



# Detecting Prognosis Risk Biomarkers for Colon Cancer Through Multi-Omics-Based Prognostic Analysis and Target Regulation Simulation Modeling

## OPEN ACCESS

### Edited by:

Claudio Sette,

Catholic University of the Sacred Heart Rome, Italy

### Reviewed by:

Arsheed A. Ganaie,

University of Minnesota Twin Cities, United States

Parvin Mehdipour,

Tehran University of Medical Sciences, Iran

Enrico Capobianco,

University of Miami, United States

Javad Zahiri,

Tarbiat Modares University, Iran

### \*Correspondence:

Zhiwei Cao

zwcao@tongji.edu.cn

Tianyi Qiu

ty\_qiu@126.com

### Specialty section:

This article was submitted to

Cancer Genetics,

a section of the journal

Frontiers in Genetics

Received: 18 October 2019

Accepted: 29 April 2020

Published: 26 May 2020

### Citation:

Yin Z, Yan X, Wang Q, Deng Z,

Tang K, Cao Z and Qiu T (2020)

Detecting Prognosis Risk Biomarkers

for Colon Cancer Through

Multi-Omics-Based Prognostic

Analysis and Target Regulation

Simulation Modeling.

Front. Genet. 11:524.

doi: 10.3389/fgene.2020.00524

Zuojing Yin<sup>1</sup>, Xinmiao Yan<sup>1</sup>, Qiming Wang<sup>1</sup>, Zeliang Deng<sup>1</sup>, Kailin Tang<sup>1</sup>, Zhiwei Cao<sup>1\*</sup> and Tianyi Qiu<sup>2\*</sup>

<sup>1</sup> Department of Gastroenterology, Shanghai Tenth People's Hospital, College of Life Science and Technology, Tongji University, Shanghai, China, <sup>2</sup> Shanghai Public Health Clinical Center, Fudan University, Shanghai, China

**Background:** Colon cancer is one of the most common health threats for humans since its high morbidity and mortality. Detecting potential prognosis risk biomarkers (PRBs) is essential for the improvement of therapeutic strategies and drug development. Currently, although an integrated prognostic analysis of multi-omics for colon cancer is insufficient, it has been reported to be valuable for improving PRBs' detection in other cancer types.

**Aim:** This study aims to detect potential PRBs for colon adenocarcinoma (COAD) samples through the cancer genome atlas (TCGA) by integrating multi-omics.

**Materials and Methods:** The multi-omics-based prognostic analysis (MPA) model was first constructed to systemically analyze the prognosis of colon cancer based on four-omics data of gene expression, exon expression, DNA methylation and somatic mutations on COAD samples. Then, the essential features related to prognosis were functionally annotated through protein-protein interaction (PPI) network and cancer-related pathways. Moreover, the significance of those essential prognostic features were further confirmed by the target regulation simulation (TRS) model. Finally, an independent testing dataset, as well as the single cell-based expression dataset were utilized to validate the generality and repeatability of PRBs detected in this study.

**Results:** By integrating the result of MPA modeling, as well the PPI network, integrated pathway and TRS modeling, essential features with gene symbols such as EPB41, PSMA1, FGFR3, MRAS, LEP, C7orf46, LOC285000, LBP, ZNF35, SLC30A3, LECT2, RNF7, and DYNC111 were identified as PRBs which provide high potential as drug targets for COAD treatment. Validation on the independent testing dataset demonstrated that these PRBs could be applied to distinguish the prognosis of COAD patients. Moreover, the prognosis of patients with different clinical conditions could also be distinguished by the above PRBs.

**Conclusions:** The MPA and TRS models constructed in this paper, as well as the PPI network and integrated pathway analysis, could not only help detect PRBs as potential therapeutic targets for COAD patients but also make it a paradigm for the prognostic analysis of other cancers.

**Keywords:** colon cancer, prognostic analysis, multi-omics analysis, *in silico* simulation, pathway integration

## INTRODUCTION

As one of the most common cancer types and the second leading cause of cancer mortality (Hernandez et al., 2014), colorectal cancer (CRC) is highly prevalent worldwide, with more than 1.2 million new cases and over 600 thousand deaths each year (Li et al., 2015). Even though nearly 60% of CRC patients can be treated through therapeutic surgical resection and adjuvant chemotherapy, approximately 20–30% of patients will eventually suffer from disease recurrence and experience poor prognosis (O’Connell et al., 2008; Andre et al., 2009). The diagnosis and prognosis of CRC, especially its branch colon cancer (Marley and Nan, 2016), has received much attention in recent researches. Thus, approaches which could efficiently identify the PRBs for colon cancer with diagnosis, monitoring, and prognosis are highly desired to improve the cure rate and overall survival (OS) (Melichar, 2013; Zhou et al., 2018a,b).

With the development of next-generation sequencing (NGS), essential PRBs for colon cancer from sequencing data such as gene expression (Calon et al., 2015; Okugawa et al., 2017), exon expression (Kato et al., 2015), DNA methylation status (Kandimalla et al., 2017), mutational profile (Yu et al., 2015; Taieb et al., 2016) and others (Zheng et al., 2001; Ozawa et al., 2017) were determined. For example, it was reported that CDX2 could be used as PRBs for stage II and stage III colon cancer (van den Braak et al., 2018). And, mutations on BRAF (V600E) and KRAS were significantly associated with disease-free survival (DFS) and OS in CRC patients with microsatellite-stable tumors (Taieb et al., 2016). Additionally, it was reported that high expression of hsa-mir-155 and low expression of hsa-let-7a-2 were correlated with poor survival in lung cancer (Yanaiharu et al., 2006). Moreover, protein biomarkers such as CA19-9, CA 72-4 and carcinoembryonic antigen (CEA), can be used as PRBs of colorectal carcinoma (Zheng et al., 2001), and plasma vascular endothelial growth factor-A (VEGF-A) can be used as PRBs for colon cancer (Luo and Xu, 2014). Despite all the above efforts, no non-invasive, specific, sensitive, and economical methods are reported to identify the PRBs for all types of CRC patients in clinical (Das et al., 2017). Existing PRBs are only sensitive for limited patients and fail to be extended for large-scale populations (Xie et al., 2018). Considering that the omics information from different patients are not consistent, it is necessary to apply multi-omics information in large-scale populations to detect general PRBs. PRBs from multi-omics rather than single one cannot only help the diagnosis of colon cancer but also increase sensitivity to conventional therapies and improve prognosis.

By taking advantage of The Cancer Genome Atlas (TCGA) program (Tomczak et al., 2015), multi-omics molecular profiles including transcriptome, exon expression, DNA methylation,

mutations, etc. are collated along with clinical annotations for patients. In that case, it is possible to discover the PRBs with multi-omics information across large-scale populations by machine learning techniques (Cruz and Wishart, 2007; Kourou et al., 2015). In this study, prognostic analysis of COAD patients was performed by integrating multi-omics data which were closely associated with the expression or regulation of genes including gene expression, exon expression, DNA methylation and mutations derived from UCSC Xena database (Liu et al., 2017; Qu et al., 2017; Zhang X. et al., 2017), as well as clinical survival information of patients. Firstly, a MPA model was generated to identify essential features that significantly affect the prognosis of COAD patients. Then, the function of the above features was analyzed through the PPI network and pathway integration analysis. Moreover, the TRS model was provided to validate the significance of those essential features that alteration could increase the OS of COAD patients. By integrating the result of MPA modeling, as well as the PPI network, integrated pathway and TRS modeling, essential features with gene symbols of EPB41, PSMA1, FGFR3, MRAS, LEP, C7orf46, LOC285000, LBPZNF35, SLC30A3, LECT2, RNF7, and DYNC1I1 were detected as PRBs for COAD. The validation of the independent dataset showed that these detected PRBs could not only distinguish the prognosis of colon cancer patients from other data sources, but also reflect significant difference between tumor and normal cells from single-cell based expression profile. Moreover, these PRBs were also effectively distinguish the prognosis of patients with different clinical conditions. With the accumulation of multi-omics data and clinical information, it is possible for us to comprehensively investigate PRBs and perform therapeutic targets for future drug development.

## MATERIALS AND METHODS

### Data Source

The overall survival (OS) of 551 COAD samples as the survival information is derived from the TCGA module of Public Xena Hubs in the UCSC Xena database (Liu et al., 2017; Qu et al., 2017; Zhang X. et al., 2017). Besides, the clinical information of COAD samples including age, weight, person neoplasm cancer status, number of first degree relatives with cancer diagnosis, etc. were downloaded from the phenotype section of the TCGA module in the UCSC Xena database. Four omics data were downloaded followed the same process, including gene expression profiles of 329 COAD samples, exon expression profiles of 329 COAD samples, DNA methylation profiles of 337 COAD samples and somatic mutation information of 217 COAD

samples (**Supplementary Table S1**). For each patient, a tumor sample was selected as the research object by filtering out the samples from normal tissue according to the nomenclature of TCGA sample IDs.

All the profile dataset of four omics were downloaded from the UCSC Xena database. Briefly, level 3 gene expression profiles with 20,530 gene features recorded in the UCSC Xena database were experimentally generated using the Illumina HiSeq 2000 RNA sequencing platform (Prego-Faraldo et al., 2018) from the University of North Carolina TCGA genome characterization center. The stored exon expression profiles were generated using the same platform as the gene expression profiles, with 239,322 exon features. The downloaded DNA methylation profiles were obtained from the platform of Illumina Infinium HumanMethylation450 (Hong et al., 2019), which consists of 375,066 methylation features. Moreover, for somatic mutations, the recorded sequencing data were generated on the Illumina GA system containing 239,322 mutations. After obtained the original mutation profiles, the information for somatic mutations was integrated into a binary matrix, in which mutations at the corresponding position were marked as 1 or 0. Besides, the gene annotation information of exon and methylation were also derived from the TCGA module of Public Xena Hubs in the UCSC Xena database.

The background Protein-protein interaction (PPI) network used in this project contained 10,462 nodes and 55,317 interactions was constructed mainly based on three database, including HPRD version 9 (Stelzl et al., 2005), Mint version 2012 (Zanzoni et al., 2002) and IntAct version 4.2.12 (Hermjakob et al., 2004). And, biological pathways for enrichment and analysis were integrated from KEGG version 87.0 (Kanehisa et al., 2017) and GeneCards version 4.12 (Safran et al., 2010). Targets of drugs were retrieved from DrugBank version 5.0 (Wishart et al., 2018) and the TTD version 2018 (Li et al., 2018).

The independent dataset was obtained from the NCBI GEO database with the accession number of GSE17538, in which the expression profile including 54,675 probes in 177 colon cancer patients with survival information from Moffitt Cancer Center (Smith et al., 2010). The single-cell based RNA-seq dataset of colon cancer was downloaded from the NCBI GEO database with the accession number of GSE81861 (Li et al., 2017). In this dataset, single-cell sequencing data containing 11 primary colorectal tumors and matched normal mucosa (NM) cell with 57,240 genes are selected. Among them, four groups including all cell count (266 cells for NM and 375 cells for tumor), all cell FPKM (215 cells for NM and 375 cells in tumor), epithelial cell count (160 cells for NM and 272 cells in tumor), and epithelial cell FPKM (160 cells for NM and 272 cells for tumor) were collected.

## Determining Prognosis-Related Features or MPA Model Construction

For each omics dataset, the intersected tumor samples with the survival records were selected as the patient samples and further divided into high-OS group (positive) and low-OS group (negative) by setting the threshold of OS as 5 years (1,825 days) (Gustafsson et al., 2016). Further, two-tailed *T*-tests

were used to evaluate the different features between positive and negative samples. For each omics profile, the top 1,000 features with *P*-values in ascending order were first screened and further filtered with conditions of  $P < 0.01$  and fold change (FC)  $> 1.5$  or  $FC < 2/3$ .

Then, to reduce the feature dimensionality of multi-omics profiles combined by single-omics, exploratory factor analysis (EFA) (Cole et al., 2018) was performed on the profile of the above detected differential features on single omics by using the psych package of R software (Lorenzo-Seva and Van Ginkel, 2016) to obtain the weight matrix between factors and original features, as well as the scoring matrix of factors. For multi-omics, including double-omics, triple-omics, and quadruple-omics, the factor scoring matrix was obtained by combining the corresponding factor scoring matrix in single-omics. Furthermore, the scoring matrix of factors in each omics dataset was integrated with logarithmically transformed OS, and unsupervised hierarchical clustering could be performed by using the pheatmap package of R software (Xu et al., 2018) to verify the classification performance of factors.

Then, to detect the essential features that might be closely associated with the prognosis of COAD in each single-omics profile, the weight matrix of factors obtained from the above EFA process was normalized from 0 to 1 and the weights of different features could be sorted in descending orders for each factor. Essential features with the maximum weight for each factor, which prompted the performance of distinguishing prognosis, were selected as the essential features in each single-omics for subsequent prognostic modeling. For multi-omics, the profiles of essential features were produced by integrating the corresponding ones in single-omics. The boxplots of above essential features in single-omics including gene expression, exon expression, and DNA methylation were generated by using the ggpubr package of R software (Jiang et al., 2018) to illustrate the distribution of above essential features in positive and negative groups. Since the features of somatic mutation are binary, they were annotated by the online tool of cBioPortal (Ricketts and Linehan, 2015).

## MPA Modeling for Colon Cancer

The MPA modeling requires three elements: (1) profiles of prognosis-related features, (2) samples with classification indicators and (3) appropriate machine learning methods. Here, 15 MPA models were created based on different combinations of descriptors including (1) four single-omics, which including 12, 39, 22, and 32 features, respectively, (2) six combinations of double-omics data, which including 51, 34, 44, 61, 71, and 54 features, respectively, (3) four combinations of triple-omics data, which including 93, 66, 83, and 73 features, respectively, and (4) combination of all quadruple-omics data, which including all 105 features. Further, training and testing datasets for the MPA model were obtained through the Diverse Subset sampling method (Yuan et al., 2017). Typically, the first sample A was randomly selected as the seed for the training dataset. Secondly, sample B with the farthest spatial distance toward sample A (in here, represents the spatial distance between omics profiles of two samples) was selected to put into the training dataset. Thirdly, the

third sample with the farthest average distance from both samples A and B were extracted. Finally, sampling was repeated until two-thirds of the positive and negative samples were extracted as the training set, and the remaining samples were defined as the testing set. Here, the above essential features were taken as the potential prognosis-related features, of which the profile was set as the feature profiles of the model training and testing. For machine learning approaches, Support Vector Machines (SVM), Neural Network (NN), Naïve Bayes (NB), Logistic Regression (LR), Random Forest Classifier (RF), Linear Regression (LiR), Keras Deep Learning (Keras) were implemented by using the python 3.7 package of sklearn, TensorFlow and keras to generate the MPA model.

### Target Regulation Simulation (TRS) Process for Essential Features

To generate the TRS process, the profiles of total 105 essential features in quadruple-omics, as well as the corresponding OS in the overlapped TCGA samples of four omics were utilized to the MPA model training and testing. After model testing, all the TCGA samples in the testing set were pre-clustered as prognosis positive and negative ones according to the OS days of 1,825. For each true negative sample, the expression profile of 105 individual features and 5,460 two-feature combinations were retrieved for simulation. Each selected individual feature or features in combinations of the true negative samples were down-regulated to the minimum value of those in positive patients, then these samples were re-evaluated through the MPA model to obtain a new classification label. The process of TRS was illustrated in **Figure 1**.

### Survival Analysis of Samples

Survival analysis (Schlumberger et al., 2017) was performed based on the classification results of testing samples by the LR method. Then, Kaplan–Meier survival curves of different types of samples were evaluated using the survival and survminer packages of R software (Modhukur et al., 2018). Besides, the log-rank test (Katai et al., 2018) was employed to test the difference between the two compared sample groups.

### Protein–Protein Interaction (PPI) Network of Essential Biomarkers

Essential features were used to construct a PPI network. The essential features of each single-omics dataset derived above were transformed into gene symbols based on the annotation information downloaded from the UCSC Xena database. Further, gene symbols of the quadruple-omics datasets were integrated and annotated into the background PPI network using Cytoscape software version 3.4.0 (Kohl et al., 2011). Different colors were used to distinguish different omics types.

### Pathway Integration Analysis of Essential Features in the PPI Network

For essential features in the PPI network, relationship between the corresponding gene symbol and colon cancer was investigated using literature search and annotated into

biological pathways, including KEGG version 87.0 (Kanehisa et al., 2017) and GeneCards version 4.12 (Rappaport et al., 2017). Pathway integration analysis was performed by Edraw max version 8.6 (Deng et al., 2019).

### Feature Comparison From Different Aspects

To evaluate the performance of features derived from different aspects, 33 features including 13 from the PPI network, 8 from the integrated pathway and 12 from the TRS process were analyzed and compared. All samples in the dataset were grouped by the median value of these individual features, respectively, and then survival analysis was performed to compare the potential PRBs from the above three different aspects. Further, all detected 13 PRBs from these three aspects were used to generate the Linear Regression model by sklearn package of python 3.7.

### Prognostic Evaluation of Samples With Different Clinical Information

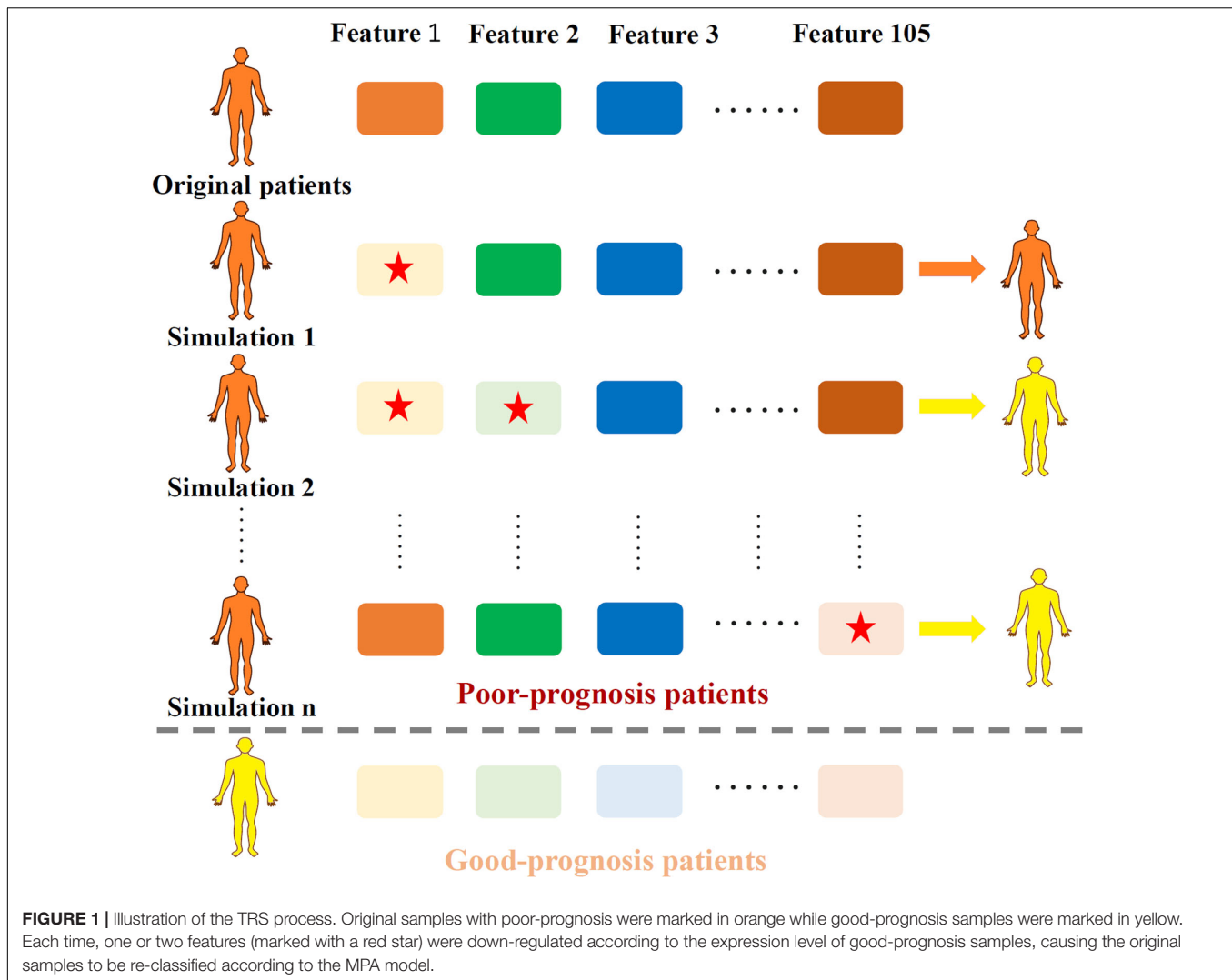
To evaluate the clinical information of patients, 11 clinical features including age, gender, weight, histological type, history of colon polyps, person neoplasm cancer status, lymphatic invasion, pathologic stage, pathologic T stage, venous invasion and number of first degree relatives with cancer diagnosis were firstly evaluated through the Cox proportional hazard (PH) model. Then, the samples were classified by the above prognostic risk features. Here, 13 PRBs were individually evaluated to estimate the survival difference by setting the median value of each PRBs as the cutoff for prognostic classifications.

## RESULTS

### Differential Expression Profiles of COAD Patient Based on Multi-Omics Analysis

To determine the essential features that closely related to the prognosis of COAD patients, differential expression features for four omics data were initially derived by setting appropriate conditions with  $P < 0.01$  and  $FC > 1.5$  or  $FC < 2/3$  (see section “Materials and Methods”). Thus, 146 features for gene expression, 1,000 features for exon expression, 362 features for DNA methylation, and 968 features for somatic mutations were selected. After factor analysis, 19 factors for gene expression, 45 factors for exon expression, 39 factors for DNA methylation and 37 factors for mutations were determined. Thus, 140 factors were used to analyze patient samples based on quadruple-omics profiles (**Figure 2**). The expression profiles of 202 overlapped patient samples included in all quadruple-omics profiles are illustrated in **Figure 2A**, in which samples with high-OS were mostly clustered into one branch (marked with blue dotted box). In that case, the overall expression profiles contained all quadruple-omics data that could significantly distinguish high-OS and low-OS patients.

Moreover, by removing redundancy, 105 unique essential features including 12 for gene expression (**Figure 2B**), 39



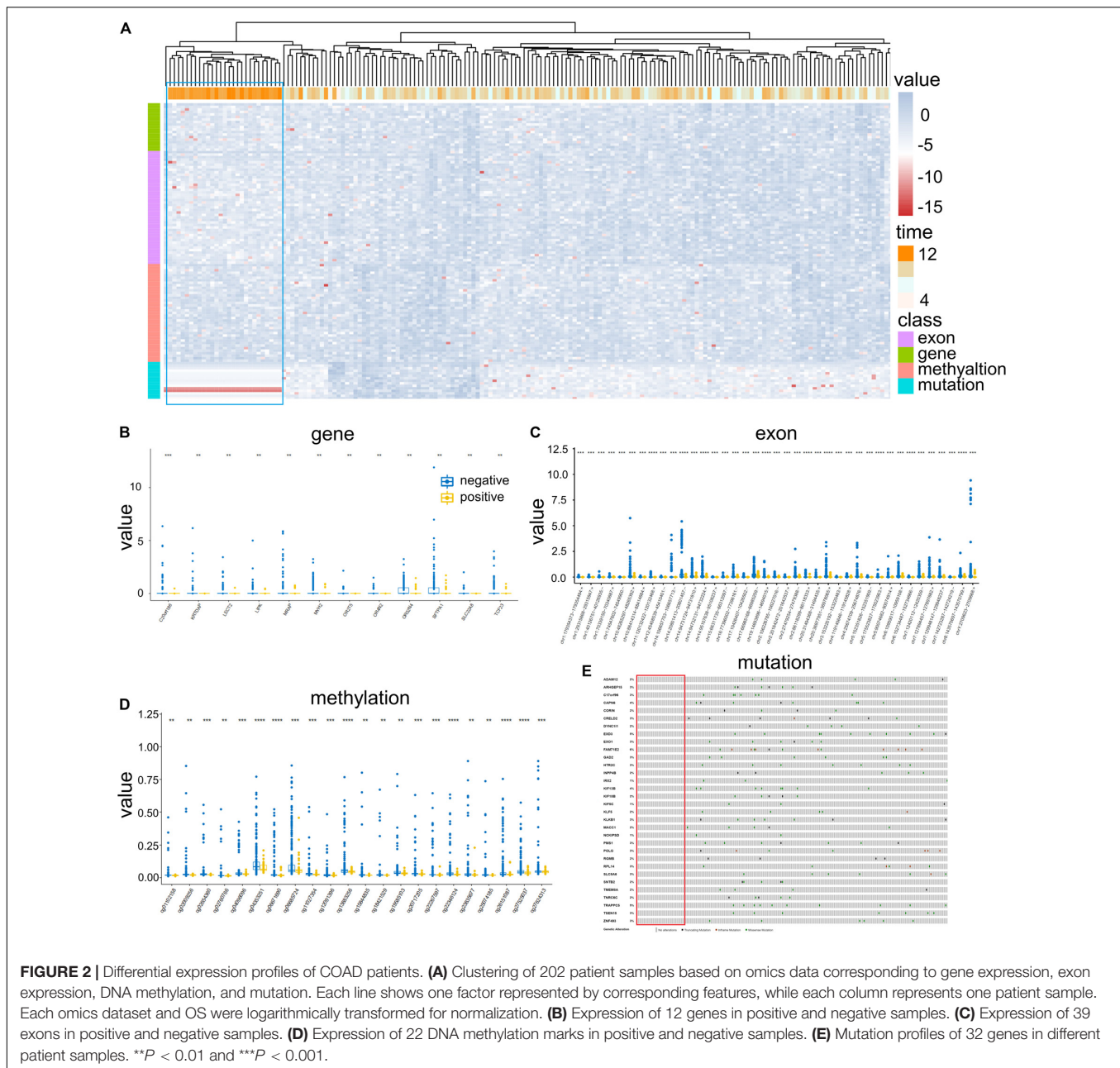
for exon expression (Figure 2C), 22 for DNA methylation (Figure 2D) and 32 for somatic mutations (Figure 2E) were retained. The further analysis illustrated that the expression levels in negative groups were generally higher than those in positive ones for both gene expression, exon expression and DNA methylation. For somatic mutation, none of the positive samples contained alterations on those 32 essential features (Figure 2E), while truncating mutations, in-frame mutations and missense mutations frequently occurred in negative groups (Supplementary Table S2). Thus, all the above essential features could be considered as prognosis-related features that were differentially expressed between the positive and negative samples from TCGA.

## Performance of MPA Modeling

The MPA modeling was established based on 15 different combinations of omics profiles, and further been evaluated through four machine learning approaches including SVM, NN, NB, and LR for comparison. The receiver operating characteristic (ROC) curves of all 15 models were illustrated

in Supplementary Figure S1. Results illustrated that the best MPA model for single-omics data could achieve the AUC value of 0.945 as the baseline, which could be further increased to 0.959 for the combination of double-omics data and 0.980 for triple-omics data. By integrating all four omics data, the classification performance could reach to the AUC of 0.998 for MPA modeling on LR (Figure 3A), followed by 0.963 for SVM, 0.936 for NN and 0.911 for NB (Figure 3B). Since the LR model revealed the best prediction performance among other approaches, it was chosen for MPA modeling and further prognosis analysis.

Further, we evaluated our MPA model through survival analysis based on the predictions of the LR model. The KM survival curves indicated that the combination of triple-omics data was still unable to ideally distinguish the high-OS and low-OS samples (Figure 3C), even though the prediction performance is distinguishable. Remarkably, the KM survival curves for positive and negative groups predicted by the MPA model, which consisted of all four omics data could be perfectly distinguishable (Figure 3D). In that case, the MPA model based on 105

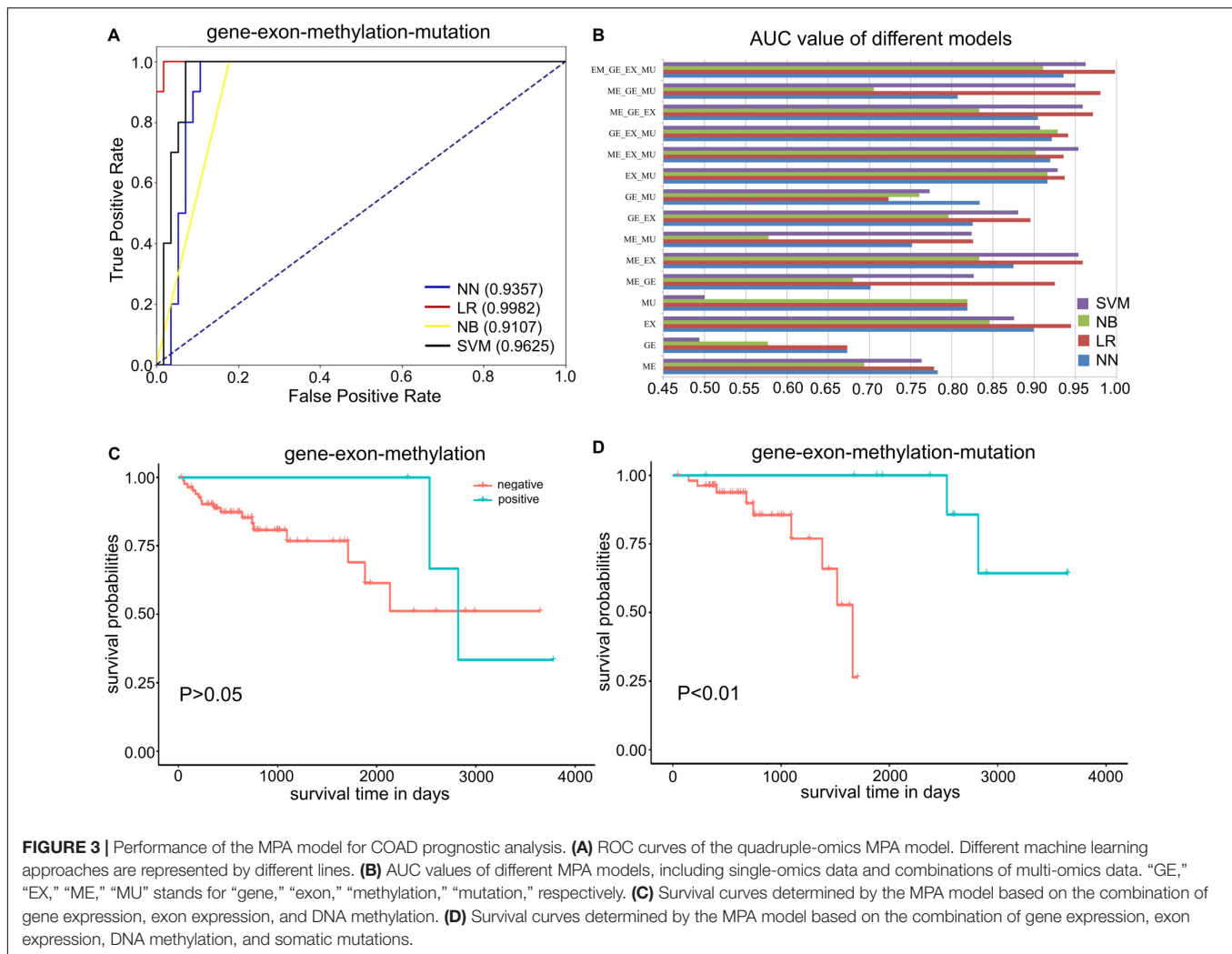


essential features derived from four omics data could be used for prognostic analysis of COAD patients.

## PPI Network and Pathway Integration Analysis of Survival-Associated Omics Features

To detect the PRBs, all prognosis-related features were transformed into non-redundant gene symbols including 12 for gene expression, 31 for exon expression, 26 for DNA methylation and 32 for somatic mutations (**Supplementary Table S3**). Since ZNF493 and MYH2 could be transformed from multi-omics data, 99 unique gene symbols were obtained

(**Supplementary Table S4**). Among them, 45 genes could be annotated into an integrated PPI network. In particular, 30 genes were defined as hub nodes with a degree over or equal to 5, which indicates those genes might participate in crucial biological functions. Further analysis showed that all 45 genes mapped in the PPI network were associated with biological processes or pathways related to the pathogenesis and development of cancer (**Table 1**). The refined PPI network which contains 30 genes with the degree over 5 were illustrated in **Figure 4A**, and canonical cancer-related pathways of those features such as cell signaling pathways, cell cycle, apoptosis, and diabetes pathways which derived from KEGG database were illustrated in **Figure 4B**.



**FIGURE 3 |** Performance of the MPA model for COAD prognostic analysis. **(A)** ROC curves of the quadruple-omics MPA model. Different machine learning approaches are represented by different lines. **(B)** AUC values of different MPA models, including single-omics data and combinations of multi-omics data. “GE,” “EX,” “ME,” “MU” stands for “gene,” “exon,” “methylation,” “mutation,” respectively. **(C)** Survival curves determined by the MPA model based on the combination of gene expression, exon expression, and DNA methylation. **(D)** Survival curves determined by the MPA model based on the combination of gene expression, exon expression, DNA methylation, and somatic mutations.

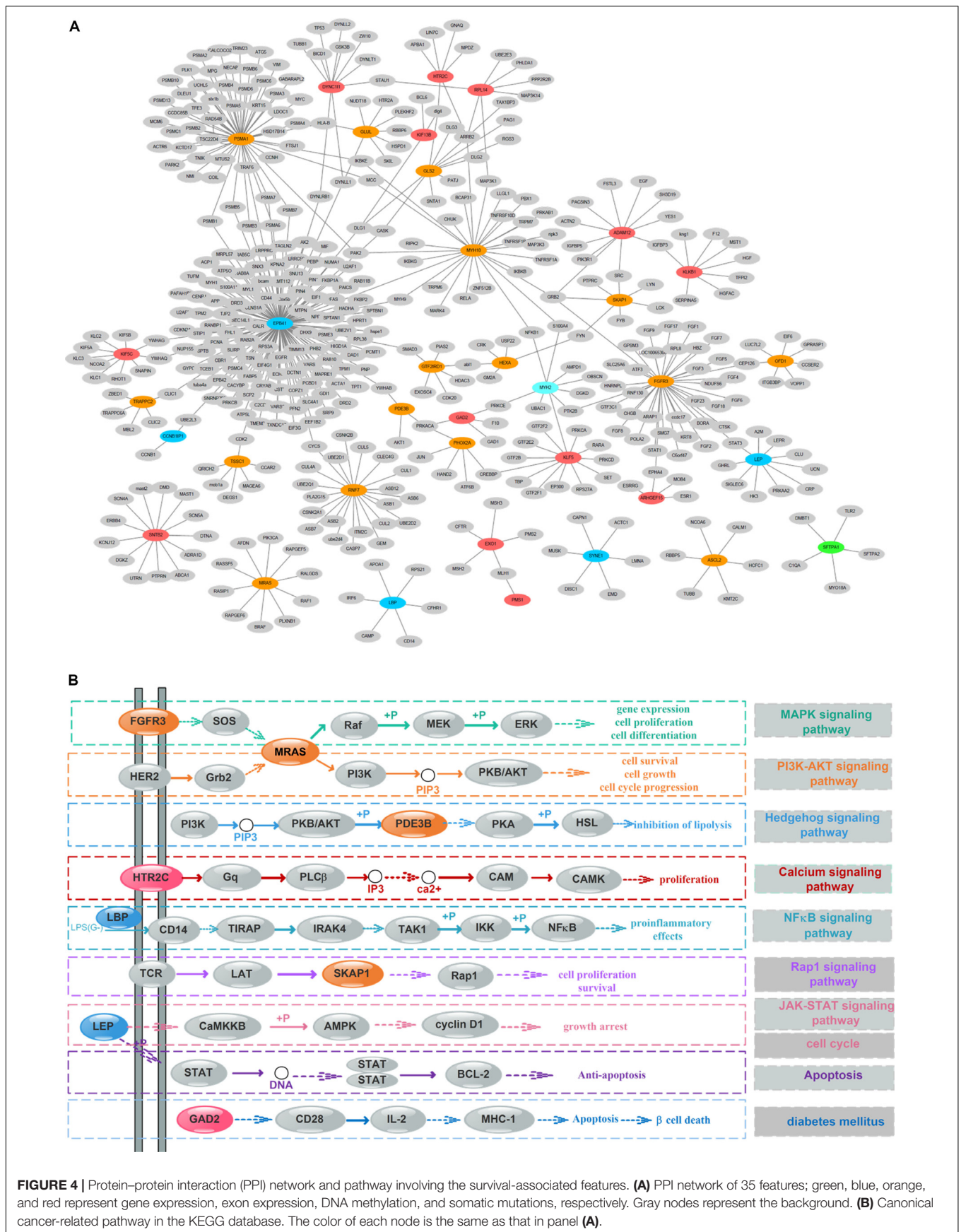
Previous researches reported that signaling pathways are frequently altered in cancers (Sanchez-Vega et al., 2018). Here, FGFR3 and MRAS in the MAPK signaling pathway were found to participate in gene expression and cellular processes including cell proliferation and cell differentiation. Additionally, MRAS is involved in the PI3K-AKT signaling pathway related to cell survival, cell growth, and cell cycle progression, while PDE3B is found in the Hedgehog signaling pathway, which is related to the inhibition of lipolysis. Besides, HTR2C which participated in the calcium signaling pathway is potentially associated with cell proliferation. Moreover, LBP involved in the NF- $\kappa$ B signaling pathway is associated with pro-inflammatory effects, and SKAP1 in the Rap1 signaling pathway is involved in cell proliferation and survival.

Besides, genetic alterations that control cell cycle progression and apoptosis are considered to be common hallmarks of multiple cancer types (Sanchez-Vega et al., 2018). LEP is found to be involved in the cell cycle, JAK-STAT signaling pathway and apoptosis to regulate tumor growth arrest and apoptosis. Moreover, emerging evidence from observational studies and meta-analyses suggest that diabetes mellitus is associated with an

increased risk of cancer, as well as cancer incidence or prognosis (Noto, 2018). In particular, diabetes has been validated as a prognostic factor in stages I to III colorectal cancer patients (Croft et al., 2018), in which GAD2 involved could induce  $\beta$ -cell death. Other cancer-associated processes or pathways and literature evidence for the occurrence or development of colon cancer can be found in **Table 1**. Thus, features such as EPB41, PSMA1, FGFR3, MRAS, and LEP which were involved in cancer-related pathways and with high PPI degrees ( $\geq 10$ ) were considered as PRBs for further analysis.

## Detection and Evaluation of Prognosis Risk Biomarkers for COAD

To further explore how the above prognosis-related features affect the prognosis of COAD patients, we performed *in silico* TRS modeling (see section “Materials and Methods”). The expression value of features in the negative group (low-OS patients) was individually or assembly adjusted to normal levels as those in the positive group (high-OS patients). Then, the prognosis level of adjusted patients was simulated through





**TABLE 1** | The information of genes in the integrated PPI network.

Gene	Class	Color	Degree	Pathway	Database	PMID
GAD2* <sup>#</sup>	Mutation	Red	4	Type I diabetes mellitus*	KEGG	3003660 <sup>§</sup>
KIF5C	Mutation	Red	10	Dopaminergic synapse	KEGG	18413843 <sup>§</sup>
CCNB1IP1	Exon	Blue	2	Cell cycle	PubMed	19437480
OFD1	Methylation	Orange	7	Cell cycle	Genecards	20835237
KIF13B	Mutation	Red	3	~	~	23142292
GTF2IRD1	Methylation	Orange	6	cGMP-PKG signaling pathway	KEGG	23804703 <sup>§</sup>
PDE3B*	Methylation	Orange	3	Hedgehog signaling pathway*	KEGG	24282571
RPL14	Mutation	Red	7	Ribosome	KEGG	24824907 <sup>§</sup>
KLF5	Mutation	red	13	Glucose/energy metabolism	Genecards	25037223
LEP*	Exon	Blue	10	JAK-STAT signaling pathway, cell cycle, apoptosis*	KEGG	25250132
HTR2C*	Mutation	Red	6	Calcium signaling pathway*	KEGG	25323578 <sup>§</sup>
FGFR3*	Methylation	Orange	37	MAPK signaling pathway*	KEGG	25623536
HEXA	Methylation	Orange	5	Metabolic pathways	Genecards	26401073
EXO1	Mutation	Red	5	Mismatch repair	KEGG	26714599
CRELD2 <sup>#</sup>	Mutation	Red	2	~	~	27392906
KLKB1	Mutation	Red	8	Complement and coagulation cascades	KEGG	27779706 <sup>§</sup>
LBP*	Exon	Blue	6	NF-kappa B signaling pathway*	KEGG	27876571
GLS2	Methylation	Orange	10	Alanine, aspartate and glutamate metabolism	KEGG	28007957
SYNE1	Exon	Blue	6	Cell cycle	Genecards	28058013
EPB41	Exon	Blue	128	Cell cycle	Genecards	28193906 <sup>§</sup>
DYNC111	Mutation	Red	10	Cell cycle	Genecards	28193906 <sup>§</sup>
GLUL	Methylation	Orange	8	Alanine, aspartate and glutamate metabolism	KEGG	28207045 <sup>§</sup>
ARHGEF15	Mutation	Red	4	GPCR Pathway	Genecards	29073728
CAPN6 <sup>#</sup>	Mutation	Red	2	Apoptosis Pathway	Genecards	29202800 <sup>§</sup>
MRAS*	Methylation	Orange	10	MAPK/PI3K-AKT signaling pathway*	KEGG	29305742 <sup>§</sup>
PSMA1	Methylation	Orange	54	wnt pathway	Genecards	29423100
ADAM12	Mutation	Red	11	Insulin growth factor-related pathway	PubMed	29731694
ASCL2	Methylation	Orange	6	~	~	29886802
POLG <sup>#</sup>	Mutation	Red	2	Metabolic pathways	Genecards	30002826
SKAP1*	Methylation	Orange	7	Rap1 signaling pathway*	KEGG	30183087 <sup>§</sup>
TSSC1	Methylation	Orange	6	~	~	30231249
DAP <sup>#</sup>	Methylation	Orange	1	Apoptosis and autophagy	Genecards	30552554
MYH10	Methylation	Orange	28	~	~	~
RNF7	Methylation	Orange	21	~	~	~
SNTB2	Mutation	Red	13	~	~	~
MYH2	Gene_exon	Lightblue	7	~	~	~
PHOX2A	Methylation	Orange	5	~	~	~
SFTPA1	Gene	Green	5	~	~	~
TRAPPC2	Methylation	Orange	5	~	~	~
SLC30A3 <sup>#</sup>	Exon	Blue	3	~	~	~
EXD3 <sup>#</sup>	Mutation	Red	2	~	~	~
FAM71E2 <sup>#</sup>	Mutation	Red	1	~	~	~
OPN4 <sup>#</sup>	Exon	Blue	1	~	~	~
PMS1	Mutation	Red	1	~	~	~
PPP4R4 <sup>#</sup>	Exon	Blue	1	~	~	~

<sup>a</sup>Columns 1–4 stands for the gene name, corresponding omics data type, color and degree of genes illustrated in the PPI network, respectively. Columns 5 and 6 represent the pathway name and database of gene annotations. Column 7 represents the PMID of the corresponding literature. \*Genes and pathways involved in the integrated pathway in **Figure 4B**. The first column represents all genes that could be mapped to the PPI network. <sup>#</sup>Genes not illustrated in **Figure 4**. <sup>§</sup>Pathway in which the gene located is associated with colon cancer.

the MPA model. Here, 105 single prognosis-related features and 5,460 combinations of two features were systemically adjusted for simulation. Results showed that 12 out of 105 single features and 1,210 out of 5,460 combined features could

change the prognosis of TCGA patients. As been illustrated in **Figure 5**, each node represents 1 of 105 single prognosis-related features and each line links two nodes represents one feature pair. Detailed information of nodes can be found in

**Supplementary Table S5.** Since the regulation of single or combined prognosis-related features could change the prognosis of TCGA patients, it is possible to detect the PRBs for COAD, and thus, drugs that targeting the corresponding PRBs might be helpful for COAD patients. It can be found that the most essential nodes include me8 (cg06685724), e20 (chr2:106226785–106227016:–), e25 (chr20:36977951–36978065:+) representing C7orf46, LOC285000 and LBP, respectively. More importantly, the above three features were involved in 104 out of 1,210 feature combinations, which exceeding other features. In that case, C7orf46, LOC285000, and LBP were defined as essential prognosis biomarkers detected by TRS modeling.

Finally, we evaluate the performance of features derived from different aspects including 13 from the PPI network, 8 from the integrated pathway and 12 from the TRS modeling. The survival analysis was performed for each feature based on the median value as the classification indicator to evaluate the performance. Results showed that the 4 features including ZNF35 (cg20717205), LOC285000 (chr2:106226785–106227016:–), SLC30A3 (chr2:27479254–27479388:–) and LECT2 derived from TRS modeling could distinguish patients with different OS (**Supplementary Figure S2**). For PPI, 3 features that hold the potential to distinguish patients with different OS were detected, including LEP (chr7:127894457–127897682), RNF7 (cg06671690) and DYNC1I1 (**Supplementary Figure S2**). Among them, LEP was also involved in the JAK-STAT signaling pathway.

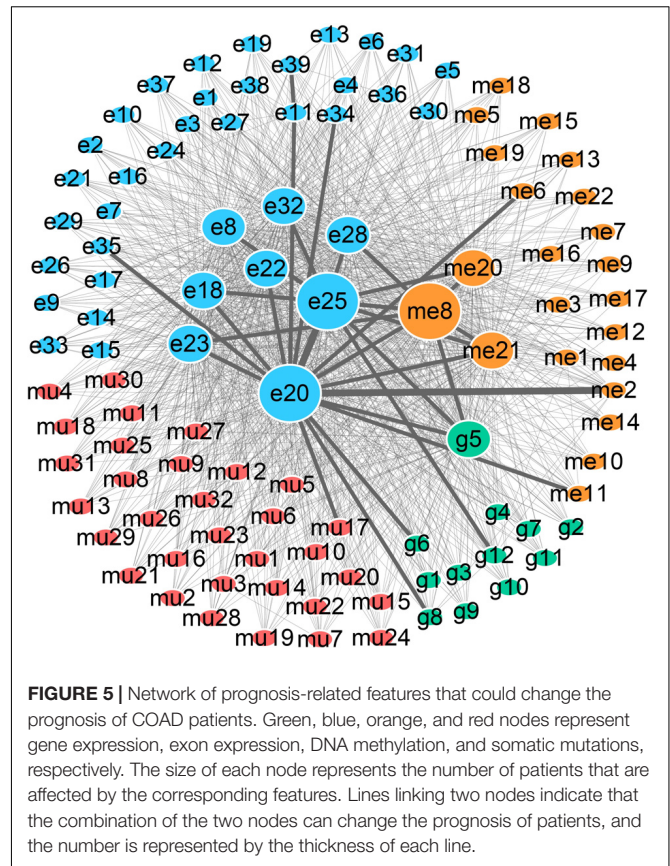
Thus, through PPI network and pathway integration analysis of all prognosis-related features detected by MPA modeling, 7 essential PRBs including EPB41, PSMA1, FGFR3, MRAS, LEP, RNF7, and DYNC1I1 were identified. Further, by combining with TRS modeling, C7orf46, LOC285000, LBP, ZNF35, SLC30A3, and LECT2 were also added and a total of 13 PRBs were eventually detected by integrating MPA and TRS modeling. Further, a prognosis risk scoring (PRS) model based on above 13 PRBs were constructed to evaluate whether those markers could distinguish high-OS and low-OS patients. By using linear regression, the PRS could be described based on the following equation (1):

$$\begin{aligned} \text{PRS} = & (-1.021 * \text{EPB41}) + (-0.364 * \text{PSMA1}) + \\ & (-0.046 * \text{FGFR3}) + (-0.113 * \text{MRAS}) + (-0.013 * \text{LEP}) + \\ & (-0.219 * \text{C7orf46}) + (-0.379 * \text{LOC285000}) + \\ & (-0.085 * \text{LBP}) + (-0.539 * \text{ZNF35}) + \\ & (-0.150 * \text{SLC30A3}) + (-0.057 * \text{LECT2}) + \\ & (-0.404 * \text{RNF7}) + (-0.143 * \text{DYNC1I1}) + 0.325 \quad (1) \end{aligned}$$

The performance of the above PRS model could reach to 0.825 for AUC value, and by setting the best threshold of 0.254, the sensitivity of 0.900 and specificity of 0.625 could be achieved.

## Validation of PRBs on Independent Testing Dataset

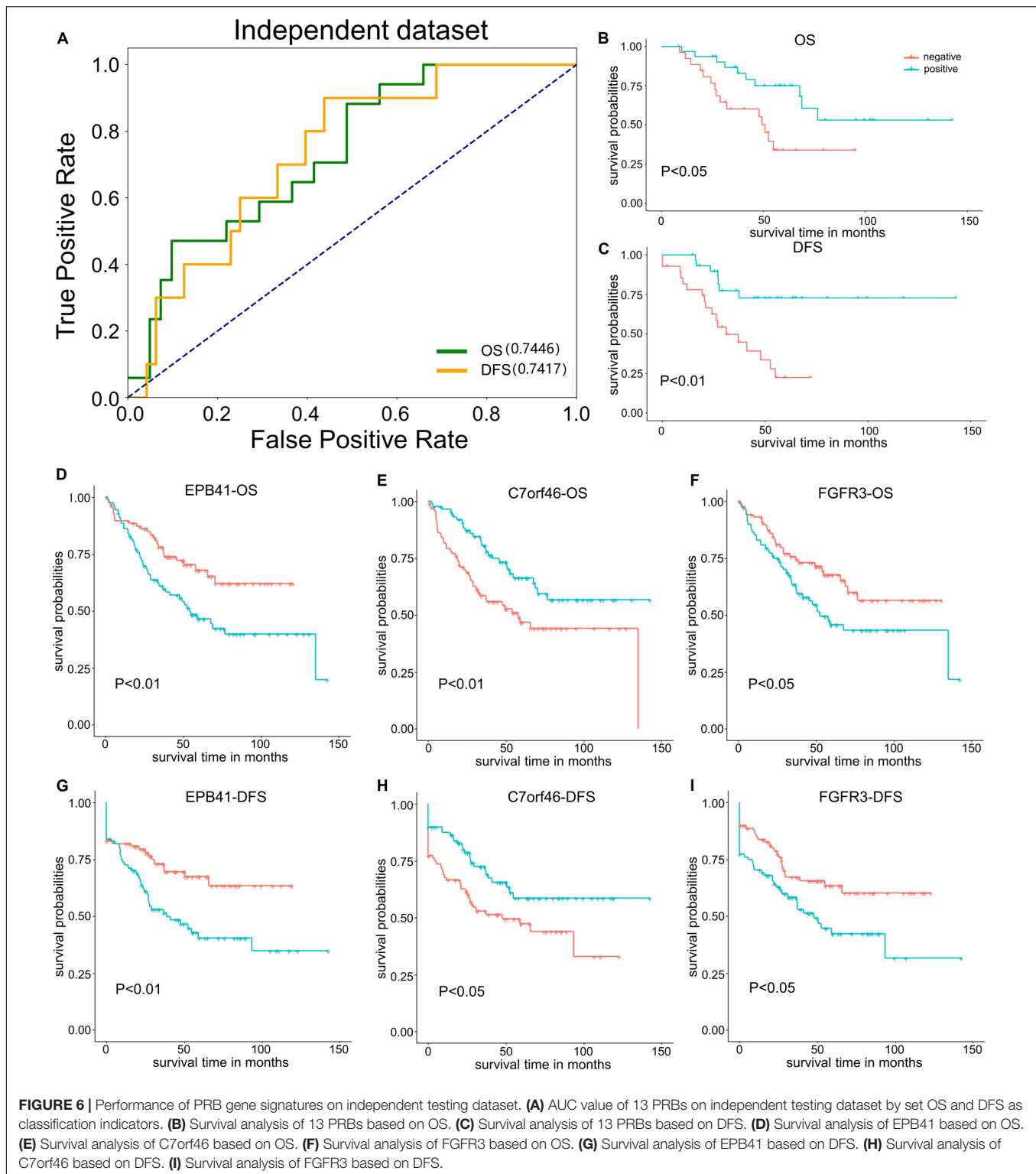
To evaluate the practicability and scalability of the PRBs mentioned above, 13 PRBs explored in this study were used to validate the performance in independent datasets



**FIGURE 5 |** Network of prognosis-related features that could change the prognosis of COAD patients. Green, blue, orange, and red nodes represent gene expression, exon expression, DNA methylation, and somatic mutations, respectively. The size of each node represents the number of patients that are affected by the corresponding features. Lines linking two nodes indicate that the combination of the two nodes can change the prognosis of patients, and the number is represented by the thickness of each line.

downloaded from the NCBI GEO dataset (see section “Materials and Methods”), which included 177 patients with survival information from Moffitt Cancer Center. Here, all potential PRBs detected above were first translated into gene signatures. Then, the expression profiles of the above gene signatures were used to generate the classification model for patient samples by setting the cutoff of OS and disease-free survival (DFS) as 5 years. Results indicated that the prediction model constructed on the above gene signatures could achieve the AUC value of 0.745 for OS and 0.742 for DFS, respectively (**Figure 6A**). Moreover, the survival analysis based on the prediction of the gene signature-based model indicated that by integrating all the above features, the classification results could successfully distinguish the positive and negative samples for both OS (**Figure 6B**) and DFS (**Figure 6C**). Furthermore, by setting the median expression value as the cutoff in the independent testing dataset, gene signatures of EPB41, C7orf46, and FGFR3 could also distinguish the positive and negative samples classified by both OS and DFS, has been illustrated in **Figures 6D–6I**.

Moreover, the single cell-based expression dataset of colon cancer was obtained from the NCBI GEO database with the accession number of GSE18161 (Li et al., 2017), which included the count and FPKM of tumor/NM cells, as well as the count and FPKM of tumor/NM epithelial cells. Further, the Wilcoxon Test and Fold change were used to evaluate whether



the above PRBs were differentially expressed between tumor and normal cells (**Supplementary Table S6**). Results showed that gene signatures such as PSMA1, FGFR3, C7orf46, RNF7, and ZNF35 were differentially expressed in normal and tumor samples with the  $P$ -value  $< 0.05$ . Other gene signatures including

EPB41, MRAS, LEP, LOC285000, LBP, LECT2, SLC30A3, and DYNC1T1 were also differentially expressed in normal and tumor samples with  $|FC| > 2$  (**Supplementary Figure S3**). Thus, the evaluation of the corresponding gene signatures for PRBs through the independent testing dataset illustrated that

above 13 PRBs detected by MPA and TRS modeling could be defined as classification indicators to predict the prognosis of COAD patients.

## PRBs Illustrated Different Affections for Patients With Different Conditions

It should be noticed that clinical information which reflect different conditions of patient samples might affect the prognosis of colon cancer (Vergo and Benson, 2012; Dienstmann et al., 2015; Karvinen and Vallance, 2015). Thus, 11 personalized clinical features including age, gender, weight, histological type, history of colon polyps, person neoplasm cancer status, lymphatic invasion, pathologic stage, pathologic T stage, venous invasion and number of first degree relatives with cancer diagnosis, which may closely related with cancer prognosis were selected to the prognostic risk assessment with Cox PH regression model. Results showed that age, weight, person neoplasm cancer status, lymphatic invasion, pathologic stage, pathologic T stage, and venous invasion were detected as the colon cancer prognosis-related risk factors (**Supplementary Table S7**). Further, samples in our testing dataset were grouped by personalized prognostic risk features as well as the number of first degree relatives with cancer diagnosis. For example, for the risk factor of age, all patients were separated into two groups, age over 65 or age less than or equal to 65. Then, all 13 PRB-related gene signatures were individually analyzed by setting the median value as the cutoff to classify the prognosis difference in each group through survival analysis.

The co-occurrence and exclusivity of PRBs could be detected in patients with different personal conditions. For example, SLC30A3 could significantly distinguish positive and negative samples for both patients older than 65 or younger than 65. However, PRBs such as LOC285000 and LEP could only distinguish those patients older than 65, while DYNC111 only illustrate significance in patients younger than 65 (**Supplementary Figure S4**). For person neoplasm cancer status, LOC285000 and LEP could distinguish the prognosis of patients free from neoplasm, while SLC30A3 and DYNC111 were significant for patients with neoplasm (**Supplementary Figure S5**). Thus, PRBs illustrated different affections for patients with different personalized clinical conditions which were closely associated with the prognosis of COAD. Detailed information of different sample groups classified by personalized prognostic risk features (**Supplementary Table S8**) and the corresponding significant PRBs (**Supplementary Table S9**) were provided.

## DISCUSSION

Identification of PRBs in colon cancer is essential for the diagnosis, monitoring, and treatment of patients. By taking advantage of next-generation sequencing technologies, large-scale data could be obtained for *in silico* analysis to reveal PRBs. Here, we presented the MPA and TRS modeling to detect the PRBs for COAD. First, by integrating multi-omics data from gene expression, exon expression, DNA

methylation, and somatic mutations. Then, features selection were obtained through dimensionality reduction based on factor analysis. After that, all 105 essential features from quadruple-omics data were integrated to generate the MPA model. Furthermore, 45 prognosis-related features were obtained through the analysis of PPI networks and mapped into multiple cancer-related pathways. Among them, some prognosis-related features were directly related with the occurrence, development or prognosis of cancer, such as PSMA1 (degree = 54) was identified as colon cancer markers by proteomic profiling (Yang et al., 2018), FGFR3 (degree = 37) was related with multi-regional colon cancer through inter- and intra-tumor profiling (Kogita et al., 2015), ALPi is selectively induced by HDACi in colon cancer cells in a KLF5 (degree = 13) dependent manner (Shin et al., 2014), GLS2 (degree = 10) was validated as differential expression gene in colon cancer cells (Alix-Panabieres et al., 2017) and LEP (degree = 10) was examined to be associated with the development of colorectal cancer (Rezaei-Tavirani et al., 2013). Furthermore, several pathways were provided to be colon cancer-related pathways, such as wnt pathway, PI3K-AKT signaling pathway and cell cycle signaling pathway were reported as common oncogenic signaling pathways (Sanchez-Vega et al., 2018), which could be regulated by PSMA1 (degree = 54), MRAS (degree = 10), and LEP (degree = 10), respectively. Detailed information of all prognosis-related features including degrees, corresponding pathways, and literature evidence was listed in **Table 1**.

Elaborate investigation of TCGA samples indicated that the above prognosis-related features were mostly overexpressed in negative samples. To further illustrate the therapeutic actionability, TRS modeling was processed to detect potential targets for inhibitors. Generally, TRS could simulate the prognosis classifications of TCGA patients with expression levels altered for individual or combined prognosis-related features. After scanning all single and combinations among 105 prognosis-related features in low-OS patients, results indicated that alter the expression level of features such as chr20:36977951–36978065:+ (LBP), chr2:106226785–106227016:– (LOC285000), and cg06685724 (C7orf46), the classification could be switched from low-OS to high-OS. Thus, by integrating MPA and TRS modeling, 8 features including chr1:29315868–29315947:+ (EPB41), cg02654360 (PSMA1), cg23835677 (FGFR3), cg18421529 (MRAS), chr7:127894457–127897682:+ (LEP), cg06685724 (C7orf46), chr2:106226785–106227016:– (LOC285000) and chr20:36977951–36978065:+ (LBP) were detected as potential PRBs. Among them, LOC28500 and LBP were also been identified as prognostic risk factors by both univariate and multivariate analysis of the Cox PH regression model (**Supplementary Table S10**). Moreover, FGFR3 has already been proved to be an essential drug target for multiple cancer types. For example, XL999, which targeting FGFR3, has the potential to prevent tumor growth and has been investigated for the treatment of unspecified cancer/tumors (**Supplementary Table S11**). In addition, to compare the prognostic performance of the prognosis-related features of colon cancer from different criteria, we evaluated the performance of the features derived

from the PPI network, integrated pathway, and TRS modeling. The results of survival analysis showed that features extracted from TRS modeling reflected better performance by comparing with those derived from the PPI network and integrated pathway, indicating that the TRS modeling might be an efficient strategy to explore PRBs for cancer prognosis. By combining 13 PRBs from different aspects, the linear regression model could reach an AUC value of 0.825. Thus, the strategy of screening PRBs from different aspects might better reflect the prognostic features of cancer patients as previous studies reported (Chen S. et al., 2019; Chen Y. H. et al., 2019; Wang et al., 2019).

Also, the above 13 PRBs were also reported to affect the progression and prognosis of different cancers. For example, EPB41, PSMA1, LEP, LECT2, and ZNF35 were associated with breast cancer (Kao et al., 2005; Deng et al., 2006; Andres et al., 2015; Feng et al., 2019). PSMA1, MRAS, LEP, SLC30A3, and RNF7 were found closely related with prostate cancer (Singh et al., 2016; Sun et al., 2016; Wang et al., 2016; Xiao et al., 2017; Zhu et al., 2017). FGFR3, LBP, LECT2, and DYNC111 were distinguishable markers in lung cancer (Wang et al., 2017, 2018; Zhang Y. et al., 2017; Hung et al., 2018). Besides intra-validation, the generality and repeatability of PRBs were evaluated through an independent dataset from the GEO database. The results of both classification and survival analysis indicated that the PRBs and corresponding gene signatures determined here could effectively distinguish the samples with different prognostic independent dataset, which could be used as prognostic classification indicator for COAD patients.

Furthermore, it is noted that among the above 13 PRBs, five PRBs of chr1:29315868–29315947:+ (EPB41), chr7:127894457–127897682:+ (LEP), chr2:106226785–106227016:– (LOC285000) chr20:36977951–36978065:+ (LBP), and chr2:27479254–27479388:– (SLC30A3), were from exon expression, while other six PRBs including cg02654360 (PSMA1), cg23835677 (FGFR3), cg18421529 (MRAS), cg06685724 (C7orf46), cg20717205 (ZNF35), and cg06671690 (RNF7) were from DNA methylation. And one of the PRBs LECT2 was from gene expression, and the left one DYNC111 was from somatic mutation. This means the expression level of exon expression and DNA methylation might be more important for the prognosis of COAD patients rather than gene expression and somatic mutations. Meanwhile, by integrating quadruple-omics data with appropriate machine learning approaches such as logistic regression, the prognosis prediction performance could be further increased to 0.998 based on 105 essential features. Thus, with the accumulation of multi-omics data and improvement of machine learning approaches, the PRBs for multiple cancer types could be detected and accelerate the development of cancer therapeutics.

## CONCLUSION

In this paper, we constructed the MPA model to comprehensively reveal the PRBs for COAD patients based on gene expression,

exon expression, DNA methylation, and somatic mutations. Besides the high performance of the MPA model for prognostic classification, 105 essential features that were closely related to COAD prognosis were detected. Furthermore, by screening through the criteria of the PPI network, cancer-related pathway and TRS modeling, essential features with gene symbols of EPB41, PSMA1, FGFR3, MRAS, LEP, C7orf46, LOC285000, LBP, ZNF35, SLC30A3, LECT2, RNF7, and DYNC111 were identified as PRBs for COAD patients. In addition, evaluation of the independent testing dataset and single-cell based RNA-seq dataset illustrated the PRBs and corresponding gene symbols detected in this study could successfully distinguish COAD patients with different prognosis. Finally, some of the PRBs were demonstrated to hold the potential to distinguish different prognosis in patients with different clinical conditions. The MPA and TRS modeling, as well as the PPI network and integrated pathway analysis presented here could only detect the PRBs to predict the prognosis of COAD patients, but also provide new perspectives for novel drug development and therapeutic applications for COAD treatment.

## DATA AVAILABILITY STATEMENT

The raw omics and phenotypic data of COAD were obtained from the TCGA module of Public Xena Hubs in the UCSC Xena database.

## AUTHOR CONTRIBUTIONS

TQ and ZY designed the study and wrote the manuscript. ZY collected the corresponding datasets and completed *in silico* analyses. XY and QW assisted in model construction. ZD and KT assisted in model validation. TQ and ZC supervised the whole project and edited the manuscript.

## FUNDING

This work was supported in part by the National Key R&D Program of China [grant numbers 2017YFC1700200 and 2017YFC0908405], the National Natural Science Foundation of China [31900483], the Fundamental Research Funds for the Central Universities [1350219165], and the Shanghai Sailing Program [19YF1441100].

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00524/full#supplementary-material>

## REFERENCES

- Alix-Panabieres, C., Cayrefourcq, L., Mazard, T., Maudelonde, T., Assenat, E., and Assou, S. (2017). Molecular portrait of metastasis-competent circulating tumor cells in colon cancer reveals the crucial role of genes regulating energy metabolism and DNA repair. *Clin. Chem.* 63, 700–713. doi: 10.1373/clinchem.2016.263582
- Andre, T., Boni, C., Navarro, M., Tabernero, J., Hickish, T., Topham, C., et al. (2009). Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *J. Clin. Oncol.* 27, 3109–3116. doi: 10.1200/jco.2008.20.6771
- Andres, S. A., Bickett, K. E., Alatoum, M. A., Kalbfleisch, T. S., Brock, G. N., and Wittliff, J. L. (2015). Interaction between smoking history and gene expression levels impacts survival of breast cancer patients. *Breast Cancer Res. Treat.* 152, 545–556. doi: 10.1007/s10549-015-3507-z
- Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Mombalona, X., Iglesias, M., et al. (2015). Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* 47, 320–362. doi: 10.1038/ng.3225
- Chen, S., Shan, T., Chen, X., Yang, W., Wu, T., Sun, X., et al. (2019). Association of treRNA with lymphatic metastasis and poor prognosis in colorectal cancer. *Int. J. Clin. Exp. Pathol.* 12, 1770–1774.
- Chen, Y. H., Lin, T. T., Wu, Y. P., Li, X. D., Chen, S. H., Xue, X. Y., et al. (2019). Identification of key genes and pathways in seminoma by bioinformatics analysis. *Oncotargets Ther.* 12, 3683–3693. doi: 10.2147/OTT.S199115
- Cole, M., Bandeen-Roche, K., Hirsch, A. G., Kuiper, J. R., Sundaresan, A. S., Tan, B. K., et al. (2018). Longitudinal evaluation of clustering of chronic sinonasal and related symptoms using exploratory factor analysis. *Allergy* 73, 1715–1723. doi: 10.1111/all.13470
- Croft, B., Reed, M., Patrick, C., Kovacevich, N., and Voutsadakis, I. A. (2018). Diabetes, obesity, and the metabolic syndrome as prognostic factors in stages I to III colorectal cancer patients. *J. Gastrointest. Cancer* 50, 221–229. doi: 10.1007/s12029-018-0056-9
- Cruz, J. A., and Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2, 59–77.
- Das, V., Kalita, J., and Pal, M. (2017). Predictive and prognostic biomarkers in colorectal cancer: a systematic review of recent advances and challenges. *Biomed. Pharmacother.* 87, 8–19. doi: 10.1016/j.biopha.2016.12.064
- Deng, S., Rong, H. B., Tu, H., Zheng, B. X., Mu, X. Y., Zhu, L. Y., et al. (2019). Molecular basis of neurophysiological and antioxidant roles of Szechuan pepper. *Biomed. Pharmacother.* 112:108696. doi: 10.1016/j.biopha.2019.108696
- Deng, S. S., Xing, T. Y., Zhou, H. Y., Xiong, R. H., Lu, Y. G., Wen, B., et al. (2006). Comparative proteome analysis of breast cancer and adjacent normal breast tissues in human. *Genomics Proteomics Bioinformatics* 4, 165–172. doi: 10.1016/s1672-0229(06)60029-6
- Dienstmann, R., Salazar, R., and Tabernero, J. (2015). Personalizing colon cancer adjuvant therapy: selecting optimal treatments for individual patients. *J. Clin. Oncol.* 33, 1787–1796. doi: 10.1200/JCO.2014.60.0213
- Feng, G., Guo, K., Yan, Q., Ye, Y., Shen, M., Ruan, S., et al. (2019). Expression of protein 4.1 family in breast cancer: database mining for 4.1 family members in malignancies. *Med. Sci. Monit.* 25, 3374–3389. doi: 10.12659/msm.914085
- Gustafsson, U. O., Ooppelstrup, H., Thorell, A., Nygren, J., and Ljungqvist, O. (2016). Adherence to the ERAS protocol is associated with 5-year survival after colorectal cancer surgery: a retrospective cohort study. *World J. Surg.* 40, 1741–1747. doi: 10.1007/s00268-016-3460-y
- Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–D455.
- Hernandez, V., Cubiella, J., Gonzalez-Mao, M. C., Iglesias, F., Rivera, C., Iglesias, M. B., et al. (2014). Fecal immunochemical test accuracy in average-risk colorectal cancer screening. *World J. Gastroenterol.* 20, 1038–1047.
- Hong, S. R., Shin, K. J., Jung, S. E., Lee, E. H., and Lee, H. Y. (2019). Platform-independent models for age prediction using DNA methylation data. *Forensic Sci. Int. Genet.* 38, 39–47. doi: 10.1016/j.fsigen.2018.10.005
- Hung, W. Y., Chang, J. H., Cheng, Y., Chen, C. K., Chen, J. Q., Hua, K. T., et al. (2018). Leukocyte cell-derived chemotaxin 2 retards non-small cell lung cancer progression through antagonizing MET and EGFR activities. *Cell. Physiol. Biochem.* 51, 337–355. doi: 10.1159/000495233
- Jiang, T. M., Liu, B., Li, J. F., Dong, X. Y., Lin, M., Zhang, M. H., et al. (2018). Association between sn-2 fatty acid profiles of breast milk and development of the infant intestinal microbiome. *Food Funct.* 9, 1028–1037. doi: 10.1039/c7fo00088j
- Kandimalla, R., Linnekamp, J. F., van Hooff, S., Castells, A., Llor, X., Andreu, M., et al. (2017). Methylation of WNT target genes AXIN2 and DKK1 as robust biomarkers for recurrence prediction in stage II colon cancer. *Oncogenesis* 6:e308. doi: 10.1038/oncsis.2017.9
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.
- Kao, C. L., Chiou, S. H., Chen, Y. J., Singh, S., Lin, H. T., Liu, R. S., et al. (2005). Increased expression of osteopontin gene in atypical teratoid/rhabdoid tumor of the central nervous system. *Mod. Pathol.* 18, 769–778. doi: 10.1038/modpathol.3800270
- Karvinen, K., and Vallance, J. (2015). Breast and colon cancer survivors' expectations about physical activity for improving survival. *Oncol. Nurs. Forum* 42, 527–533. doi: 10.1188/15.ONF.527-533
- Katai, H., Ishikawa, T., Akazawa, K., Isobe, Y., Miyashiro, I., Oda, I., et al. (2018). Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese Gastric Cancer Association (2001–2007). *Gastric Cancer* 21, 144–154. doi: 10.1007/s10120-017-0716-7
- Katoh, S., Goi, T., Naruse, T., Ueda, Y., Kurebayashi, H., Nakazawa, T., et al. (2015). Cancer stem cell marker in circulating tumor cells: expression of CD44 variant exon 9 is strongly correlated to treatment refractoriness, recurrence and prognosis of human colorectal cancer. *Anticancer Res.* 35, 239–244.
- Kogita, A., Yoshioka, Y., Sakai, K., Togashi, Y., Sogabe, S., Nakai, T., et al. (2015). Inter- and intra-tumor profiling of multi-regional colon cancer and metastasis. *Biochem. Biophys. Res. Commun.* 458, 52–56. doi: 10.1016/j.bbrc.2015.01.064
- Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.* 696, 291–303. doi: 10.1007/978-1-60761-987-1\_18
- Kourou, K., Exarchos, T. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005
- Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49, 708–718. doi: 10.1038/ng.3818
- Li, J. L., Liu, Y., Wang, C., Deng, T., Liang, H. W., Wang, Y. F., et al. (2015). Serum miRNA expression profile as a prognostic biomarker of stage II/III colorectal adenocarcinoma. *Sci. Rep.* 5:12921.
- Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* 46, D1121–D1127.
- Liu, J., Li, T., and Liu, X. L. (2017). DDA1 is induced by NR2F6 in ovarian cancer and predicts poor survival outcome. *Eur. Rev. Med. Pharmacol.* 21, 1206–1213.
- Lorenzo-Seva, U., and Van Ginkel, J. R. (2016). Multiple imputation of missing values in exploratory factor analysis of multidimensional scales: estimating latent trait scores. *Anal. Psicol.* 32, 596–608.
- Luo, H. Y., and Xu, R. H. (2014). Predictive and prognostic biomarkers with therapeutic targets in advanced colorectal cancer. *World J. Gastroenterol.* 20, 3858–3874.
- Marley, A. R., and Nan, H. M. (2016). Epidemiology of colorectal cancer. *Int. J. Mol. Epidemiol.* 7, 105–114.
- Melichar, B. (2013). Biomarkers in the treatment of cancer: opportunities and pitfalls. *Clin. Chem. Lab. Med.* 51, 1329–1333.
- Modhukur, V., Iljasenko, T., Metsalu, T., Lokk, K., Laisk-Podar, T., and Vilo, J. (2018). MethSurv: a web tool to perform multivariable survival analysis

- using DNA methylation data. *Epigenomics* 10, 277–288. doi: 10.2217/epi-2017-0118
- Noto, H. (2018). “Diabetes and cancers,” in *Diabetes and Aging-Related Complications*, ed. S.-I. Yamagishi (Berlin: Springer), 113–126. doi: 10.1007/978-981-10-4376-5\_9
- O’Connell, M. J., Campbell, M. E., Goldberg, R. M., Grothey, A., Seitz, J. F., Benedetti, J. K., et al. (2008). Survival following recurrence in stage II and III colon cancer: findings from the ACCENT data set. *J. Clin. Oncol.* 26, 2336–2341. doi: 10.1200/JCO.2007.15.8261
- Okugawa, Y., Toiyama, Y., Toden, S., Mitoma, H., Nagasaka, T., Tanaka, K., et al. (2017). Clinical significance of SNORA42 as an oncogene and a prognostic biomarker in colorectal cancer. *Gut* 66, 107–117. doi: 10.1136/gutjnl-2015-309359
- Ozawa, T., Matsuyama, T., Toiyama, Y., Takahashi, N., Ishikawa, T., Uetake, H., et al. (2017). CCAT1 and CCAT2 long noncoding RNAs, located within the 8q. 24.21 ‘gene desert’, serve as important prognostic biomarkers in colorectal cancer. *Ann. Oncol.* 28, 1882–1888. doi: 10.1093/annonc/mdx248
- Prego-Faraldo, M. V., Martinez, L., and Mendez, J. (2018). RNA-Seq analysis for assessing the early response to DSP toxins in *Mytilus galloprovincialis* digestive gland and gill. *Toxins* 10:417. doi: 10.3390/toxins10100417
- Qu, L. P., Zhong, Y. M., Zheng, Z., and Zhao, R. X. (2017). CDH17 is a downstream effector of HOXA13 in modulating the Wnt/beta-catenin signaling pathway in gastric cancer. *Eur. Rev. Med. Pharmacol.* 21, 1234–1241.
- Rappaport, N., Fishilevich, S., Nudel, R., Twik, M., Belinky, F., Plaschkes, I., et al. (2017). Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect. *Biomed. Eng. Online* 16(Suppl. 1):72.
- Rezaei-Tavirani, M., Safaei, A., and Zali, M. R. (2013). The association between polymorphisms in insulin and obesity related genes and risk of colorectal cancer. *Iran. J. Cancer Prev.* 6, 179–185.
- Ricketts, C. J., and Linehan, W. M. (2015). Gender specific mutation incidence and survival associations in clear cell renal cell carcinoma (CCRCC). *PLoS One* 10:e0140257. doi: 10.1371/journal.pone.0140257
- Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the human gene integrator. *Database* 2010:baq020. doi: 10.1093/database/baq020
- 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, 321–337.e10.
- Schlumberger, M., Elisei, R., Muller, S., Schoffski, P., Brose, M., Shah, M., et al. (2017). Overall survival analysis of EXAM, a phase III trial of cabozantinib in patients with radiographically progressive medullary thyroid carcinoma. *Ann. Oncol.* 28, 2813–2819. doi: 10.1093/annonc/mdx479
- Shin, J., Carr, A., Corner, G. A., Togel, L., Davaos-Salas, M., Tran, H., et al. (2014). The intestinal epithelial cell differentiation marker intestinal alkaline phosphatase (ALPI) is selectively induced by histone deacetylase inhibitors (HDACi) in colon cancer cells in a kruppel-like factor 5 (KLF5)-dependent manner. *J. Biol. Chem.* 289, 25306–25316. doi: 10.1074/jbc.m114.557546
- Singh, C. K., Malas, K. M., Tydrick, C., Siddiqui, I. A., Iczkowski, K. A., and Ahmad, N. (2016). Analysis of zinc-exporters expression in prostate cancer. *Sci. Rep.* 6:36772. doi: 10.1038/srep36772
- Smith, J. J., Deane, N. G., Wu, F., Merchant, N. B., Zhang, B., Jiang, A., et al. (2010). Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 138, 958–968. doi: 10.1053/j.gastro.2009.11.005
- Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029
- Sun, Y., Jia, X., Hou, L., and Liu, X. (2016). Screening of differently expressed miRNA and mRNA in prostate cancer by integrated analysis of transcription data. *Urology* 94, 313.e1–313.e6. doi: 10.1016/j.urology.2016.04.041
- Taieb, J., Zaanani, A., Le Malicot, K., Julie, C., Blons, H., Mineur, L., et al. (2016). Prognostic effect of BRAF and KRAS mutations in patients with stage III colon cancer treated with leucovorin, fluorouracil, and oxaliplatin with or without cetuximab a post hoc analysis of the PETACC-8 trial. *JAMA Oncol.* 2, 643–652. doi: 10.1001/jamaonc.2015.5225
- Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136
- van den Braak, R. R. J. C., Martens, J. W. M., and Ijzermans, J. N. M. (2018). CDX2 as a prognostic biomarker in stage II and stage III colon cancer (vol 374, pg 211, 2016). *N. Engl. J. Med.* 379, 2481–2481. doi: 10.1056/nejmc1814750
- Vergo, M. T., and Benson, A. B. III (2012). Point: treating stage II colon cancer: the quest for personalized adjuvant care. *J. Natl. Compr. Canc. Netw.* 10, 1370–1374. doi: 10.6004/jnccn.2012.0142
- Wang, D. C., Shi, L., Zhu, Z., Gao, D., and Zhang, Y. (2017). Genomic mechanisms of transformation from chronic obstructive pulmonary disease to lung cancer. *Semin. Cancer Biol.* 42, 52–59. doi: 10.1016/j.semcancer.2016.11.001
- Wang, N., Song, X., Liu, L., Niu, L., Wang, X., Song, X., et al. (2018). Circulating exosomes contain protein biomarkers of metastatic non-small-cell lung cancer. *Cancer Sci.* 109, 1701–1709. doi: 10.1111/cas.13581
- Wang, X., Tsui, B., Ramamurthy, G., Zhang, P., Meyers, J., Kenney, M. E., et al. (2016). Theranostic agents for photodynamic therapy of prostate cancer by targeting prostate-specific membrane antigen. *Mol. Cancer Ther.* 15, 1834–1844. doi: 10.1158/1535-7163.MCT-15-0722
- Wang, Z., Wang, Z., Niu, X., Liu, J., Wang, Z., Chen, L., et al. (2019). Identification of seven-gene signature for prediction of lung squamous cell carcinoma. *Oncotargets Ther.* 12, 5979–5988. doi: 10.2147/OTT.S198998
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082.
- Xiao, M., Liu, L., Zhang, S., Yang, X., and Wang, Y. (2018). Cancer stem cell biomarkers for head and neck squamous cell carcinoma: a bioinformatic analysis. *Oncol. Rep.* 40, 3843–3851. doi: 10.3892/or.2018.6771
- Xiao, Y., Jiang, Y., Song, H., Liang, T., Li, Y., Yan, D., et al. (2017). RNF7 knockdown inhibits prostate cancer tumorigenesis by inactivation of ERK1/2 pathway. *Sci. Rep.* 7:43683. doi: 10.1038/srep43683
- Xie, H. B., Ma, B., Gao, Q. J., Zhan, H. J., Liu, Y. C., Chen, Z. C., et al. (2018). Long non-coding RNA CRNDE in cancer prognosis: review and meta-analysis. *Clin. Chim. Acta* 485, 262–271. doi: 10.1016/j.cca.2018.07.003
- Xu, G. S., Mao, X. H., Wang, J. M., and Pan, H. Q. (2018). Clustering and recent transmission of *Mycobacterium tuberculosis* in a Chinese population. *Infect. Drug Resist.* 11, 323–330. doi: 10.2147/IDR.S156534
- Yanaiharu, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., et al. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* 9, 189–198. doi: 10.1016/j.ccr.2006.01.025
- Yang, Q., Roehrl, M. H., and Wang, J. Y. (2018). Proteomic profiling of antibody-inducing immunogens in tumor tissue identifies PSMA1, LAP3, ANXA3, and maspin as cancer cancer markers. *Oncotarget* 9, 3996–4019. doi: 10.18632/oncotarget.23583
- Yu, J., Wu, W. K. K., Li, X. C., He, J., Li, X. X., Ng, S. S. M., et al. (2015). Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. *Gut* 64, 636–645. doi: 10.1136/gutjnl-2013-306620
- Yuan, Y., Zheng, X. T., and Lu, X. Q. (2017). Discovering diverse subset for unsupervised hyperspectral band selection. *IEEE Trans. Image Process.* 26, 51–64. doi: 10.1109/TIP.2016.2617462
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). MINT: a molecular interaction database. *FEBS Lett.* 513, 135–140. doi: 10.1016/s0014-5793(01)03293-8
- Zhang, X., Lv, Q. L., Huang, Y. T., Zhang, L. H., and Zhou, H. H. (2017). Akt/FoxM1 signaling pathway-mediated upregulation of MYBL2 promotes progression of human glioma. *J. Exp. Clin. Cancer Res.* 36:105. doi: 10.1186/s13046-017-0573-6
- Zhang, Y., Wang, D. C., Shi, L., Zhu, B., Min, Z., and Jin, J. (2017). Genome analyses identify the genetic modification of lung cancer subtypes. *Semin. Cancer Biol.* 42, 20–30. doi: 10.1016/j.semcancer.2016.11.005
- Zheng, C. X., Zhan, W. H., Zhao, J. Z., Zheng, D., Wang, D. P., He, Y. L., et al. (2001). The prognostic value of preoperative serum levels of CEA, CA19-9 and CA72-4 in patients with colorectal cancer. *World J. Gastroenterol.* 7, 431–434.
- Zhou, Y. P., Shan, T., Ding, W. Z., Hua, Z. Y., Shen, Y. J., Lu, Z. H., et al. (2018a). Study on mechanism about long noncoding RNA MALAT1 affecting pancreatic

- cancer by regulating Hippo-YAP signaling. *J. Cell. Physiol.* 233, 5805–5814. doi: 10.1002/jcp.26357
- Zhou, Y. P., Wei, Q., Fan, J. S., Cheng, S. J., Ding, W. Z., and Hua, Z. Y. (2018b). Prognostic role of the neutrophil-to-lymphocyte ratio in pancreatic cancer: a meta-analysis containing 8252 patients. *Clin. Chim. Acta* 479, 181–189. doi: 10.1016/j.cca.2018.01.024
- Zhu, Y., Wang, L., Liu, M., Qian, C., Yousuf, A., Oto, A., et al. (2017). MRI-based prostate cancer detection with high-level representation and hierarchical classification. *Med. Phys.* 44, 1028–1039. doi: 10.1002/mp.12116

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

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