (Braidotti et al., 2004; Blackburn et al., 2007), we observed a similar trend in our analysis. Pro-apoptotic protein Bmf which regulate the death of CD8 T cells (Hubner et al., 2010), is a probable prognostic biomarker for PDAC (HR = 0.62), samples with high expression Bmf have a good prognosis.

Mucins are high molecular weight glycoproteins with oligosaccharides attached to serine or threonine residues of the mucin core protein backbone that play important roles as diagnostic and prognostic markers for carcinogenesis and tumor invasion (Hollingsworth and Swanson, 2004). We separately analyzed the promoter DNA methylation and mucin gene expression in pancreatic ductal cancer. We observed significant upregulation of MUC2, MUC5B, and MUC13 in PDAC. MUC5B and MUC13 overexpressed in pancreatic ductal cancer (Kaur et al., 2013), the MUC5B expression is highly sensitive to change in promoter methylation (Yamada et al., 2011). We observed the hypomethylation of MUC5B promoter CpG cg20911165 and cg03609102 which is negatively correlated with the gene expression (**Figure 4**). We also observed overexpression of MUC2 gene, in general, its expression goes down in PDAC but some report also suggests overexpression of MUC2 (Niv, 2017). Survival analysis of PDAC data reveals that patients which have higher expression of MUC21 have low survival rate (Cox-*P*-value = 0.04, HR = 1.6).

Pathways analysis didn't observe any significantly enriched pathways for the differentially expressed genes in pathway enrichment analysis, as number of genes is not enough for analysis. But, pathway analysis of loci with dm-CpGs suggested that MAPK signaling, Rap1 signaling, cAMP signaling, cancer signaling, and mucin type O-glycan biosynthesis pathways were enriched. We conjecture that the nicotine and morphine addiction pathway showed up in our analysis because these PDAC patients are current or past smokers (**Table 3**). Many other cancer-related genes showed up differentially expressed in PDAC, including MUC2, MUC5B, MUC13, ALDH3A1, CDCA7, and CCL2. Several histone core proteins were overexpressed in PDAC. Our current study also indicated that HIST1H2BC, HIST1H2BJ, and HIST1H3H were associated with poor survival of PDAC patients (**Table 4**).

CONCLUSIONS

To our knowledge, this study represents the first TCGA-based PDAC methylome data analysis. The DNA methylome of pancreatic ductal cancer showed significant changes from normal samples. Most of hypermethylation taking place within the promoter regions and methylation in the promoter region have a strong association with corresponding gene expression. A 10 Mb region of chromosome 7 has the highest hypermethylation density, and this region harbors a number of HOX cluster genes. MUC family genes and histone core proteins are overexpressed, expression of MUC21 and several histone core HIST1H2AC, HIST1H2BC, and HIST3H2A are also associated with patients' survival. Role of hsa-mir-196b and Nek2 in PDAC patients' survival is further reconfirmed. Our analysis reveals that protein-coding genes, ARTNTL2, CELF2, EFR3B, B3GNT3, and long non-coding genes, CASC11, GATA6-AS are potential prognostic biomarkers of PDAC. Promoter methylation of ZNF154 and ZNF382, which were previously reported as early stage urine/

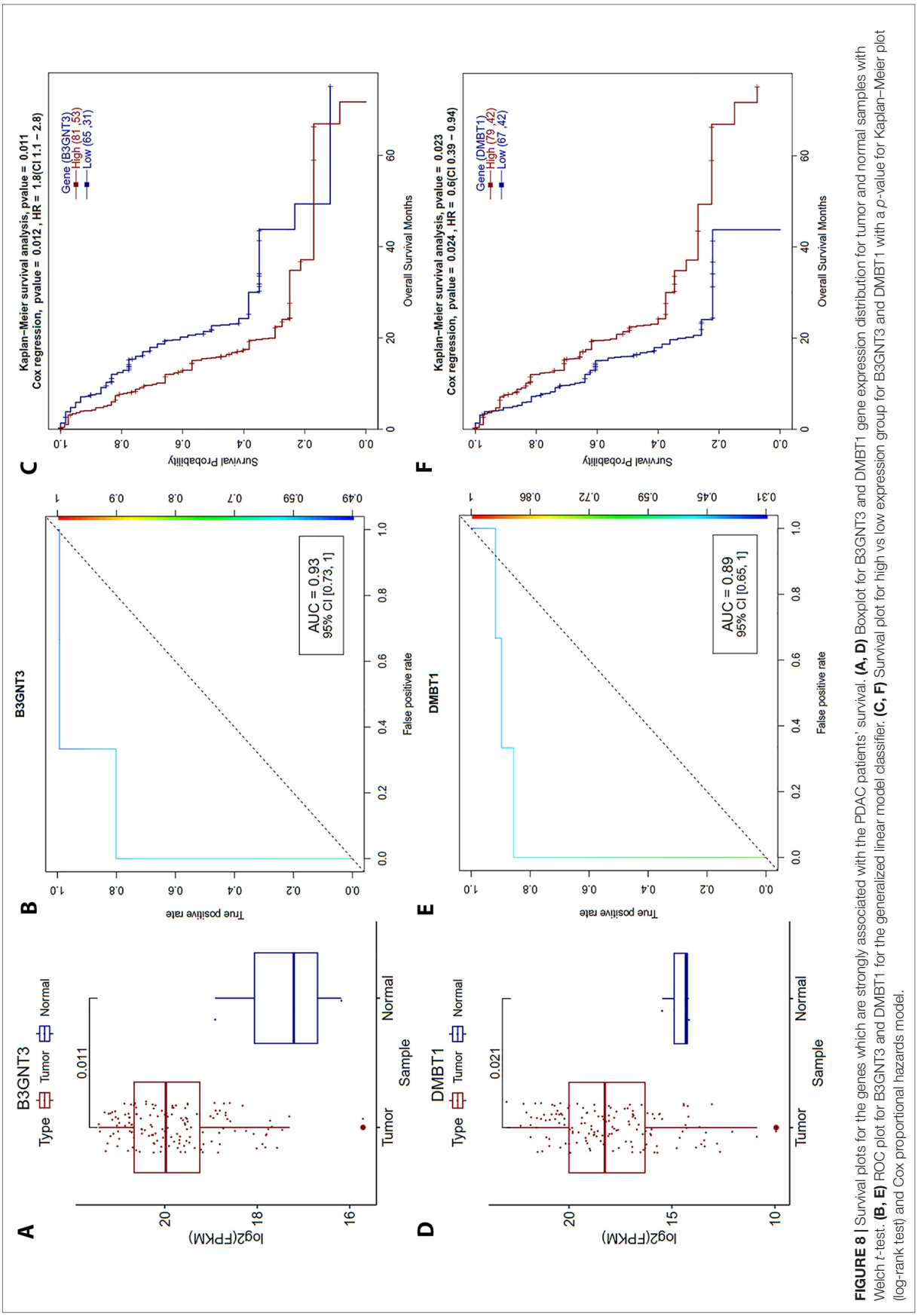

FIGURE 8 | Survival plots for the genes which are strongly associated with the PDAC patients' survival. **(A, D)** Boxplot for B3GNT3 and DMBT1 gene expression distribution for tumor and normal samples with Welch *t*-test. **(B, E)** ROC plot for B3GNT3 and DMBT1 for the generalized linear model classifier. **(C, F)** Survival plot for high vs low expression group for B3GNT3 and DMBT1 with a *p*-value for Kaplan–Meier plot (log-rank test) and Cox proportional hazards model.

blood-based biomarkers have the potential to be prognostic biomarkers for PDAC.

DATA ANALYSIS

All analyses were performed using the R version 3.5.1 (R Development Core Team 2015). We performed differential methylation/expression and survival analysis by using R/Bioconductor tools. List of tools used for this analysis are available in **Table S1**.

AUTHOR CONTRIBUTIONS

NM, SS, and CG are responsible for the study design. NM and SS performed the statistical analysis and generated figures. NM and SS drafted the manuscript and CG edited and improved the manuscript and approved it.

FUNDING

The authors thank the Bioinformatics and Systems Biology Core, which receive partial support from National Institutes of Health grants [P20GM103427, P30CA036727].

SUPPLEMENTARY MATERIALS

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00624/full#supplementary-material

REFERENCES

Adamska, A., Domenichini, A., and Falasca, M. (2017). Pancreatic ductal adenocarcinoma: current and evolving therapies. *Int. J. Mol. Sci.* 18 (7), 1338. doi: 10.3390/ijms18071338

Aine, M., Sjodahl, G., Eriksson, P., Veerla, S., Lindgren, D., Ringner, M., et al. (2015). Integrative epigenomic analysis of differential DNA methylation in urothelial carcinoma. *Genome Med.* 7 (1), 23. doi: 10.1186/s13073-015-0144-4

Akbani, R., Xhang, N., and Broom, B. M. (2010). TCGA batch effect. [http://bioinformatics.mdanderson.org/tcgambatch/].

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bert, S. A., Robinson, M. D., Strbenac, D., Statham, A. L., Song, J. Z., Hulf, T., et al. (2013). Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* 23 (1), 9–22. doi: 10.1016/j.ccr.2012.11.006

Blackburn, A. C., Hill, L. Z., Roberts, A. L., Wang, J., Aud, D., Jung, J., et al. (2007). Genetic mapping in mice identifies DMBT1 as a candidate modifier of mammary tumors and breast cancer risk. *Am. J. Pathol.* 170 (6), 2030–2041. doi: 10.2353/ajpath.2007.060512

Bojjireddy, N., Guzman-Hernandez, M. L., Reinhard, N. R., Jovic, M., and Balla, T. (2015). EFR3s are palmitoylated plasma membrane proteins that control responsiveness to G-protein-coupled receptors. *J. Cell Sci.* 128 (1), 118–128. doi: 10.1242/jcs.157495

Brady, J. J., Chuang, C. H., Greenside, P. G., Rogers, Z. N., Murray, C. W., Caswell, D. R., et al. (2016). An arntl2-driven secretome enables lung adenocarcinoma metastatic self-sufficiency. *Cancer Cell* 29 (5), 697–710. doi: 10.1016/j.ccell.2016.03.003

Braidotti, P., Nuciforo, P. G., Mollenhauer, J., Poustka, A., Pellegrini, C., Moro, A., et al. (2004). DMBT1 expression is down-regulated in breast cancer. *BMC Cancer* 4, 46. doi: 10.1186/1471-2407-4-46

Braig, S., Mueller, D. W., Rothhammer, T., and Bosserhoff, A. K. (2010). MicroRNA miR-196a is a central regulator of HOX-B7 and BMP4 expression in malignant melanoma. *Cell Mol. Life. Sci.* 67 (20), 3535–3548. doi: 10.1007/s00018-010-0394-7

Cheng, Z., Liu, F., Zhang, H., Li, X., Li, Y., Li, J., et al. (2017). miR-135a inhibits tumor metastasis and angiogenesis by targeting FAK pathway. *Oncotarget* 8 (19), 31153–31168. doi: 10.18632/oncotarget.16098

Chiaravalli, M., Reni, M., and O'Reilly, E. M. (2017). Pancreatic ductal adenocarcinoma: state-of-the-art 2017 and new therapeutic strategies. *Cancer Treat. Rev.* 60, 32–43. doi: 10.1016/j.ctrv.2017.08.007

Chong, G. O., Jeon, H. S., Han, H. S., Son, J. W., Lee, Y. H., Hong, D. G., et al. (2017). Overexpression of microRNA-196b accelerates invasiveness of cancer cells in recurrent epithelial ovarian cancer through regulation of homeobox A9. *Cancer Genomics Proteomics* 14 (2), 137–141. doi: 10.21873/cgp.20026

Chu, A., Robertson, G., Brooks, D., Mungall, A. J., Birol, I., Coope, R., et al. (2016). Large-scale profiling of microRNAs for the cancer genome atlas. *Nucleic Acids. Res.* 44 (1), e3. doi: 10.1093/nar/gkv808

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids. Res.* 44 (8), e71. doi: 10.1093/nar/gkv1507

Fantini, S., Salsi, V., and Zappavigna, V. (2018). HOX cluster-embedded microRNAs and cancer. *Biochim. Biophys. Acta. Rev. Cancer* 1869 (2), 230–247. doi: 10.1016/j.bbcan.2018.03.002

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6 (269), pl1. doi: 10.1126/scisignal.2004088

Gao, L., Zhang, H., Zhang, B., Zhu, J., Chen, C., and Liu, W. (2018). B3GNT3 overexpression is associated with unfavourable survival in non-small cell lung cancer. *J. Clin. Pathol.* 71 (7), 642–647. doi: 10.1136/jclinpath-2017-204860

Ge, J., Chen, Z., Li, R., Lu, T., and Xiao, G. (2014). Upregulation of microRNA-196a and microRNA-196b cooperatively correlate with aggressive progression and unfavorable prognosis in patients with colorectal cancer. *Cancer Cell Int.* 14 (1), 128. doi: 10.1186/s12935-014-0128-2

Gu, Z., Eils, R., and Schlesner, M. (2016). gtrellis: an R/Bioconductor package for making genome-level Trellis graphics. *BMC Bioinformatics* 17, 169. doi: 10.1186/s12859-016-1051-4

Gutschner, T., Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., et al. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* 73 (3), 1180–1189. doi: 10.1158/0008-5472.CAN-12-2850

Hollingsworth, M. A., and Swanson, B. J. (2004). Mucins in cancer: protection and control of the cell surface. *Nat. Rev. Cancer* 4 (1), 45–60. doi: 10.1038/nrc1251

Huang, L., Chen, K., Cai, Z. P., Chen, F. C., Shen, H. Y., Zhao, W. H., et al. (2017). DEPDC1 promotes cell proliferation and tumor growth *via* activation of E2F signaling in prostate cancer. *Biochem. Biophys. Res. Commun.* 490 (3), 707–712. doi: 10.1016/j.bbrc.2017.06.105

Hubner, A., Cavanagh-Kyros, J., Rincon, M., Flavell, R. A., and Davis, R. J. (2010). Functional cooperation of the proapoptotic Bcl2 family proteins Bmf and Bim *in vivo*. *Mol. Cell Biol.* 30 (1), 98–105. doi: 10.1128/MCB.01155-09

Jones, P. A., and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* 3 (6), 415–428. doi: 10.1038/nrg816

Kanno, S., Nosho, K., Ishigami, K., Yamamoto, I., Koide, H., Kurihara, H., et al. (2017). MicroRNA-196b is an independent prognostic biomarker in patients with pancreatic cancer. *Carcinogenesis* 38 (4), 425–431. doi: 10.1093/carcin/bgx013

Kaur, S., Kumar, S., Momi, N., Sasson, A. R., and Batra, S. K. (2013). Mucins in pancreatic cancer and its microenvironment. *Nat. Rev. Gastroenterol. Hepatol.* 10 (10), 607–620. doi: 10.1038/nrgastro.2013.120

Kiss, I., Mlcochova, J., Souckova, K., Fabian, P., Poprach, A., Halamkova, J., et al. (2017). MicroRNAs as outcome predictors in patients with metastatic colorectal cancer treated with bevacizumab in combination with FOLFOX. *Oncol. Lett.* 14 (1), 743–750. doi: 10.3892/ol.2017.6255

Kokuryo, T., Hibino, S., Suzuki, K., Watanabe, K., Yokoyama, Y., Nagino, M., et al. (2016). Nek2 siRNA therapy using a portal venous port-catheter system for liver metastasis in pancreatic cancer. *Cancer Sci.* 107 (9), 1315–1320. doi: 10.1111/cas.12993

Li, H., Chen, S., Liu, J., Guo, X., Xiang, X., Dong, T., et al. (2018). Long non-coding RNA PVT1-5 promotes cell proliferation by regulating miR-126/SLC7A5 axis in lung cancer. *Biochem. Biophys. Res. Commun.* 495 (3), 2350–2355. doi: 10.1016/j.bbrc.2017.12.114

Li, Z., Huang, H., Chen, P., He, M., Li, Y., Arnovitz, S., et al. (2012). miR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia. *Nat. Commun.* 3, 688. doi: 10.1038/ncomms1681

Lim, J. Y., Yoon, S. O., Seol, S. Y., Hong, S. W., Kim, J. W., Choi, S. H., et al. (2013). Overexpression of miR-196b and HOXA10 characterize a poor-prognosis gastric cancer subtype. *World J. Gastroenterol.* 19 (41), 7078–88. doi: 10.3748/wjg.v19.i41.7078

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi: 10.1186/s13059-014-0550-8

Lu, Y. C., Chang, J. T., Chan, E. C., Chao, Y. K., Yeh, T. S., Chen, J. S., et al. (2016). mir-196, an emerging cancer biomarker for digestive tract cancers. *J. Cancer* 7 (6), 650–655. doi: 10.7150/jca.13460

Mansfield, J. H., and McGlinn, E. (2012). Evolution, expression, and developmental function of Hox-embedded miRNAs. *Curr. Top. Dev. Biol.* 99, 31–57. doi: 10.1016/B978-0-12-387038-4.00002-1

Margolin, G., Petrykowska, H. M., Jameel, N., Bell, D. W., Young, A. C., and Elnitski, L. (2016). Robust detection of DNA hypermethylation of ZNF154 as a pan-cancer locus with in silico modeling for blood-based diagnostic development. *J. Mol. Diagn.* 18 (2), 283–298. doi: 10.1016/j.jmoldx.2015.11.004

Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., et al. (1995). 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.* 1 (7), 686–692. doi: 10.1038/nm0795-686

Mishra, N. K., and Guda, C. (2017). Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* 8 (17), 28990–29012. doi: 10.18632/oncotarget.15993

Mollenhauer, J., Herbertz, S., Holmskov, U., Tolnay, M., Krebs, I., Merlo, A., et al. (2000). DMBT1 encodes a protein involved in the immune defense and in epithelial differentiation and is highly unstable in cancer. *Cancer Res.* 60 (6), 1704–10.

Neumann, P., Jae, N., Knau, A., Glaser, S. F., Fouani, Y., Rossbach, O., et al. (2018). The lncRNA GATA6-AS epigenetically regulates endothelial gene expression via interaction with LOXL2. *Nat. Commun.* 9 (1), 237. doi: 10.1038/s41467-017-02431-1

Neureiter, D., Jager, T., Ocker, M. and Kiesslich, T. (2014). Epigenetics and pancreatic cancer: pathophysiology and novel treatment aspects. *World J. Gastroenterol.* 20 (24), 7830–7848. doi: 10.3748/wjg.v20.i24.7830

Ning, Z., Wang, A., Liang, J., Liu, J., Zhou, T., Yan, Q., et al. (2014). Abnormal expression of Nek2 in pancreatic ductal adenocarcinoma: a novel marker for prognosis. *Int. J. Clin. Exp. Pathol.* 7 (5), 2462–9.

Niv, Y. (2017). Mucin expression and the pancreas: a systematic review and meta-analysis. *World J. Meta-Analysis* 5 (2), 5. doi: 10.13105/wjma.v5.i2.63

Nones, K., Waddell, N., Song, S., Patch, A. M., Miller, D., Johns, A., et al. (2014). Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling. *Int. J. Cancer* 135 (5), 1110–1118. doi: 10.1002/ijc.28765

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17 (5), 510–522. doi: 10.1016/j.ccr.2010.03.017

Oh, T., Kim, N., Moon, Y., Kim, M. S., Hoehn, B. D., Park, C. H., et al. (2013). Genome-wide identification and validation of a novel methylation biomarker, SDC2, for blood-based detection of colorectal cancer. *J. Mol. Diagn.* 15 (4), 498–507. doi: 10.1016/j.jmoldx.2013.03.004

Pan, L., Zhou, L., Yin, W., Bai, J., and Liu, R. (2018). miR-125a induces apoptosis, metabolism disorder and migrationimpairment in pancreatic cancer cells by targeting Mfn2-related mitochondrial fission. *Int. J. Oncol.* 53 (1), 124–136. doi: 10.3892/ijo.2018.4380

Paradise, B. D., Barham, W., and Fernandez-Zapico, M. E. (2018). Targeting epigenetic aberrations in pancreatic cancer, a new path to improve patient Outcomes? *Cancers (Basel)* 10 (5), 128. doi: 10.3390/cancers10050128

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., Lord R, V., et al. (2015). *De novo* identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 8, 6. doi: 10.1186/1756-8935-8-6

Petrovics, G., Zhang, W., Makarem, M., Street, J. P., Connelly, R., Sun, L., et al. (2004). Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene* 23 (2), 605–611. doi: 10.1038/sj.onc.1207069

Phipson, B., Maksimovic, J., and Oshlack, A. (2016). missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* 32 (2), 286–288. doi: 10.1093/bioinformatics/btv560

R Core Team. (2019). *A language and environment for statistical computing*. Viena, Austria: R Foundation for Statistical Computing. [https://www.r-project.org/].

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74 (11), 2913–2921. doi: 10.1158/0008-5472.CAN-14-0155

Ramachandran, K., Speer, C., Nathanson, L., Claros, M., and Singal, R. (2016). Role of DNA methylation in cabazitaxel resistance in prostate cancer. *Anticancer Res.* 36 (1), 161–168.

Ramalingam, S., Ramamoorthy, P., Subramaniam, D., and Anant, S. (2012). Reduced expression of RNA binding protein CELF2, a putative tumor suppressor gene in colon cancer. *Immunogastroenterology* 1 (1), 27–33. doi: 10.7178/ig.1.1.7

Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 32 (2), 185–203 e13. doi: 10.1016/j.ccell.2017.07.007

Reinert, T., Borre, M., Christiansen, A., Hermann, G. G., Orntoft, T. F., and Dyrskjot, L. (2012). Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. *PLoS One* 7 (10), e46297. doi: 10.1371/journal.pone.0046297

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids. Res.* 43 (7), e47. doi: 10.1093/nar/gkv007

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11 (3), R25. doi: 10.1186/gb-2010-11-3-r25

Sanchez-Vega, F., Gotea, V., Petrykowska, H. M., Margolin, G., Krivak, T. C., DeLoia, J. A., et al. (2013). Recurrent patterns of DNA methylation in the ZNF154, CASP8, and VHL promoters across a wide spectrum of human solid epithelial tumors and cancer cell lines. *Epigenetics* 8 (12), 1355–1372. doi: 10.4161/epi.26701

Sato, N., Fukushima, N., Hruban, R. H., and Goggins, M. (2008). CpG island methylation profile of pancreatic intraepithelial neoplasia. *Mod. Pathol.* 21 (3), 238–244. doi: 10.1038/modpathol.3800991

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* 103 (5), 1412–1417. doi: 10.1073/pnas.0510310103

Shi, Y. X., Yin, J. Y., Shen, Y., Zhang, W., Zhou, H. H., and Liu, Z. Q. (2017). Genome-scale analysis identifies NEK2, DLGAP5 and ECT2 as promising diagnostic and prognostic biomarkers in human lung cancer. *Sci. Rep.* 7 (1), 8072. doi: 10.1038/s41598-017-08615-5

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21 (20), 3940–3941. doi: 10.1093/bioinformatics/bti623

Slater, E. P., Strauch, K., Rospleszcz, S., Ramaswamy, A., Esposito, I., Kloppel, G., et al. (2014). MicroRNA-196a and -196b as potential biomarkers for the early detection of familial pancreatic cancer. *Transl. Oncol.* 7 (4), 464–471. doi: 10.1016/j.tranon.2014.05.007

Su, Y. F., Liang, C. Y., Huang, C. Y., Peng, C. Y., Chen, C. C., Lin, M. C., et al. (2014). A putative novel protein, DEPDC1B, is overexpressed in oral cancer patients, and enhanced anchorage-independent growth in oral cancer cells that is mediated by Rac1 and ERK. *J. Biomed. Sci.* 21, 67. doi: 10.1186/s12929-014-0067-1

Subramaniam, D., Ramalingam, S., Linehan, D. C., Dieckgraefe, B. K., Postier, R. G., Houchen, C. W., et al. (2011). RNA binding protein CUGBP2/CELF2 mediates curcumin-induced mitotic catastrophe of pancreatic cancer cells. *PLoS One* 6 (2), e16958. doi: 10.1371/journal.pone.0016958

Tan, A. C., Jimeno, A., Lin, S. H., Wheelhouse, J., Chan, F., Solomon, A., et al. (2009). Characterizing DNA methylation patterns in pancreatic cancer genome. *Mol. Oncol.* 3 (5-6), 425–438. doi: 10.1016/j.molonc.2009.03.004

Tang, R. X., Chen, W. J., He, R. Q., Zeng, J. H., Liang, L., Li, S. K., et al. (2017). Identification of a RNA-Seq based prognostic signature with five lncRNAs for lung squamous cell carcinoma. *Oncotarget* 8 (31), 50761–50773. doi: 10.18632/oncotarget.17098

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., et al. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29 (2), 189–196. doi: 10.1093/bioinformatics/bts680

Thompson, M. J., Rubbi, L., Dawson, D. W., Donahue, T. R., and Pellegrini, M. (2015). Pancreatic cancer patient survival correlates with DNA methylation of pancreas development genes. *PLoS One* 10 (6), e0128814. doi: 10.1371/journal.pone.0128814

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* 19 (1A), A68–77. doi: 10.5114/wo.2014.47136

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* 17 (6), 520–5. doi: 10.1093/bioinformatics/17.6.520

Vincent, A., Omura, N., Hong, S. M., Jaffe, A., Eshleman, J., and Goggins, M. (2011). Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma. *Clin. Cancer Res.* 17 (13), 4341–4354. doi: 10.1158/1078-0432.CCR-10-3431

Wang, W. Y., Hsu, C. C., Wang, T. Y., Li, C. R., Hou, Y. C., Chu, J. M., et al. (2013). A gene expression signature of epithelial tubulogenesis and a role for ASPM in pancreatic tumor progression. *Gastroenterology* 145 (5), 1110–1120. doi: 10.1053/j.gastro.2013.07.040

Wang, Z., Yang, B., Zhang, M., Guo, W., Wu, Z., Wang, Y., et al. (2018). lncRNA epigenetic landscape analysis identifies EPIC1 as an oncogenic lncRNA that interacts with MYC and promotes cell-cycle progression in cancer. *Cancer Cell* 33 (4), 706–720 e9. doi: 10.1016/j.ccell.2018.03.006

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., et al. (2004). Epigenetic programming by maternal behavior. *Nat. Neurosci.* 7 (8), 847–854. doi: 10.1038/nn1276

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi: 10.1038/ng.2764

Yamada, N., Kitamoto, S., Yokoyama, S., Hamada, T., Goto, M., Tsutsumida, H., et al. (2011). Epigenetic regulation of mucin genes in human cancers. *Clin. Epigenetics* 2 (2), 85–96. doi: 10.1007/s13148-011-0037-3

Yang, Z., Jones, A., Widschwendter, M., and Teschendorff, A. E. (2015). An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol.* 16, 140. doi: 10.1186/s13059-015-0699-9

Yekta, S., Shih, I. H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304 (5670), 594–596. doi: 10.1126/science.1097434

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 (5), 284–7. doi: 10.1089/omi.2011.0118

Zhang, C., Xiang, T., Li, S., Ye, L., Feng, Y., Pei, L., et al. (2018). The novel 19q13 KRAB zinc-finger tumour suppressor ZNF382 is frequently methylated in oesophageal squamous cell carcinoma and antagonises Wnt/beta-catenin signalling. *Cell Death Dis.* 9 (5), 573. doi: 10.1038/s41419-018-0604-z

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., et al. (2011). International cancer genome consortium data portal–a one-stop shop for cancer genomics data. *Database (Oxford)* 2011, bar026. doi: 10.1093/database/bar026

Zhang, W., Hou, T., Niu, C., Song, L., and Zhang, Y. (2015). B3GNT3 Expression is a novel marker correlated with pelvic lymph node metastasis and poor clinical outcome in early-stage cervical cancer. *PLoS One* 10 (12), e0144360. doi: 10.1371/journal.pone.0144360

Zhang, W., Shu, P., Wang, S., Song, J., Liu, K., Wang, C., et al. (2018). ZNF154 is a promising diagnosis biomarker and predicts biochemical recurrence in prostate cancer. *Gene* 675, 136–143. doi: 10.1016/j.gene.2018.06.104

Zhang, Z., Zhou, C., Chang, Y., Zhang, Z., Hu, Y., Zhang, F., et al. (2016). Long non-coding RNA CASC11 interacts with hnRNP-K and activates the WNT/beta-catenin pathway to promote growth and metastasis in colorectal cancer. *Cancer Lett.* 376 (1), 62–73. doi: 10.1016/j.canlet.2016.03.022

Zhou, W., Laird, P. W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids. Res.* 45 (4), e22. doi: 10.1093/nar/gkw967

Zhu, Y., Zhang, J. J., Zhu, R., Zhu, Y., Liang, W. B., Gao, W. T., et al. (2011). The increase in the expression and hypomethylation of MUC4 gene with the progression of pancreatic ductal adenocarcinoma. *Med. Oncol.* 28 Suppl 1, S175–184. doi: 10.1007/s12032-010-9683-0

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

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