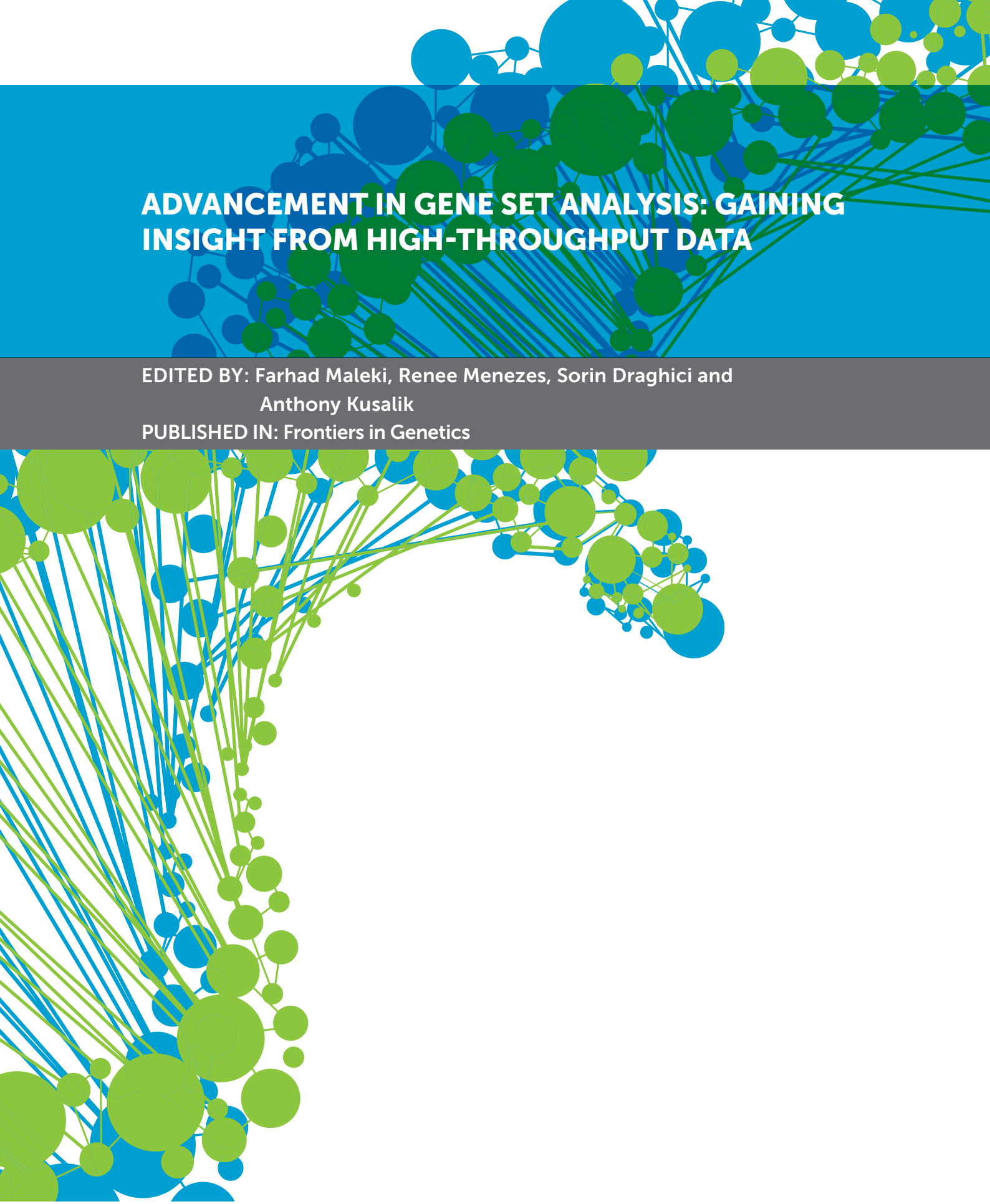

ADVANCEMENT IN GENE SET ANALYSIS: GAINING INSIGHT FROM HIGH-THROUGHPUT DATA

EDITED BY: Farhad Maleki, Renee Menezes, Sorin Draghici and Anthony Kusalik

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

Frontiers eBook Copyright Statement

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

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

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

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

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

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

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

ISSN 1664-8714
ISBN 978-2-88976-423-5
DOI 10.3389/978-2-88976-423-5

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

ADVANCEMENT IN GENE SET ANALYSIS: GAINING INSIGHT FROM HIGH-THROUGHPUT DATA

Topic Editors:
Farhad Maleki, McGill University, Canada
Renee Menezes, The Netherlands Cancer Institute (NKI), Netherlands
Sorin Draghici, Wayne State University, United States
Anthony Kusalik, University of Saskatchewan, Canada

Citation: Maleki, F., Menezes, R., Draghici, S., Kusalik, A., eds. (2022). Advancement in Gene Set Analysis: Gaining Insight From High-throughput Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-423-5

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 26 May 2022
doi: 10.3389/fgene.2022.928724

Editorial: Advancement in Gene Set Analysis: Gaining Insight From High-Throughput Data

Farhad Maleki[1]*, *Sorin Draghici*[2], *Renee Menezes*[3] and *Anthony Kusalik*[4]

[1]Augmented Intelligence & Precision Health Laboratory, Department of Radiology and Research Institute of the McGill University Health Centre, Montreal, QC, Canada, [2]Department of Computer Science, Wayne State University, Detroit, MI, United States, [3]Biostatistics Centre and Department of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, Netherlands, [4]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

Keywords: gene set analysis, pathway analysis, enrichment analysis, gene expression, next-generation sequencing

Editorial on the Research Topic

Advancement in Gene Set Analysis: Gaining Insight from High-Throughput Data

The existence of high-throughput technologies allows for the study of a large number of genes in a single experiment. However, analyzing such high-throughput data and interpreting the results are challenging (Draghici, 2016).

Phenotypes or biological conditions often result from the coordinated activity of a group of genes or biomolecules. Consequently, the study of the coordinated expression pattern of biologically related genes is essential for understanding the mechanisms underlying these conditions or phenotypes. Knowledge bases such as GO (Consortium, 2004) and KEGG (Kanehisa and Goto, 2000) aim to capture knowledge about the roles that genes play in various biological processes and locations. Such resources can be generally divided into: 1) gene set databases (e.g., GO), which include only associations between genes and annotations such as biological processes; and 2) pathway databases (e.g., KEGG), which also capture knowledge related to the interactions between the genes.

Various categories of methods have been developed over time to extract knowledge from such resources (Maleki et al., 2020). The very first methods used a simple approach to identify the gene sets that are enriched in differentially expressed genes (Khatri et al., 2002; Dennis et al., 2003; Draghici et al., 2003b). This approach has various limitations including the fact that it ignores the magnitude of the measured gene expressions. This was addressed by the second generation of methods, pioneered by GSEA (Subramanian et al., 2005), and called functional class scoring (FCS). FCS methods use the correlation between gene expression and the phenotype but still ignore all the interactions between genes. This was addressed by the third generation of methods, called topology-based, or pathway analysis methods. The first such method, impact analysis (Draghici et al., 2007; Tarca et al., 2009), was soon followed by a plethora of over 20 other approaches (Khatri et al., 2012; Mitrea et al., 2013; Nguyen et al., 2018). Many of these methods have been bench-marked recently (Nguyen et al., 2019).

Even though pathway analysis methods are very different from enrichment and FCS methods, we will use "gene set analysis" to generically refer to the entire family of methods aimed at understanding the coordinated expression pattern of known gene sets or pathways. Despite the widespread use of gene set analysis, little consensus exists in the research community regarding best practices. This Research Topic is aimed at highlighting methodological advances as well as applications of gene set analysis to improve the utility of these methods in gaining insight from high-throughput expression studies. Highlights are as follows.

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes, University of Nottingham, United Kingdom

***Correspondence:**
Farhad Maleki
farhad.maleki@mail.mcgill.ca

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

Received: *26 April 2022*
Accepted: *06 May 2022*
Published: *26 May 2022*

Citation:
Maleki F, Draghici S, Menezes R and Kusalik A (2022) Editorial: Advancement in Gene Set Analysis: Gaining Insight From High-Throughput Data. Front. Genet. 13:928724. doi: 10.3389/fgene.2022.928724

Testing for case-control gene expression differences between two groups is a common approach in studies in which researchers are interested in the "difference of differences". Weiner et al. describe a frequent methodological error in using and interpreting gene set analysis methods for such studies. The error occurs when researchers test for differential expression separately in each group and consider genes with significant expression differences in only one comparison—i.e., one group—specific to that group. Based on this assumption, a gene set enrichment analysis is used to find gene sets/pathways specific to only one group. Weiner et al. empirically show that such an approach could report differentially enriched gene sets even for scenarios with no statistically significant differences between the groups.

Marczyk et al. evaluate the effect of incorporating different approaches for integrating single-nucleotide polymorphism (SNP) information and linkage disequilibrium correction on the performance of several gene set analysis methods. They suggest that linkage disequilibrium correction and Stouffer integration could improve the performance of gene set analysis for genome-wide association studies.

Several articles focus on gene set analysis for cancer research. Luo et al. use GSEA (Subramanian et al., 2005) to study the pathways associated with DNA methylation-derived differentially expressed genes in patients with prostate cancer. Song et al. also identify a ubiquitin-related gene signature for prostate cancer prognosis. Li et al. study the association of S100 genes with well-known tumor-related pathways. Xu et al. utilize gene set analysis to identify biological functions and pathways associated with the ferroptosis-related genes in patients with skin cutaneous melanoma. Tan et al. use GSEA to identify gene sets associated with genes co-expressed with the SBSN gene. He et al. find genes differentially expressed in patients with renal cell carcinoma to be associated with autophagy-related pathways. They suggest a prognosis risk score for renal cell carcinoma based on autophagy-related genes that are differentially expressed in patients with the cancer.

The applications of gene set analysis are not limited to cancer research. Yousef et al. employ gene set analysis to validate the biological relevance of the results of their algorithm for miRNA-mRNA regulatory module detection. Du et al. identify hub genes and pathways implicated in osteoporosis. Wu et al. explore potential hub genes in non-alcoholic fatty liver disease and gene sets associated with these genes.

Due to the complex nature of gene set analysis, developing tools that conduct gene set analysis and facilitate interpreting its results is valuable. Among tools commonly used for gene set analysis are DAVID (Dennis et al., 2003), Enrichr (Kuleshov et al., 2016), WebGestalt (Liao et al., 2019), iPathwayGuide (Ahsan and Draghici, 2017), and Onto-Tools (Draghici et al., 2003a). In this Research Topic, Yue et al. present "PAGER Web APP" as an interactive web-based application supporting online R scripting of integrative gene set analysis, and Odom et al. develop an R Package for integrative analysis of multi-omics datasets offering the functionality to work with matched or non-matched samples.

Despite the existence of a large number of gene set analysis methods, there is little consistency among different methods when analyzing the same gene expression dataset (Maleki et al., 2019b; Nguyen et al., 2019). Although gene set overlap is a common phenomenon in gene set databases, most gene set analysis methods disregard such an overlap. This results in a lack of specificity of these methods (Maleki et al., 2020).

Evaluating gene set analysis methods is extremely important (Zyla et al., 2016, 2019) However, most gene set analysis methods have been evaluated either based on oversimplified data—which do not represent real expression datasets and real gene set knowledge bases—or based on real expression datasets with presumed enrichment status for gene sets. Maleki et al. (2021) developed Silver as a methodology for evaluating such methods without relying on oversimplifying assumptions. Besides a thorough evaluation, new gene set analysis methods need to be systematically assessed to find the minimum number of samples required to achieve reproducible results (Maleki et al., 2019a).

The papers published in this Research Topic indicate that the development of gene set analysis methods and tools remains an active research area.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Ahsan, S., and Draghici, S. (2017). Identifying Significantly Impacted Pathways and Putative Mechanisms with iPathwayGuide. *Curr. Protoc. Bioinforma.* 57, 7–30. doi:10.1002/cpbi.24

Consortium, G. O. (2004). The Gene Ontology (GO) Database and Informatics Resource. *Nucleic Acids Res.* 32, D258–D261. doi:10.1093/nar/gkh036

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, P3. doi:10.1186/gb-2003-4-5-p3

Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003b). Global Functional Profiling of Gene Expression. *Genomics* 81, 98–104. doi:10.1016/s0888-7543(02)00021-6

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., and Tainsky, M. A. (2003a). Onto-Tools, the Toolkit of the Modern Biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.* 31, 3775–3781. doi:10.1093/nar/gkg624

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., et al. (2007). A Systems Biology Approach for Pathway Level Analysis. *Genome Res.* 17, 1537–1545. doi:10.1101/gr.6202607

Draghici, S. (2016). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. Boca Raton, FL: CRC Press.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput. Biol.* 8, e1002375. doi:10.1371/journal.pcbi.1002375

Khatri, P., Draghici, S., Ostermeier, G. C., and Krawetz, S. A. (2002). Profiling Gene Expression Using Onto-Express. *Genomics* 79, 266–270. doi:10.1006/geno.2002.6698

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

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205. doi:10.1093/nar/gkz401

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2019a). Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. *Hum. Genomics* 13, 42. doi:10.1186/s40246-019-0226-2

Maleki, F., Ovens, K., Hogan, D. J., and Kusalik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* 11, 654. doi:10.3389/fgene.2020.00654

Maleki, F., Ovens, K. L., Hogan, D. J., Rezaei, E., Rosenberg, A. M., and Kusalik, A. J. (2019b). Measuring Consistency Among Gene Set Analysis Methods: A Systematic Study. *J. Bioinform. Comput. Biol.* 17, 1940010. doi:10.1142/s0219720019400109

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2021). Silver: Forging Almost Gold Standard Datasets. *Genes* 12, 1523. doi:10.3390/genes12101523

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., et al. (2013). Methods and Approaches in the Topology-Based Analysis of Biological Pathways. *Front. Physiol.* 4, 278. doi:10.3389/fphys.2013.00278

Nguyen, T., Mitrea, C., and Draghici, S. (2018). Network-Based Approaches for Pathway Level Analysis. *Curr. Protoc. Bioinforma.* 61, 8–24. doi:10.1002/cpbi.42

Nguyen, T. M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Correction to: Identifying Significantly Impacted Pathways: a Comprehensive Review and Assessment. *Genome Biol.* 20, 234. doi:10.1186/s13059-019-1882-1

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics* 25, 75–82. doi:10.1093/bioinformatics/btn577

Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S. H. E., Polanska, J., and Weiner, J., 3rd (2019). Gene Set Enrichment for Reproducible Science: Comparison of CERNO and Eight Other Algorithms. *Bioinformatics* 35, 5146–5154. doi:10.1093/bioinformatics/btz447

Zyla, J., Marczyk, M., and Polanska, J. (2016). "Sensitivity, Specificity and Prioritization of Gene Set Analysis When Applying Different Ranking Metrics," in International Conference on Practical Applications of Computational Biology & Bioinformatics, June 13, 2016, Seville, Spain. (Berlin: Springer), 61–69. doi:10.1007/978-3-319-40126-3_7

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

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

Copyright © 2022 Maleki, Draghici, Menezes and Kusalik. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 October 2021
doi: 10.3389/fgene.2021.772487

Bioinformatics Analysis Explores Potential Hub Genes in Nonalcoholic Fatty Liver Disease

Chutian Wu[1†], *Yun Zhou*[1,2†], *Min Wang*[1†], *Guolin Dai*[1], *Xiongxiu Liu*[1], *Leizhen Lai*[1] *and Shaohui Tang*[1*]

[1]Department of Gastroenterology, The First Affiliated Hospital, Jinan University, Guangzhou, China, [2]Department of Gastroenterology, The First Affiliated Hospital, Gannan Medical University, Ganzhou, China

Background: Nonalcoholic fatty liver disease (NAFLD) is now recognized as the most prevalent chronic liver disease worldwide. However, the dysregulated gene expression for NAFLD is still poorly understood.

Material and methods: We analyzed two public datasets (GSE48452 and GSE89632) to identify differentially expressed genes (DEGs) in NAFLD. Then, we performed a series of bioinformatics analyses to explore potential hub genes in NAFLD.

Results: This study included 26 simple steatosis (SS), 34 nonalcoholic steatohepatitis (NASH), and 13 healthy controls (HC). We observed 6 up- and 19 down-regulated genes in SS, and 13 up- and 19 down-regulated genes in NASH compared with HC. Meanwhile, the overlapping pathways between SS and NASH were PI3K-Akt signaling pathway and pathways in cancer. Then, we screened out 10 hub genes by weighted Gene Co-Expression Network Analysis (WGCNA) and protein-protein interaction (PPI) networks. Eventually, we found that *CYP7A1/GINS2/PDLIM3* were associated with the prognosis of hepatocellular carcinoma (HCC) in the TCGA database.

Conclusion: Although further validation is still needed, we provide useful and novel information to explore the potential candidate genes for NAFLD prognosis and therapeutic options.

Keywords: nonalcoholic fatty liver disease, nonalcoholic steatohepatitis, differentially expressed genes, hepatocellular carcinoma, bioinformatics analysis

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Celal Ulaşoğlu,
Okan University, Turkey
Negin Forouzesh,
California State University, Los Angeles, United States
Hamid Ceylan,
Atatürk University, Turkey

***Correspondence:**
Shaohui Tang
tangshaohui206@163.com

†*These authors have contributed equally to this work*

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

Received: *08 September 2021*
Accepted: *18 October 2021*
Published: *29 October 2021*

Citation:
Wu C, Zhou Y, Wang M, Dai G, Liu X, Lai L and Tang S (2021) Bioinformatics Analysis Explores Potential Hub Genes in Nonalcoholic Fatty Liver Disease. Front. Genet. 12:772487. doi: 10.3389/fgene.2021.772487

INTRODUCTION

Nonalcoholic fatty liver disease (NAFLD) is now recognized as the most prevalent chronic liver disease worldwide, with a prevalence ranging from 13% in Africa to 42% in southeast Asia, and it may become the major cause of end-stage liver diseases by 2025 (Zarrinpar et al., 2016; Younossi, 2019; Huang et al., 2021). NAFLD represents a spectrum of disease severity, ranging from simple steatosis (SS) termed as nonalcoholic fatty liver (NAFL) to nonalcoholic steatohepatitis (NASH), cirrhosis, and hepatocellular carcinoma (HCC) (Natarajan et al., 2014). It has been well-recognized that obesity, insulin resistance, and type 2 diabetes mellitus are the strongest risk factors for NAFLD (Chen and Tian, 2020). The cause of NAFLD is multifactorial, including genetic and environmental factors (Chen and Tian, 2020). However, possible effects and underlying mechanisms for NAFLD are still not understood. Meanwhile, NAFLD-related HCC usually lacks symptoms and tends to be

TABLE 1 | The data are shown as median (interquartile range, IQR). HC, healthy control; SS, simple steatosis; NASH, nonalcoholic steatohepatitis; BMI, body mass index; NAS, NAFLD activity score.

Dataset	HC	SS	NASH
GSE48452 (n)	5	9	17
Gender (male: female)	0:5	2:7	4:13
Age (years)	45.0 (35.0–62.0)	37.0 (32.0–46.5)	47 (36–50.5)
BMI (kg/m^2)	21.0 (18.8–23.5)	51.9 (45.7–55.7)	47.8 (33.4–55.7)
Steatosis (%)	0 (0–2.0)	30.0 (15.0–70.0)	75 (70.0–85.0)
NAS	0.5 (0–1.0)	1.0 (1.0–3.0)	5.0 (5.0–5.5)
GSE89632 (n)	8	17	17
Gender (male: female)	4:4	12:5	9:8
Age (years)	42.5 (26.5–54.3)	45.0 (35–51.5)	44.0 (35.5–52.5)
BMI (kg/m^2)	21.2 (19.9–23.1)	28.9 (27.5–31.3)	32.0 (29.65–33.6)
Steatosis (%)	0 (0–0.8)	40.0 (15.0–55.0)	40.0 (17.5–70.0)
NAS	0 (0–0)	2.0 (1.0–2.0)	5.0 (4.0–6.0)

The data are shown as mean and median (interquartile range, IQR). HC, healthy control; SS, simple steatosis; NASH, nonalcoholic steatohepatitis; BMI, body mass index; NAS, NAFLD activity score.

diagnosed at a later stage and is related to poorer survival than viral hepatitis-related HCC (Younossi et al., 2015; Huang et al., 2021). In addition, NAFLD-related HCC is now proliferating and will increase in parallel with the obesity epidemic (Desai et al., 2019). Therefore, it is essential to investigate in detail the mechanism in the pathogenesis of NAFLD to find new potential targets for prognosis and therapy, especially in obese population.

Many genome-wide association studies have indicated that *PNPLA3*, *HNF1A*, *NCAN*, *GCKR*, *MBOTAT*, *FADS1*, *PPAR*, *TNF*, and *TM6SF2* are important genetic and epigenetic modifiers played important roles in the pathogenesis and progression of NAFLD (Choudhary and Duseja, 2021). Meanwhile, some bioinformatics researches offer new ideas for exploring potential targets of NAFLD. Zeng et al. (2020) found that *AKR1B10* and *SPP1* were related to immune cell infiltrations and associated with NAFLD progression. Liu et al. (2020) reported that *TOP2A*, *NHP2L1*, *PCNA*, *CHEK1*, *ACACA*, *CCS*, *ACACB* had a significant impact on NAFLD progression and were associated with HCC progression. What's more, Wang et al. (2016) indicated that *Lp1*, *Ces2*, *Fasn*, *Hmgcs1*, *Sc4mol*, *Fads1*, and *Mup1* were associated with lipid metabolism, and *Cbr3*, *Trib3*, *Nfe212* were related to oxidative stress in NAFLD mouse model. Obviously, there is significant heterogeneity between studies in both animal and human experiments. Although many studies have been devoted to exploring the pathogenesis and progression of NAFLD, there are still no effective drugs for the treatment of NAFLD except for lifestyle changes (Leoni et al., 2018). Thus, combination bioinformatics analysis with public microarray data will contribute to explore novel pathways and genes regulating NAFLD.

Therefore, we analyzed two public datasets to identify differentially expressed genes (DEGs) among healthy controls (HC), SS, and NASH. Then, Weighted Gene Co-Expression Network Analysis (WGCNA) and protein-protein interaction (PPI) networks were performed to explore the impact of DEGs on NAFLD. This study aimed to screen potential genes for NAFLD development.

MATERIAL AND METHODS

Data Retrieving and Processing

The gene expression profiles of GSE48452 (Ahrens et al., 2013) and GSE89632 (Arendt et al., 2015) were downloaded from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo). To prevent the effects of overweight in the evaluation, healthy obesity with body mass index (BMI) over 24 kg/m^2 were excluded from the HC group. Besides, due to NAFLD commonly happened to the obese population, NALFD patients with BMI less than 24 kg/m^2 were also excluded from the experimental group. What is more, individuals with bariatric surgery or severely missing data at baseline were also ruled out. Finally, 9 SS samples, 17 NASH samples, and 5 HC samples in the GSE48452, and 17 SS samples, 17 NASH samples, and 8 HC samples in the GSE89632 were included in this study (**Table 1**). HCC data were obtained from The Cancer Genome Atlas (TCGA) database, including 374 HCC samples and 50 normal samples.

For the analysis of DEGs, we used the GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/) to generate the R script, which used two R packages (GEOquery and limma). The threshold for the DEGs was set as p-value <0.05 and $|\log_2$ fold change (FC)$| \geq 1$. Heat maps were drawn using R package "pheatmap". Venn diagram was performed using the jvenn tool (http://jvenn.toulouse.inra.fr/app/example.html), and the overlaps represented the intersection between the two datasets. Figure 1 illustrated the overall research design.

Diagnostic Methods of Different States of NAFLD

All the samples in GSE48452 and GSE89632 were validated using histological examination by a board-certified pathologist before molecular analysis, and hematoxylin and eosin (H&E) and chromotrope aniline blue (CAB) stained sections were used for histological analysis. The different states of NAFLD were diagnosed using criteria from NAFLD Activity Score (NAS) (Kleiner et al., 2005).

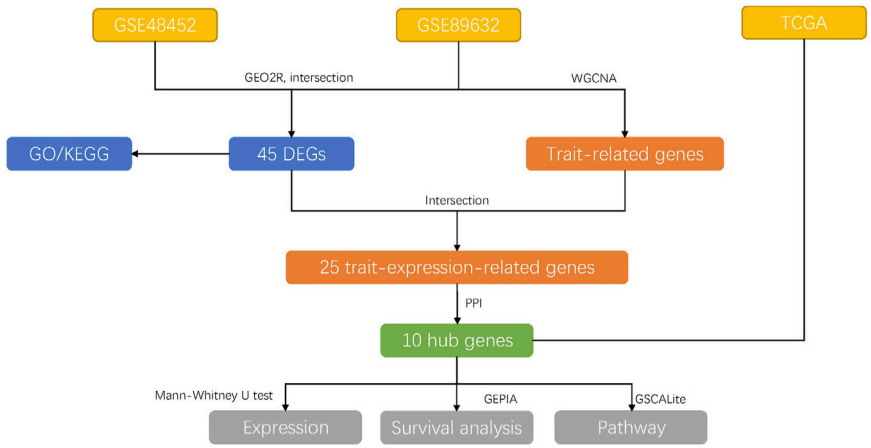

FIGURE 1 | The overall research designs. The data were downloaded from GEO and TCGA databases. Then, the DEGs were explored among the groups using GEO2R and were performed GO and KEGG analysis later, respectively. Subsequently, the GSE89632 dataset was used for WGCNA analysis to explore the trait-related genes, which were intersected with the DEGs to find trait-expression-related genes. PPI network analysis was performed to detect hub genes. Then, the expression, survival rate, and pathway of 10 hub genes were explored in the TCGA database.

Gene Ontology Analysis and Kyoto Encyclopedia of Genes and Genomes Pathway Enrichment Analysis

GO is a commonly used bioinformatics tool that supply comprehensive information on gene function of individual genomic products based on defined features and is primarily divided into three parts, molecular function (MF), biological process (BP), and cellular component (CC). KEGG is a database resource for understanding high-level biological functions and utilities. To identify the function of DEGs, GO and KEGG analysis were performed using Metascape (metascape.org) database with default settings. We determined that results were statistically significant at a level of less than 0.05 using a *p*-value. Then, histograms and bubble plots were generated with R package "ggplot2".

Weighted Gene Co-Expression Network Analysis

WGCNA is a well-established method for studying biological networks and diseases (Rasmussen et al., 2020). Considering that GSE89632 had more comprehensive and complete data, we used GSE89632 to detect modules highly correlated with NAFLD, and WGCNA was performed using R package "WGCNA" and carried out on all genes. The scale-free topology of the networks was assessed for various values of the β shrinkage parameter, and we chose $\beta = 5$ based on scale-free topology criterion. Finally, the dynamic tree cut algorithm was applied to the dendrogram for module identification with the mini-size of module gene numbers set as 50, and similar modules were merged following a height cutoff of 0.05. In the module-trait analysis, gene-trait significance (GS) value >0.3 and module membership (MM) value >0.55 were defined as a threshold (Zeng et al., 2020).

Protein-Protein Interaction Network Construction

Metascape (metascape.org) database was used to construct a protein-protein interaction (PPI) network with default settings. Disconnected nodes in the network were deleted. Then, the Cytoscape software (v3.8.2) was utilized to visualize the PPI network. We used CytoHubba plugin to identify the hub genes through molecular complex detection (MCC) (Chin et al., 2014).

Relationship Between Hub Gens in NAFLD and Hepatocellular Carcinoma Prognosis

The pathway activity was acquired from GSCALite: A Web Server for Gene Set Cancer Analysis (http://bioinfo.life.hust.edu.cn/web/GSCALite/), the survival analysis was collected from Gene Expression Profiling Interactive Analysis (GEPIA, http://gepia.cancer-pku.cn/), and the immunohistochemical pictures were collected from the Human Protein Atlas (HPA, https://www.proteinatlas.org/) database.

Statistical Analysis

Statistical analysis was performed using R software (Version 4.1.0). Statistical comparisons between groups of normalized data were performed using the t-test or Mann-Whitney U-test according to the test condition. A difference with $p < 0.05$ was considered significant.

RESULTS

Identification of DEGs in the NAFLD Patients

The DEGs among HC, SS, and NASH in GSE48452 and GSE89632 datasets were identified, respectively (**Figures 2A,B**

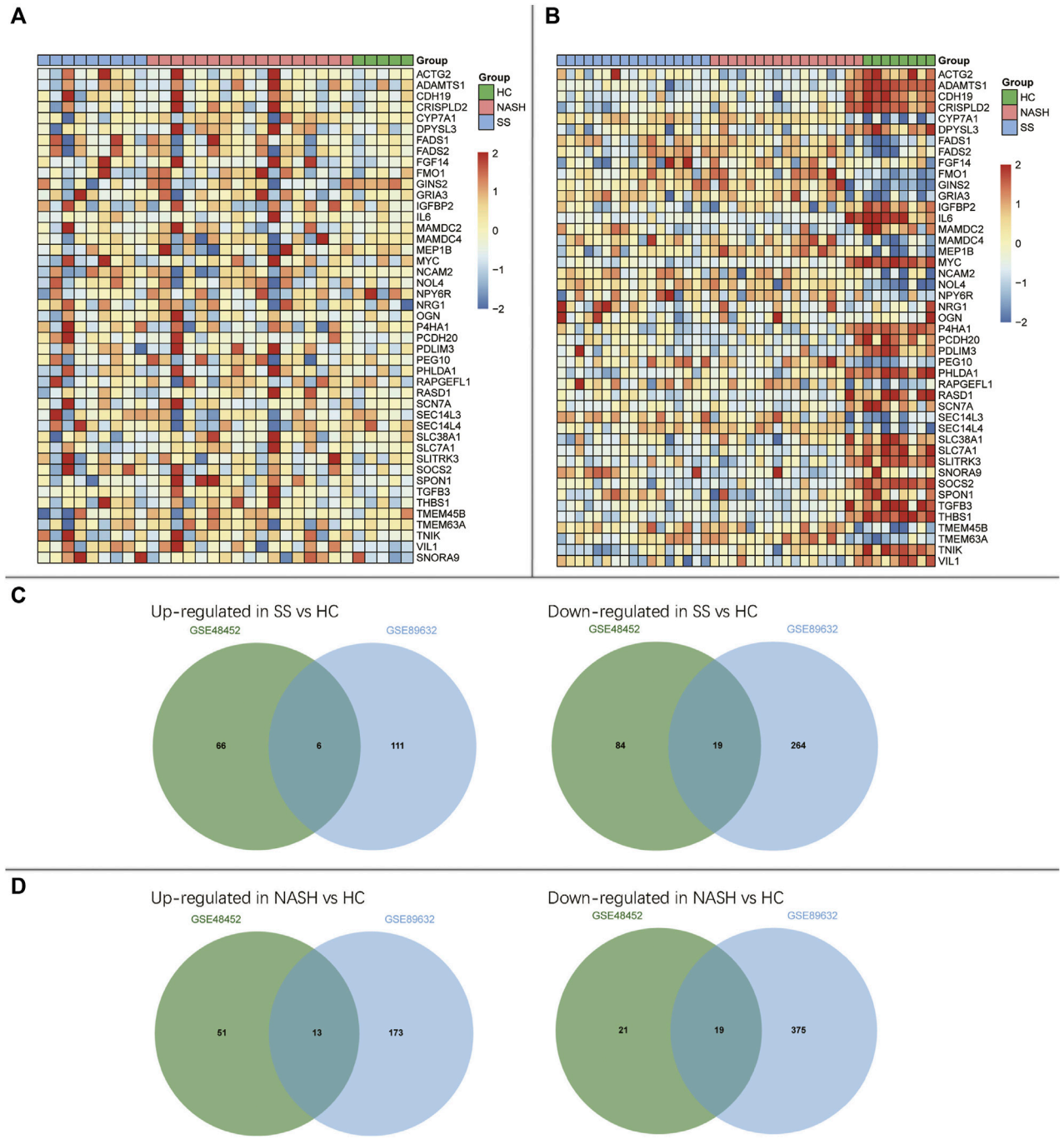

FIGURE 2 | Identification of differentially expressed genes (DEGs) among HC, SS, and NASH. **(A)** Heatmap of overlapping DEGs in GSE48452; **(B)** Heatmap of overlapping DEGs in GSE89632; **(C)** Venn diagrams displayed the overlapping DEGs of up- and down-regulated genes between HC and SS; **(D)** Venn diagrams displayed the overlapping DEGs of up- and down-regulated genes between HC and NASH.

and **Supplementary Tables S1–S2**). Then, we sought for the overlapping DEGs between the two datasets. We observed 6 up- and 19 down-regulated genes in SS compared with HC (**Figure 2C**). We also found 13 up- and 19 down-regulated genes in NASH compared with HC (**Figure 2D**).

GO and KEGG Pathway Enrichment Analysis

To explore the potential roles of DEGs among HC, SS, and NASH, GO and KEGG pathway enrichment analysis were performed. The up-regulated genes between HC and SS were

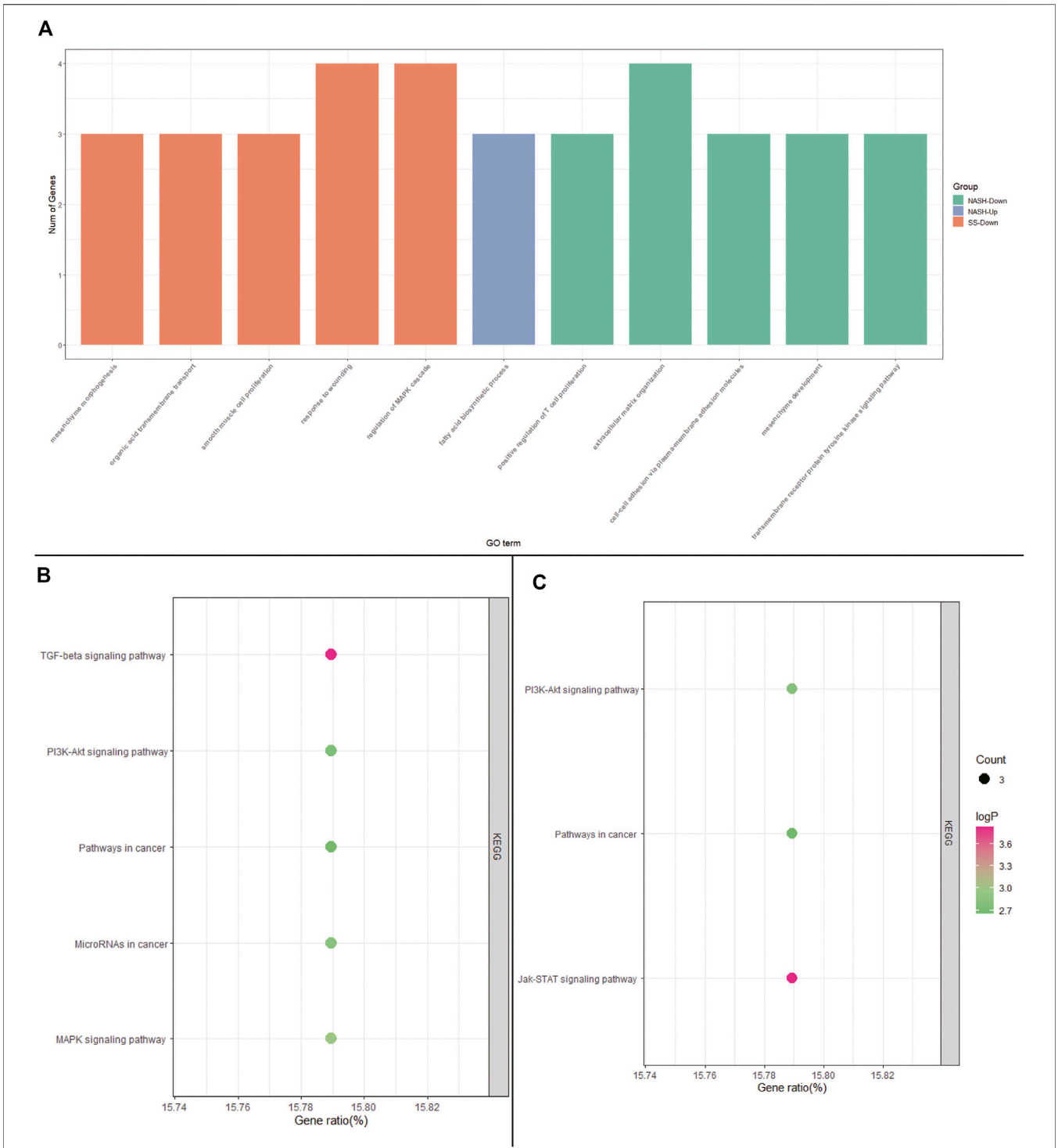

FIGURE 3 | GO and KEGG pathway enrichment analysis. **(A)** GO analysis of DEGs among HC, SS, and NASH; **(B)** KEGG analysis of down-regulated DEGs between HC and SS; **(C)** KEGG analysis of down-regulated DEGs between HC and NASH.

too few to allow identification of GO and KEGG pathway enrichment analysis, and the up-regulated genes between HC and NASH failed to enrich pathway in KEGG.

GO analysis showed that the down-regulated genes between HC and SS were mainly involved in biological processes (BP) associated with the mesenchyme morphogenesis, organic acid

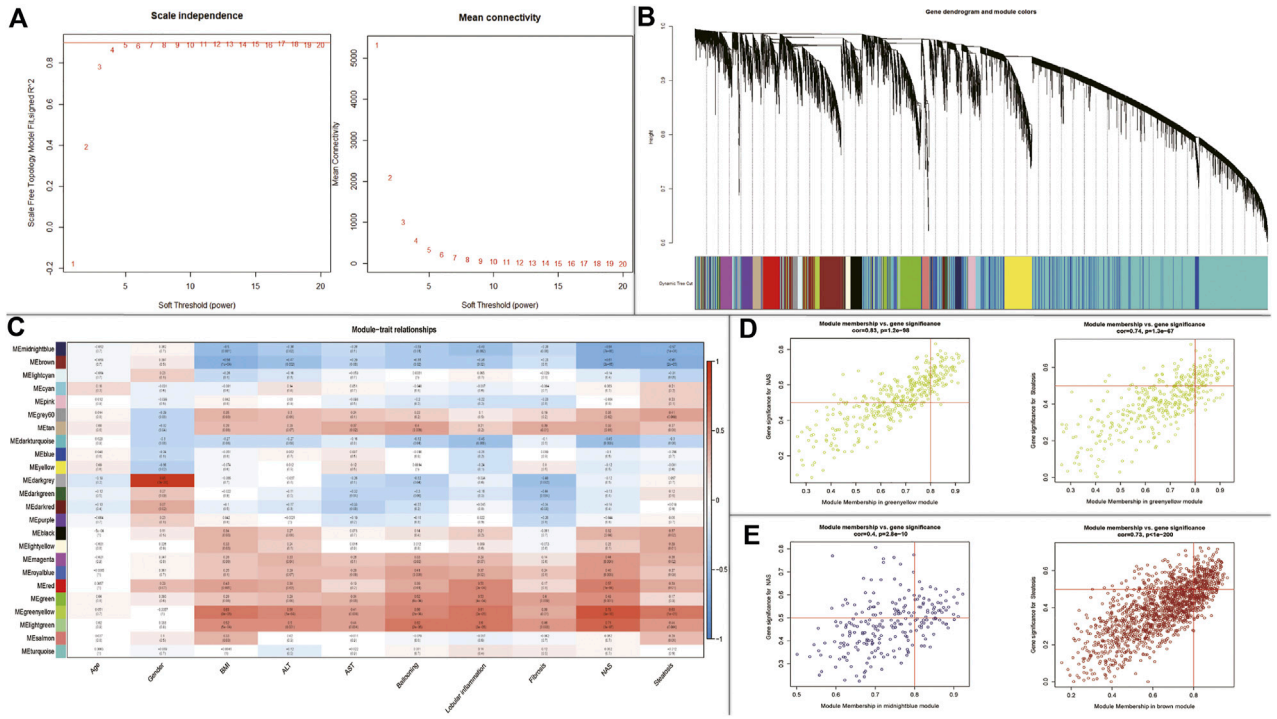

FIGURE 4 | WGCNA to identify trait-related modules and genes. **(A)** Calculating soft-thresholding power; **Left:** scale-free fit indices using different soft-thresholding powers; **Right:** mean connectivity using different soft-thresholding powers; **(B)** The dendrogram clustered by Dynamic Tree Cut algorithm; **(C)** The heatmap profiling the correlations between module eigengenes and the clinical characteristics; **(D)** Scatter plot of gene significance for NAS and steatosis (Up-regulated); **(E)** Scatter plot of gene significance for NAS and steatosis (Down-regulated).

transmembrane transport, smooth muscle cell proliferation, response to wounding, and regulation of MAPK cascade (**Figure 3A** and **Supplementary Table S3**). KEGG analysis indicated that the down-regulated genes between HC and SS primarily enriched in TGF-beta signaling pathway, MAPK signaling pathway, MicroRNAs in cancer, PI3K-Akt signaling pathway, and pathways in cancer (**Figure 3B** and **Supplementary Table S4**).

The DEGs between HC and NASH were mainly involved in biological processes (BP) associated with fatty acid biosynthetic process, positive regulation of T cell proliferation, extracellular matrix organization, cell-cell adhesion via plasma-membrane adhesion molecules, mesenchyme development, and transmembrane receptor protein tyrosine kinase signaling pathway (**Figure 3C** and **Supplementary Table S5**). KEGG analysis indicated that the DEGs between HC and NASH were primarily enriched in Jak-STAT signaling pathway, PI3K-Akt signaling pathway, and pathways in cancer (**Figure 3D** and **Supplementary Table S6**).

Identification of Key Modules by WGCNA

WGCNA was performed to identify key modules related to clinical traits by using GSE89632 dataset. The power of $\beta = 5$ (scale-free $R^2 = 0.89$) was selected as the soft thresholding parameter to construct a scale-free network (**Figure 4A**). A total of 24 modules were identified (**Figure 4B**). Similar module clustering was constructed by using dynamic hybrid cutting (threshold = 0.05). The results in **Figure 4C** showed that the greenyellow module was the highest positive module correlated to NAFLD activity score (NAS, $R^2 = 0.79$, $p = 9e^{-10}$) and steatosis ($R^2 = 0.63$, $p = 1e^{-5}$). In addition, the midnightblue module was highly negative correlated to NAS ($R^2 = 0.64$, $p = 7e^{-6}$), and the brown module was highly negative correlated to steatosis ($R^2 = 0.61$, $p = 2e^{-5}$). **Figures 4D,E** showed the positive and negative modules.

In the module-trait analysis, we intersected the trait-related genes highly associated with NAS and steatosis and 45 DEGs generated from expression difference analysis, and finally extracted 25 trait-expression-related genes for the following analysis (**Supplementary Table S7–S8**).

Identification of Hub Genes and Construction of Protein-Protein Interaction Network

Subsequently, we construct a PPI network with 25 trait-expression-related genes in the Metascape database. Then, 15 filtered genes were identified (**Figure 5A**) and later imported into CytoHubba plugin to explore the hub genes by "MCC" methods. The results showed that *MYC*, *TGFB3*, *ADAMTS1*, *THBS1*, *RASD1*, *PCDH20* (Down-regulated genes), *MAMDC4*, *CYP7A1*, *GINS2*, and *PDLIM3* (Up-regulated genes) were the top 10 hub genes (**Figure 5B**).

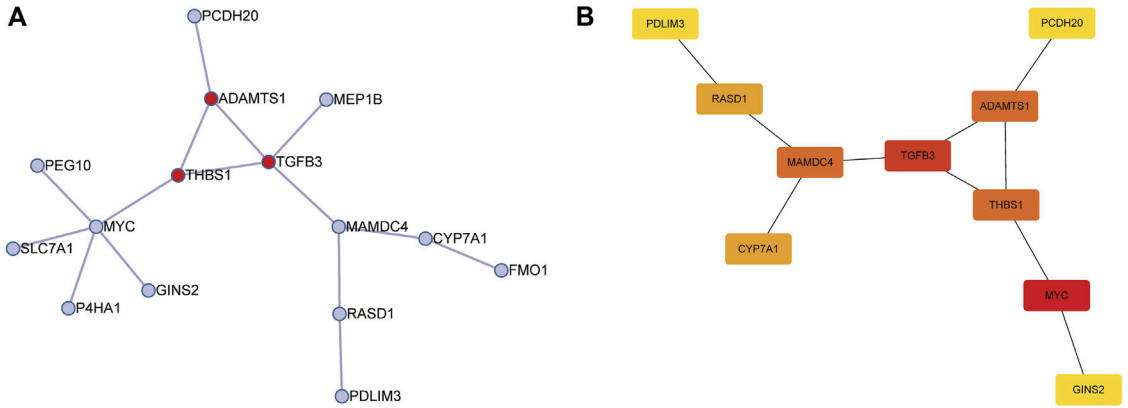

FIGURE 5 | Construction of protein-protein interaction (PPI) networks for 25 trait-expression-related genes in the Metascape database. **(A)** PPI network of 15 trait-expression-related genes. Red node represented the Molecular Complex Detection (MCODE) algorithm applied to identify densely connected network components. **(B)** Results of the CytoHubba plugin; the color changed from yellow to red was indicative of the rank of protein, and the deeper the red staining, the higher rank of protein was.

Hub Genes in NAFLD Were Associated With Hepatocellular Carcinoma Prognosis

Afterwards, the possible relationship between hub genes and hepatocellular carcinoma (HCC) was explored. We found that *CYP7A1*, *GINS2*, and *PDLIM3* were significantly up-regulated, and *MYC*, *MAMDC4*, *ADAMTS1*, *THBS1*, and *RASD1* were significantly down-regulated in HCC tumor samples compared with normal samples using the TCGA dataset (**Figure 6A**). Moreover, we found that the 8 genes above were enriched in tumor-related pathways, such as apoptosis, cell cycle, and epithelial-mesenchymal transition (EMT) (**Figure 6B**). Subsequently, we performed survival analysis in the genes above. As demonstrated in **Figure 6C**, *CYP7A1*-high (using quartile cutoff points) patients showed higher overall survival (OS) rates compared to *CYP7A1*-low patients but had no effects on disease free survival rate (DFS). What is more, compared to *GINS2*- high (using quartile cutoff points) and *PDLIM3*-high (using median cutoff points) patients, the OS rates were higher in low expression patients. In addition, *GINS2*-low patients showed a higher DFS rate compared to *GINS2*-high patients (**Figures 6D,E**). In the HPA database, the expression of *CYP7A1*/*GINS2*/*PDLIM3* was also abnormally elevated in HCC, but the immunohistochemical picture of *CYP7A1* was missing. (**Figure 6F**).

DISCUSSION

Currently, the pathogenesis of NAFLD is still unclear, and the therapeutic treatments are also limited. In the present study, we identified 45 intersected DEGs between HC-SS group and HC-NASH group, and respectively performed GO and KEGG pathway enrichment analysis to explore the potential effects of these DEGs in NAFLD. The results showed that the GO enrichments were involved in fatty acid metabolism, mesenchyme, extracellular matrix, cell adhesion, and inflammatory and immune response, which also played important roles in tumorigenesis. KEGG analysis showed that the DEGs were primarily enriched in TGF-beta signaling pathway, PI3K-Akt signaling pathway, pathways in cancer, MicroRNAs in cancer, MAPK signaling pathway, and Jak-STAT signaling pathway. Both the results of GO and KEGG analysis all pointed to tumorigenesis. Meanwhile, the overlapping pathways between SS and NASH were PI3K-Akt signaling pathway and pathways in cancer, suggesting that the two pathways could be an important therapeutic target for NAFLD. The PI3K-AKT signaling pathway is known for regulating metabolism, cell growth, and cell survival. The active form of PI3K is an oncogene; thus, amplification and mutations of PI3K are usually found in many kinds of cancers (Matsuda et al., 2013). However, in this study, the PI3K-AKT signaling pathway was down-regulated in NAFLD patients. Previous studies had shown that the inhibition of PI3K-AKT signaling pathway increased hepatic insulin resistance, which exacerbated the accumulation of fat in the liver (Ntandja et al., 2020); what's more, a restoration of PI3K-AKT pathway improved the liver injury and fat accumulation (Li et al., 2013). Long-duration effects of lipotoxicity aggravated the inflammatory reaction in the liver, leading to dysregulation of the PI3K-AKT signaling pathway, which might finally result in HCC (Asgharpour et al., 2016). Our findings were also consistent with previous reports (Wang et al., 2019; Liu et al., 2020).

Due to NAS and steatosis were the two main pathologic indicators in the estimation of NAFLD, we tried to find out the DEGs related to the NAS and steatosis. We identified 25 DEGs related to the NAS and steatosis, and PPI network analysis was performed to explore the hub genes in the pathogenesis and progression of NAFLD. Eventually, we determined 10 hub genes (Down-regulated genes: *MYC*, *TGFB3*, *ADAMTS1*, *THBS1*, *RASD1*, *PCDH20*; Up-regulated genes: *MAMDC4*, *CYP7A1*, *GINS2*, and *PDLIM3*) related to NAS and steatosis.

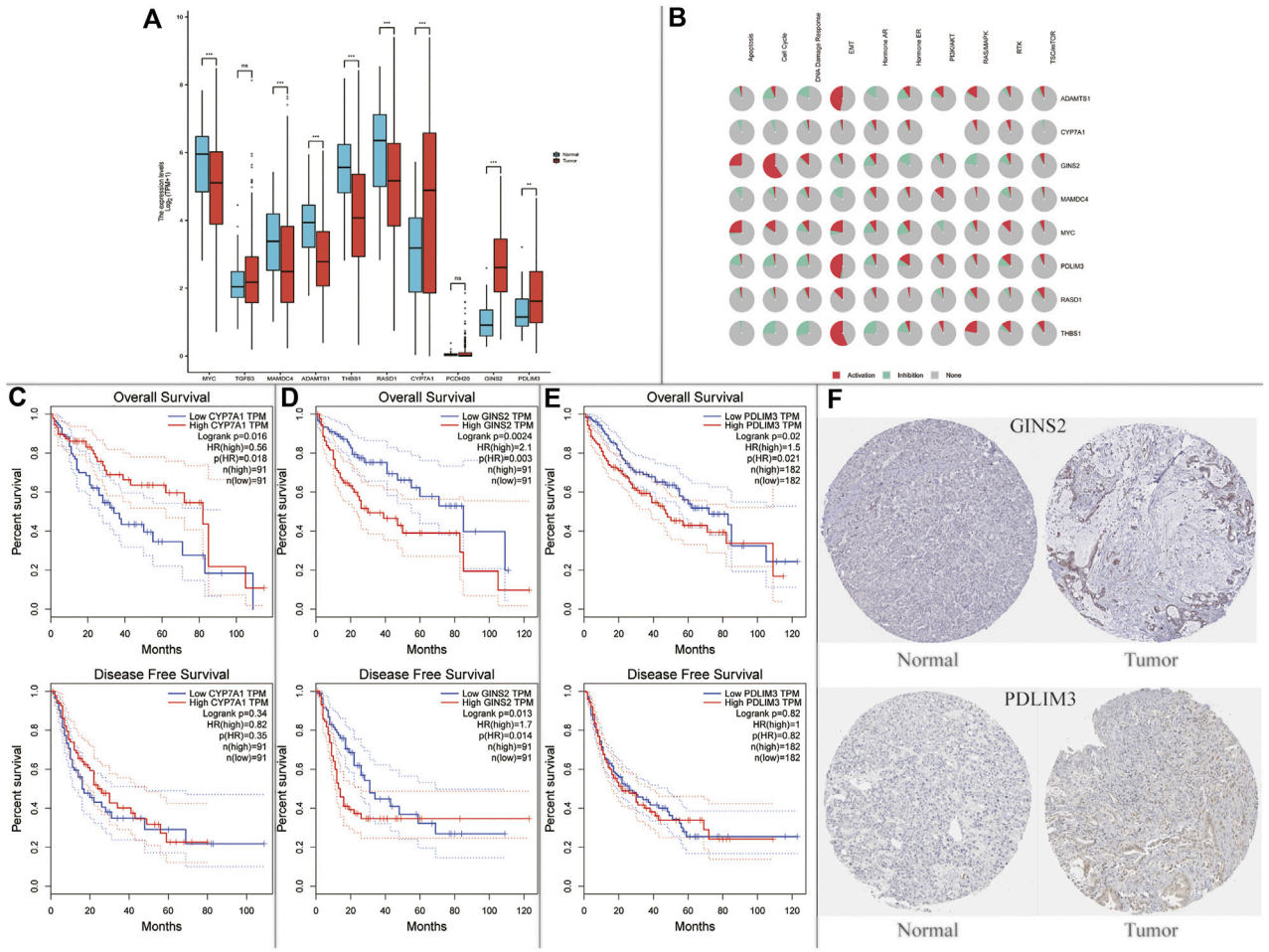

FIGURE 6 | Expression and survival analysis of the NAFLD's hub genes in hepatocellular carcinoma (HCC). **(A)** Hub genes in NAFLD were dysregulated in hepatocellular carcinoma (ns, not significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$); **(B)** Enriched pathways of 10 hub genes in the TCGA database. **(C–E)** Survival plots of *CYP7A1*, *GINS2*, and *PDLIM3*; **(F)** Protein expression of *GINS2* and *PDLIM3* between normal patients and HCC patients in the HPA database.

HCC is the fourth-leading cause of cancer death worldwide, and the morbidity of NAFLD-related HCC is predicted to increase dramatically by 2030, with increases of 82, 117, and 122% from 2016 in China, France, and the USA, respectively (Yang et al., 2019; Huang et al., 2021). Therefore, we explore whether these ten hub genes were associated with the progression in HCC in the TCGA database. We found that *CYP7A1*, *GINS2*, and *PDLIM3* were significantly up-regulated, and *MYC*, *MAMDC4*, *ADAMTS1*, *THBS1*, and *RASD1* were significantly down-regulated in HCC tumor samples compared to normal samples. Surprisingly, we also found that *CYP7A1*/*GINS2*/*PDLIM3* were correlated with HCC prognosis.

CYP7A1, catalyzing the first and rate-limiting step in the classic bile acid synthesis pathway, has been shown to be involved in lipid metabolism (Wang et al., 2020). Deficiency of *CYP7A1* caused by homozygous deletion mutations can inhibit the production of bile acids, leading to the accumulation of cholesterol in the liver, reducing LDL receptors and elevating LDL cholesterol (Pullinger et al., 2002). However, *CYP7A1* was up-regulated in SS and NASH group compared with HC group in our study. Previous studies have shown that *CYP7A1* and its associated cholesterol processes were adversely regulated in NAFLD (Wruck and Adjaye, 2017), and glucose stimulates *CYP7A1* transcription in human hepatocytes (Chiang and Ferrell, 2020). Therefore, up-regulating CYP7A1 in NAFLD may be the consequence rather than the cause of disease (Jia and Zhai, 2019). In addition, increased *CYP7A1* expression and bile acid synthesis ameliorated hepatic inflammation and fibrosis, proving its anti-tumor effects (Liu et al., 2016).

GINS2, a member of the *GINS* family, plays a crucial role in DNA duplication and is highly expressed in various types of cancer (Kubota et al., 2003; Tian et al., 2020). However, very little research can be found about *GINS2* in the liver, especially in NAFLD. Previous bioinformatics studies indicated that *GINS2* might be the hub genes in the development of NASH to HCC and predicted poor prognosis in HCC, but there was no further experiment to verify its effects on NAFLD (Lian et al., 2018; Zhang et al., 2020).

PDLIM3, highly expressed in skeletal and cardiac muscle, has been suggested to play a pivotal role in myocyte stability, signal transduction, and mechanical signaling, especially in growth and remodeling processes (Zheng et al., 2010). Interestingly, *PDLIM3* was firstly screened out for a new hub gene in the pathogenesis of NAFLD and was associated with the prognosis of HCC. *PDLIM3* was highly related to EMT in the GSCALite database, which might partially reveal its effects in the pathogenesis in NAFLD and HCC. More future studies are needed to gain more insights about *PDLIM3*.

In the present study, more attention was applied to the pathogenesis of NAFLD in obesity, which was rare in other studies. However, the present study had several limitations. Firstly, further experiments were required to verify these results. Secondly, it was hard to identify HCC patients caused by NAFLD in the TCGA database, which might impact the outcomes.

In conclusion, we analyzed two public datasets to identify DEGs among HC, SS and NASH. GO and KEGG pathway analysis revealed that the pathogenesis and progression of NAFLD were highly associated with tumorigenesis. Finally, we screened out 10 hub genes related to NAS and steatosis, and three of them were correlated with HCC prognosis. Although further validation is still needed, we provide useful and novel information to explore the potential candidate genes for NAFLD prognosis and therapeutic options.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

CW, YZ, and MW contributed equally to this paper. CW, YZ, and MW analyzed the study data, helped draft the manuscript, made critical revisions of the manuscript. GD, XL and LL assisted with data collection and the analysis. ST supervised the research and edited the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors appreciate study investigators and staff who participated in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.772487/full#supplementary-material

REFERENCES

Ahrens, M., Ammerpohl, O., von Schönfels, W., Kolarova, J., Bens, S., Itzel, T., et al. (2013). DNA Methylation Analysis in Nonalcoholic Fatty Liver Disease Suggests Distinct Disease-specific and Remodeling Signatures after Bariatric Surgery. *Cel Metab.* 18 (2), 296–302. doi:10.1016/j.cmet.2013.07.004

Arendt, B. M., Comelli, E. M., Ma, D. W. L., Lou, W., Teterina, A., Kim, T., et al. (2015). Altered Hepatic Gene Expression in Nonalcoholic Fatty Liver Disease Is Associated with Lower Hepatic N-3 and N-6 Polyunsaturated Fatty Acids. *Hepatology* 61 (5), 1565–1578. doi:10.1002/hep.27695

Asgharpour, A., Cazanave, S. C., Pacana, T., Seneshaw, M., Vincent, R., Banini, B. A., et al. (2016). A Diet-Induced Animal Model of Non-alcoholic Fatty Liver Disease and Hepatocellular Cancer. *J. Hepatol.* 65 (3), 579–588. doi:10.1016/j.jhep.2016.05.005

Chen, Y., and Tian, Z. (2020). Roles of Hepatic Innate and Innate-like Lymphocytes in Nonalcoholic Steatohepatitis. *Front. Immunol.* 11, 1500. doi:10.3389/fimmu.2020.01500

Chiang, J. Y. L., and Ferrell, J. M. (2020). Up to Date on Cholesterol 7 Alpha-Hydroxylase (CYP7A1) in Bile Acid Synthesis. *Liver Res.* 4 (2), 47–63. doi:10.1016/j.livres.2020.05.001

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., and Lin, C.-Y. (2014). cytoHubba: Identifying Hub Objects and Sub-networks from Complex Interactome. *BMC Syst. Biol.* 8 (Suppl. 4), S11. doi:10.1186/1752-0509-8-S4-S11

Choudhary, N. S., and Duseja, A. (2021). Genetic and Epigenetic Disease Modifiers: Non-alcoholic Fatty Liver Disease (NAFLD) and Alcoholic Liver Disease (ALD). *Transl Gastroenterol. Hepatol.* 6, 2. doi:10.21037/tgh.2019.09.06

Desai, A., Sandhu, S., Lai, J.-P., and Sandhu, D. S. (2019). Hepatocellular Carcinoma in Non-cirrhotic Liver: A Comprehensive Review. *Wjh* 11 (1), 1–18. doi:10.4254/wjh.v11.i1.1

Huang, D. Q., El-Serag, H. B., and Loomba, R. (2021). Global Epidemiology of NAFLD-Related HCC: Trends, Predictions, Risk Factors and Prevention. *Nat. Rev. Gastroenterol. Hepatol.* 18 (4), 223–238. doi:10.1038/s41575-020-00381-6

Jia, X., and Zhai, T. (2019). Integrated Analysis of Multiple Microarray Studies to Identify Novel Gene Signatures in Non-alcoholic Fatty Liver Disease. *Front. Endocrinol.* 10, 599. doi:10.3389/fendo.2019.00599

Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., et al. (2005). Design and Validation of a Histological Scoring System for Nonalcoholic Fatty Liver Disease. *Hepatology* 41 (6), 1313–1321. doi:10.1002/hep.20701

Kubota, Y., Takase, Y., Komori, Y., Hashimoto, Y., Arata, T., Kamimura, Y., et al. (2003). A Novel Ring-like Complex of Xenopus Proteins Essential for the Initiation of DNA Replication. *Genes Dev.* 17 (9), 1141–1152. doi:10.1101/gad.1070003

Leoni, S., Tovoli, F., Napoli, L., Serio, I., Ferri, S., and Bolondi, L. (2018). Current Guidelines for the Management of Non-alcoholic Fatty Liver Disease: A Systematic Review with Comparative Analysis. *Wjg* 24 (30), 3361–3373. doi:10.3748/wjg.v24.i30.3361

Li, Y., Hai, J., Li, L., Chen, X., Peng, H., Cao, M., et al. (2013). Administration of Ghrelin Improves Inflammation, Oxidative Stress, and Apoptosis during and after Non-alcoholic Fatty Liver Disease Development. *Endocrine* 43 (2), 376–386. doi:10.1007/s12020-012-9761-5

Lian, Y.-F., Li, S.-S., Huang, Y.-L., Wei, H., Chen, D.-M., Wang, J.-L., et al. (2018). Up-regulated and Interrelated Expressions of GINS Subunits Predict Poor Prognosis in Hepatocellular Carcinoma. *Biosci. Rep.* 38 (6), BSR20181178. doi:10.1042/BSR20181178

Liu, H., Pathak, P., Boehme, S., and Chiang, J. L. (2016). Cholesterol 7α-Hydroxylase Protects the Liver from Inflammation and Fibrosis by Maintaining Cholesterol Homeostasis. *J. Lipid Res.* 57 (10), 1831–1844. doi:10.1194/jlr.M069807

Liu, J., Lin, B., Chen, Z., Deng, M., Wang, Y., Wang, J., et al. (2020). Identification of Key Pathways and Genes in Nonalcoholic Fatty Liver Disease Using Bioinformatics Analysis. *aoms.* 16 (2), 374–385. doi:10.5114/aoms.2020.93343

Matsuda, S., Kobayashi, M., and Kitagishi, Y. (2013). Roles for PI3K/AKT/PTEN Pathway in Cell Signaling of Nonalcoholic Fatty Liver Disease. *ISRN Endocrinol.* 2013, 1–7. doi:10.1155/2013/472432

Natarajan, S. K., Ingham, S. A., Mohr, A. M., Wehrkamp, C. J., Ray, A., Roy, S., et al. (2014). Saturated Free Fatty Acids Induce Cholangiocyte Lipoapoptosis. *Hepatology* 60 (6), 1942–1956. doi:10.1002/hep.27175

Ntandja Wandji, L. C., Gnemmi, V., Mathurin, P., and Louvet, A. (2020). Combined Alcoholic and Non-alcoholic Steatohepatitis. *JHEP Rep.* 2 (3), 100101. doi:10.1016/j.jhepr.2020.100101

Pullinger, C. R., Eng, C., Salen, G., Shefer, S., Batta, A. K., Erickson, S. K., et al. (2002). Human Cholesterol 7α-Hydroxylase (CYP7A1) Deficiency Has a Hypercholesterolemic Phenotype. *J. Clin. Invest.* 110 (1), 109–117. doi:10.1172/JCI1538710.1172/jci0215387

Rasmussen, A. H., Kogelman, L. J. A., Kristensen, D. M., Chalmer, M. A., Olesen, J., and Hansen, T. F. (2020). Functional Gene Networks Reveal Distinct Mechanisms Segregating in Migraine Families. *Brain* 143 (10), 2945–2956. doi:10.1093/brain/awaa242

Tian, W., Yang, X., Yang, H., and Zhou, B. (2020). GINS2 Functions as a Key Gene in Lung Adenocarcinoma by WGCNA Co-expression Network Analysis. *Ott* Vol. 13, 6735–6746. doi:10.2147/OTT.S255251

Wruck, W., and Adjaye, J. (2017). Meta-analysis Reveals Up-Regulation of Cholesterol Processes in Non-alcoholic and Down-Regulation in Alcoholic Fatty Liver Disease. *Wjh* 9 (8), 443–454. doi:10.4254/wjh.v9.i8.443

Wang, C., Tao, Q., Wang, X., Wang, X., and Zhang, X. (2016). Impact of High-Fat Diet on Liver Genes Expression Profiles in Mice Model of Nonalcoholic Fatty Liver Disease. *Environ. Toxicol. Pharmacol.* 45, 52–62. doi:10.1016/j.etap.2016.05.014

Wang, H., Liu, Y., Wang, D., Xu, Y., Dong, R., Yang, Y., et al. (2019). The Upstream Pathway of mTOR-Mediated Autophagy in Liver Diseases. *Cells* 8 (12), 1597. doi:10.3390/cells8121597

Wang, Y., Gunewardena, S., Li, F., Matye, D. J., Chen, C., Chao, X., et al. (2020). An FGF15/19-TFEB Regulatory Loop Controls Hepatic Cholesterol and Bile Acid Homeostasis. *Nat. Commun.* 11 (1), 3612. doi:10.1038/s41467-020-17363-6

Yang, J. D., Hainaut, P., Gores, G. J., Amadou, A., Plymoth, A., and Roberts, L. R. (2019). A Global View of Hepatocellular Carcinoma: Trends, Risk, Prevention and Management. *Nat. Rev. Gastroenterol. Hepatol.* 16 (10), 589–604. doi:10.1038/s41575-019-0186-y

Younossi, Z. M. (2019). Non-alcoholic Fatty Liver Disease - A Global Public Health Perspective. *J. Hepatol.* 70 (3), 531–544. doi:10.1016/j.jhep.2018.10.033

Younossi, Z. M., Otgonsuren, M., Henry, L., Venkatesan, C., Mishra, A., Erario, M., et al. (2015). Association of Nonalcoholic Fatty Liver Disease (NAFLD) with Hepatocellular Carcinoma (HCC) in the United States from 2004 to 2009. *Hepatology* 62 (6), 1723–1730. doi:10.1002/hep.28123

Zarrinpar, A., Gupta, S., Maurya, M. R., Subramaniam, S., and Loomba, R. (2016). Serum microRNAs Explain Discordance of Non-alcoholic Fatty Liver Disease in Monozygotic and Dizygotic Twins: a Prospective Study. *Gut* 65 (9), 1546–1554. doi:10.1136/gutjnl-2015-309456

Zeng, F., Zhang, Y., Han, X., Zeng, M., Gao, Y., and Weng, J. (2020). Predicting Non-alcoholic Fatty Liver Disease Progression and Immune Deregulations by Specific Gene Expression Patterns. *Front. Immunol.* 11, 609900. doi:10.3389/fimmu.2020.609900

Zhang, D., Liu, J., Xie, T., Jiang, Q., Ding, L., Zhu, J., et al. (2020). Oleate Acid-Stimulated HMMR Expression by CEBPα Is Associated with Nonalcoholic Steatohepatitis and Hepatocellular Carcinoma. *Int. J. Biol. Sci.* 16 (15), 2812–2827. doi:10.7150/ijbs.49785

Zheng, M., Cheng, H., Banerjee, I., and Chen, J. (2010). ALP/Enigma PDZ-LIM Domain Proteins in the Heart. *J. Mol. Cel Biol.* 2 (2), 96–102. doi:10.1093/jmcb/mjp038

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

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

Copyright © 2021 Wu, Zhou, Wang, Dai, Liu, Lai and Tang. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fgene.2021.758725

Prognostic Values and Clinical Significance of S100 Family Member's Individualized mRNA Expression in Pancreatic Adenocarcinoma

Xiaomin Li[1], *Ning Qiu*[2]* *and Qijuan Li*[3]

[1]Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China, [2]Key Laboratory of Ocean and Marginal Sea Geology, Guangdong Southern Marine Science & Engineering Laboratory (Guangzhou), South China Sea Institute of Oceanology, Innovation Academy of South China Sea Ecology and Environmental Engineering, Chinese Academy of Sciences, Guangzhou, China, [3]Department of Clinical Laboratory, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, China

Objective: Pancreatic adenocarcinoma (PAAD) is a common malignant tumor worldwide. S100 family (S100s) is wildly involved in regulating the occurrence, development, invasion, metastasis, apoptosis, and drug resistance of many malignant tumors. However, the expression pattern, prognostic value, and oncological role of individual S100s members in PAAD need to be elucidated.

Methods: The transcriptional expression levels of S100s were analyzed through the Oncomine and GEPIA, respectively. The protein levels of S100s members in PAAD were studied by Human Protein Atlas. The correlation between S100 mRNA expression and overall survival and tumor stage in PAAD patients was studied by GEPIA. The transcriptional expression correlation and gene mutation rate of S100s members in PAAD patients were explored by cBioPortal. The co-expression networks of S100s are identified using STRING and Gene MANIA to predict their potential functions. The correlation of S100s expression and tumor-infiltrating immune cells was tested by TIMER. Pathway activity and drug target analyzed by GSCALite.

Results: 13 S100s members were upregulated in PAAD tissues. 15 S100s members were associated with TP53 mutation. Expression levels of S100A3/A5/A6/A10/A11/A14/A16/B/P/Z were significantly correlated with the pathological stage. Prognosis analysis demonstrated that PAAD patients with low mRNA levels of S100A1/B/Z or high levels of S100A2/A3/A5/A10/A11/A14/A16 had a poor prognosis. Immuno-infiltration analysis showed that the mRNA levels of S100A10/A11/A14/A16 were correlated with the infiltration degree of macrophages in PAAD. Drug sensitivity analysis showed that

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Amina Laham,
McGill University, Canada
Alireza Khodadadi-Jamayran,
NYU Grossman School of Medicine, United States

***Correspondence:**
Ning Qiu
ningqiu@scsio.ac.cn

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

Received: *14 August 2021*
Accepted: *14 October 2021*
Published: *03 November 2021*

Citation:
Li X, Qiu N and Li Q (2021) Prognostic Values and Clinical Significance of S100 Family Member's Individualized mRNA Expression in Pancreatic Adenocarcinoma. Front. Genet. 12:758725. doi: 10.3389/fgene.2021.758725

Abbreviations: AKT, protein kinase B; AR, androgen receptor; CPTAC, clinical proteomic tumor analysis consortium; CTRP, clinical trials reporting program; DEG, differentially expressed gene; EMT, epithelial-mesenchymal transition; ER, estrogen receptor; GDSC, genomics of drug sensitivity in cancer; GEPIA, gene expression profiling interactive; GO, gene ontology; GTEx, genotype-Tissue expression; HPA, the human protein atlas; IHC, Immunohistochemistry; KEGG, kyoto encyclopedia of genes and genomes; MAPK, mitogen-activated protein kinase; mTOR, mechanistic target of rapamycin; OS, overall survival; PAAD, pancreatic adenocarcinoma; PCA, principal component analysis; PI3K, phosphoinositide 3-kinase; RTK, receptor tyrosine kinase; S100s, S100 family; TCGA, the cancer genome atlas; TPM, transcripts per million; TSC, Tuberous sclerosis complex.

PAAD expressing high levels of S100A2/A6/A10/A11/A13/A14/A16 maybe resistant to small molecule drugs.

Conclusion: This study identifies the clinical significance and biological functions of the S100s in PAAD, which may provide novel insights for the selection of prognostic biomarkers.

Keywords: mRNA expression, S100 family, pancreatic cancer, biomarker, prognosis

INTRODUCTION

Pancreatic adenocarcinoma (PAAD) is a highly lethal disease and has become the seventh leading cause of cancer-related deaths, accounting for about 4.7% of global mortality (Sung et al., 2021). PAAD is associated with a very poor prognosis, with a 5-years survival rate of as low as 8% after diagnosis (Mishra et al., 2019). The low survival rate is attributed to various factors, which is mainly caused by the high rate of advanced PAAD since over 50% of PAAD patients are diagnosed at an advanced stage (Ilic and Ilic, 2016). Moreover, PAAD is characterized not only by early recurrence and invasion but also the resistance to chemotherapy and radiotherapy (Adamska et al., 2018). Although great strides have been made on the screening, diagnosis and comprehensive therapy combining surgery, chemotherapy, and radiotherapy, PAAD remains a highly malignant tumor with limited treatment options (Kamisawa et al., 2016). Conventional treatments, such as surgery, have poor clinical outcomes, and only about 20% of patients benefiting from radical surgery (Lai et al., 2019). At present, there are no reported cases of good efficacy of targeted therapy for pancreatic cancer driver genes (Zhang et al., 2021). The prognosis of PAAD is largely determined by early diagnosis and treatment. Therefore, seeking for key genes and proteins related to the occurrence, development, and metastasis of PAAD is of great significance in improving the prognosis.

Invasion and metastasis are typical events during the malignant progression of tumors, alongside tumor cell proliferation, shedding, dissemination, angiogenesis, implantation, and other aspects (Marchesi et al., 2010). Various proteins found to be implicated in tumorigenesis and tumor development. Calcium-binding proteins are a large family, which are responsible for mediating the cell cycle progression, cell differentiation, enzyme activation, muscle contraction, etc. (Donato, 1999). The S100 family (S100s) is one of the largest subfamilies of calcium-binding proteins, which plays a key role in cell proliferation, apoptosis, differentiation, and inflammation. So far, at least 20 members (S100A1-A14/A7A/A16/B/P/G/Z, etc) of the S100s have been reported (Allgöwer et al., 2020a). However, these members have some commonalities and differences in their respective organizational structures, which may make them play different roles in the occurrence and development of tumors. The chromosomal regions encoding S100s genes have poor stability and they are closely associated with the occurrence and development of tumors. Once tumorigenesis initiates, the gene in this region is easy to recombine and interfere with the S100s gene (Engelkamp et al., 1993; Schäfer et al., 1995; Marenholz et al., 2004; Marenholz et al., 2006; Goh et al., 2017). Thus, S100s are usually dysregulated in human malignant tumors, including PAAD (Bresnick et al., 2015; Xue et al., 2015). S100s have distinct expression and function patterns in tumorigenesis and tumor development. For example, S100A4/A7/A8/A9/A13 was found to have tumorigenesis role, however, S100A6/A8/A9 found to have anti-tumor activity (Salama et al., 2008). S100s members can also participate in the regulation of various biological functions associated with the progress of PAAD. The expression of S100P/A4 have been proven to be related to the differentiation, metastasis, prognosis, and drug resistance of pancreatic cancer. S100A2/A10 have recently been suggested as negative prognostic biomarkers for pancreatic cancer (Bachet et al., 2013; Bydoun et al., 2018a). S100P/A11 are unfavorable to the prognosis of PAAD patients undergoing surgical resection (Xiao et al., 2012; Camara et al., 2020). Collectively, the clinical significance of S100s members and their potential application in the development of PAAD have been highlighted, although their predictive potential and biological characteristics need to be further validated. Moreover, the relationship between S100s and immune cell infiltration and drug resistance of PAAD remains unclear (Chen et al., 2021).

In our study, with the help of GEPIA, we systematically evaluated the transcriptional expressions of the S100s and their relationship with tumor stage and prognostic signature in PAAD. Based on data mining and analyses, we also clarified the gene mutation, potential biological roles, and drug resistance of S100s members in PAAD. Besides, we also assessed the correlation between S100s mRNA expression and immune cell infiltration in PAAD using the tool TIMER.

MATERIALS AND METHODS

Oncomine

Oncomine gene expression array database (Adamska et al., 2018; Ma et al., 2019) was used to analyze the mRNA level of S100s members in various cancers. The student's t-test was used to generate a *p*-value for the comparison between clinical cancer specimens and pair control tissues. The thresholds for each S100s member were set at fold change:2; *p*-value: 0.001; gene rank: 10%.

Gene expression profile Interactive Analysis

Gene expression profile Interactive Analysis (GEPIA) is a bioinformatics analysis tool for evaluating RNA expression, which contains 9,736 tumors and 8,587 normal samples from the Cancer Genome Atlas and Genotype-tissue Expression dataset (Tang et al., 2017). The database delivers rapid and

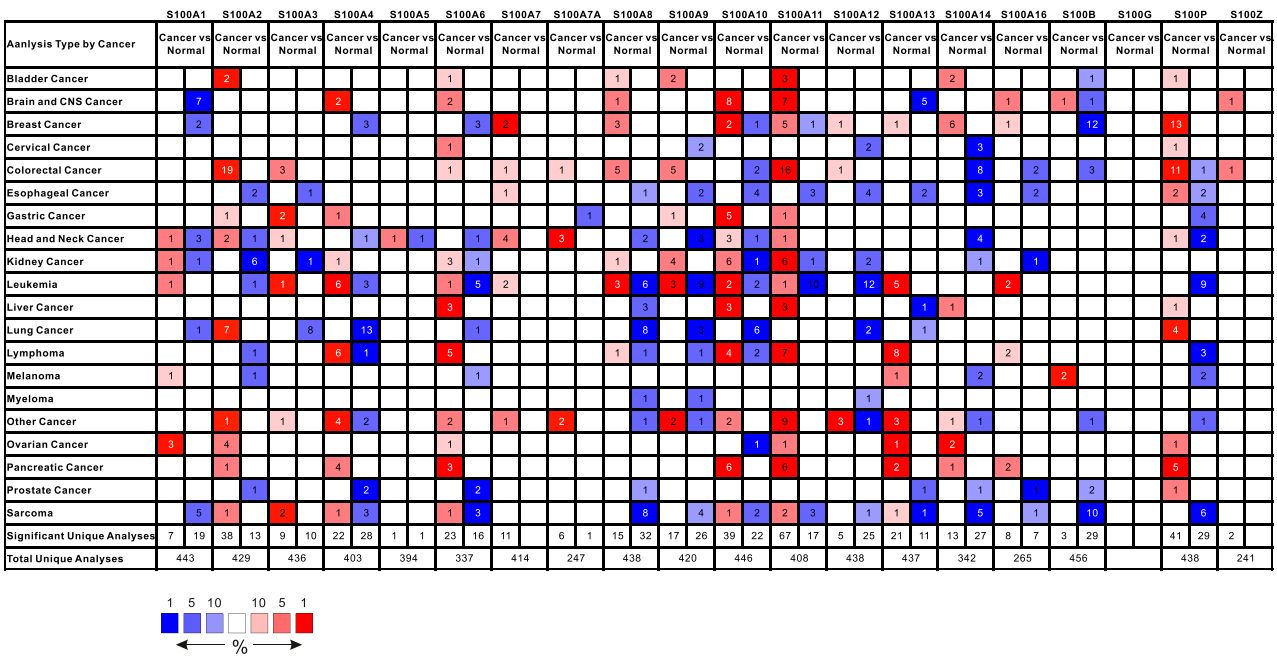

FIGURE 1 | The mRNA expression of the S100 family members (S100s) in 20 different cancer types using ONCOMINE. Color was determined by the best gene rank percentile gene for analyses within the cell; blue represented down-expression and red represented over-expression. The numbers in colored cells represents the quantities of datasets which satisfies the threshold: gene rank percentile (10%), P-value (0.001), and fold change (2.0). The mRNA expression of S100A2/A4/A6/A10/A11A13/14/A16/P were higher in tumor than in normal pancreatic tissues.

customizable functionalities, including differential expression analysis, profiling plotting, correlation analysis, patient survival analysis, similar gene detection, and dimensionality reduction analysis.

cBioPortal

CBioPortal is a bioinformatics analysis tool, providing a comprehensive analysis of complex cancer genomics and clinical characteristics (Gao et al., 2013). The pancreatic adenocarcinoma (TCGA, Firehose Legacy) dataset including data from 186 cases with pathology reports was chosen for further analyses of the S100s. We used it to analyze the genetic mutation, co-expression, and pathway of S100s.

STRING and GeneMANIA

STRING is a comprehensive publicly available bioinformatics database, providing network prediction of protein-protein interactions based on physical and functional correlations (Szklarczyk et al., 2019). We used it to construct the S100s interaction relationship network and explore the interaction among the S100s to predict the core proteins and key candidate genes. GeneMANIA is a flexible prediction platform, which can construct gene interaction networks by predicting co-expression, physical interactions, interactions, shared protein domains, and pathways (Warde-Farley et al., 2010). We use it to establish S100s gene co-expression networks to predict their potential function.

TIMER

TIMER is a useful and flexible web interface, providing six main analysis modules to systematically evaluate the infiltration and clinical effects of distinct immune populations in the tumor microenvironment (Li et al., 2020a). We used TCGA_PAAD datasets to evaluate the correlation between S100s expression and the abundance of immune infiltrating cells by Spearman correlation analysis.

GSCALite

GSCALite is a comprehensive publicly available bioinformatics database for genomic set cancer analysis, including expression, single nucleotide variation (SNV), copy number variation (CNV), methylation, small molecular drug targets, and cancer pathway activity analysis. The pathway activity module represents the correlation between gene expression and the pathway activity group (inhibition and activation) determined by the pathway score. Genome aberration not only affects the clinical therapeutic response but also provides a large number of research targets for the study of potential drug targets (Liu et al., 2018). GSCALite database integrates the drug sensitivity and gene expression profile data of cancer cells in GDSC and CTRP. Researchers can use this database to mine potential biomarkers and predict valuable small drugs, which is conducive to better research design and clinical trials in the future. We used it to analyze the pathway activity and drug targets of the S100s.

TABLE 1 | Significant changes in the transcriptional expression of the S100 gene family members between PAAD and pancreatic normal samples (OnCOMine database).

Gene	Type of pancreatic cancer vs normal samples	Fold change	t-test	*P* Value	Reference	Tumor samples	Normal samples
S100A2	Pancreatic carcinoma vs normal	7.68	6.38	2.97E-08	Pei	36	16
S100A4	Pancreatitis vs normal	2.57	5.70	4.45E-04	Logsdon	10	5
	Pancreatic adenocarcinoma vs normal	4.44	5.74	6.52E-05	Logsdon	10	5
	Pancreatic ductal adenocarcinoma vs normal	4.37	7.61	1.54E-10	Badea	36	36
	Pancreatic carcinoma vs normal	4.86	5.57	3.49E-06	Pei	36	16
S100A6	Pancreatic carcinoma vs normal	9.15	9.15	4.98E-12	Pei	36	16
	Pancreatic ductal adenocarcinoma vs normal	5.92	9.02	1.32E-12	Badea	36	36
	Pancreatic carcinoma vs normal	4.76	4.84	2.31E-04	Segara	12	6
S100A10	Pancreatic carcinoma vs normal	4.30	8.15	8.32E-07	Segara	12	6
	Pancreatic adenocarcinoma vs normal	7.58	10.08	2.79E-06	Logsdon	10	5
	Pancreatitis vs normal	2.97	5.13	4.89E-04	Logsdon	10	5
	Pancreatic ductal adenocarcinoma vs normal	3.10	8.38	1.31E-11	Badea	36	36
	Pancreatic adenocarcinoma vs normal	5.54	6.07	7.36E-04	Iacobuzio-Donahue	17	5
	Pancreatic carcinoma vs normal	3.54	6.06	2.05E-06	Pei	36	16
S100A11	Pancreatic carcinoma vs normal	7.52	7.65	7.48E-07	Segara	12	6
	Pancreatic ductal adenocarcinoma vs normal	4.43	10.45	1.19E-14	Badea	36	36
	Pancreatitis vs normal	5.45	5.51	7.60E-04	Logsdon	10	5
	Pancreatic adenocarcinoma vs normal	18.29	10.08	9.25E-05	Logsdon	10	5
	Pancreatic adenocarcinoma vs normal	7.20	6.01	8.54E-04	Iacobuzio-Donahue	17	5
	Pancreatic carcinoma vs normal	4.95	5.83	6.35E-06	Pei	36	16
S100A13	Pancreatic carcinoma vs normal	2.68	6.00	1.85E-05	Segara	12	6
	Pancreatic ductal adenocarcinoma vs normal	2.19	7.26	3.32E-10	Badea	36	36
S100A14	Pancreatic carcinoma vs normal	6.46	6.50	4.85E-07	Pei	36	16
S100A16	Pancreatic Carcinoma vs Normal	4.40	7.26	9.46E-08	Pei	36	16
	Pancreatic Ductal Adenocarcinoma vs Normal	2.33	6.94	1.05E-09	Badea	36	36
S100P	Pancreatic Adenocarcinoma vs Normal	24.02	11.30	2.13E-08	Iacobuzio-Donahue	17	5
	Pancreatic Adenocarcinoma vs Normal	20.31	15.50	4.88E-10	Logsdon	10	5
	Pancreatic Carcinoma vs Normal	77.93	10.19	1.61E-12	Pei	36	16
	Pancreatic Carcinoma vs Normal	17.73	6.94	1.23E-05	Segara	12	6
	Pancreatic Ductal Adenocarcinoma vs Normal	13.18	8.44	8.53E-13	Badea	36	36

Human Protein Atlas

The Human Protein Atlas (HPA) is a valuable platform for studying the localization and expression of proteins, as it contains more than 10 million immunohistochemistry images and 82,000 high-resolution immunofluorescence images (Thul and Lindskog, 2018). The protein levels of S100s members in normal and PAAD pancreatic tissues were compared using representative immunohistochemical images of HPA.

UALCAN

UALCAN is a user-friendly online website for analyzing cancer OMICS data (TCGA, MET500, and CPTAC) (Sighoko et al., 2011). We used it to analyze the relationships between S100s mRNA expression and TP53 mutation.

RESULTS

Expression Level of the S100s Members in Pancreatic Adenocarcinoma

Transcriptional levels of the S100s members between tumor and normal samples in twenty types of cancers were assessed by the tool ONCOMINE. As shown in **Figure 1**, nine S100s members, including S100A2/A4/A6/A10/A11/A13/A14/A16/P were significantly upregulated in pancreatic cancer (fold change = 2, $p < 0.001$). Then, through the oncology database, we further studied the significant differences in S100s transcription levels between distinct subtypes of pancreatic cancer and normal control tissues, as demonstrated in **Table 1**. In Pei's dataset, the mRNA expression level of S100A2 was overexpressed in pancreatic carcinoma versus normal samples with a fold change of 7.68 (**Table 1**). In Logsdon's datasets, S100A4 was found to be higher expressed in pancreatitis (fold change = 2.57), and pancreatic adenocarcinoma (fold change = 4.44) versus normal tissues. Badea et al. reported that S100A4 was increased in pancreatic ductal adenocarcinoma (fold change = 4.37), and Pei et al. showed that S100A4 also overexpressed in Pancreatic carcinoma (fold change = 4.86) compared to normal samples. In Badea's dataset, S100A6 was found higher expressed in pancreatic ductal adenocarcinoma compared to normal tissues (fold change = 5.92). Overexpression of S100A6 in pancreatic carcinoma was also found in Pei's dataset (fold change = 9.15) and Segara's dataset (fold change = 4.76). Additionally, in Logsdon's datasets, S100A10/A11 were overexpressed in pancreatic adenocarcinoma (fold change = 7.58 and 18.29), and pancreatitis (fold change = 2.97 and 5.45) compared to normal samples. According to Segara's dataset, S100A10/A11 were also found higher expressed in pancreatic carcinoma (fold change = 4.3 and 7.52). In Badea's datasets, higher expressed S100A10/A11 were found in pancreatic ductal adenocarcinoma ((fold change = 3.1 and

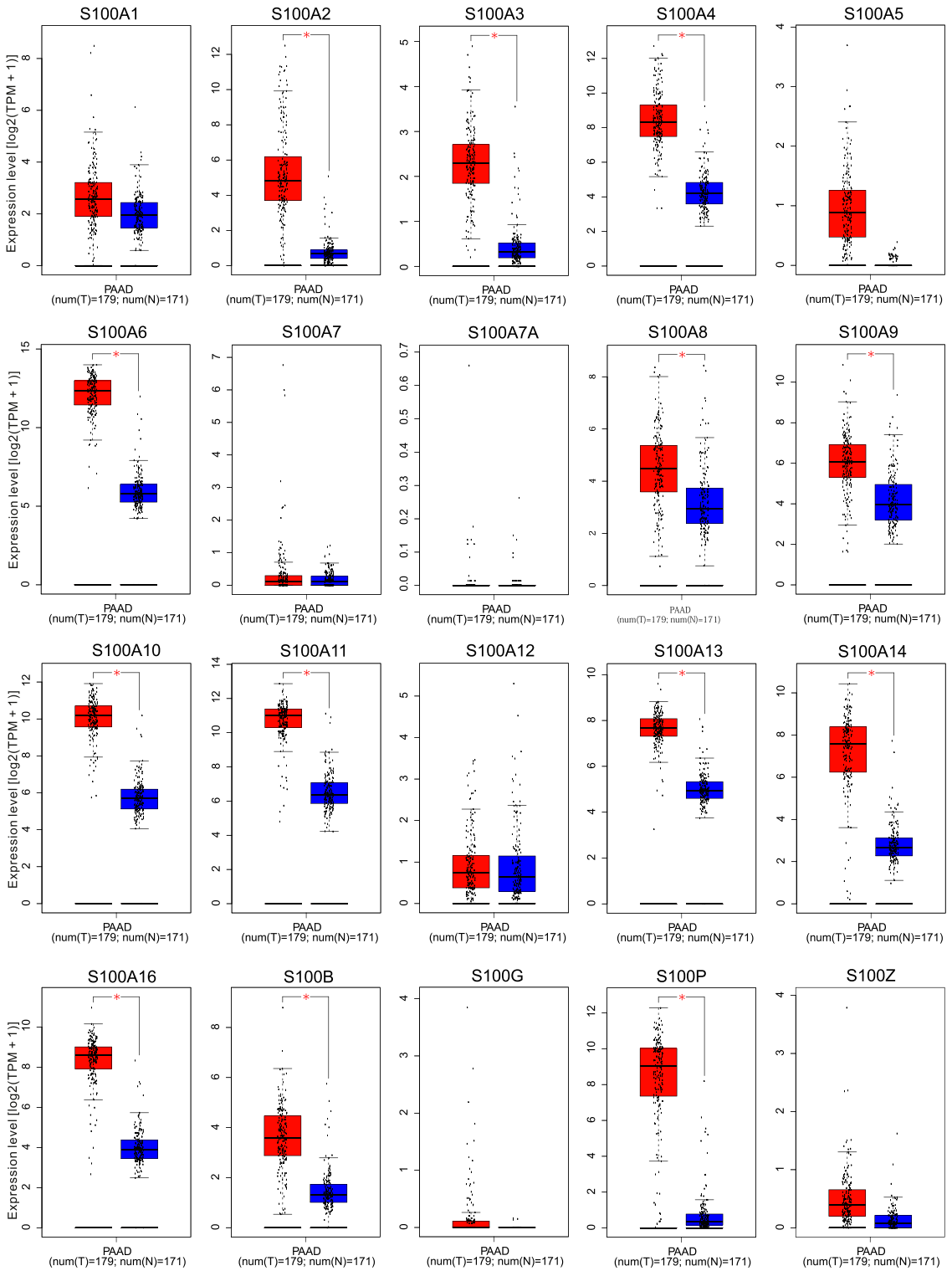

FIGURE 2 | The transcription levels of the S100 family members in Pancreatic adenocarcinoma patients using GEPIA. Box plots from GEPIA gene expression data compared the expression of specific S100 family members in PAAD and normal tissues. Select log2 (TPM + 1) transformed expression data for plotting. Blue color represents normal tissue and red color represents tumor tissue. The levels of S100A2/A3/A4/A6/A8/A9/A10/A11/A13/A14/A16/B/P in PAAD tissues were higher than that in normal paired samples as determined by Student's t-test. $p < 0.05$ was considered to be statistically significant. **Abbreviation:** num, sample number; T, tumor; N, normal; TPM, Transcripts per million. * Demonstrate that the results were statistically significant.

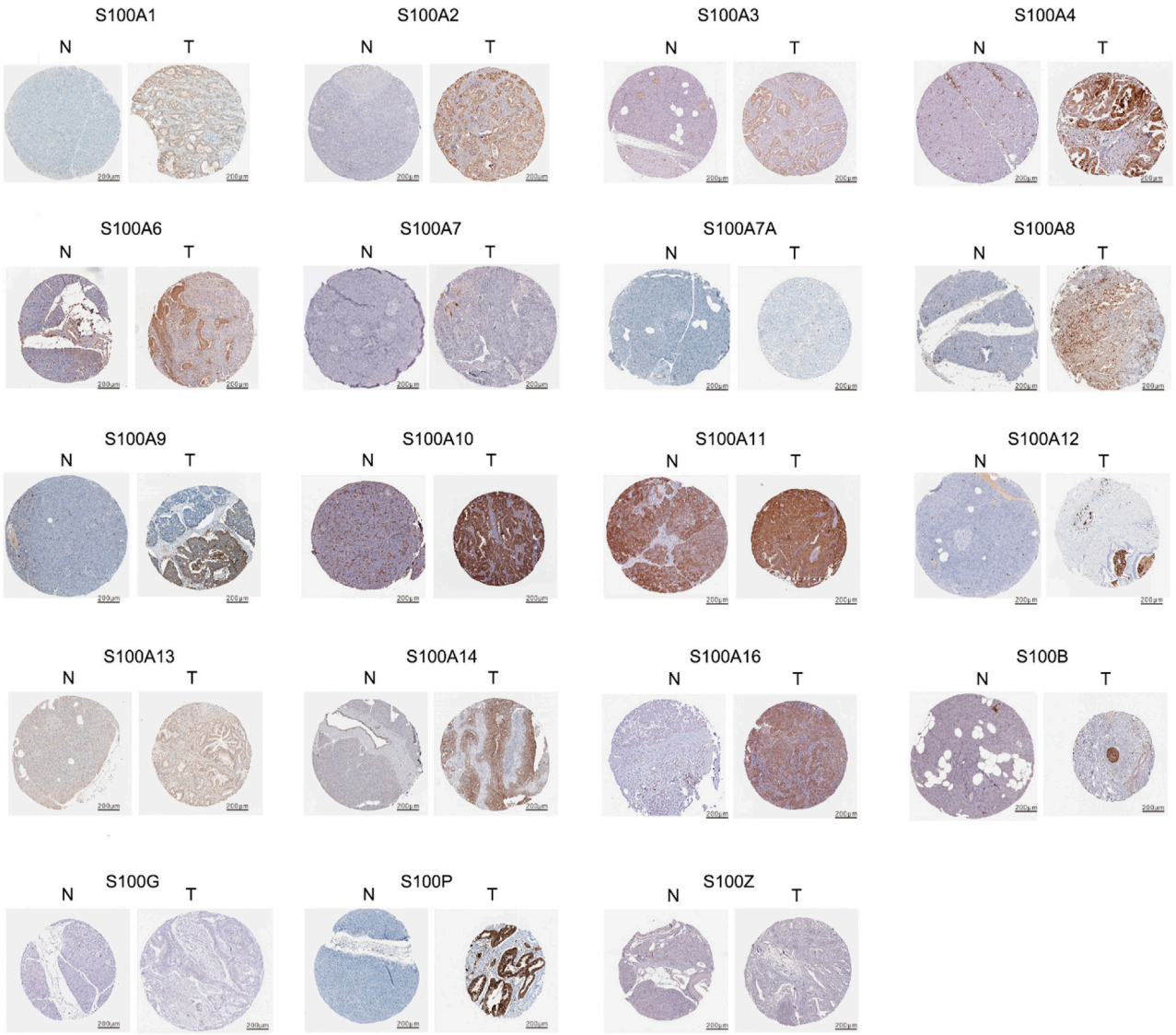

FIGURE 3 | Representative immunohistochemistry images (IHC) of the S100 family members in PAAD and normal pancreatic tissues using HPA database. The expression of S100A5 protein was not shown in the database. The expression of S00A7/A7A/A8/A9/A12/A14/A16/B/P/G/Z proteins were not detected in normal pancreatic tissues. The remain of S100 protein was expressed at low to moderate levels in some normal pancreatic tissues. Low protein expression of S100A2/A3/A7/A7A/A8/A13 were detected in PAAD tissues. Medium protein expression of S100A1/A4/A6/A9/A11/A16/B were detected in PAAD tissues. High protein expression of S100A10/A14/P were observed in PAAD tissues.

4.43). In Iacobuzio-Donahue's dataset, S100A10/A11 were showed overexpressed in pancreatic adenocarcinoma (fold change = 5.54 and 7.20). Additionally, in Pei's datasets, S100A10/A11/A14/A16/P were found higher expressed in pancreatic carcinoma (fold change = 3.54, 4.95, 6.46, 4.40, 77.93) in comparison with normal samples. S100A13/P were significantly upregulated in pancreatic carcinoma, with fold changes of 2.68 and 17.73 in Segara's dataset. Higher expressed S100A13/16/P were found in pancreatic ductal adenocarcinoma (fold change = 2.19, 2.33, 13.18) of Badea's datasets. S100P in pancreatic adenocarcinoma showed a similar trend, with 24.02 and 20.3-fold changes in Iacobuzio-Donahue's and Logsdon's datasets, respectively. According to ONCOMINE analysis, there was no significant difference in mRNA expression of S100A1/A3/A5/A7/A7A/A8/A9/A12/B/G/Z between patients with PAAD and normal controls.

Next, we utilized the Gene Expression Profiling Interactive (GEPIA) dataset to confirm the mRNA expression levels of differentially expressed S100s factors in PAAD and corresponding control tissues shown in the Oncomine database. We found that the transcription levels of S100A2/A3/A4/A6/A8/A9/A10/A11/A13/A14/A16/B/P were all highly expressed in PAAD tissues as compared with normal pancreatic tissues. Other S100 gene family members including S100A1/A5/A7/A7A/A12/G/Z have indicated no significant

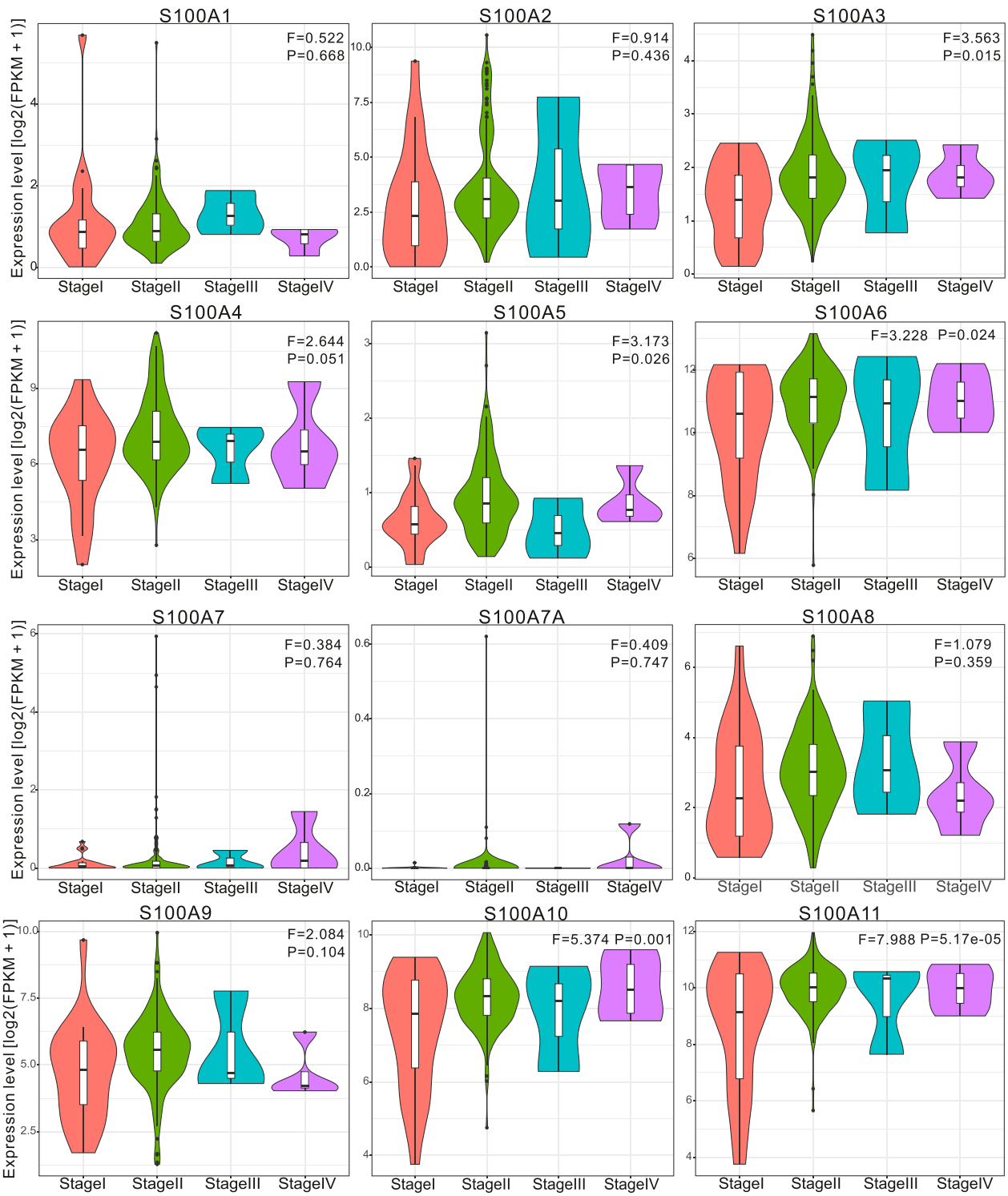

FIGURE 4 | Violin plot demonstrated the correlation between S100s transcription level and tumor stages in patients with PAAD (calculated data form TCGA-PAAD). The difference of individual S100s gene expression in each stage was analyzed by one-way ANOVA, in which Pr (>F) < 0.05 was considered to be statistically significant. The number of pancreatic cancer samples at each tumor stage was 21 cases in stage I, 146 in stage II, 3 in stage III, and 4 in stage IV. **Abbreviation:** F value, the statistical value of F test; Pr (>F), *p*-value.

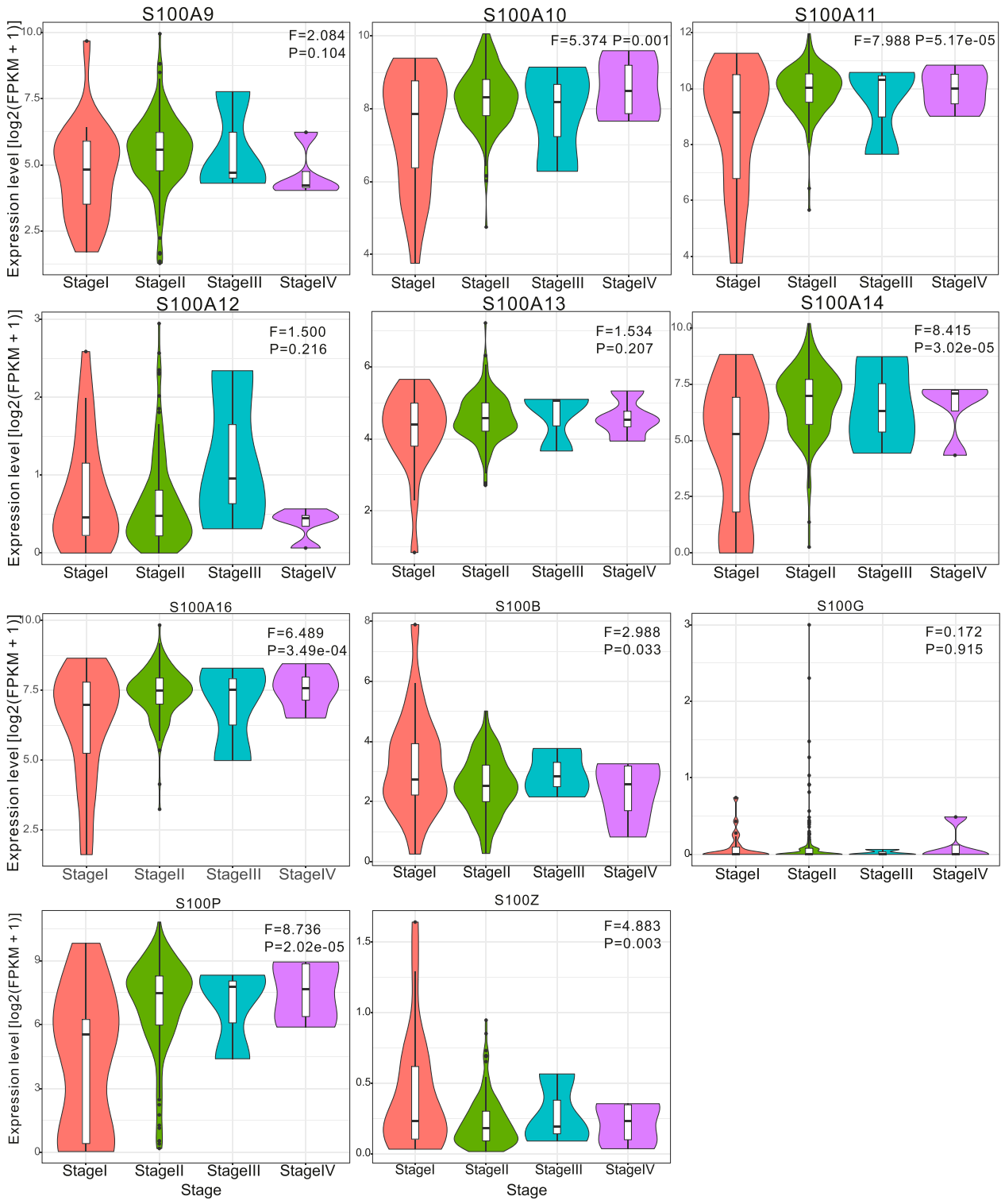

FIGURE 4 | Continued.

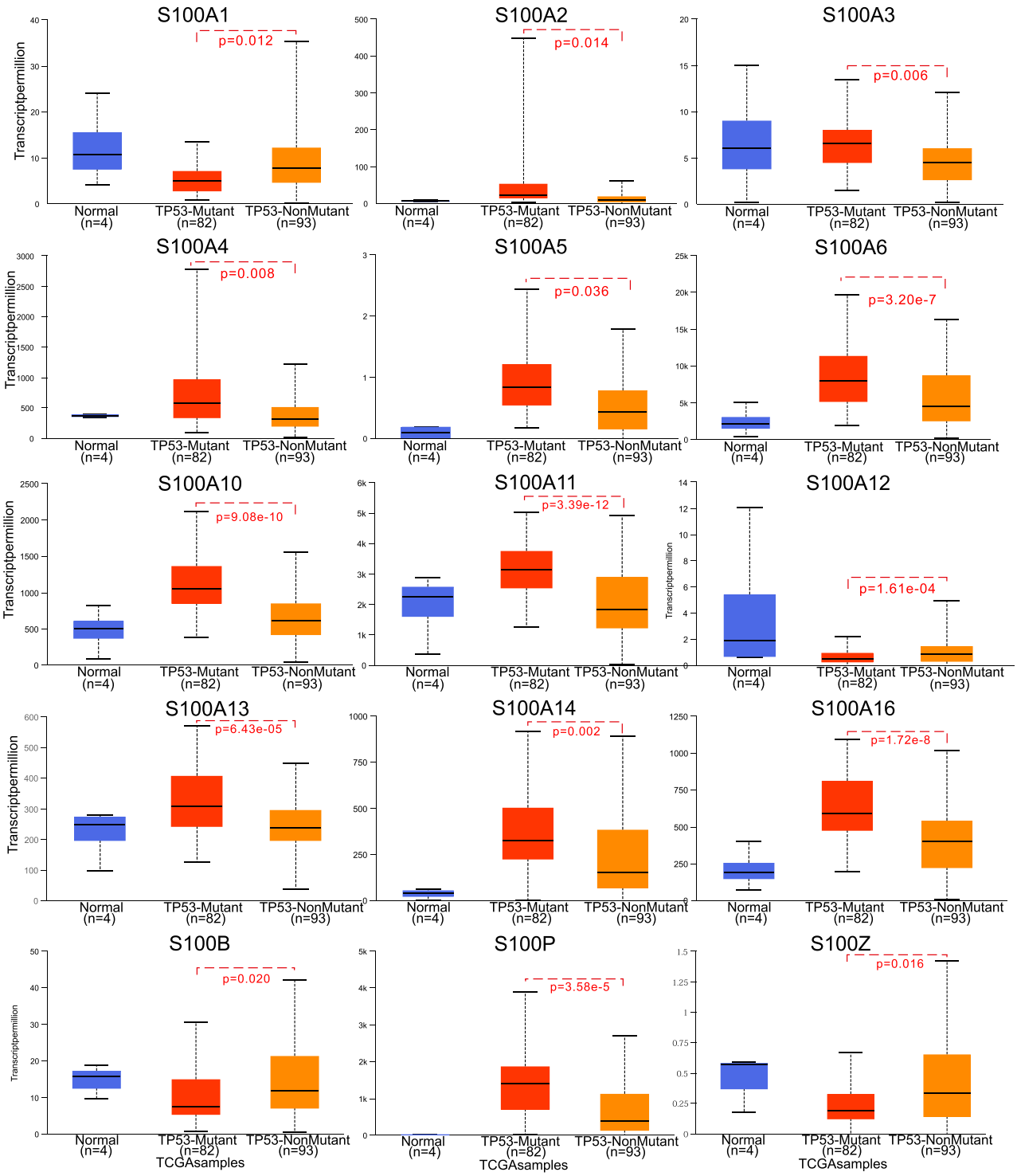

FIGURE 5 | Correlation between S100s mRNA expression and TP53 mutation in PAAD (UALCAN). The significance of difference between TP53-mutant and TP53-nonmutant estimated by Student's *t*-test with p value. The sample number N and the Student's *t*-test p value are marked on the figure, respectively.

differences between PAAD and normal tissues (**Figure 2** and **Supplementary Figure S1**).

After detecting the transcriptional level of the S100s in pancreatic adenocarcinoma tissues, we used the HPA to study its protein expression level. As shown in **Figure 3**, the expression of S00A7/A7A/A8/A9/A12/A14/A16/B/P/G/Z protein was not detected in normal pancreatic tissues, while the remaining S100 protein was expressed at low to moderate levels in some normal

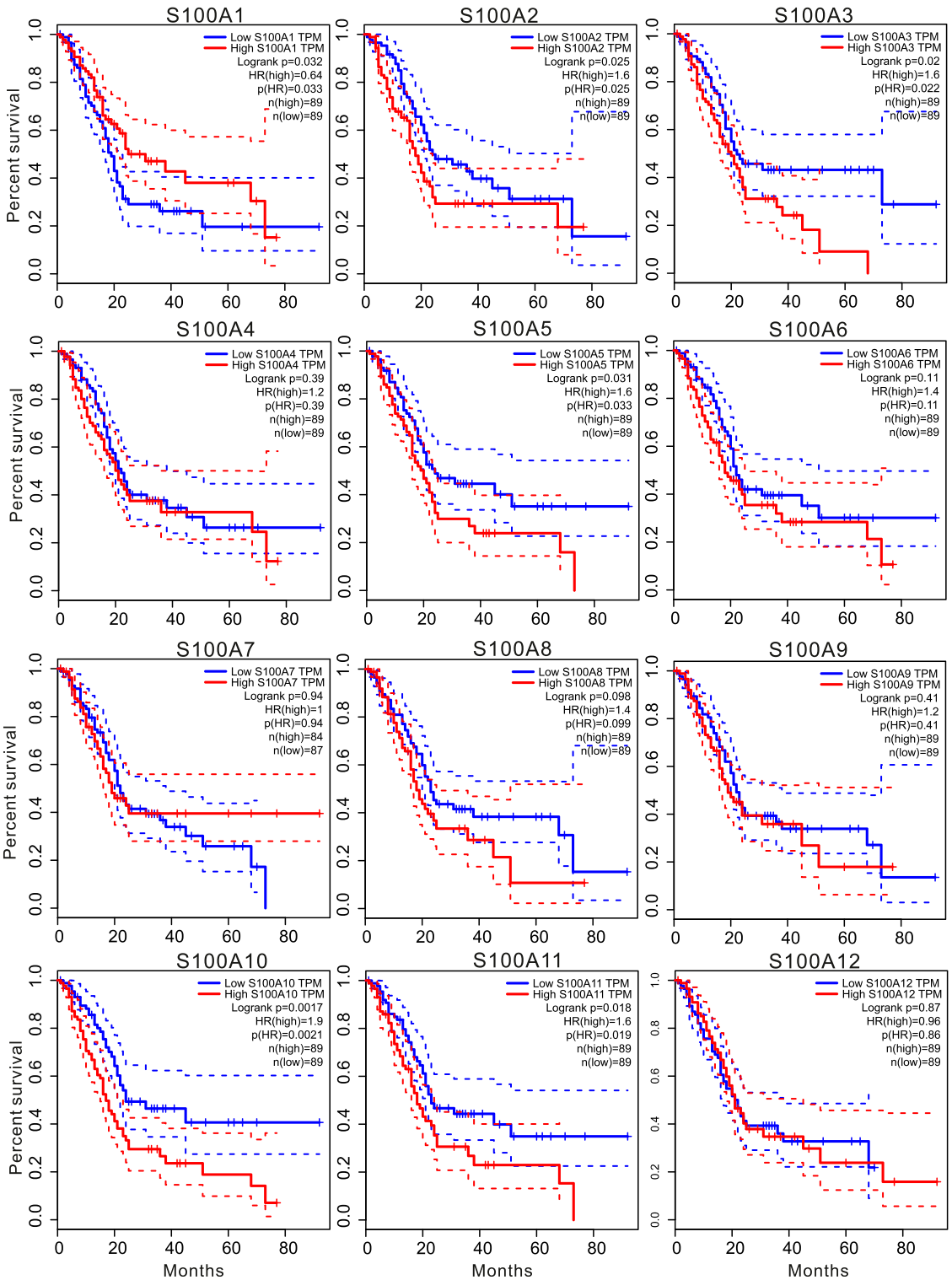

FIGURE 6 | Kaplan-Meier survival analysis showed that higher mRNA expression of S100A2/A3/A5/A10/A11/A14/A16 was significantly associated with shorter survival time, while higher expression of S100A1/B/Z was associated with longer survival time (GEPIA,n = 178). The *p*-value was less than 0.05.

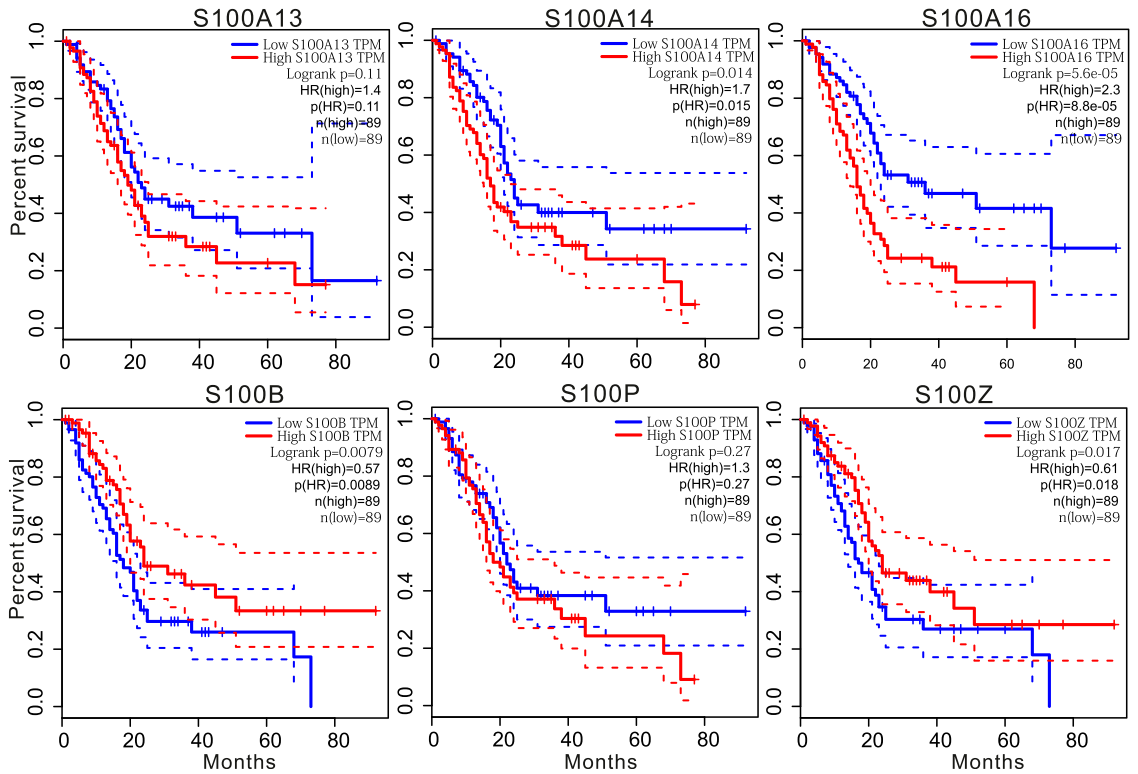

FIGURE 6 | Continued.

pancreatic tissues. The expression of S100A5 protein was not shown in the database. Low expression (protein expression scored <25%) of S100A2/A3/A7/A7A/A8/A13 protein was detected in PAAD tissues. Medium expression (protein expression scored ranged from 25 to 75%) of S100A1/A4/A6/A9/A11/A16/B protein was detected in PAAD samples. High expression (protein expression scored >75%) of S100A10/A14/P protein was observed in PAAD tissues. However, due to the small size of the PAAD immunohistochemical results in the HPA database, the conclusion remains to be further verified.

The Relationship Between the mRNA Expression Levels of the S100s Members and Clinical Characteristics in Pancreatic Adenocarcinoma

To evaluate the clinical significance of the differentially expressed S100 gene in the progression of PAAD patients, we analyzed the correlation between the transcriptional expression level of S100s members and clinicopathological features. The original data we used were derived from TCGA databases. We found that the transcriptional levels of S100A3/A5/A6/A10/A11/A14/A16/B/P/Z were correlated with tumor stage (**Figure 4**). There was no significant correlation between the remaining S100s members and the tumor stage.

We then explored the prognostic value of S100 family members in pancreatic cancer of different TP53 status. We found that the high expression of S100A2/A3/A4/A5/A6/A10/A11/A13/A14/A16/P in pancreatic cancer was positively correlated with TP53 mutation, while the high expression of S100A1/A12/B/Z was negatively correlated with TP53 mutation (**Figure 5**).

Prognostic Values of S100s in Patients With Pancreatic Adenocarcinoma

To further explore the key role of S100s members in the survival of PAAD, we used the GEPIA database to analyze the relationship between the expression of S100 factors and the overall survival (OS) of patients with PAAD. As displayed in **Figure 6**, we found that the high transcriptional levels of S100A2 (Hazard ratio (HR) = 1.6, $p = 0.025$), S100A3 (HR = 1.6, $p = 0.02$), S100A5 (HR = 1.6, $p = 0.031$), S100A10 (HR = 1.9, $p = 0.0017$), S100A11 (HR = 1.6, $p = 0.018$), S100A14 (HR = 1.7, $p = 0.014$) and S100A16 (HR = 2.3, $p =$ 5.6e-05) were significantly associated with poor OS in PAAD, while low mRNA expression of S100A1 (HR = 0.64, $p = 0.032$), S100B (HR = 0.57, $p = 0.0079$) and S100Z (HR = 0.61, $p = 0.017$) were associated with worse OS. However, the statistical significance of S100A4 (HR = 1.2, $p = 0.39$), S100A6 (HR = 1.4, $p = 0.11$), S100A7 (HR = 1, $p = 0.94$), S100A8 (HR = 1.4, $p = 0.098$), S100A9 (HR = 1.2, $p = 0.41$), S100A12 (HR = 0.96, $p = 0.87$), S100A13 (HR = 1.4, $p = 0.11$), S100P (HR = 1.3, $p = 0.27$) was not detected. The GEPIA databases did not provide the survival analysis results of S100A7A and S100G in the majority of PAAD molecular subtypes. Thus, the PAAD patients with low mRNA levels of the S100A1/B/Z or high

mRNA levels of S100A2/A3/A5/A10/A11/A14/A16 were predicted to have worse OS.

Genetic Mutation and Interaction Networks Analysis of S100s Members in Pancreatic Adenocarcinoma

The S100 alterations and correlations were analyzed by the tool cBioPortal for Pancreatic Adenocarcinoma (TCGA, Firehose Legacy). S100s were altered in 70 samples of 149 patients with PAAD, accounting for 47% (**Figure 7A**). As shown in **Figure 7B**, the genetic alteration rates of the S100 gene family members in PAAD ranges from 0 to 16% individually (S100A1, 6%; S100A2, 9%; S100A3, 9%; S100A4, 13%; S100A5, 11%; S100A6, 10%; S100A7, 5%; S100A7A, 5%; S100A8, 6%; S100A9, 7%; S100A10, 14%; S100A11, 9%; S100A12, 6%; S100A13, 10%; S100A14, 16%; S100A16, 13%; S100B, 0%; S100G, 2.7%; S100P, 8%; S100Z, 3%).

Meanwhile, the transcriptional expression correlation of S100s members individually in patients with PAAD (TCGA, Firehose Legacy) was calculated by the cBioPortal database, and Pearson correlation analysis was carried out. The statistical significance is as follows: S100A6 with S100A10/A11/A13/A14/A16/P; S100A7 with S100A7A/A8; S100A7A with S100A8; S100A8 with S100A9; S100A10 with S100A11/A14/A16; S100A11 with S100A13/A14/A16; S100A12 with S100A13; S100A13 with S100A16; S100A14 with S100A16 (**Figure 7C**).

The interaction relationship and potential regulatory mechanism of S100s factors in PAAD were mined by using the GeneMANIA database, and the gene interaction network was constructed. The network consists of 40 genes, including 20 members of the S100s and another 20 genes extracted from the GeneMANIA. The analysis illustrated that there was a close genetic relationship among the members of the S100s. The results displayed that the S100s was co-expressed interactively with TCHHL1, S100A7L2, HRNR, FLG2, RPTN, TCHH, FLG, CRNN, SNTN, CABP7, CALN1, KCNIP4, GUCA1C, KCNIP2, CALML5, SRI, SDF4, EFCAB11, CIB4, EFCAB7 (**Figure 7D**). The functions of the S100s were mainly associated with granulocyte chemotaxis, sequestering of metal ions, regulation of water loss via the skin, granulocyte migration, and response to fungus (**Figure 7D**).

We used the tool STRING to construct a protein-protein interaction network of the S100s to explore their potential interactions. The networks contained 30 nodes and 122 edges (**Figure 7E**). The top five proteins most associated with S100s were ANXA2, AGER, AHNAK, HLTF, and S100PBP. The main biological processes involved in the PPI network were positive regulation of response to external stimulus, regulated exocytosis, astrocyte differentiation, myeloid leukocyte activation, and cell activation involved in the immune response. An IL-17 signaling pathway is the main pathway of KEGG.

Cancer Pathway Activity and Drug Sensitivity Analysis of the S100s in Pancreatic Adenocarcinoma

We further analyzed the role of the S100s in canonical cancer-related pathways using GSCALite and Cbioportal databases. Firstly, we used the platform GSCALite to analyze the activities of ten well-known tumor-related pathways, such as RTK, RAS/MAPK, TSC/mTOR, cell cycle, DNA damage response, EMT, hormone ER, hormone AR, PI3K/AKT, and apoptosis pathway. We found that most members of the S100s were associated with the activation of EMT, apoptosis, RTK, and RAS/MAPK pathway; and the inhibition of Hormone AR, DNA Damage Response, and RTK pathway (**Figure 8A**). Furthermore, we also analyzed the cancer pathway of the S100s based on cBioPortal. Through the analysis of the proportion of gene changes in ten typical carcinogenic signaling pathways, we found that the S100 alterations were closely related to the carcinogenic changes of key gene loci such as KRAS, CDKN2A, TP53, MYC, SMAD4 (with more than 10% mutation) (**Supplementary Figure S2**).

Gene mutations may affect clinical therapeutic responses and become potential drug targets. GSCALite was an publicly available platform that integrated the drug sensitivity and gene expression profile data of cancer cell lines in GDSC and CTRP, which was helpful for researchers to mine and predict valuable small drugs. Drug sensitivity indicated that the high expression of S100A2/A6/A10/A11/A13/A14/A16 was resistant to small molecule drugs or high expression of S100A1/A7/A9/B/Z were sensitive to small molecule drugs (**Figure 8B**).

Correlation Between S100s and Immune Cell Infiltration in Patients With Pancreatic Adenocarcinoma

In recent years, studies have shown that the enrichment of immune cells in the tumor microenvironment is closely related to tumor proliferation and development (de Visser et al., 2006; Lei et al., 2020). To evaluate the effect of the S100s on the degree of immune cell infiltration in the tumor microenvironment, we used TIMER to analyze the correlation between the expression level of the S100s members and immune infiltrating cells in PAAD. As shown in **Figure 9**, S100A3 was positively correlated with the abundance of $CD4^+$ T cells, neutrophils, and dendritic cells. S100A4 showed a significant correlation with the abundance of neutrophils, and dendritic cells. S100A5 was negatively correlated with the infiltration of $CD8^+$ T cells and macrophages. S100A14 was also negatively associated with the infiltration of CD4+cells, macrophage, and dendritic cell. Except for B cells and $CD4^+$ T cells, S100A6 was negatively correlated with the abundance of the other immune cells ($CD8^+$ T cells, macrophages, neutrophils, and dendritic cells. For S100A7/A10/A11, the expression of these genes was negatively associated with the abundance of macrophages. Meanwhile, the expression of S100A13/P was associated with the infiltration of $CD4^+$ T cells and macrophages. Besides, S100A1/A16 showed a significant correlation with the infiltration of $CD8^+$ T cells, $CD4^+$ T cells, macrophages, and dendritic cells. Interestingly, the expression of S100A12/B/Z was associated with the infiltration of these six immune cells (B cells, CD8 + T cells,

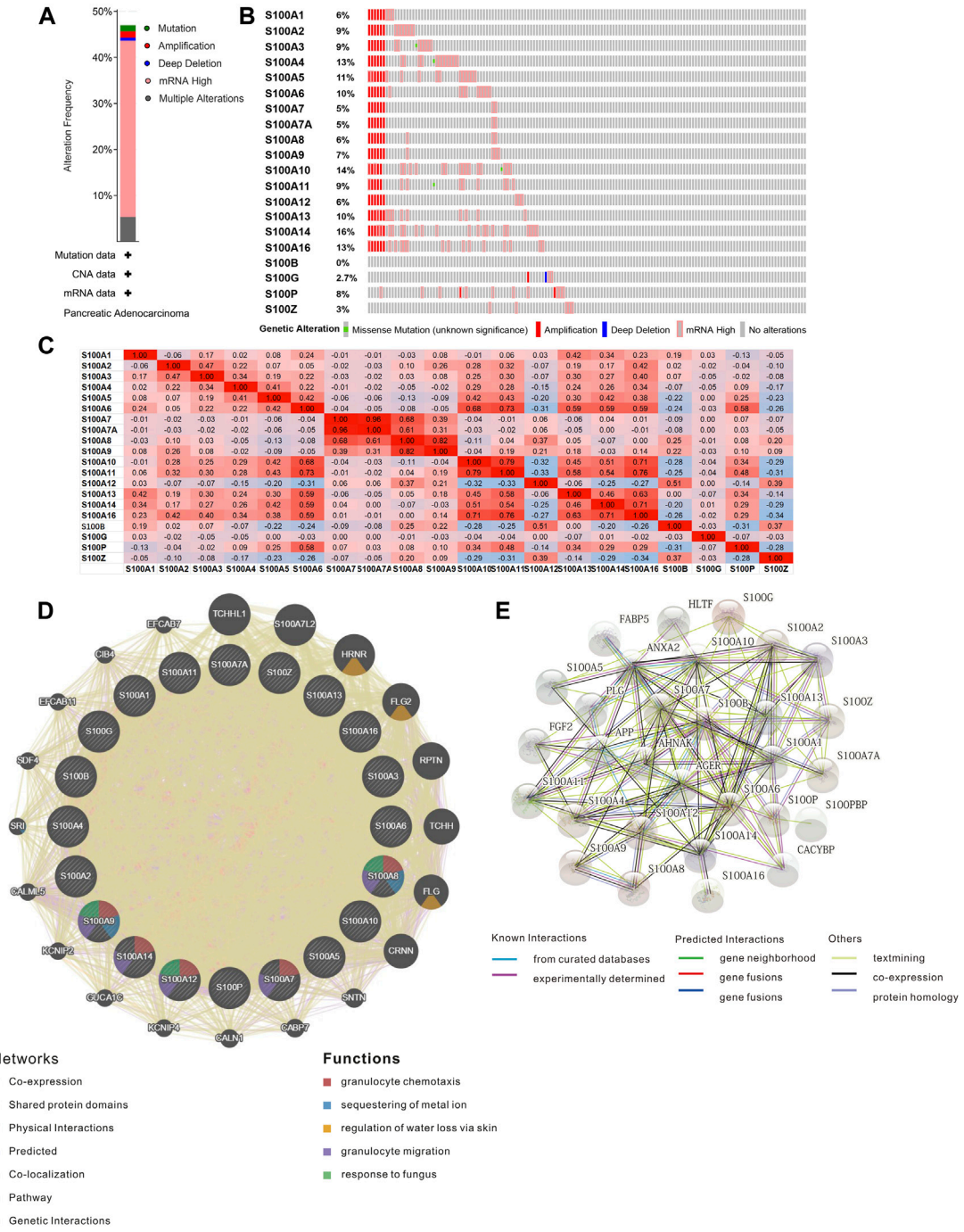

FIGURE 7 | Genetic mutation and interaction networks analysis of the S100s members in PAAD (cBioPortal, GeneMINIA, and STRING). **(A)** Summary of mutation frequency in S100 factors in PAAD (n=149), **(B)** Genetic alteration of each S100s in PAAD (n=149), **(C)** Correlation among different S100s factors in PAAD by Pearson's correlation coefficients based on cBioportal databases (positive correlation is red, negative correlation is blue, n=149), **(D)** Gene interaction networks among S100s in the GeneMANIA database. **(E)** Protein-protein interaction network of the S100s members in PAAD.

CD4 + T cells, macrophages, neutrophil, and dendritic cells). Except for B cells, S100A8 and S100A9 was positively correlated with the infiltration of the other immune cells ($CD8^+$ T cells, $CD4^+$ T cells, macrophages, neutrophil, and dendritic cells). In conclusion, the above results suggested that the S100s may play an important role in the immune infiltration of PAAD.

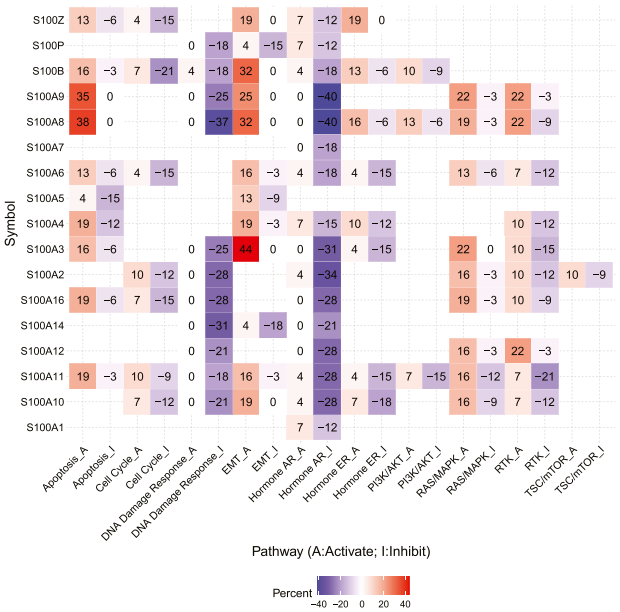

FIGURE 8 | Cancer pathway activity and drug sensitivity analysis of the S100s in PAAD (GSCALite, n=178). **(A)**. The role of the S100s in the famous cancer-related pathways. **(B)–(C)**. Spearman correlation showed that S100 gene expression was related to small drugs. Negative correlation means that high gene expression is sensitive to the drug, and vice versa.

Differential Expression Analysis of Samples with High Expression of the Significant S100 Genes (top 25 percentile) vs Low Expression (bottom 25 percentile) Samples

We selected five S100 genes (S100A2/A10/A11/A14/A16 by using the above analysis, which were highly expressed in PAAD and associated with poor prognosis and chemoresistance of PAAD, for further differential expression analysis between the high (top 25 percentile) and low samples (bottom 25 percentile). We downloaded pancreatic cancer samples and normal samples from the TCGA and the GEXT database respectively. We used Wilcox. test to analyze differentially expressed genes in samples. The screening condition was FDR = 0.05 and logFC = 1.5. We found that the differentially expressed genes of these five S100 genes were 1,048, 1858, 2,492, 1,668, and 2,482, respectively (**Figure 10A** and **Figure 10B**). We enriched the GO and KEGG signal pathways of these differential genes respectively. We found that there were many similarities in the signal pathways enriched by these genes (**Figure 11A** and **Figure 11B**). We also found that there were significant differences between the high and low expression samples by using the principal component analysis (PCA). (**Figure 12**).

DISCUSSION

The transcription and protein expression levels of multiple S100s members have been confirmed to have changed in various of tumors (Allgöwer et al., 2020b). Emerging evidence displayed that the biological characteristics of most S100s members are complex and multifactorial, and these members are actively involved in the process of tumorigenesis and progression, such as tumor cell proliferation, angiogenesis, metastasis, immune escape and drug resistance (Donato, 2001; Chen et al., 2014; Bresnick et al., 2015; Moravkova et al., 2016; Zhang et al., 2017; Ma et al., 2019). In this study, we evaluated the expression, genomic mutation, prognostic value, potential function, relationship with immune infiltration, and involvement in the regulation of drug resistance of the S100s in PAAD. We hope that our results will help to reveal the role of the S100s in cancer and provide new theoretical knowledge for the identification of tumor prognostic biomarkers and therapeutic targets. We found that the transcription levels of 13 S100s members were overexpressed in PAAD patients. The high expression of S100A2/A3/A4/A5/A6/A10/A11/A12/A13/A14/A16/P was positively associated with TP53 mutation. Combined with the transcriptional expression profile, we concluded that high mRNA expression of S100A2/A3/A10/A11/A14/A16 was correlated with poor OS, while high expression of S100B was related to better OS. Moreover, the high expression of S100A2/A6/A10/A11/A13/A14/A16 was resistant to small drugs. These findings suggested that S100A2/A10/A11/A14/A16 may have the potential to become a biomarker for prognosis or treatment of PAAD.

S100A2 is a well-studied S100s member in cancer, which is closely related to the occurrence of various human tumorigenesis (Wolf et al., 2011). S100A2 seems to have both carcinogenic and anticancer effects in cancer. Previous studies have shown that S100A2 plays an anticancer role in oral cancer, prostate cancer, breast cancer, and lung cancer, while it acts as a cancer promoter

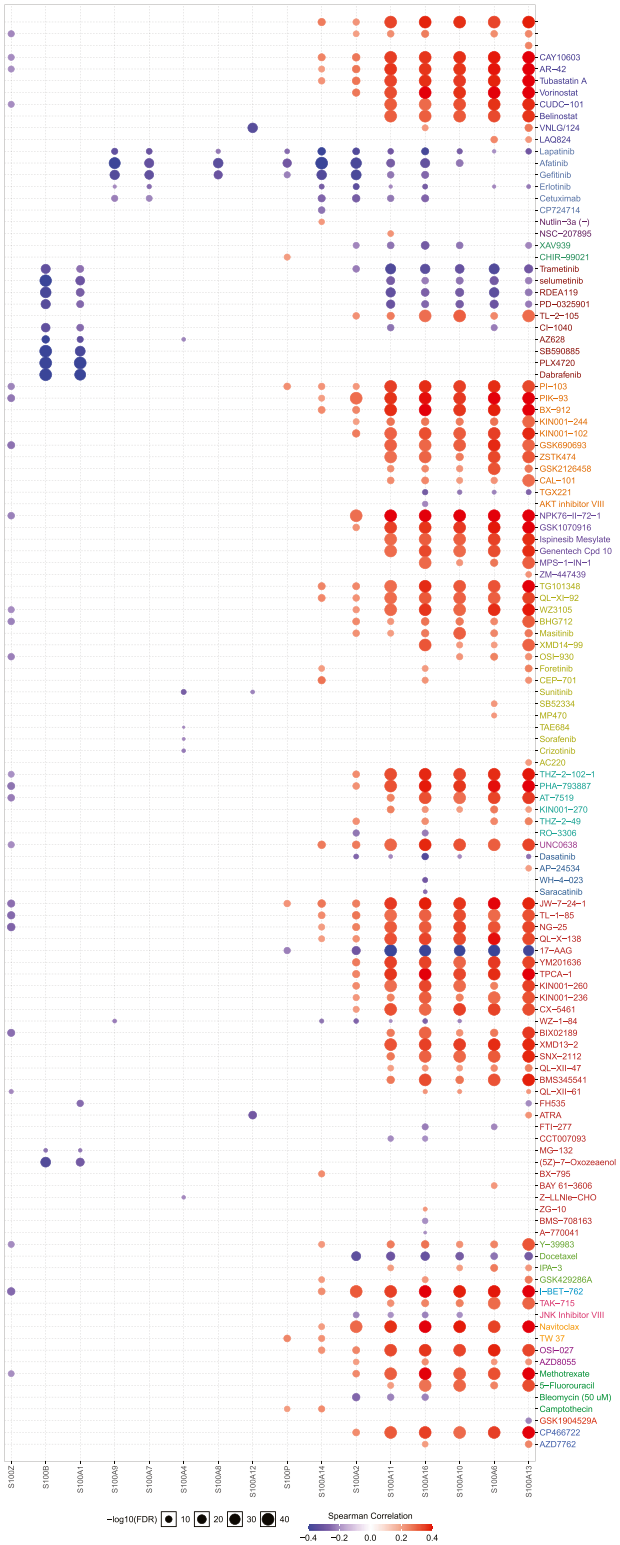

FIGURE 8 | Continued.

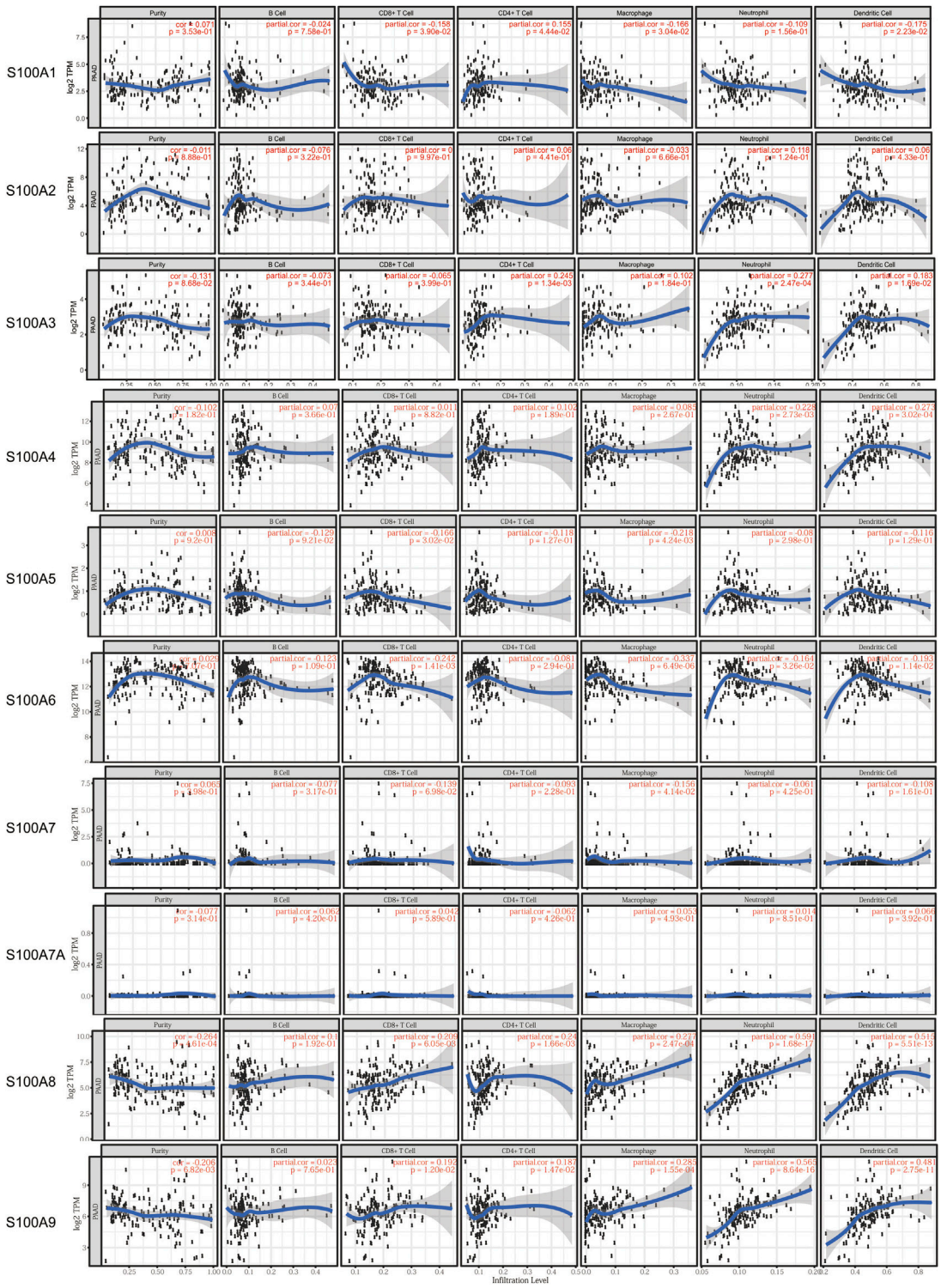

FIGURE 9 | Correlations between S100 gene mRNA expression and immune infiltration abundance in PAAD (TIMER database, n=179). The mRNA expression of some S100s factors was correlated with the infiltration abundance of B cells, CD8+ T cells, CD4+ T cells, macrophages, neutrophils and dendritic cells.

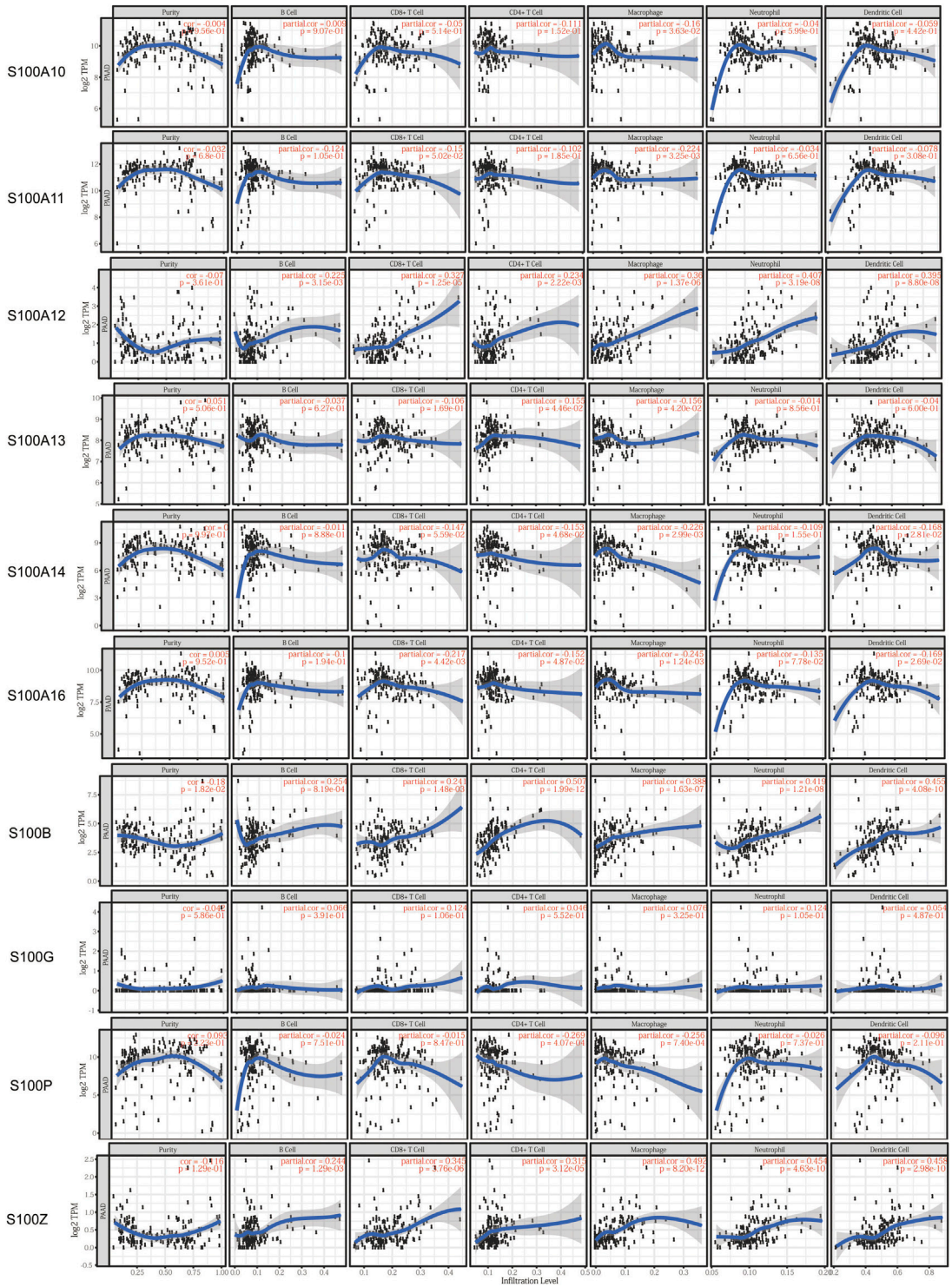

FIGURE 9 | Continued.

in ovarian cancer, gastric cancer, and esophageal squamous carcinoma (Liu et al., 2000; Rehman et al., 2005; Bulk et al., 2009; Donato et al., 2013; Gross et al., 2014; Bresnick et al., 2015). Overexpression of S100A2 was related to advanced histological grade, high T stage, and poor prognosis in pancreatic cancer (Zhuang et al., 2021). A large retrospective study suggests that

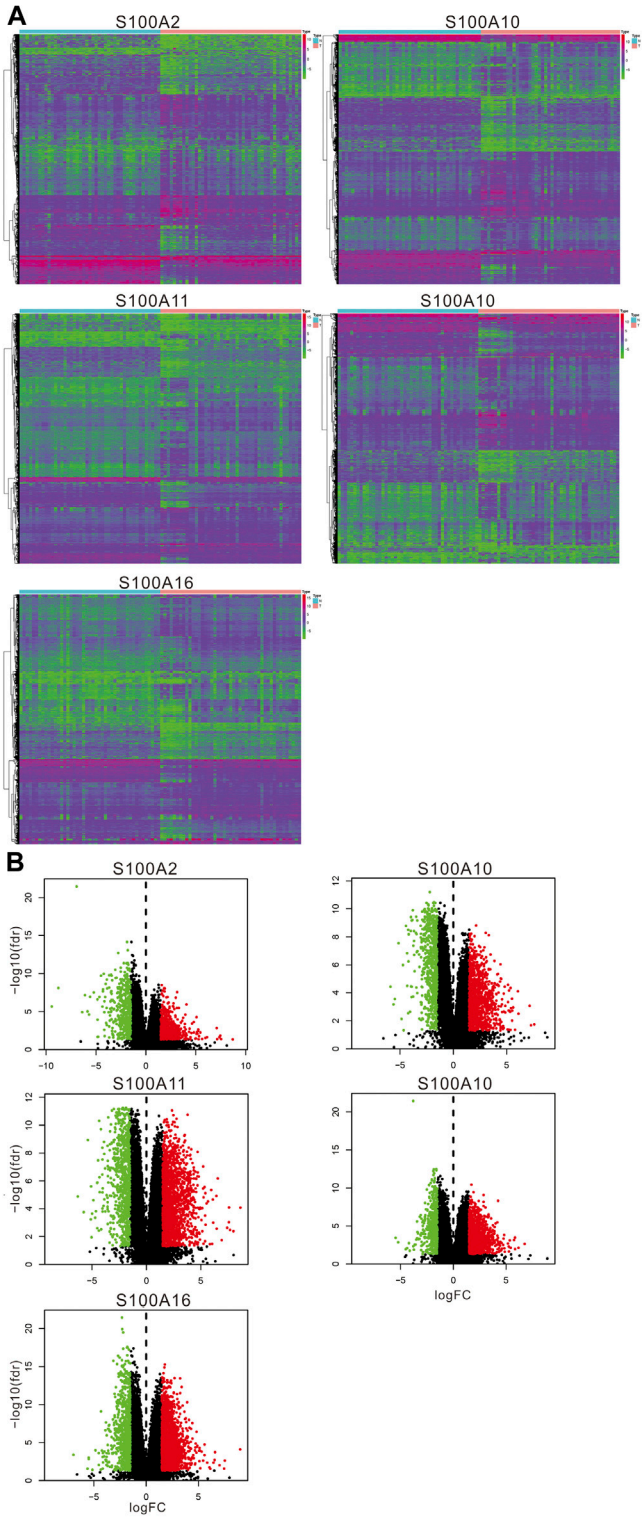

FIGURE 10 | Differential expression analysis of samples with high expression of the significant S100 genes **(top 25%)** vs low expression **(bottom 25%)** samples (calculated data form TCGA-PAAD database, n=178). **(A)** Differentially expressed genes were displayed by heatmap. Wilcox test was used to analyze differentially expressed genes. The screening conditions are FDR=0.05 and logFC=1.5. The green on the heatmap indicates low expression and red indicates high expression. **(B)** Volcano plot. The green dot on the volcano plot represents down-regulated expression, and the red dot represents up-regulated expression.

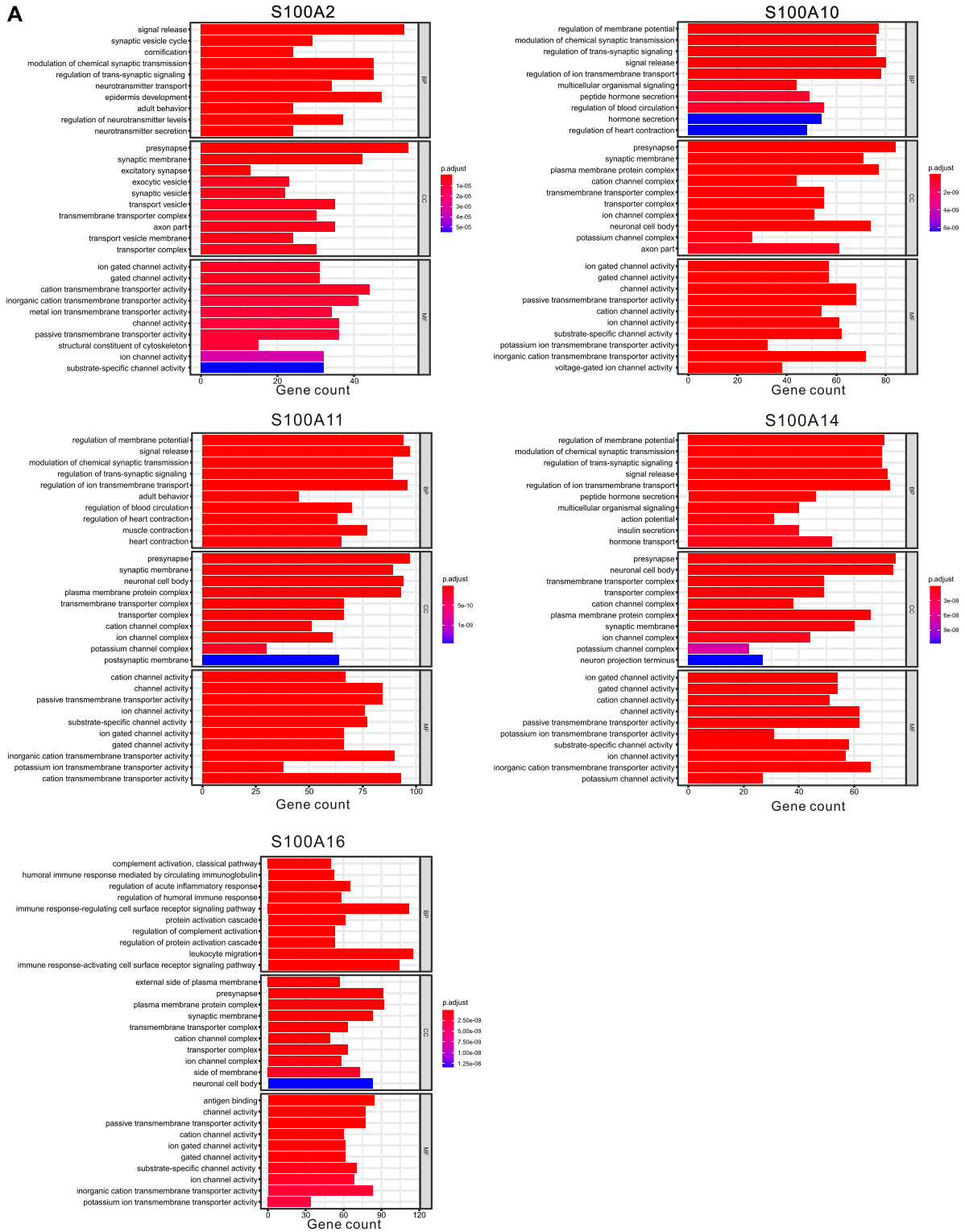

FIGURE 11 | Functional enrichment analysis of differential genes in samples with significant S100 gene high expression **(top 25%)** and low expression **(bottom 25%)** samples (calculated data from TCGA-PAAD, n=178). **(A)** GO enrichment analysis (BP: biological process; CC: cellular component; MF: molecular function). **(B)** Top 30 most significant KEGG pathways.

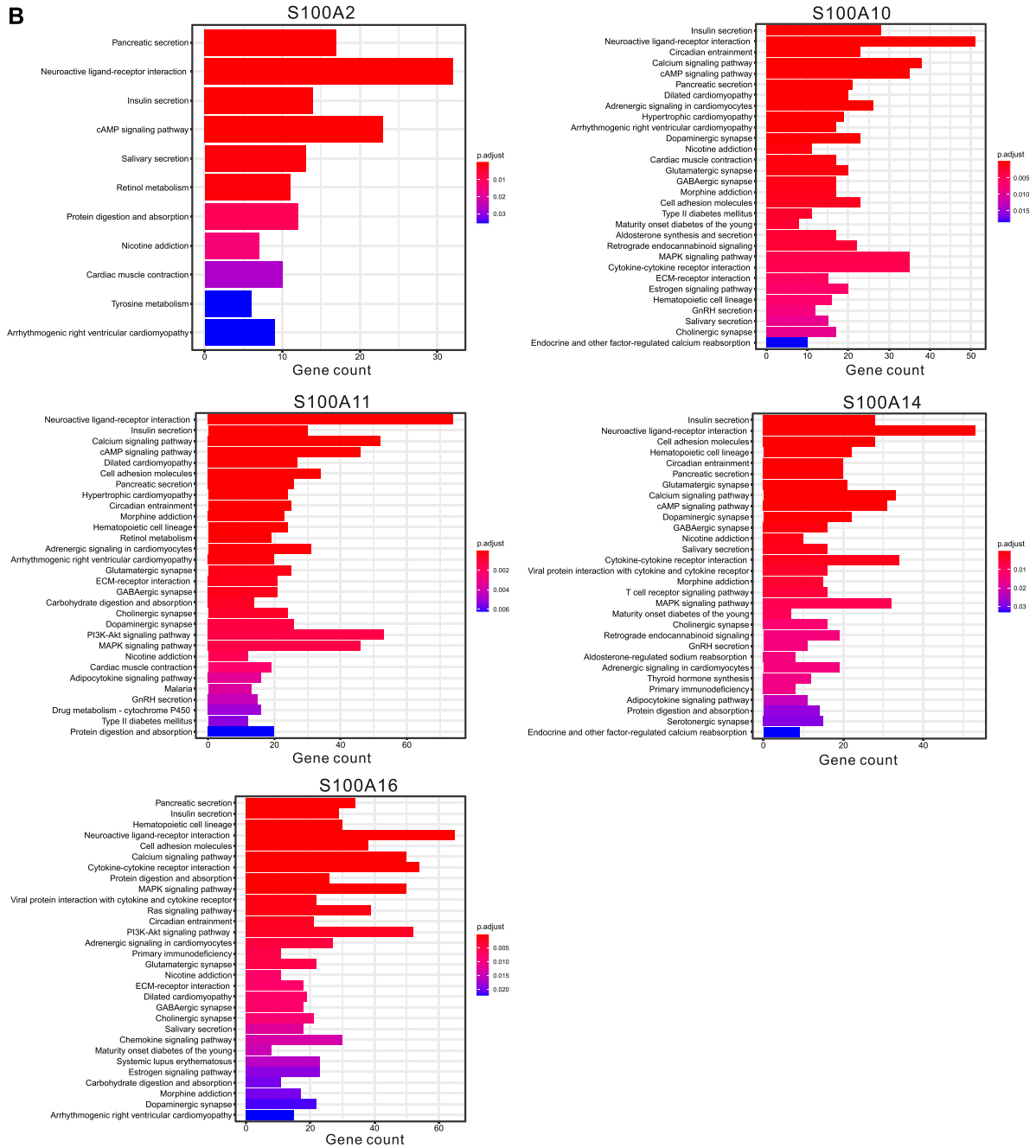

FIGURE 11 | Continued.

S100A2 may have the potential to become a biomarker for predicting pancreatectomy or replication therapy in patients with pancreatic cancer (Jamieson et al., 2011; Bachet et al., 2013). S100A2 is regulated by cell cycle progression and tumor suppressor gene p53, and it can interact with RAGE *in vitro* (Leclerc et al., 2009; Donato et al., 2013). According to the signaling pathway activated by S100A2/RAGE in pancreatic cancer cells, S100A2 can be used as either a tumor suppressor or a tumor promoter (Leclerc et al., 2009). Our results showed that S100A2 was significantly up-regulated in PAAD. Moreover, high S100A2 mRNA expression was related to worse OS, drug resistance, and high TP53 mutation.

S100A10 plays an important role in the progression of various cancers. S100A10 is up-regulated by the Smad4-dependent transforming growth factor-β 1 signaling pathway (Bydoun et al., 2018a). Previous studies have shown that the

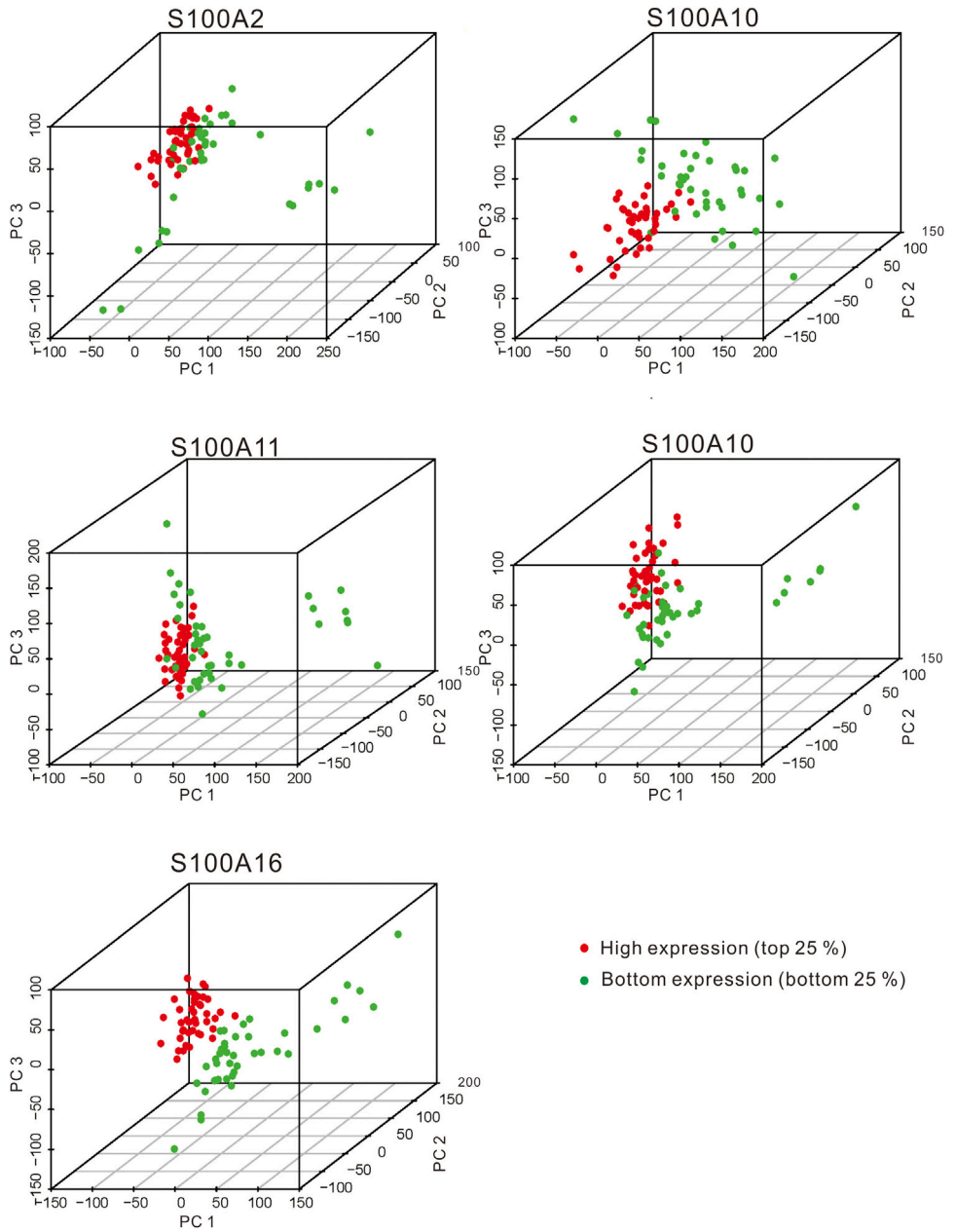

FIGURE 12 | The PCA plots showed the genes in samples with significant S100 gene high expression **(top 25%)** and low expression **(bottom 25%)** samples (calculated data from TCGA-PAAD, n=178). The red dot and green dot indicate the smaples with the high expression of the significant S100 genes **(top 25 percentile)** and the low expression **(bottom 25 percentile)** samples, respectively.

overexpression of S100A10 is usually related to tumor size, pathological TNM stage, lymphovascular invasion, lymph node metastasis, poor prognosis, and drug resistance in many malignancies (Zhang et al., 2004; Saiki and Horii, 2019). Overexpression of S100A10 is associated with a poor prognosis of pancreatic ductal adenocarcinoma (Bydoun et al., 2018b). Meanwhile, Bydoun et al. (2018) demonstrated that S100A10 is involved in a new mechanism of plasminogen activation during the epithelial-mesenchymal transition (EMT) (Bydoun et al., 2018a). In our study, we found that the mRNA level of S100A10 was highly expressed in pancreatic cancer and was related to the high tumor stage. Overexpression of S100A10 is associated with poor prognosis, high TP53 mutation, and drug resistance. In addition, S100A10 may be related to the activation of RAS/MAPK, EMT, and SMAD4-dependent TGF-β1 pathways.

S100A11 has been shown to play a dual role in tumors, such as inhibiting cancer of the bladder and kidney or promoting cancer

of the pancreas (Yao et al., 2007; Donato et al., 2013; Gross et al., 2014; Bresnick et al., 2015). Previous studies have shown that S100A11 increases in the early stages of pancreatic cancer, but decreases as cancer progress (Ohuchida et al., 2006). Overexpression of S100A11 is associated with an unfavorable prognosis of patients with pancreatic ductal carcinoma undergoing surgical resection (Xiao et al., 2012). S100A11 can also promote the proliferation and viability of pancreatic cancer cells by up-regulating the PI3K/AKT signaling pathway, which is considered to be a promising new drug target for targeted therapy of pancreatic cancer (Xiao et al., 2018). In addition, S100A11 is considered as a molecular marker for early diagnosis of pancreatic cancer or for screening patients with high-risk lesions that have progressed to pancreatic cancer (Ohuchida et al., 2006). Our results displayed that S100A11 mRNA expression was significantly higher than that in normal pancreatic tissues. Overexpression of S100A11 is associated with tumor stage, drug resistance, high TP 53 mutation, and shorter overall survival.

S100A14 plays an important role in the occurrence and development of various human cancers (Zhao et al., 2013; Tanaka et al., 2015; Al-Ashkar and Zetoune, 2020; Li et al., 2020b; Zhu et al., 2021). Gene expression microarray showed that the S100A14 expression in pancreatic cancer tissues was significantly higher than that in corresponding non-tumor tissues (Ohuchida et al., 2006; Zhuang et al., 2021). Studies show that S100A14 protein is often overexpressed in pancreatic ductal adenocarcinoma (PDAC) cell lines and tissues. High expression of S100A14 was significantly correlated with advanced tumor stage and shorter overall survival in patients with pancreatic cancer (Ohuchida et al., 2006; Zhuang et al., 2021). Transient silencing of S100A14 can inhibit the proliferation, clone formation, migration, and invasion of high-level endogenous S100A14 cells (Ohuchida et al., 2006). Continuous knockout of S100A14 by transduction of lentivirus-carrying shRNAs inhibited the formation of subcutaneous tumors in nude mice and made PDAC cells more sensitive to gemcitabine treatment (Ohuchida et al., 2006). Our results indicate that S100A14 is highly expressed both in mRNA and protein levels of pancreatic cancer. High expression of S100A14 was associated with poor overall survival, tumor stage, high TP53 mutation, and drug resistance. In addition, S100A14 was negatively correlated with the infiltration of CD4 + T cells, macrophages, and dendritic cells.

S100A16 is a new member of the S100 protein family, which is functionally expressed in various tumors. The expression of S100A16 in PDAC was up-regulated, and it was negatively correlated with the prognosis of patients with PDAC (Zhuang et al., 2021). S100A16 can promote the proliferation, migration, and invasion of PDAC cells through AKT and ERK1/2 signaling pathways mediated by fibroblast growth factor 19 (FGF19) (Fang et al., 2021). S100A16 can also regulate the cell cycle and apoptosis of pancreatic cancer cells (Fang et al., 2021). Moreover, previous studies have shown that S100A16 can induce EMT and promote the metastasis of human PDAC cells by enhancing the expression of TWIST1 and activating the STAT3 signaling pathway (Li et al., 2021). Similar to these results, our results suggest that S100A16 is highly expressed in pancreatic cancer. High expression of S100A16 was associated with worse overall survival, tumor stage, high TP 53 mutation, and drug resistance.

In the current study, we also found that the expression of the S100s was correlated with the infiltration of certain immune cells in PAAD. Limited studies have been used to elucidate the role of the S100s in immune infiltration. Previous studies have shown that the higher expression of S1000A6/A10/A11/A14/A16 may damage the infiltration and cytotoxicity of $CD8^+$ T cells by stimulating the adhesion-Ras signal transduction pathway in pancreatic cancer (Zhuang et al., 2021). S100A10 expressed on the surface of macrophages plays an important role in the interaction with tumor microenvironment and tumor growth (O'Connell et al., 2010). S100A16 is negatively correlated with immune activity (T cells, cytokines, chemokines, cell adhesion molecules, co-receptors, signal adapters, JAK/STAT pathway) and infiltration (macrophages and T cells), resulting in extensive immunosuppression (Chen et al., 2021). Consequently, our results contain an unconventional function of the S100s and provide new insights into the infiltration of immune cells in PAAD.

It is undeniable that our research has certain shortcomings. First, all the results are mined from the published database, so the quality of the database determines the reliability of our results. Second, our conclusions need to be verified by further experiments and multicenter clinical cohort samples. Third, this study did not explore the detailed mechanism of the predicted S100s members in pancreatic cancer, especially the impact on tumor immune microenvironment and drug sensitivity. To solve these problems, further *in vivo* and *in vitro* experiments are needed to verify these results.

CONCLUSION

In this study, we systemically investigated the transcription and protein level, genetic mutations, prognostic value, enriched signal pathways, and the correlation with immune infiltration cells of each S100s member in PAAD patients. We evaluated the expression, genomic mutation, prognostic value, potential function, relationship with immune infiltration, and involvement in the regulation of drug resistance of the S100s in PAAD. We found that 13 S100s members were overexpressed in PAAD patients. Combined with the transcriptional expression profile, we concluded that high mRNA expression of S100A2/A3/A10/A11/A14/A16 were significantly correlated with the poor OS of PAAD, while a high expression of S100B is favorable to the prognosis of PAAD. Overexpression of S100A2/A3/A4/A5/A6/A10/A11/A13/A14/A16/P in pancreatic cancer is positively correlated with TP53 mutation, while the high expression of S100A1/B/Z is negatively correlated with TP53 mutation. Immuno-infiltration analysis showed that the mRNA levels of S100A10/A11/A14/A16 were significantly correlated with the infiltration degree of macrophages in PAAD. Moreover, PAAD patients expressing high levels of S100A2/A6/A10/A11/A13/A14/A16 were resistant to small molecule drugs. These findings

suggested that S100A2/A10/A11/A14/A16 may be prognostic and therapeutic targets of PAAD. We hope that our results will help to reveal the role of the S100s in cancers and provide new theoretical knowledge for identifying tumor prognostic biomarkers and therapeutic targets.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

XL and NQ designed and performed the research study, analyzed the data. XL, NQ, and LJ critical revision of the manuscript. All authors contributed to manuscript revision, read, and approved the final manuscript.

FUNDING

This project was supported by the Key Special Project for Introduced Talents Team of Southern Marine Scienece and Engineering Guangdong Laboratory (GML2019ZD0204, 2019BT02H594), Foundation of Guangzhou Women and Children's Medical Center (1600074), and the Natural Science Foundation of Guangdong Province (2021A1515011526).

ACKNOWLEDGMENTS

The authors thank the members of participants for taking part in the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.758725/full#supplementary-material

REFERENCES

Adamska, A., Elaskalani, O., Emmanouilidi, A., Kim, M., Abdol Razak, N. B., Metharom, P., et al. (2018). Molecular and Cellular Mechanisms of Chemoresistance in Pancreatic Cancer. *Adv. Biol. Regul.* 68, 77–87. doi:10.1016/j.jbior.2017.11.007

Al-Ashkar, N., and Zetoune, A. B. (2020). S100A14 Serum Level and its Correlation with Prognostic Factors in Breast Cancer. *J. Egypt. Natl. Canc Inst.* 32 (1), 37. doi:10.1186/s43046-020-00048-y

Allgöwer, C., Kretz, A. L., von Karstedt, S., Wittau, M., Henne-Bruns, D., and Lemke, J. (2020). Friend or Foe: S100 Proteins in Cancer. *Cancers* 12 (8), 2037. doi:10.3390/cancers12082037

Allgöwer, C., Kretz, A.-L., von Karstedt, S., Wittau, M., Henne-Bruns, D., and Lemke, J. (2020). Friend or Foe: S100 Proteins in Cancer. *Cancers* 12 (8), 2037. doi:10.3390/cancers12082037

Bachet, J. B., Maréchal, R., Demetter, P., Bonnetain, F., Cros, J., Svrcek, M., et al. (2013). S100A2 Is a Predictive Biomarker of Adjuvant Therapy Benefit in Pancreatic Adenocarcinoma. *Eur. J. Cancer* 49 (12), 2643–2653. doi:10.1016/j.ejca.2013.04.017

Bresnick, A. R., Weber, D. J., and Zimmer, D. B. (2015). S100 Proteins in Cancer. *Nat. Rev. Cancer* 15 (2), 96–109. doi:10.1038/nrc3893

Bulk, E., Sargin, B., Krug, U., Hascher, A., Jun, Y., Knop, M., et al. (2009). S100A2 Induces Metastasis in Non-small Cell Lung Cancer. *Clin. Cancer Res.* 15 (1), 22–29. doi:10.1158/1078-0432.CCR-08-0953

Bydoun, M., Sterea, A., Liptay, H., Uzans, A., Huang, W. Y., Rodrigues, G. J., et al. (2018). S100A10, a Novel Biomarker in Pancreatic Ductal Adenocarcinoma. *Mol. Oncol.* 12 (11), 1895–1916. doi:10.1002/1878-0261.12356

Bydoun, M., Sterea, A., Weaver, I. C. G., Bharadwaj, A. G., and Waisman, D. M. (2018). A Novel Mechanism of Plasminogen Activation in Epithelial and Mesenchymal Cells. *Sci. Rep.* 8 (1), 14091. doi:10.1038/s41598-018-32433-y

Camara, R., Ogbeni, D., Gerstmann, L., Ostovar, M., Hurer, E., Scott, M., et al. (2020). Discovery of Novel Small Molecule Inhibitors of S100P with *In Vitro* Anti-metastatic Effects on Pancreatic Cancer Cells. *Eur. J. Med. Chem.* 203, 112621. doi:10.1016/j.ejmech.2020.112621

Chen, H., Xu, C., Jin, Q., and Liu, Z. (2014). S100 Protein Family in Human Cancer. *Am. J. Cancer Res.* 4 (2), 89–115.

Chen, T., Xia, D. M., Qian, C., and Liu, S. R. (2021). Integrated Analysis Identifies S100A16 as a Potential Prognostic Marker for Pancreatic Cancer. *Am. J. Transl Res.* 13 (5), 5720–5730.

de Visser, K. E., Eichten, A., and Coussens, L. M. (2006). Paradoxical Roles of the Immune System during Cancer Development. *Nat. Rev. Cancer* 6 (1), 24–37. doi:10.1038/nrc1782

Donato, R. (1999). Functional Roles of S100 Proteins, Calcium-Binding Proteins of the EF-Hand Type. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1450 (3), 191–231. doi:10.1016/S0167-4889(99)00058-0

Donato, R., R. Cannon, B., Sorci, G., Riuzzi, F., Hsu, K., Weber, D. J., et al. (2013). Functions of S100 Proteins. *Curr. Mol. Med.* 13 (1), 24–57. doi:10.2174/156652413804486214

Donato, R. (2001). S100: a Multigenic Family of Calcium-Modulated Proteins of the EF-Hand Type with Intracellular and Extracellular Functional Roles. *Int. J. Biochem. Cel Biol.* 33 (7), 637–668. doi:10.1016/s1357-2725(01)00046-2

Engelkamp, D., Schafer, B. W., Mattei, M. G., Erne, P., and Heizmann, C. W. (1993). Six S100 Genes Are Clustered on Human Chromosome 1q21: Identification of Two Genes Coding for the Two Previously Unreported Calcium-Binding Proteins S100D and S100E. *Proc. Natl. Acad. Sci.* 90 (14), 6547–6551. doi:10.1073/pnas.90.14.6547

Fang, D., Zhang, C., Xu, P., Liu, Y., Mo, X., Sun, Q., et al. (2021). S100A16 Promotes Metastasis and Progression of Pancreatic Cancer through FGF19-Mediated AKT and ERK1/2 Pathways. *Cell Biol Toxicol* 37, 555–571. doi:10.1007/s10565-020-09574-w

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). doi:10.1126/scisignal.2004088

Goh, J. Y., Feng, M., Wang, W., Oguz, G., Yatim, S. M. J. M., Lee, P. L., et al. (2017). Chromosome 1q21.3 Amplification Is a Trackable Biomarker and Actionable Target for Breast Cancer Recurrence. *Nat. Med.* 23 (11), 1319–1330. doi:10.1038/nm.4405

Gross, S. R., Sin, C. G. T., Barraclough, R., and Rudland, P. S. (2014). Joining S100 Proteins and Migration: for Better or for Worse, in Sickness and in Health. *Cell. Mol. Life Sci.* 71 (9), 1551–1579. doi:10.1007/s00018-013-1400-7

Ilic, M., and Ilic, I. (2016). Epidemiology of Pancreatic Cancer. *Wjg* 22 (44), 9694–9705. doi:10.3748/wjg.v22.i44.9694

Jamieson, N. B., Denley, S. M., Logue, J., MacKenzie, D. J., Foulis, A. K., Dickson, E. J., et al. (2011). A Prospective Comparison of the Prognostic Value of Tumor- and Patient-Related Factors in Patients Undergoing Potentially Curative Surgery for Pancreatic Ductal Adenocarcinoma. *Ann. Surg. Oncol.* 18 (8), 2318–2328. doi:10.1245/s10434-011-1560-3

Kamisawa, T., Wood, L. D., Itoi, T., and Takaori, K. (2016). Pancreatic Cancer. *The Lancet* 388 (10039), 73–85. doi:10.1016/S0140-6736(16)00141-0

Lai, E., Puzzoni, M., Ziranu, P., Pretta, A., Impera, V., Mariani, S., et al. (2019). New Therapeutic Targets in Pancreatic Cancer. *Cancer Treat. Rev.* 81, 101926, 2019 . Artn 101926. doi:10.1016/J.Ctrv.2019.101926

Leclerc, E., Fritz, G., Vetter, S. W., and Heizmann, C. W. (2009). Binding of S100 Proteins to RAGE: an Update. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1793 (6), 993–1007. doi:10.1016/j.bbamcr.2008.11.016

Lei, X., Lei, Y., Li, J. K., Du, W. X., Li, R. G., Yang, J., et al. (2020). Immune Cells within the Tumor Microenvironment: Biological Functions and Roles in Cancer Immunotherapy. *Cancer Lett.* 470, 126–133. doi:10.1016/j.canlet.2019.11.009

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48 (W1), W509–W514. doi:10.1093/nar/gkaa407

Li, T., Ren, T., Huang, C., Li, Y., Yang, P., Che, G., et al. (2021). S100A16 Induces Epithelial-Mesenchymal Transition in Human PDAC Cells and Is a New Therapeutic Target for Pancreatic Cancer Treatment that Synergizes with Gemcitabine. *Biochem. Pharmacol.* 189, 114396. doi:10.1016/j.bcp.2020.114396

Li, X., Wang, M., Gong, T., Lei, X., Hu, T., Tian, M., et al. (2020). A S100A14-Ccl2/cxcl5 Signaling axis Drives Breast Cancer Metastasis. *Theranostics* 10 (13), 5687–5703. doi:10.7150/thno.42087

Liu, C. J., Hu, F. F., Xia, M. X., Han, L., Zhang, Q., and Guo, A. Y. (2018). GSCALite: a Web Server for Gene Set Cancer Analysis. *Bioinformatics* 34 (21), 3771–3772. doi:10.1093/bioinformatics/bty411

Liu, D., Rudland, P. S., Sibson, D. R., Platt-Higgins, A., and Barraclough, R. (2000). Expression of Calcium-Binding Protein S100A2 in Breast Lesions. *Br. J. Cancer* 83 (11), 1473–1479. doi:10.1054/bjoc.2000.1488

Ma, N., Zhu, L., Yang, L., Cui, Y., and Zhan, Y. (2019). Prognostic Values of S100 Family mRNA Expression in Ovarian Cancer. *Cbm* 25 (1), 67–78. doi:10.3233/CBM-182276

Marchesi, F., Piemonti, L., Mantovani, A., and Allavena, P. (2010). Molecular Mechanisms of Perineural Invasion, a Forgotten Pathway of Dissemination and Metastasis. *Cytokine Growth Factor. Rev.* 21 (1), 77–82. doi:10.1016/j.cytogfr.2009.11.001

Marenholz, I., Heizmann, C. W., and Fritz, G. (2004). S100 Proteins in Mouse and Man: from Evolution to Function and Pathology (Including an Update of the Nomenclature). *Biochem. Biophysical Res. Commun.* 322 (4), 1111–1122. doi:10.1016/j.bbrc.2004.07.096

Marenholz, I., Lovering, R. C., and Heizmann, C. W. (2006). An Update of the S100 Nomenclature. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1763 (11), 1282–1283. doi:10.1016/j.bbamcr.2006.07.013

Mishra, N. K., Southekal, S., and Guda, C. (2019). Survival Analysis of Multi-Omics Data Identifies Potential Prognostic Markers of Pancreatic Ductal Adenocarcinoma. *Front. Genet.* 10, 10. doi:10.3389/fgene.2019.00624

Moravkova, P., Kohoutova, D., Rejchrt, S., Cyrany, J., and Bures, J. (20162016). Role of S100 Proteins in Colorectal Carcinogenesis. *Gastroenterol. Res. Pract.* 2016, 1–7. doi:10.1155/2016/2632703

O'Connell, P. A., Surette, A. P., Liwski, R. S., Svenningsson, P., and Waisman, D. M. (2010). S100A10 Regulates Plasminogen-dependent Macrophage Invasion. *Blood* 116 (7), 1136–1146. doi:10.1182/blood-2010-01-264754

Ohuchida, K., Mizumoto, K., Ohhashi, S., Yamaguchi, H., Konomi, H., Nagai, E., et al. (2006). S100A11, a Putative Tumor Suppressor Gene, Is Overexpressed in Pancreatic Carcinogenesis. *Clin. Cancer Res.* 12 (18), 5417–5422. doi:10.1158/1078-0432.CCR-06-0222

Rehman, I., Cross, S. S., Catto, J. W. F., Leiblich, A., Mukherjee, A., Azzouzi, A.-R., et al. (2005). Promoter Hyper-Methylation of Calcium Binding Proteins S100A6 and S100A2 in Human Prostate Cancer. *Prostate* 65 (4), 322–330. doi:10.1002/pros.20302

Saiki, Y., and Horii, A. (2019). Multiple Functions of S100A10, an Important Cancer Promoter. *Pathol. Int.* 69 (11), 629–636. doi:10.1111/pin.12861

Salama, I., Malone, P. S., Mihaimeed, F., and Jones, J. L. (2008). A Review of the S100 Proteins in Cancer. *Eur. J. Surg. Oncol. (Ejso)* 34 (4), 357–364. doi:10.1016/j.ejso.2007.04.009

Schäfer, B. W., Wicki, R., Engelkamp, D., Mattei, M.-g., and Heizmann, C. W. (1995). Isolation of a YAC Clone Covering a Cluster of Nine S100 Genes on Human Chromosome 1q21: Rationale for a New Nomenclature of the S100 Calcium-Binding Protein Family. *Genomics* 25 (3), 638–643. doi:10.1016/0888-7543(95)80005-7

Sighoko, D., Curado, M. P., Bourgeois, D., Mendy, M., Hainaut, P., and Bah, E. (2011). Increase in Female Liver Cancer in the Gambia, West Africa: Evidence from 19 Years of Population-Based Cancer Registration (1988-2006). *PLoS One* 6 (4), e18415. doi:10.1371/journal.pone.0018415

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–d613. doi:10.1093/nar/gky1131

Tanaka, M., Ichikawa-Tomikawa, N., Shishito, N., Nishiura, K., Miura, T., Hozumi, A., et al. (2015). Co-expression of S100A14 and S100A16 Correlates with a Poor Prognosis in Human Breast Cancer and Promotes Cancer Cell Invasion. *BMC Cancer* 15, 53. doi:10.1186/s12885-015-1059-6

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a Web Server for Cancer and normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res.* 45 (W1), W98–W102. doi:10.1093/nar/gkx247

Thul, P. J., and Lindskog, C. (2018). The Human Protein Atlas: A Spatial Map of the Human Proteome. *Protein Sci.* 27 (1), 233–244. doi:10.1002/pro.3307

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 (Web Server issue), W214–W220. doi:10.1093/nar/gkq537

Wolf, S., Haase-Kohn, C., and Pietzsch, J. (2011). S100A2 in Cancerogenesis: a Friend or a Foe? *Amino Acids* 41 (4), 849–861. doi:10.1007/s00726-010-0623-2

Xiao, M.-B., Jiang, F., Ni, W.-K., Chen, B.-Y., Lu, C.-H., Li, X.-Y., et al. (2012). High Expression of S100A11 in Pancreatic Adenocarcinoma Is an Unfavorable Prognostic Marker. *Med. Oncol.* 29 (3), 1886–1891. doi:10.1007/s12032-011-0058-y

Xiao, M., Li, T., Ji, Y., Jiang, F., Ni, W., Zhu, J., et al. (2018). S100A11 Promotes Human Pancreatic Cancer PANC-1 Cell Proliferation and Is Involved in the PI3K/AKT Signaling Pathway. *Oncol. Lett.* 15 (1), 175–182. doi:10.3892/ol.2017.7295

Xue, T. C., Zhang, B. H., Ye, S. L., and Ren, Z. G. (2015). Differentially Expressed Gene Profiles of Intrahepatic Cholangiocarcinoma, Hepatocellular Carcinoma, and Combined Hepatocellular-Cholangiocarcinoma by Integrated Microarray Analysis. *Tumor Biol.* 36 (8), 5891–5899. doi:10.1007/s13277-015-3261-1

Yao, R., Davidson, D. D., Lopez-Beltran, A., MacLennan, G. T., Montironi, R., and Cheng, L. (2007). The S100 Proteins for Screening and Prognostic Grading of Bladder Cancer. *Histol. Histopathol* 22 (9), 1025–1032. doi:10.14670/HH-22.1025

Zhang, C., Zou, Y., Zhu, Y., Liu, Y., Feng, H., Niu, F., et al. (2021). Three Immune-Related Prognostic mRNAs as Therapeutic Targets for Pancreatic Cancer. *Front. Med.* 8. , 2021 Artn 649326. doi:10.3389/Fmed.2021.649326

Zhang, L., Fogg, D. K., and Waisman, D. M. (2004). RNA Interference-Mediated Silencing of the S100A10 Gene Attenuates Plasmin Generation and Invasiveness of Colo 222 Colorectal Cancer Cells. *J. Biol. Chem.* 279 (3), 2053–2062. doi:10.1074/jbc.M310357200

Zhang, S., Wang, Z., Liu, W., Lei, R., Shan, J., Li, L., et al. (2017). Distinct Prognostic Values of S100 mRNA Expression in Breast Cancer. *Sci. Rep.* 7, 39786. doi:10.1038/srep39786

Zhao, F.-T., Jia, Z.-S., Yang, Q., Song, L., and Jiang, X. J. (2013). S100A14 Promotes the Growth and Metastasis of Hepatocellular Carcinoma. *Asian Pac. J. Cancer Prev.* 14 (6), 3831–3836. doi:10.7314/apjcp.2013.14.6.3831

Zhu, H., Gao, W., Li, X., Yu, L., Luo, D., Liu, Y., et al. (2021). S100A14 Promotes Progression and Gemcitabine Resistance in Pancreatic Cancer. *Pancreatology* 21 (3), 589–598. doi:10.1016/j.pan.2021.01.011

Zhuang, H., Chen, X., Dong, F., Zhang, Z., Zhou, Z., Ma, Z., et al. (2021). Prognostic Values and Immune Suppression of the S100A Family in Pancreatic Cancer. *J. Cel Mol Med* 25 (6), 3006–3018. doi:10.1111/jcmm.16343

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

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fgene.2021.758981

Ferroptosis-Related Gene Signature Predicts the Prognosis of Skin Cutaneous Melanoma and Response to Immunotherapy

Ziqian Xu[1†], Yihui Xie[1†], Yaqi Mao[1], Juntao Huang[2], Xingyu Mei[1], Jun Song[1], Yue Sun[1], Zhixian Yao[3] and Weimin Shi[1]**

[1]Department of Dermatology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]Department of Otolaryngology Head and Neck Surgery, Ningbo Medical Center (Ningbo Lihuili Hospital), The Affiliated Lihuili Hospital of Ningbo University, Ningbo, China, [3]Department of Urology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Yun Hak Kim,
Pusan National University, South Korea
Haider H. Dar,
University of Pittsburgh, United States

***Correspondence:**
Zhixian Yao
yzxbrooklyn@sjtu.edu.cn
Weimin Shi
swm666042@163.com

[†]These authors have contributed equally to this work

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

Received: 15 August 2021
Accepted: 14 October 2021
Published: 03 November 2021

Citation:
Xu Z, Xie Y, Mao Y, Huang J, Mei X, Song J, Sun Y, Yao Z and Shi W (2021) Ferroptosis-Related Gene Signature Predicts the Prognosis of Skin Cutaneous Melanoma and Response to Immunotherapy. Front. Genet. 12:758981. doi: 10.3389/fgene.2021.758981

Ferroptosis is a non-apoptotic regulated cell death process, and much research has indicated that ferroptosis can induce the non-apoptotic death of tumor cells. Ferroptosis-related genes are expected to become a biological target for cancer treatment. However, the regulation of ferroptosis-related genes in skin cutaneous melanoma (SKCM) has not been well studied. In the present study, we conducted a systematic analysis of SKCM based on RNA sequencing data and clinical data obtained from The Cancer Genome Atlas (TCGA) database and the FerrD database. SKCM patients from the GSE78220 and MSKCC cohorts were used for external validation. Applying consensus clustering on RNA sequencing data from TCGA the generated ferroptosis subclasses of SKCM, which were analyzed based on the set of differentially expressed ferroptosis-related genes. Then, a least absolute shrinkage and selection operator (LASSO)-Cox regression was used to construct an eight gene survival-related linear signature. The median cut-off risk score was used to divide patients into high- and low-risk groups. The time-dependent receiver operating characteristic curve was used to examine the predictive power of the model. The areas under the curve of the signature at 1, 3, and 5 years were 0.673, 0.716, and 0.746, respectively. Kaplan-Meier survival analysis showed that the prognosis of high-risk patients was worse than that of low-risk patients. Univariate and multivariate Cox regression analyses showed that the risk signature was a robust independent prognostic indicator. By incorporating risk scores with tumor staging, a nomogram was constructed to predict prognostic outcomes for SKCM patients. In addition, the immunological analysis showed different immune cell infiltration patterns. Programmed-death-1 (PD-1) immunotherapy showed more significant benefits in the low-risk group than in the high-risk group. In summary, a model based on ferroptosis-related genes can predict the prognosis of SKCM and could have a potential role in guiding targeted therapy of SKCM.

Keywords: melanoma, ferroptosis, signature, prognosis, risk score

INTRODUCTION

Human skin cutaneous melanoma (SKCM) is a highly malignant tumor derived from melanocytes, that is prone to occur in adults. The mortality rate of melanoma is up to 75% (Davis et al., 2019). Early-stage, localized melanoma can be curable if appropriate and sufficient treatment is administered (Coit et al., 2013) ; however, the tumor tends to metastasize and spread to other parts of the body (Aris and Barrio, 2015), once metastasized, the 5-years survival rate decreases to only 15% (Coit et al., 2013). SKCM is now the third most commonly diagnosed cancer in the United States, with an estimated 192,000 new cases in 2019. The incidence is six times higher than 40 years ago (Welch et al., 2021). Surgical resection is associated with a satisfactory prognosis for melanoma in the early stages, while treatment of metastatic melanoma mainly relies on immunotherapy (Leonardi et al., 2020). Many studies have explored the relationship between cancer cells, the tumor microenvironment (TME), and the immune system. However, not all treatments that block immune suppression control points are effective for all patients (Rodríguez-Cerdeira et al., 2017). Therefore, identifying robust predictive biomarkers for both clinical prognosis and treatment response is essential.

In recent years, tumor ferroptosis has gathered much interest. Iron is involved in many biochemical processes in the human body, including oxygen transport, various biosyntheses, and the electron transport chain (as a cofactor), playing crucial roles in cell survival (Bogdan et al., 2016). In mitochondrial oxidative phosphorylation, cells produce reactive oxygen species (ROS) while generating ATP. Excessive ROS levels will lead to oxidative stress, directly or indirectly damaging macromolecules, such as proteins, nucleic acids, and lipids, leading to cell damage or death (Yu et al., 2017). Ferroptosis is an iron-dependent form of programmed cell death, which differs from apoptosis, cell necrosis, and autophagy.

The mechanism of ferroptosis is based on influencing iron metabolism in cells, resulting in intracellular ROS production and excessive oxidation of polyunsaturated fatty acids (Dixon et al., 2012; Battaglia et al., 2020). Ferroptosis is mainly regulated by system X_C^- and glutathione peroxidase 4(GPX4). System X_C^- is a Na^+-dependent cysteine—glutamic acid exchange transporter in the membrane, which completes the intracellular and extracellular glutamate—cysteine exchange (Sato et al., 1999; Bridges et al., 2012). Cell uptake of cysteine is a crucial step in glutathione (GSH) synthesis, and the generation and maintenance of GSH is the key to protecting cells from ROS damage (Yu et al., 2017). GPX4 is an enzyme that decomposes H_2O_2 and organic peroxides into water or the corresponding alcohols, and GSH is an indispensable cofactor in its activation (Ursini et al., 1995). The ferroptosis inducers, erastin and buthionine sulfoximine, reduce the activity of GPX4 and increase the level of ROS in the cytoplasm and lipid (Yang et al., 2014), leading to cell ferroptosis. Ferroptosis is considered as an adaptive process to eliminate malignant cells damaged by nutrient deficiency, infection, or other stress from the body (Mou et al., 2019). A number of studies have demonstrated that ferroptosis plays a role in ischemia-reperfusion injury, cancer, and other diseases. At present, sorafenib and other ferroptosis inducers are used for the treatment of cancer in clinical practice (Fearnhead et al., 2017).

Non-cellular components in the TME may reprogram tumor initiation and invasiveness, but the relationship between iron metabolism and TME is still unclear (Sacco et al., 2021). Tumor-associated macrophages loaded with iron can promote the production of ROS and pro-inflammatory cytokines (tumor necrosis factor-α and interleukin-6), thereby inducing tumor cell death in lung cancer (Costa da Silva et al., 2017). Moreover, new evidence suggests that immune checkpoint blockade decreases tumor growth in a ferroptosis-dependent manner in animal models (Tang et al., 2020), which is considered to be related to antitumor immunity. In addition, the immunotherapy-activated $CD8^+$ T lymphocytes induce ferroptosis in cancer cells by downregulating genes (*SLC7A11* and *SLC3A2*) that encode two subunits of the X_C^- system, and the molecular basis behind this phenomenon may be related to interferon (IFN)-γ-mediated transcriptional repression of *SLC7A11* and *SLC3A2* (Wang et al., 2019). Consequently, identifying biomarkers of iron metabolism within the TME may aid the development of effective cancer treatment strategies.

Focusing on melanoma, many studies have found ferroptosis regulators, as well as traditional ferroptosis-inducing agents. It was found that miR-9 reduced erastin- and RSL3-induced ferroptosis in melanoma cells, and knockout of miR-9 could cause ferroptosis in melanoma cells (Zhang et al., 2018). In addition, the inactivation of miR-137 enhances the anti-melanoma activity of erastin by increasing ferroptosis (Luo et al., 2018). However, it is still unknown whether ferroptosis-related genes are related to the prognosis of SKCM patients.

In the present study, we conducted a comprehensive analysis based on transcript and clinical data obtained from The Cancer Genome Atlas (TCGA) and the FerrD databases. We constructed a predictive model on account of eight ferroptosis-related genes. The model can be regarded as an independent predictor of overall survival (OS) of SKCM. A nomogram was established to further explore the prognosis of SKCM based on risk score and tumor stage. Furthermore, we analyzed the mutational differences of ferroptosis-related genes in high-risk and low-risk groups and the associations between the ferroptosis-related risk score and immune cell infiltration patterns and immunotherapy. External validation cohorts were used to verify the ferroptosis-related risk score predicting the response to immunotherapy of the two subgroups.

MATERIALS AND METHODS

Data Collection

The RNA-sequencing (RNA-Seq) and genomic data for SKCM were downloaded from UCSC Xena (http: //xenabrowser.net/), and the clinical data were downloaded via the R package "TCGAbiolinks". The samples were screened to retain those, including survival status and survival time. A total of 457 samples were obtained. Ferroptosis-related genes were downloaded from the FerrDb database (www.zhounan.org/ferrdb/operations/download.html) and published

literature (Liang et al., 2020) and merged. Thus, 268 ferroptosis-related genes were obtained in the analysis (**Supplementary Table S1**), from which an expression matrix of the ferroptosis-related genes in the SKCM RNA-Seq data was obtained as a candidate gene set expression matrix.

Additionally, the gene expression profile and clinical data of the two independent cohorts was obtained, GSE78220 (Snyder et al., 2014), from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) and MSKCC cohorts (Newman et al., 2015), from cbioportal (http://www.cbioportal.org/study/summary?id=skcm_mskcc_2014). These two independent cohorts were used as the external validation cohorts.

SKCM Subclass Identification

The R package "ConsensusClusterPlus" was used to apply consensus clustering analysis to the candidate gene set and to classify results by dividing the samples into two clusters. Survival analysis showed that the prognosis based on the two clusters was significantly different ($p < 0.001$). The differences in ferroptosis-related genes between the two clusters were analyzed using R package "DESeq2" ($p < 0.05$ and $|logFC| > 0.5$). There were 116 differentially expressed genes (DEGs) identified as a new candidate gene set, containing 63 downregulated genes and 53 upregulated genes. Subsequently, the DEGs were annotated by gene ontology (GO) and using Kyoto Encyclopedia of Genes and Genomes (KEGG) data in R package "clusterProfiler". Finally, univariate cox regression analysis was performed on the 116 DEGs ($p \leq 0.01$), for which a total of 28 prognostic-related genes were identified for further analysis.

Construction and Validation of a Ferroptosis-Related Risk Signature

Using the R package "survival", univariate Cox regression analysis was performed on the prognostic-related genes, where $p \leq 0.001$ was regarded as statistically significant. The least absolute shrinkage and selection operator (LASSO) method was used to eliminate overfitting with the R package "glmnet". Then, multivariate Cox regression was used to select the independent prognostic factors. Meanwhile, the correlation coefficients of these genes were calculated. Using the these genes and corresponding correlation coefficients, a prognosis-related model was constructed. Then, the samples were divided into high-risk and low-risk groups by the median of the risk scores. The receiver operating characteristic (ROC) curve was used to evaluate the survival prediction ability of the model, and the Kaplan–Meier (K-M) method was used to analyze the survival difference between the two risk subgroups. Next, a nomogram was constructed based on the prognostic signature using R package "RMS", and a calibration plot was drawn to evaluate the consistency between the prognostic model's actual and predicted survival rates.

Analysis of Somatic Mutation and Immunotherapy Differences Between High- and Low-Risk Groups

"CIBERSORT" (Wilkerson and Hayes, 2010) in R software was used to analyze the differences in immune cells infiltration between high- and low-risk groups. Meanwhile, the correlation between immune cells and risk score was analyzed. Concomitantly, the difference in response to PD-1 immunotherapy between high- and low-risk groups was verified in GSE78220 and MSKCC cohorts. The mutation differences of ferroptosis-related genes in high- and low-risk groups were analyzed using the R package "maftools".

Statistical Analysis

Student's *t*-tests and Wilcoxon rank-sum tests were used to compare the differences between the high- and low-risk groups. The Kruskal-Wallis test was used for comparisons of prognoses between groups. The K-M method was used to generate survival curves for the subgroups in each data set. The log-rank test determines the statistical significance of differences. Univariate and multivariate Cox regression analyses were applied to define independent prognostic-related factors. The ROC curve was visualized using the R package "pROC", and the areas under the curve (AUC) and confidence intervals were calculated to evaluate the model's accuracy in predicting prognosis. All statistical analyzses were performed using SPSS (version 23.0) and R software (version 3.6.1). For each analysis, statistical significance was set at $p < 0.05$.

RESULTS

Identifying Two Subtypes and Their Distinct Ferroptosis Patterns

To describe our research systematically and comprehensively, a flow chart is shown in **Figure 1A**. After filtering out normal samples and samples without survival data, we obtained 457 gene expression profiles of SKCM samples from the TCGA dataset. From the FerrDb database and previous literature, 268 ferroptosis-related genes were identified. A total of 84 ferroptosis-related genes were selected. Consensus clustering (CC) was used to divide melanoma samples into two clusters (cluster 1 and cluster 2). In cancer research, unsupervised class discovery classifies intrinsic populations with common biological characteristics, which may exist but are unknown. The CC method is a type of unsupervised class discovery, which provides quantitative and visual stability evidence for estimating the number in a dataset (Wilkerson and Hayes, 2010). After CC, the optimal total cluster number was set to $k = 2$ (the two subclasses were designated as cluster 1 and cluster 2). When $k = 2$, the consensus matrix heat map maintained the clearest cluster partition, indicating the clustering having the highest consensus (**Figure 1B**; **Supplementary Figure S1A–C**). The k value determined by a cumulative density function (CDF) plot means maximum stability, at which the distribution reaches an approximate maximum (**Supplementary**

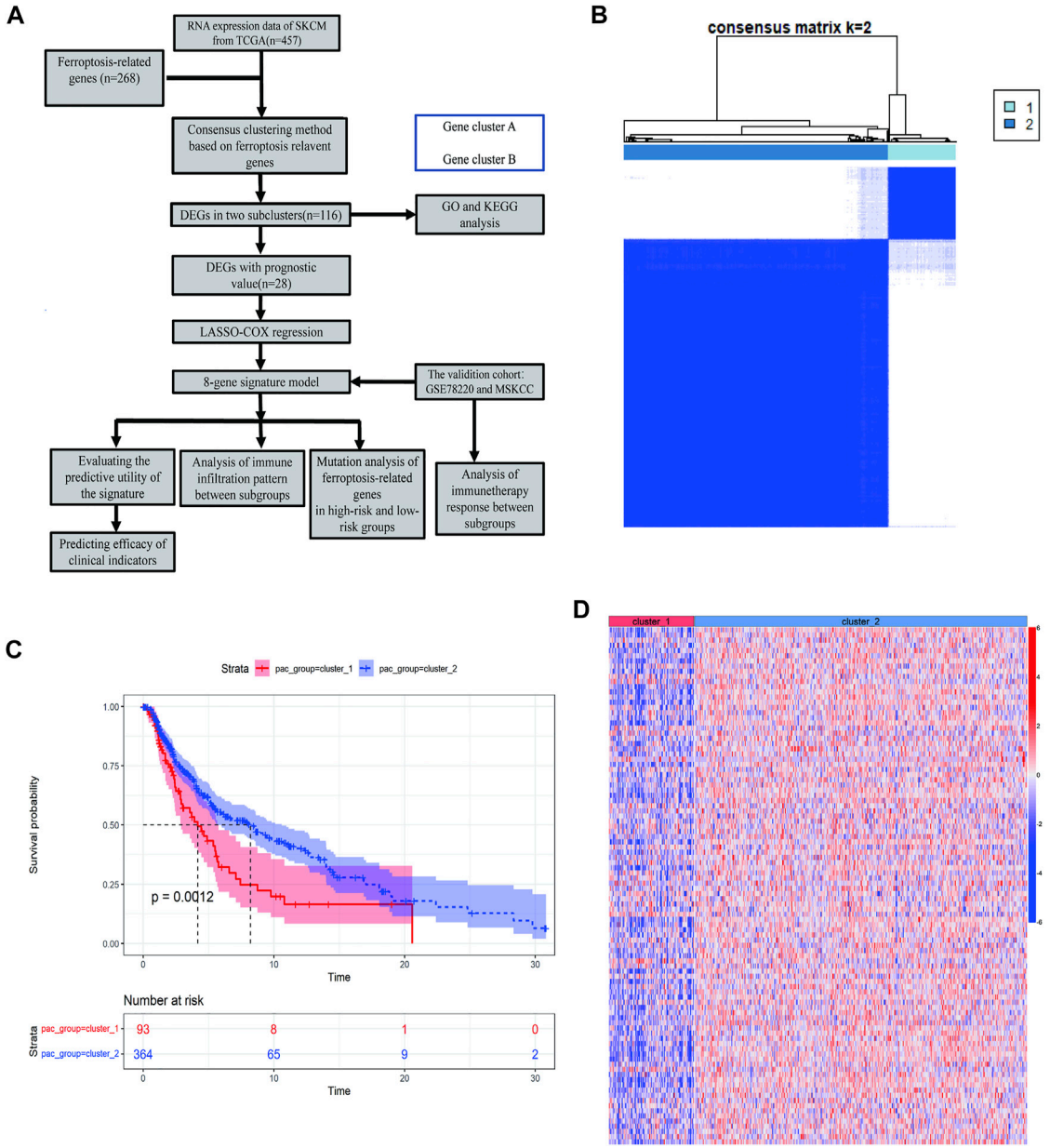

FIGURE 1 | Identification of melanoma subtypes using consensus clustering in the ferroptosis set. **(A)** Flow chart of the study. **(B)** Consensus clustering applied on 268 ferroptosis-related genes. Samples were divided into cluster 1 and cluster 2. **(C)** Survival analysis of samples in Clusters 1 and 2 in TCGA cohort. **(D)** Heat map of ferroptosis-related genes expression in different clusters.

Figure S1D). The survival curve showed a significant difference between cluster 1 and cluster 2 (**Figure 1C**). The gene expression heat map shows the expression of ferroptosis-related genes between the two clusters (**Figure 1D**).

We then analyzed the differences in the expression of ferroptosis-related genes between the two clusters and identified 116 DEGs. GO enrichment and KEGG pathway analyses were performed on the DEGs to help clarify the related biological functions and pathways of the ferroptosis-related genes. The most abundant biological processes (BP) involved include response to oxidative stress, cellular response to chemical stress, cellular response to oxidative stress, response to nutrient levels, cellular response to metal ions, ROS metabolic process, multicellular organismal homeostasis, response to reactive oxygen species, and cellular response to extracellular stimulus (**Supplementary Figure S2A**). In terms of molecular functions (MF), the DEGs were mainly enriched in iron ion binding and oxidoreductase

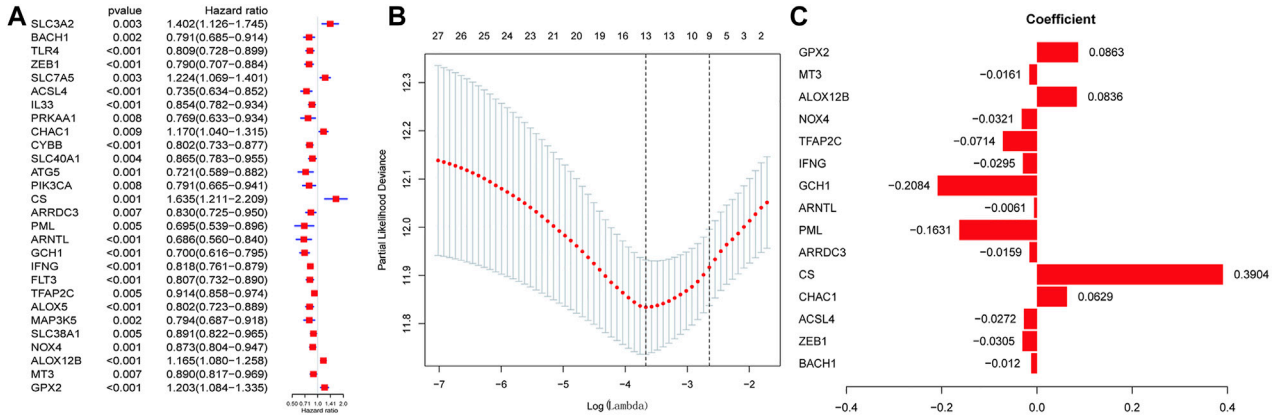

FIGURE 2 | Discovering ferroptosis-related genes with the most significant prognostic values. **(A)** Univariate Cox Analysis of the candidate ferroptosis-related gene set. **(B)** Cross-validation for tuning parameter selection in the proportional hazards model. **(C)** Lasso coefficient spectrum of ferroptosis-related genes related to prognosis in SKCM.

activity (**Supplementary Figure S2B**). The main cellular component terminologies identified were protein kinase complex, autophagosome, phagophore assembly site, secondary lysosome, nucleotide-activated protein kinase complex, and promyelocytic leukemia nuclear (PML) body (**Supplementary Figure S2C**). Interestingly, KEGG enriched ferroptosis and some tumorigenesis-related pathways, such as melanoma (**Supplementary Figure S2D**). Thus, these findings indicate that the occurrence of ferroptosis is related to tumorigenesis.

Discovering Ferroptosis-Related Genes With the Most Significant Prognostic Values

Next, the prognostic effect of ferroptosis-related genes in SKCM was examined. Among the SKCM patients in the expression matrix of the previous candidate gene set, 28 prognostic-related genes were selected through the univariate Cox regression analysis ($p < 0.001$) for further analysis (**Supplementary Figure S3**). There was a statistically significant difference in the expression of prognostic-related factors between the two clusters, where *ACSL1 ALOXS ARNTL CTBB FLT3 GCH1 IFNG IL33 TLR4*, and *ZEB1* were strongly significantly different ($p \leq 0.001$) (**Supplementary Figure S3A**). Overall, prognostic-related factors impact on the survival and prognosis in the two clusters, especially *GCH1 CTBB TLR4 ALOX5 FLT3*, and *IFNG* ($p < 0.0001$) (**Supplementary Figure S3B**).

Construction of Predictive Utility Evaluation of the Ferroptosis-Related Gene Signature

We then performed LASSO regression analysis to establish an optimal survival-related linear risk assessment model (**Figures 2A–C**), which included eight genes. The purpose of LASSO regression analysis is to minimize the risk of overfitting. The risk scores of samples were calculated from regression coefficients generated by multivariate Cox regression analysis.

Through the further screening, eight genes were finally selected, where the formula for calculating the risk score was as follows: risk score = [*CHAC1* × 0.125855936071576 + *CS* × 0.560660638020698 + *PML* × (−0.30765625698674) + *GCH1* × (−0.308787668222114) + *TFAP2C* × (−0.110257841756473) + *NOX4* × (−0.06854182999606) + *ALOX12B* × 0.107841778270567 + *GPX* × 0.125495051893846]

The results obtained were used to stratify patients into low-risk ($n = 228$) and high-risk ($n = 228$) groups based on the median risk score. Clinical information for samples is shown in **Table 1**. We then evaluated the predictive utility of the ferroptosis-related gene signature. The K-M analysis showed that the high-risk group had a worse survival probability than the low-risk group in the TCGA cohort (**Figure 3A**). ROC curves were used to evaluate the sensitivity of the risk model prediction, showing AUCs in the TCGA cohort for 1, 3, and 5 years, which were 0.673, 0.716, and 0.746, respectively (**Figure 3B**). GSE78220 and MSKCC cohorts were used to verify the validation of TGGA database risk scores. K-M survival analysis indicated that the low-risk subgroup had better overall survival (**Figures 3C,D**; **Table 2**, **3**). Univariate and multivariate Cox regression analyses were performed on the TCGA cohort to test whether the eight-gene signature was a suitable independent prognostic indicator. Univariate Cox regression revealed that high-risk scores, T stage, N stage, and age were associated with poor survival prognosis ($p < 0.001$; **Figure 3E**). Through multivariate Cox regression, T stage, N stage, and risk score were independent predictors of melanoma ($p < 0.05$; **Figure 3F**). The patients were ranked from left to right according to the increasing risk scores, and a scatter plot shows the distribution of patients according to their risk scores (**Figure 3G**). A heat map presented differentially expressed ferroptosis-related genes between the high-risk and low-risk groups (**Figure 3H**).

TABLE 1 | Characteristics of patients in low- and high-risk scores in TGGA cohort.

Characteristic	High, *N* = 228[a]	Low, *N* = 228[a]
Age	60 (51, 71)	56 (45, 68)
Gender		
Female	90 (39%)	82 (36%)
Male	138 (61%)	146 (64%)
Bmi	26.1 (23.0, 30.2)	28.0 (24.9, 33.3)
Unknown	97	121
M.stage		
M0	206 (95%)	200 (93%)
M1	10 (4.6%)	14 (6.5%)
Unknown	12	14
N.stage		
N0	117 (53%)	109 (50%)
N1	31 (14%)	42 (19.8%)
N2	23 (10.4%)	25 (11.6%)
N3	29 (13%)	27 (12%)
NX	21 (9.5%)	13 (6.0%)
Unknown	7	12
T.stage		
T0	6 (2.8%)	17 (8.1%)
T1	13 (5.9%)	27 (12.8%)
T2	37 (17%)	40 (19%)
T3	42 (19.3%)	48 (22.8%)
T4	97 (44%)	51 (24.1%)
Tis	6 (2.8%)	1 (0.5%)
TX	17 (7.8%)	27 (13%)
Unknown	10	17
TCGA_subtype		
-	12 (7.6%)	6 (3.6%)
BRAF_Hotspot_Mutants	58 (37%)	88 (53%)
NF1_Any_Mutants	13 (8.2%)	12 (7.3%)
RAS_Hotspot_Mutants	47 (30%)	44 (27%)
Triple_WT	28 (18%)	15 (9.1%)
Unknown	70	63
Sample_type		
Additional Metastatic	1 (0.4%)	0 (0%)
Metastatic	151 (66%)	206 (90%)
Primary Tumor	76 (33%)	22 (9.6%)
Braf		
False	112 (49%)	83 (36%)
True	116 (51%)	145 (64%)

[a]*Median (IQR);* n *(%).*

Incorporating Ferroptosis Risk Scores into the Nomogram and Validation of its Clinical Benefit

Subsequently, a nomogram was created, which predicted the probability of specific clinical outcomes or events based on the values of multiple variables. The factors for the establishment of the nomogram included the risk scores and tumor stages. In the nomogram, columnar height represents the distribution and number of patients. For instance, for a patient in the high-risk group, the tumor stage was N0 and T4, and their total score was 1.93. The probability of their survival time being <1 year was 0.091, <3 years was 0.473, and <5 years was 0.645 (**Figure 4A**). The ROC curve indicates that the N-stage, T-stage, and risk scores result in better predictions than other clinical futures (**Figure 4B**). The calculated C index was 0.70. The calibration curve results for the 1-, 3-, and 5-years survival rates showed that the predicted survival rate was closely related to the actual rate (**Figure 4C**).

Associations Between the Ferroptosis-Related Risk Scores and Immune Cell Infiltration Patterns and Immunotherapy

We analyzed the difference in immune infiltration between high-risk and low-risk groups. In addition, the correlation between immune cells and model genes or risk scores were analyzed. The results showed that 12 of the 22 immune cells had significantly difference in the proportion between the high-risk and low-risk groups. Macrophages M0 and M2, mast cells, monocytes, and $CD4^+$ T cells accounted for a relatively high proportion of cells in the high-risk group. In contrast, B-native cells, macrophages M1, and $CD8^+$ cells accounted for a relatively high proportion of cells in the low-risk group, indicating a correlation between the prognosis difference of risk score and the immune infiltration of cancer tissue (**Figure 5A**). Subsequently, the correlation between immune cells and model genes was analyzed. The results showed that the expression of *GCH1* significantly correlated with macrophage M1 cells, $CD8^+$ T cells, and activated $CD4^+$ memory T cells (**Figure 5B**). The correlation analysis between immune cell infiltration patterns and risk scores showed a specific correlation between risk scores and dendritic cells (**Figure 5C**).

Then, we analyzed the relationship between risk score and response to immunotherapy in the external validation cohorts, to further evaluate the effect of immunotherapy further. In both independent cohorts, complete response (CR) or partial response (PR) to immunotherapy drugs accounted for more samples in the low-risk group than in the high-risk group. Almost all the samples in the high-risk group responded to PD-1 chemotherapeutic drugs with progressive disease (PD)/stable disease (SD), especially in the MSKCC cohort, indicating that the low-risk group received more benefits in terms of the risk score than the high-risk group (**Figures 5D,E**).

The Relationship Between Ferroptosis-Related Risk Scores and Somatic Mutations

In addition, mutational differences in the ferroptosis-related genes between the high-risk group and low-risk group were analyzed. **Figure 6** shows the top 20 genes in terms of ferroptosis-related gene mutations; *ALB* and *ABCC1* had the highest mutation rates in both the high- and low-risk groups.

DISCUSSION

Cell death depends on special regulated molecular mechanisms regulation called regulated cell death (RCD) (Galluzzi et al., 2018). Modern medical researches uses these unique biological processes to treat cancer by allowing cancer cells to die, relying on

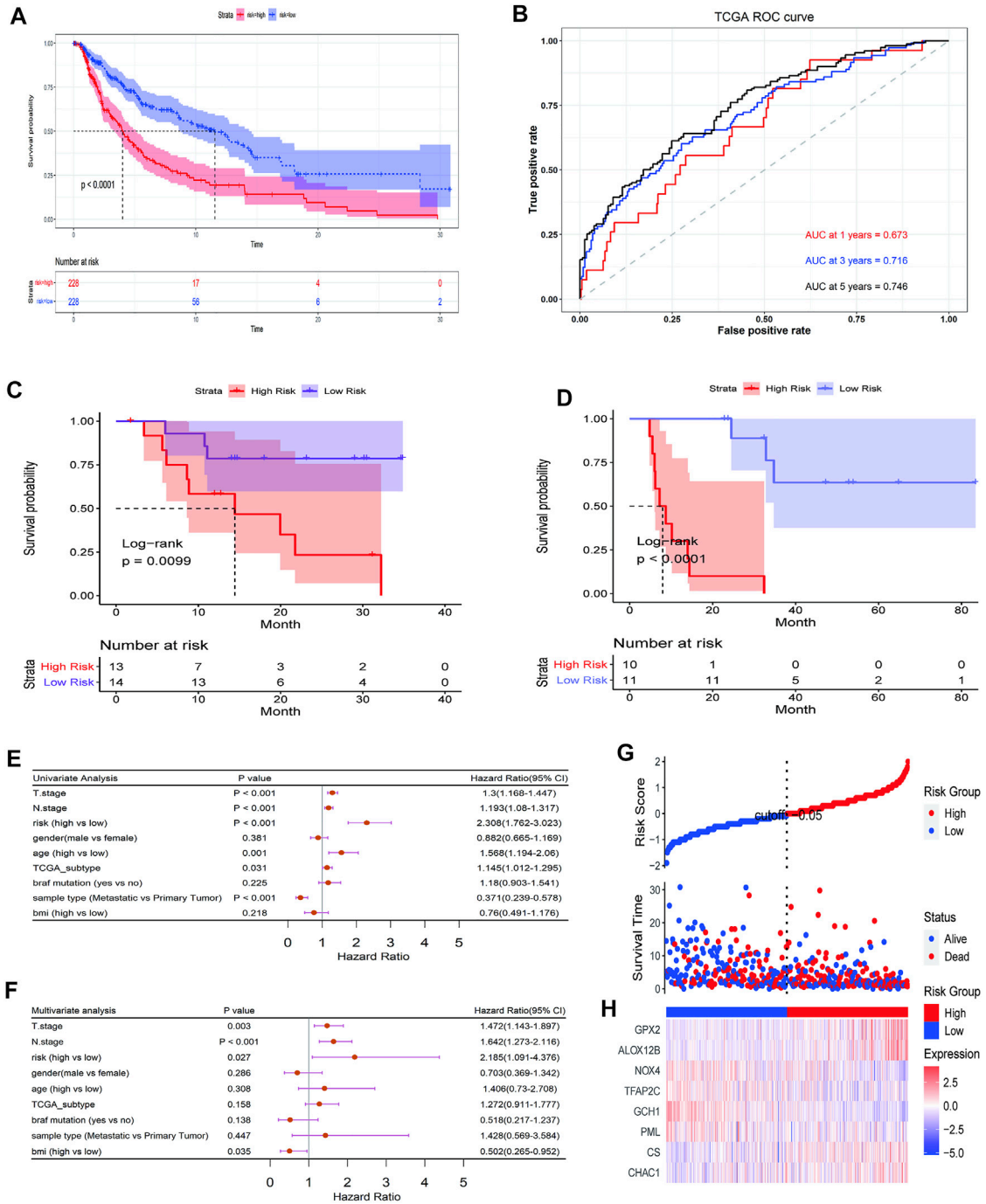

FIGURE 3 | Prognostic significance of the ferroptosis-related gene signature derived risk scores in TGGA cohort. **(A)** Kaplan-Meier analysis of melanoma patients was stratified by median risk score in TCGA cohort. **(B)** Receiver operating characteristic (ROC) curve of risk score signature. **(C, D)** Kaplan-Meier analysis of melanoma patients was stratified by median risk score in GSE78220 cohort **(C)** and MSKCC cohort **(D)**. **(E, F)** Univariate and multivariate Cox regression analysis showed the predictive utility of the risk-score signature possesses excellent prognostic independence. **(G)** The curve shows the distribution of patient risk scores, survival status, and survival time. **(H)** Heatmap shows the expression of each gene in the risk-score signature.

dedicated molecular machinery pharmacologically or genetically. Caspase-dependent apoptosis is a generally recognized RCD process against cancer (Liang et al., 2019). However, tumor cells have the characteristics of resistance to apoptosis, and drug resistance will also occur in the process of chemotherapy-induced apoptosis of cancer (Gottesman et al., 2002). Therefore, it is necessary to find some new forms of RCD to develop anticancer drugs.

TABLE 2 | Characteristics of patients in low- and high-risk scores in GSE78220 cohort.

Characteristic	High, *N* = 13[a]	Low, *N* = 14[a]
Gender		
Female	4 (31%)	4 (29%)
Male	9 (69%)	10 (71%)
Age	63 (55, 70)	58 (54, 64)
OS_time	12 (6, 20)	18 (14, 30)
OS_Status	9 (69%)	3 (21%)
ICI Response		
CR/PR	4 (31%)	10 (71%)
SD/PD	9 (69%)	4 (29%)

[a]*n (%); Median (IQR).*

Ferroptosis is an iron-dependent form of RCD characterized by the overwhelming accumulation of lethal lipid peroxidation; it is different from apoptosis, necrosis, and autophagy (Dixon et al., 2012). Ferroptosis can be triggered by exogenous small molecules (such as erastin, sorafenib, or sulfasalazine) or regulating physiological conditions (such as the high extracellular glutamate concentration) to block the X_C^- system. Other ferroptosis inducers can directly inhibit GPX4, ultimately leading to lipid peroxide accumulation (Liang et al., 2019). Increasing oxidizable polyunsaturated phospholipids or interfering with the iron balance to destroy the balance of lipid metabolism balance can also sensitize cells to ferroptosis, Strong iron dependence can make cancer cells more susceptible to iron overload and ROS accumulation, enabling tumor microenvironment-targeted, ferroptosis-mediated cancer therapy (Vigil et al., 2010; Hassannia et al., 2019). Studies have found that clear cell carcinomas, highly aggressive malignancies, are usually not susceptible to conventional anticancer treatment. However, their unique metabolic state has been identified to be susceptible to ferroptosis (Zou et al., 2019). In addition, a large number of studies have confirmed the crucial role of ferroptosis in inducing cancer cells death and inhibiting tumor growth. The metabolic status of breast cancer, liver cancer, colorectal cancer, and other malignant tumors is closely related to ferroptosis (Sun et al., 2015). However, most current studies focus on the role of iron metabolism in cancer occurrence and treatment, and the relationship between genes related to ferroptosis and cancer prognosis remains to be explored.

Melanoma cells have a highly mutagenic nature and an immune escape mechanism (Davis et al., 2019), including downregulation of the expression of tumor-associated antigens and melanoma differentiation antigens to inhibit cytotoxic T cell recognition and clearance of tumor cells and secretion of immune inhibitory molecules such as transforming growth factor-beta (TGF-β) and prostaglandin E2 to escape immunity (Palmer et al., 2011; Pitcovski et al., 2017). In addition, melanoma can close the immune response by expressing programmed cell death protein 1/2 (PD-1/2) to avoid immune destruction (Passarelli et al., 2017). So far, the most effective treatment for metastatic melanoma is immune checkpoint inhibitors, such as anti-PD1, PD-L1/2, and CTLA4 antibodies. Still, the complications of these treatments are serious, and many of patients have inadequate treatment responses. So, the combination of ferroptosis and immunotherapy may have good prospects for melanoma therapy. Studies have found that immunotherapy activated $CD8^+$ T cells enhance iron-specific lipid peroxidation in tumor cells, and increased ferroptosis contributes to the antitumor effect of immunotherapy (Wang et al., 2019, 8). At present, the relationship between ferroptosis and tumor immunotherapy is still not very clear. Nevertheless, the establishment of some prognostic signatures can be established that use the ferroptosis-related genes to evaluate different tumor immune characteristics to guide individualized immunotherapy (Zhuo et al., 2020; Jiang et al., 2021).

This study found that the ferroptosis-related genes can classify melanoma patients into two classes that exhibit significant differences in clinical and molecular features. Patients were classified into high-risk and low-risk groups by LASSO regression analysis. We established a prognostic model based on eight genes, which were composed of the risk-related genes (*CHAC1*, *CS*, *GPX2*, and *ALOX12B*) and the protective genes *(PML*, *GCH1*, *TFAP2C*, and *NOX4*). The GCH1-BH4 pathway is a new pathway that is independent of the GPX4/glutathione system and regulates ferroptosis. GCH1 is a rate-limiting enzyme for the synthesis of tetrahydrobiopterin. Overexpression of *GCH1* can eliminate lipid peroxidation and almost completely inhibit ferroptosis (Wei et al., 2020). In high-grade serous ovarian cancers, due to the silencing of *PML*, ROS content, lipid peroxidation, and lysosomes, and the lysosomal $Fe2^+$levels are reduced, which can result in potential ferroptosis and improved sensitivity to immunotherapy (Gentric et al., 2019). *NOX4* encodes ROS-producing enzymes enriched in the kidney, where high expression is an essential source of renal ROS. Meanwhile, inhibition of *NOX4* reduces the cystine deprivation-induced cell death and lipid ROS, suggesting its vital role in ferroptosis (Yang et al., 2019). *TFAP2C* plays an essential role in cell differentiation, tissue development, and tumor biology. *TFAP2C* upregulates the *GPX4* gene in tumor cells and regulates some ferroptosis regulators, such as epidermal growth factor receptor (EGFR), CDKN1A, and YAP1, thereby

TABLE 3 | Characteristics of patients in low- and high-risk scores in MSKCC cohort.

Characteristic	High, *N* = 10[a]	Low, *N* = 11[a]
Age	60 (56, 63)	54 (46, 64)
Sex		
Female	6 (60%)	6 (55%)
Male	4 (40%)	5 (45%)
OS_time	8 (6, 13)	35 (28, 53)
OS_Status		
Alive	0 (0%)	8 (73%)
Dead	10 (100%)	3 (27%)
ICI Response		
CR/PR	0 (0%)	8 (73%)
SD/PD	10 (100%)	3 (27%)

[a]*Median (IQR);* n *(%).*

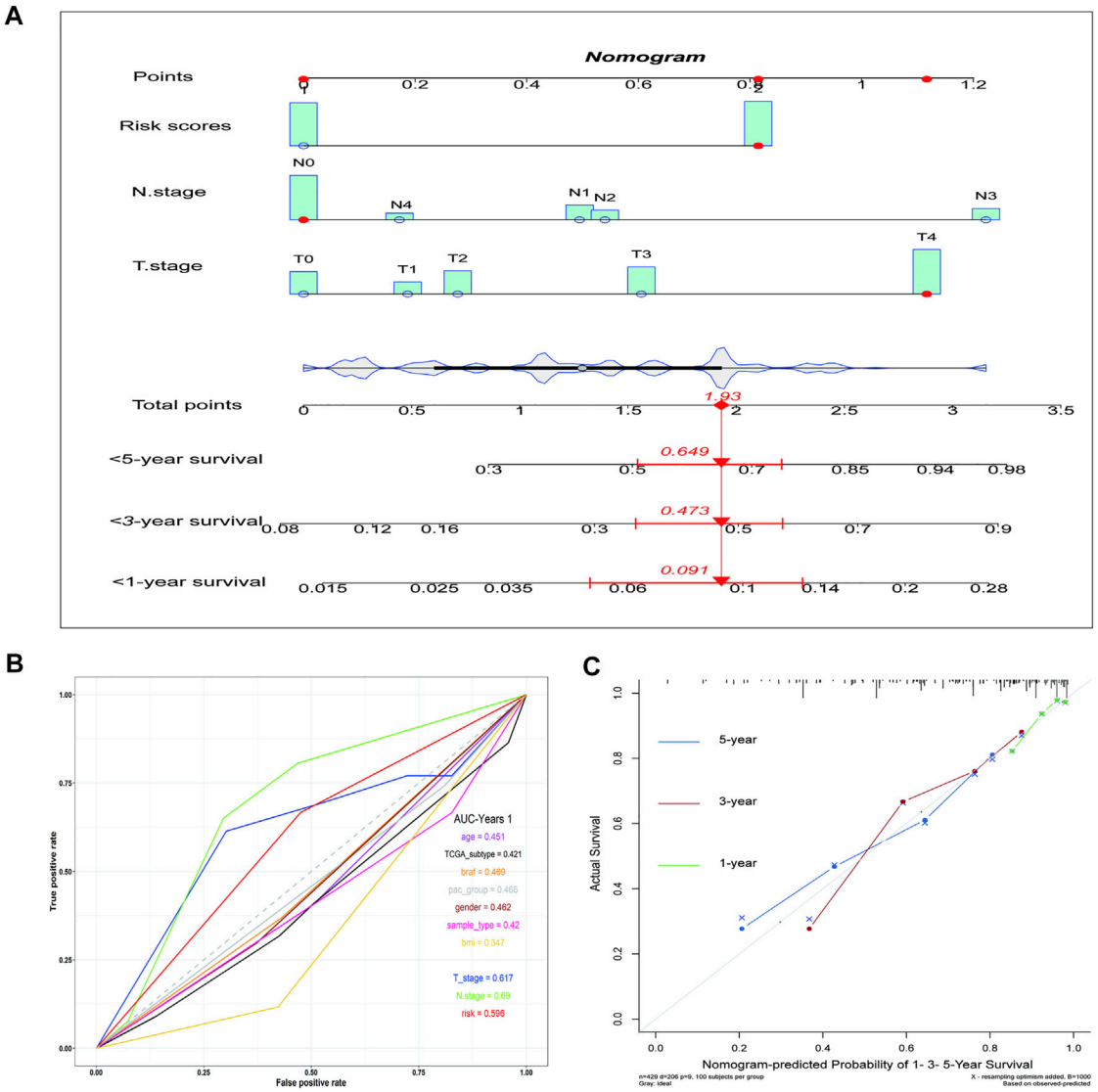

FIGURE 4 | The nomogram shows the impact of various clinical features on survival prognosis in melanoma. **(A)** A nomogram of the melanoma cohort used to indicate the clinical information and overall survival. **(B)** ROC curve of clinical features. **(C)** Calibration maps used to predict the 1, 3, and 5-years survival. The *x*-axis and *y*-axis represent the expected and actual survival rates of the nomogram. The solid line represents the predicted nomogram, and the vertical line represents the 95% confidence interval.

negatively regulating ferroptosis (Dai et al., 2020). Mitochondria play a central role in fatty acid metabolism and provide specific lipid precursors for lipid oxidation. *CS* participates in mitochondrial lipid metabolism and regulates the activation and synthesis of fatty acids. Silencing *CS* can rescue cell viability from erastin-induced ferroptosis (Wang et al., 2020). *ALOX12B* encodes lipoxygenase, which is associated with autosomal dominant fish scale disease and proliferation of epidermoid carcinoma cells (Jiang et al., 2020). Glutathione-specific g-glutamyl cyclotransferase (*CHAC1*) is a downstream target of the eIF2alpha-*ATF4* pathway, and *CHAC1* upregulation may be useful as a Pharmacodynamic marker for cystine or cysteine-starved cells (Dixon et al., 2014). Overexpression of *CHAC1* led to a robust depletion of glutathione, which was alleviated in a CHAC1 catalytic mutant. On activating the expression of *ATF4* and *CHAC1,* the initial glutathione depletion by inhibiting cystine transport leads to ferroptosis (Ratan, 2020). In summary, numerous studies have shown that the above genes are related to ferroptosis, providing theoretical support for our risk model.

After establishing the ferroptosis-related risk model, our samples were divided into high-risk and low-risk groups according to the median risk score. We used statistical methods such as univariate and multivariate Cox regression analysis to show that the risk score model is an excellent independent prognostic indicator for evaluating the overall survival of melanoma patients. Subsequently, we developed a nomogram, combined with multiple clinical

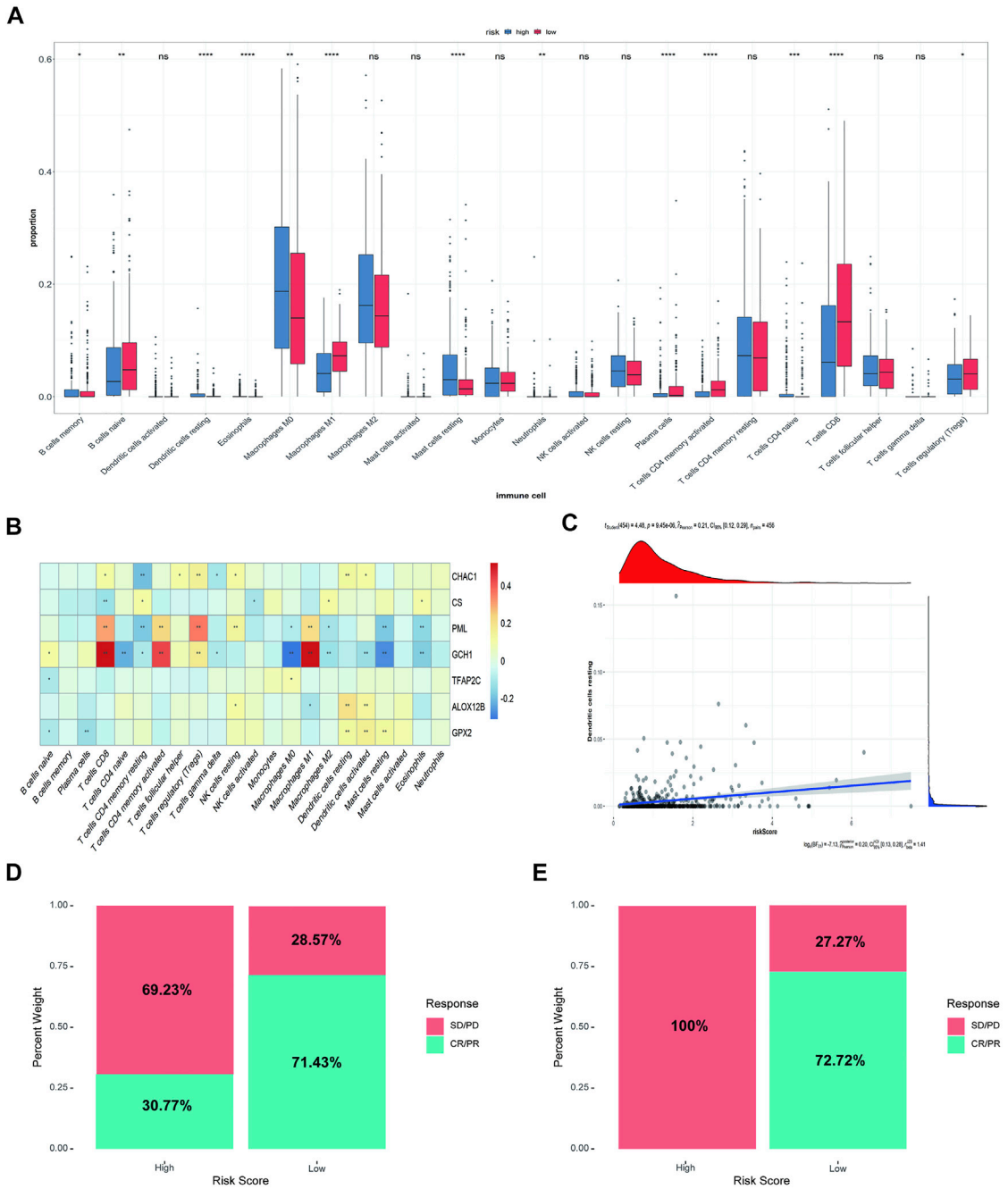

FIGURE 5 | Immune correlation analysis between high-risk and low-risk groups and Comparison of rick scores and chemotherapy drug therapeutic effect. **(A)** The immune infiltration analysis between risk groups. **(B)** The correlation analysis between immune cells and 8 ferroptosis-related genes (ns: not significant, *: $p \leq 0.05$, **: $p \leq 0.01$; ***: $p \leq 0.001$). **(C)** The correlation analysis between risk score and dendritic cells. **(D)** The distribution of the drug response of samples in the high-risk group and low-risk group in GSE78220 cohort. **(E)** The distribution of the drug response of samples in the high-risk group and low-risk group in MSKCC cohort.

indicators and biological attributes, set up a personalized prognostic prediction scale, and quantified the risk of different individuals with numerous risk factors. In addition, we also analyzed the immune correlation between high- and low-risk groups and found that the risk score could indicate the relationship between prognosis difference and immune tissue infiltration, which could better guide the immunotherapy of SKCM.

Finally, our research also has certain limitations. As a retrospective study, it cannot fully contain all clinical data, so there will be some limitations in variable selection, which must be verified by selecting as many cohorts as possible. In addition,

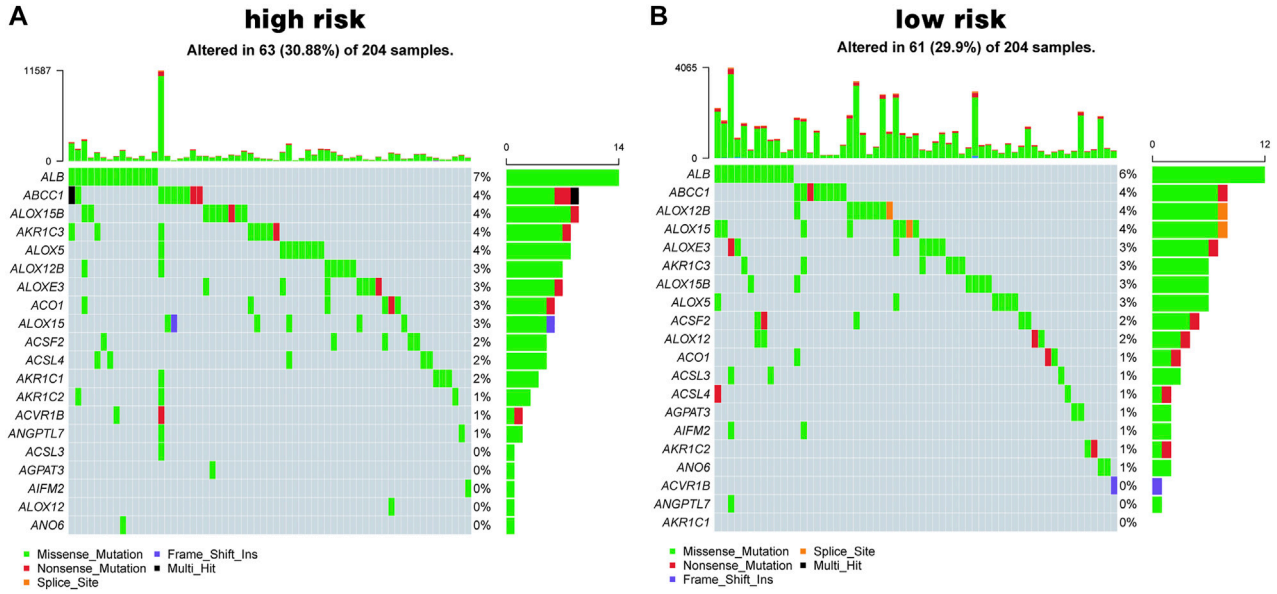

FIGURE 6 | Ferroptosis mutation map. **(A, B)** The top 20 mutated genes in high-risk group **(A)** and in low-risk group **(B)**.

there should be prospective studies to further evaluate the clinical value of our model. Finally, a series of experiments should be carried out to explore the further evaluate the prognostic value of the eight ferroptosis-related gene signature.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

WS and ZY conceived and designed the study. ZX and YX provided equal contributions to research design, data analysis and article writing. YM and JH revised the article. XM, JS, and YS helped to perform the statistical analysis and polish the article. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors gratefully acknowledge contributions from the public database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.758981/full#supplementary-material

REFERENCES

Aris, M., and Barrio, M. a. M. (2015). Combining Immunotherapy with Oncogene-Targeted Therapy: A New Road for Melanoma Treatment. *Front. Immunol.* 6. doi:10.3389/fimmu.2015.00046

Battaglia, A. M., Chirillo, R., Aversa, I., Sacco, A., Costanzo, F., and Biamonte, F. (2020). Ferroptosis and Cancer: Mitochondria Meet the "Iron Maiden" Cell Death. *Cells* 9, 1505. doi:10.3390/cells9061505

Bogdan, A. R., Miyazawa, M., Hashimoto, K., and Tsuji, Y. (2016). Regulators of Iron Homeostasis: New Players in Metabolism, Cell Death, and Disease. *Trends Biochem. Sci.* 41, 274–286. doi:10.1016/j.tibs.2015.11.012

Bridges, R. J., Natale, N. R., and Patel, S. A. (2012). System Xc- Cystine/glutamate Antiporter: an Update on Molecular Pharmacology and Roles within the CNS. *Br. J. Pharmacol.* 165, 20–34. doi:10.1111/j.1476-5381.2011.01480.x

Coit, D. G., Andtbacka, R., Anker, C. J., Bichakjian, C. K., Carson, W. E., Daud, A., et al. (2013). Melanoma, Version 2.2013. *J. Natl. Compr. Canc Netw.* 11, 395–407. doi:10.6004/jnccn.2013.0055

Costa da Silva, M., Breckwoldt, M. O., Vinchi, F., Correia, M. P., Stojanovic, A., Thielmann, C. M., et al. (2017). Iron Induces Anti-tumor Activity in Tumor-Associated Macrophages. *Front. Immunol.* 8, 1479. doi:10.3389/fimmu.2017.01479

Dai, C., Chen, X., Li, J., Comish, P., Kang, R., and Tang, D. (2020). Transcription Factors in Ferroptotic Cell Death. *Cancer Gene Ther.* 27, 645–656. doi:10.1038/s41417-020-0170-2

Davis, L. E., Shalin, S. C., and Tackett, A. J. (2019). Current State of Melanoma Diagnosis and Treatment. *Cancer Biol. Ther.* 20, 1366–1379. doi:10.1080/15384047.2019.1640032

Dixon, S. J., Lemberg, K. M., Lamprecht, M. R., Skouta, R., Zaitsev, E. M., Gleason, C. E., et al. (2012). Ferroptosis: an Iron-dependent Form of Nonapoptotic Cell Death. *Cell* 149, 1060–1072. doi:10.1016/j.cell.2012.03.042

Dixon, S. J., Patel, D. N., Welsch, M., Skouta, R., Lee, E. D., Hayano, M., et al. (2014). Pharmacological Inhibition of Cystine-Glutamate Exchange Induces

Endoplasmic Reticulum Stress and Ferroptosis. *eLife* 3, e02523. doi:10.7554/eLife.02523

Fearnhead, H. O., Vandenabeele, P., and Vanden Berghe, T. (2017). How Do We Fit Ferroptosis in the Family of Regulated Cell Death? *Cell Death Differ* 24, 1991–1998. doi:10.1038/cdd.2017.149

Galluzzi, L., Vitale, I., Aaronson, S. A., Abrams, J. M., Adam, D., Agostinis, P., et al. (2018). Molecular Mechanisms of Cell Death: Recommendations of the Nomenclature Committee on Cell Death 2018. *Cel Death Differ* 25, 486–541. doi:10.1038/s41418-017-0012-4

Gentric, G., Kieffer, Y., Mieulet, V., Goundiam, O., Bonneau, C., Nemati, F., et al. (2019). PML-regulated Mitochondrial Metabolism Enhances Chemosensitivity in Human Ovarian Cancers. *Cel Metab.* 29, 156–173. doi:10.1016/j.cmet.2018.09.002

Gottesman, M. M., Fojo, T., and Bates, S. E. (2002). Multidrug Resistance in Cancer: Role of ATP-dependent Transporters. *Nat. Rev. Cancer* 2, 48–58. doi:10.1038/nrc706

Hassannia, B., Vandenabeele, P., and Vanden Berghe, T. (2019). Targeting Ferroptosis to Iron Out Cancer. *Cancer Cell* 35, 830–849. doi:10.1016/j.ccell.2019.04.002

Jiang, P., Yang, F., Zou, C., Bao, T., Wu, M., Yang, D., et al. (2021). The Construction and Analysis of a Ferroptosis-Related Gene Prognostic Signature for Pancreatic Cancer. *Aging* 13, 10396–10414. doi:10.18632/aging.202801

Jiang, T., Zhou, B., Li, Y., Yang, Q., Tu, K., and Li, L. (2020). ALOX12B Promotes Carcinogenesis in Cervical Cancer by Regulating the PI3K/ERK1 Signaling Pathway. *Oncol. Lett.* 20, 1360–1368. doi:10.3892/ol.2020.11641

Leonardi, G., Candido, S., Falzone, L., Spandidos, D., and Libra, M. (2020). Cutaneous Melanoma and the Immunotherapy Revolution (Review). *Int. J. Oncol.* 57, 609–618. doi:10.3892/ijo.2020.5088

Liang, C., Zhang, X., Yang, M., and Dong, X. (2019). Recent Progress in Ferroptosis Inducers for Cancer Therapy. *Adv. Mater.* 31, 1904197. doi:10.1002/adma.201904197

Liang, J.-Y., Wang, D.-S., Lin, H.-C., Chen, X.-X., Yang, H., Zheng, Y., et al. (2020). A Novel Ferroptosis-Related Gene Signature for Overall Survival Prediction in Patients with Hepatocellular Carcinoma. *Int. J. Biol. Sci.* 16, 2430–2441. doi:10.7150/ijbs.45050

Luo, M., Wu, L., Zhang, K., Wang, H., Zhang, T., Gutierrez, L., et al. (2018). miR-137 Regulates Ferroptosis by Targeting Glutamine Transporter SLC1A5 in Melanoma. *Cel Death Differ* 25, 1457–1472. doi:10.1038/s41418-017-0053-8

Mou, Y., Wang, J., Wu, J., He, D., Zhang, C., Duan, C., et al. (2019). Ferroptosis, a New Form of Cell Death: Opportunities and Challenges in Cancer. *J. Hematol. Oncol.* 12, 34. doi:10.1186/s13045-019-0720-y

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Palmer, S. R., Erickson, L. A., Ichetovkin, I., Knauer, D. J., and Markovic, S. N. (2011). Circulating Serologic and Molecular Biomarkers in Malignant Melanoma. *Mayo Clinic Proc.* 86, 981–990. doi:10.4065/mcp.2011.0287

Passarelli, A., Mannavola, F., Stucci, L. S., Tucci, M., and Silvestris, F. (2017). Immune System and Melanoma Biology: a Balance between Immunosurveillance and Immune Escape. *Oncotarget* 8, 106132–106142. doi:10.18632/oncotarget.22190

Pitcovski, J., Shahar, E., Aizenshtein, E., and Gorodetsky, R. (2017). Melanoma Antigens and Related Immunological Markers. *Crit. Rev. Oncology/Hematology* 115, 36–49. doi:10.1016/j.critrevonc.2017.05.001

Ratan, R. R. (2020). The Chemical Biology of Ferroptosis in the Central Nervous System. *Cel Chem. Biol.* 27, 479–498. doi:10.1016/j.chembiol.2020.03.007

Rodríguez-Cerdeira, C., Carnero Gregorio, M., López-Barcenas, A., Sánchez-Blanco, E., Sánchez-Blanco, B., Fabbrocini, G., et al. (20172017). Advances in Immunotherapy for Melanoma: A Comprehensive Review. *Mediators Inflamm.* 2017, 1–14. doi:10.1155/2017/3264217

Sacco, A., Battaglia, A. M., Botta, C., Aversa, I., Mancuso, S., Costanzo, F., et al. (2021). Iron Metabolism in the Tumor Microenvironment-Implications for Anti-cancer Immune Response. *Cells* 10, 303. doi:10.3390/cells10020303

Sato, H., Tamba, M., Ishii, T., and Bannai, S. (1999). Cloning and Expression of a Plasma Membrane Cystine/glutamate Exchange Transporter Composed of Two Distinct Proteins. *J. Biol. Chem.* 274, 11455–11458. doi:10.1074/jbc.274.17.11455

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* 371, 2189–2199. doi:10.1056/NEJMoa1406498

Sun, X., Ou, Z., Xie, M., Kang, R., Fan, Y., Niu, X., et al. (2015). HSPB1 as a Novel Regulator of Ferroptotic Cancer Cell Death. *Oncogene* 34, 5617–5625. doi:10.1038/onc.2015.32

Tang, R., Xu, J., Zhang, B., Liu, J., Liang, C., Hua, J., et al. (2020). Ferroptosis, Necroptosis, and Pyroptosis in Anticancer Immunity. *J. Hematol. Oncol.* 13, 110. doi:10.1186/s13045-020-00946-7

Ursini, F., Maiorino, M., Brigelius-Flohé, R., Aumann, K. D., Roveri, A., Schomburg, D., et al. (1995). Diversity of Glutathione Peroxidases. *Methods Enzymol.* 252, 38–53. doi:10.1016/0076-6879(95)52007-4

Vigil, D., Cherfils, J., Rossman, K. L., and Der, C. J. (2010). Ras Superfamily GEFs and GAPs: Validated and Tractable Targets for Cancer Therapy? *Nat. Rev. Cancer* 10, 842–857. doi:10.1038/nrc2960

Wang, H., Liu, C., Zhao, Y., and Gao, G. (2020). Mitochondria Regulation in Ferroptosis. *Eur. J. Cel Biol.* 99, 151058. doi:10.1016/j.ejcb.2019.151058

Wang, W., Green, M., Choi, J. E., Gijón, M., Kennedy, P. D., Johnson, J. K., et al. (2019). CD8+ T Cells Regulate Tumour Ferroptosis during Cancer Immunotherapy. *Nature* 569, 270–274. doi:10.1038/s41586-019-1170-y

Wei, X., Yi, X., Zhu, X.-H., and Jiang, D.-S. (20202020). Posttranslational Modifications in Ferroptosis. *Oxidative Med. Cell Longevity* 2020, 1–12. doi:10.1155/2020/8832043

Welch, H. G., Mazer, B. L., and Adamson, A. S. (2021). The Rapid Rise in Cutaneous Melanoma Diagnoses. *N. Engl. J. Med.* 8. doi:10.1056/nejmsb2019760

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* 26, 1572–1573. doi:10.1093/bioinformatics/btq170

Yang, W.-H., Ding, C.-K. C., Sun, T., Rupprecht, G., Lin, C.-C., Hsu, D., et al. (2019). The Hippo Pathway Effector TAZ Regulates Ferroptosis in Renal Cell Carcinoma. *Cel Rep.* 28, 2501–2508. e4. doi:10.1016/j.celrep.2019.07.107

Yang, W. S., SriRamaratnam, R., Welsch, M. E., Shimada, K., Skouta, R., Viswanathan, V. S., et al. (2014). Regulation of Ferroptotic Cancer Cell Death by GPX4. *Cell* 156, 317–331. doi:10.1016/j.cell.2013.12.010

Yu, H., Guo, P., Xie, X., Wang, Y., and Chen, G. (2017). Ferroptosis, a New Form of Cell Death, and its Relationships with Tumourous Diseases. *J. Cel. Mol. Med.* 21, 648–657. doi:10.1111/jcmm.13008

Zhang, K., Wu, L., Zhang, P., Luo, M., Du, J., Gao, T., et al. (2018). miR-9 Regulates Ferroptosis by Targeting Glutamic-oxaloacetic Transaminase GOT1 in Melanoma. *Mol. Carcinogenesis* 57, 1566–1576. doi:10.1002/mc.22878

Zhuo, S., Chen, Z., Yang, Y., Zhang, J., Tang, J., and Yang, K. (2020). Clinical and Biological Significances of a Ferroptosis-Related Gene Signature in Glioma. *Front. Oncol.* 10, 590861. doi:10.3389/fonc.2020.590861

Zou, Y., Palte, M. J., Deik, A. A., Li, H., Eaton, J. K., Wang, W., et al. (2019). A GPX4-dependent Cancer Cell State Underlies the clear-cell Morphology and Confers Sensitivity to Ferroptosis. *Nat. Commun.* 10, 1617. doi:10.1038/s41467-019-09277-9

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

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

Copyright © 2021 Xu, Xie, Mao, Huang, Mei, Song, Sun, Yao and Shi. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 November 2021
doi: 10.3389/fgene.2021.759596

TGF-Beta Induced Key Genes of Osteogenic and Adipogenic Differentiation in Human Mesenchymal Stem Cells and MiRNA–mRNA Regulatory Networks

Genfa Du[1†], Xinyuan Cheng[2†], Zhen Zhang[1], Linjing Han[1], Keliang Wu[2], Yongjun Li[3] and Xiaosheng Lin[1]**

[1]Department of Orthopedics, Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine, Guangzhou University of Chinese Medicine, Shenzhen, China, [2]The Fourth Clinical Medical College, Guangzhou University of Chinese Medicine, Shenzhen, China, [3]Department of Orthopedics, Shunde Hospital Guangzhou University of Chinese Medicine, Foshan, China

OPEN ACCESS

Edited by:
Anthony Kusalik,
University of Saskatchewan, Canada

Reviewed by:
Huaming Chen,
University of Adelaide, Australia
Hifzur Rahman Ansari,
King Abdullah International Medical Research Center (KAIMRC), Saudi Arabia

***Correspondence:**
Yongjun Li
jun555@126.com
Xiaosheng Lin
lxshengtcm@126.com

[†]These authors have contributed equally to this work and share first authorship

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

Received: 16 August 2021
Accepted: 28 October 2021
Published: 25 November 2021

Citation:
Du G, Cheng X, Zhang Z, Han L, Wu K, Li Y and Lin X (2021) TGF-Beta Induced Key Genes of Osteogenic and Adipogenic Differentiation in Human Mesenchymal Stem Cells and MiRNA–mRNA Regulatory Networks. Front. Genet. 12:759596. doi: 10.3389/fgene.2021.759596

Background: The clinical efficacy of osteoporosis therapy is unsatisfactory. However, there is currently no gold standard for the treatment of osteoporosis. Recent studies have indicated that a switch from osteogenic to adipogenic differentiation in human bone marrow mesenchymal stem cells (hMSCs) induces osteoporosis. This study aimed to provide a more comprehensive understanding of the biological mechanisms involved in this process and to identify key genes involved in osteogenic and adipogenic differentiation in hMSCs to provide new insights for the prevention and treatment of osteoporosis.

Methods: Microarray and bioinformatics approaches were used to identify the differentially expressed genes (DEGs) involved in osteogenic and adipogenic differentiation, and the biological functions and pathways of these genes were analyzed. Hub genes were identified, and the miRNA–mRNA interaction networks of these hub genes were constructed.

Results: In an optimized microenvironment, transforming growth factor-beta (TGF-beta) could promote osteogenic differentiation and inhibit adipogenic differentiation of hMSCs. According to our study, 98 upregulated genes involved in osteogenic differentiation and 66 downregulated genes involved in adipogenic differentiation were identified, and associated biological functions and pathways were analyzed. Based on the protein–protein interaction (PPI) networks, the hub genes of the upregulated genes (CTGF, IGF1, BMP2, MMP13, TGFB3, MMP3, and SERPINE1) and the hub genes of the downregulated genes (PPARG, TIMP3, ANXA1, ADAMTS5, AGTR1, CXCL12, and CEBPA) were identified, and statistical analysis revealed significant differences. In addition, 36 miRNAs derived from the upregulated hub genes were screened, as were 17 miRNAs derived from the downregulated hub genes. Hub miRNAs (hsa-miR-27a/b-3p, hsa-miR-128-3p, hsa-miR-1-3p, hsa-miR-98-5p, and hsa-miR-130b-3p) coregulated both osteogenic and adipogenic differentiation factors.

Conclusion: The upregulated hub genes identified are potential targets for osteogenic differentiation in hMSCs, whereas the downregulated hub genes are potential targets for adipogenic differentiation. These hub genes and miRNAs play important roles in adipogenesis and osteogenesis of hMSCs. They may be related to the prevention and treatment not only of osteoporosis but also of obesity.

Keywords: TGF-beta, osteoporosis, obesity, osteogenesis, adipogenesis, mesenchymal stem cell

INTRODUCTION

Osteoporosis is one of the most common chronic aged-related diseases in the world, and it is especially common in postmenopausal women (Zhi et al., 2021). It is characterized by loss of bone mass, degeneration of bone microstructure, and reduction of bone strength (Black and Roen, 2016). It is a highly prevalent disease that affects an estimated 200 million people worldwide (Vellucci et al., 2014). It has been reported that approximately 50% of women and 20% of men over the age of 50 years will have osteoporotic fractures in their remaining years (Sambrook and Cooper, 2006). This inevitably leads to higher mortality, high medical costs, and social burden. Notably, however, there is still no gold standard for the treatment of osteoporosis (McClung et al., 2019). Drugs such as bisphosphonates, calcitonin, and estrogen can delay the progression of osteoporosis (Chang et al., 2019; Eastell et al., 2019), but these drugs must be taken for a long time and may cause serious side effects (Ensrud and Crandall, 2017). Thus, it is very important to understand the pathogenesis of osteoporosis, and further investigation of novel osteoporosis targets is imperative.

Bone marrow mesenchymal stem cells (BMSCs) have the capacity to differentiate into many cell types, including osteoblasts and adipocytes (Yu et al., 2021), which are closely associated with osteoporosis (Haasters et al., 2014). Previous studies have indicated that the connection between fat and bone is a significant factor in the pathology of senile bone loss and that fat in bone marrow could be used as a diagnostic and therapeutic tool in osteoporosis (Duque, 2008). In recent years, there has been increasing interest in the interaction between fat and bone cells in bone marrow (van Zoelen et al., 2016). It has been reported that an imbalance between bone formation and bone loss occurs with aging and that the bone marrow component shifts to adipocytes, osteoclast activity enhances, and osteoblast function declines (Rosen and Bouxsein, 2006; Hu et al., 2018). Osteogenic differentiation is inhibited and bone formation is reduced, leading to bone that is filled with adipocytes instead of osteoblasts, thereby inducing osteoporosis (Rosen and Bouxsein, 2006; Hu et al., 2018). Therefore, there is a negative correlation between bone formation and fat accumulation in bone marrow (Souza et al., 2021). Although there is a competitive relationship between adipogenesis and osteogenesis during the differentiation of human BMSCs (hMSCs), the adipo-osteogenic signaling pathway could be altered to favor osteoblasts for the prevention of osteoporosis (Wu et al., 2020a). Notably, however, the specific mechanisms of hMSC differentiation into osteogenesis and adipogenesis in osteoporosis remain unclear. These considerations indicate that it is necessary to elucidate the relationship between adipogenic and osteogenic differentiation and to develop new drugs to prevent the differentiation of hMSCs into fat cells.

Bone morphogenetic proteins (BMPs) are multifunctional growth factors that belong to the transforming growth factor-beta (TGF-beta) superfamily (Halloran et al., 2020), and they have dual roles. The microenvironment is conducive to adipogenic or osteogenic differentiation, promoting either adipogenesis or osteogenesis (Atashi et al., 2015). TGF-beta is an important factor during bone formation and remodeling, and studies have indicated that TGF-beta could stimulate early differentiation of osteoblasts, while inhibiting late differentiation of osteoblasts into osteocytes (Tang and Alliston, 2013). The aims of the present study were to reveal the potential mechanisms underlying osteogenic and adipogenic differentiation of hMSCs and to investigate new targets for use in osteoporosis treatment. The microarray dataset GSE84500 from the Gene Expression Omnibus (GEO) database was used. Two groups were selected to identify differentially expressed genes (DEGs) related to osteogenic and adipogenic differentiation in hMSCs: a BMP2+3-isobutyl-1-methylxanthine (IBMX) group and a BMP2+IBMX+TGF-beta group. HMSCs were cultured under the same adipogenic conditions induced by BMP2 and IBMX, causing some to differentiate into adipocytes and others to differentiate into osteoblasts, and then TGF-beta was added to the culture. Functional and pathway enrichment analyses of DEGs were performed, and a protein–protein interaction (PPI) network was constructed to identify hub genes, which were verified at the mRNA expression level. MiRTarBase, TargetScan, and CyTargetLinker software were used to identify microRNAs that potentially regulate hub genes, providing a basis for further studies. The results of the current study may provide insight into the mechanisms of osteogenesis and adipogenesis and facilitate new therapeutic strategies for osteoporosis or obesity.

MATERIALS AND METHODS

Microarray Data

The GEO dataset module from GEO database was selected (https://www.ncbi.nlm.nih.gov/geo/). An advanced search was then conducted as follows: ((osteoporosis) AND Bone marrow mesenchymal stem cells) AND "Expression profiling by array" [Filter]). The main purpose of this study was related to TGF-beta-induced osteogenic and adipogenic differentiation in hMSCs, and

the inclusion organism of the dataset was *Homo sapiens*. Accordingly, only the mRNA microarray dataset GSE84500, which contains sufficient samples and four time-points, was available from the GEO database. The dataset includes normal hMSC samples from three different donors (van Zoelen et al., 2016). To better evaluate the TGF-beta-induced switch from adipogenic to osteogenic differentiation, 24 samples of hMSCs were selected from a BMP2+IBMX (BI) group and a BMP2+IBMX+TGF-beta (BIT) group. The two groups included 12 samples from 1, 2, 3, and 7 days of cell culture, with six samples at each time-point. This dataset platform was GPL570 ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array).

Identification of Differentially Expressed Genes

The GEO2R function (https://www.ncbi.nlm.nih.gov/geo/geo2r/) from the GEO database was used to identify DEGs in the BI and BIT groups. The original gene expression data were log2 converted, and DEG analysis was conducted with the default setting in GEO2R. DEGs with adjusted p-values <0.05 were considered statistically significant, and logFC ≥ 1 or logFC ≤ −1 was selected as the DEG threshold. Samples at each time-point were analyzed for upregulated and downregulated genes. In order to reduce false-positive results caused by operational error or culture conditions during cell experiments and to acquire stable genes, the intersections of the upregulated and downregulated genes of four time-points were used. Lastly, TGF-beta-mediated upregulated and downregulated genes were identified. A relative log expression (RLE) diagram was used to evaluate the quality of the sample chip, and a heatmap and a volcano plot were constructed using the pheatmap and gplots packages in R language, respectively.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Functional Analysis of Differentially Expressed Genes

To analyze the functions and potential pathways of the DEGs identified, the online DAVID software (https://david.ncifcrf.gov/) was used to perform Gene Ontology (GO) functional analysis (biological process = BP, cellular component = CC, and molecular function = MF) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. The Functional Annotation module was selected from the DAVID database, and the upregulated and downregulated genes were imported into the gene list of the Functional Annotation Tool for GO and KEGG analysis, respectively. The identifier selected wasofficial_gene_symbol, and the species selected was *H. sapiens*. Gene_ontology (BP/CC/MF) and pathways (kegg_pathway) were selected for enrichment analysis. Enrichment options were chosen from the default setting from the functional annotation chart, and p-value <0.05 was defined as statistically significant. According to the enrichment options, the eligible terms were screened out. Based on the p-value, the top six terms of GO functional enrichment terms (BP/CC/MF) were visualized with bar charts, and the ordinate is represented by–log10 (p-value).

Protein–Protein Interaction Networks of Differentially Expressed Genes and Hub Gene Identification

The STRING database is an online tool designed to identify PPIs between DEGs from experiments and predictions (https://www.string-db.org/), and it was used to construct the PPI networks in the current study. All upregulated and downregulated genes were imported into the gene list. The criterion was medium confidence for selection ≥0.4, and *H. sapiens* was the selected organism. PPI networks were downloaded and deposited into Cytoscape v3.7.2 (https://cytoscape.org/), which was used to map interactions among the DEGs. The cytoHubba plugin from Cytoscape was then used to screen the hub genes of the PPI networks. To enhance data reliability, hub genes of upregulated and downregulated genes were obtained from the degree of intersection between MCC, MNC, and Degree modules.

Construction of MiRNA–mRNA Interaction Networks

The CyTargetLinker4.1 plugin from Cytoscape (https://apps.cytoscape.org/apps/cytargetlinker) was used to predict miRNA–mRNA interaction networks. The Linksets module of the CyTargetLinker tutorial presentation (https://cytargetlinker.github.io/pages/tutorials/tutorial1) was used, and then the Linksets of MiRTarBase release 8.0 and TargetScan release 7.2 were selected (https://cytargetlinker.github.io/pages/linksets). Of these, the miRTarBase release 8.0 database is dedicated to collecting microRNA–target interactions (MTIs) with experimental evidence. Mirtarbase_hsa_8.0.xgmml.zip (https://cytargetlinker.github.io/pages/linksets/mirtarbase) included 502,652 MTIs, 15,038 target genes, and 2,595 microRNAs; and it was downloaded. Additionally, targetscan_hsa_7.2.xgmml.zip (https://cytargetlinker.github.io/pages/linksets/targetscan) included 264,563 MTIs, 13,077 target genes, and 405 microRNAs; and it was also downloaded. The first step of generating a miRNA–mRNA interaction network was the creation of txt files including upregulated genes and downregulated genes. The second step was the selection "File" from Cytoscape software → then "Network from File" → then "Import txt file." The third step was importation of mirtarbase_hsa_8.0.xgmml and targetscan_hsa_7.2.xgmml into the CyTargetLinker component of Cytoscape software. The miRNA–mRNA interaction networks were thus constructed. An intersection threshold of 2 was set for the miRTarBase and TargetScan databases, and miRNA–mRNA interaction networks (hub miRNAs) were obtained and shared by the two databases.

mRNA Expression Levels of Hub Genes and Validation

To investigate hub genes in the BI group and the BIT group, the mRNA expression levels of the top seven hub genes of

TABLE 1 | Summary of the 24 samples.

Data number	Sample name
GSM2238550	hMSC, treated with BMP2+IBMX, 1-day differentiation
GSM2238551	hMSC, treated with BMP2+IBMX, 1-day differentiation
GSM2238552	hMSC, treated with BMP2+IBMX, 1-day differentiation
GSM2238553	hMSC, treated with BMP2+IBMX+TGFB, 1-day differentiation
GSM2238554	hMSC, treated with BMP2+IBMX+TGFB, 1-day differentiation
GSM2238555	hMSC, treated with BMP2+IBMX+TGFB, 1-day differentiation
GSM2238562	hMSC, treated with BMP2+IBMX, 2-day differentiation
GSM2238563	hMSC, treated with BMP2+IBMX, 2-day differentiation
GSM2238564	hMSC, treated with BMP2+IBMX, 2-day differentiation
GSM2238565	hMSC, treated with BMP2+IBMX+TGFB, 2-day differentiation
GSM2238566	hMSC, treated with BMP2+IBMX+TGFB, 2-day differentiation
GSM2238567	hMSC, treated with BMP2+IBMX+TGFB, 2-day differentiation
GSM2238574	hMSC, treated with BMP2+IBMX, 3-day differentiation
GSM2238575	hMSC, treated with BMP2+IBMX, 3-day differentiation
GSM2238576	hMSC, treated with BMP2+IBMX, 3-day differentiation
GSM2238577	hMSC, treated with BMP2+IBMX+TGFB, 3-day differentiation
GSM2238578	hMSC, treated with BMP2+IBMX+TGFB, 3-day differentiation
GSM2238579	hMSC, treated with BMP2+IBMX+TGFB, 3-day differentiation
GSM2238586	hMSC, treated with BMP2+IBMX, 7-day differentiation
GSM2238587	hMSC, treated with BMP2+IBMX, 7-day differentiation
GSM2238588	hMSC, treated with BMP2+IBMX, 7-day differentiation
GSM2238589	hMSC, treated with BMP2+IBMX+TGFB, 7-day differentiation
GSM2238590	hMSC, treated with BMP2+IBMX+TGFB, 7-day differentiation
GSM2238591	hMSC, treated with BMP2+IBMX+TGFB, 7-day differentiation

Note. hMSC, human bone marrow mesenchymal stem cell.

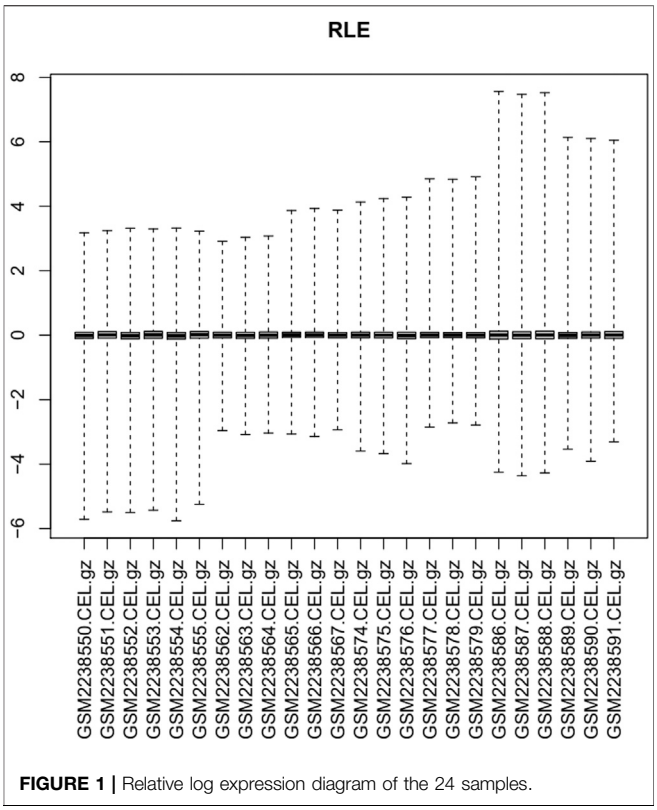

FIGURE 1 | Relative log expression diagram of the 24 samples.

upregulated and downregulated genes were identified from GSE84500. Each sample mRNA expression level for each time-point was obtained in two groups through R language and the GPL570 platform. The mRNA expression levels of 24 samples from four time-points were then combined, and they were divided into two groups of 12 samples. Lastly, the top seven hub genes in the BI group and BIT group were compared. The unpaired t-test was used for statistical analysis, and parameter testing and normality testing were conducted before the t-test. $p < 0.05$ was defined as a statistically significant difference. Statistical data are presented as the mean ± SD. GraphPad Prism (version 7.0) was used to conduct all statistical analyses and to generate graphs.

RESULTS

Identification of Differentially Expressed Genes

Via filtering by set conditions, a total of 24 hMSC samples were acquired (**Table 1**). In evaluation of the quality of the sample chip, the median of 24 samples was almost on the same line and close to 0 (**Figure 1**), indicating superior quality of standardization. At the 1-day time-point, in the BIT group, 222 genes were upregulated in comparison with the BI group, in which 148 genes were downregulated. At the 2-day time-point, in the BIT group, 328 genes were upregulated in comparison with the BI group, in which 375 genes were downregulated. At the 3-day time-point, the corresponding numbers were 533 upregulated and 515 downregulated, and at the 7-day time-point, the corresponding numbers were 786 upregulated and 754 downregulated. The DEGs from the four time-points were combined, and the overlap of the final 98 upregulated and 66 downregulated genes was visualized as a Venn diagram (**Figures 2A,B**) and a volcano map (**Figure 2C**). Meanwhile, a heatmap for 164 DEGs from the log2 mRNA expression level of this microarray is shown (**Figure 2D**).

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Functional Analysis of Differentially Expressed Genes

In GO functional analysis, upregulated and downregulated genes were enriched in various BP, CC, and MF terms (**Figures 3A,B**). In the BP category, the upregulated genes were mainly involved in the negative regulation of TGF-beta receptor pathway, skeletal system development, negative regulation of cell migration, and bone mineralization; the downregulated genes were mainly involved in the response to peptide hormone, Rho protein signal transduction, and response to mechanical stimulus. In the CC categories, the upregulated genes were mainly involved in extracellular matrix (ECM), extracellular space, proteinaceous ECM, and extracellular region; the downregulated genes were mainly involved in proteinaceous ECM, extracellular space, and invadopodium. Analysis of the MF category further demonstrated that the upregulated genes were mainly involved in heparin binding, growth factor activity, actin binding, and protein heterodimerization activity; the downregulated genes were mainly involved in metalloendopeptidase activity, indanol dehydrogenase activity, and protein binding bridging.

Five KEGG signaling pathways were identified (**Tables 2**, **3**). The upregulated genes were primarily involved in three pathways, and the downregulated genes were primarily involved in two

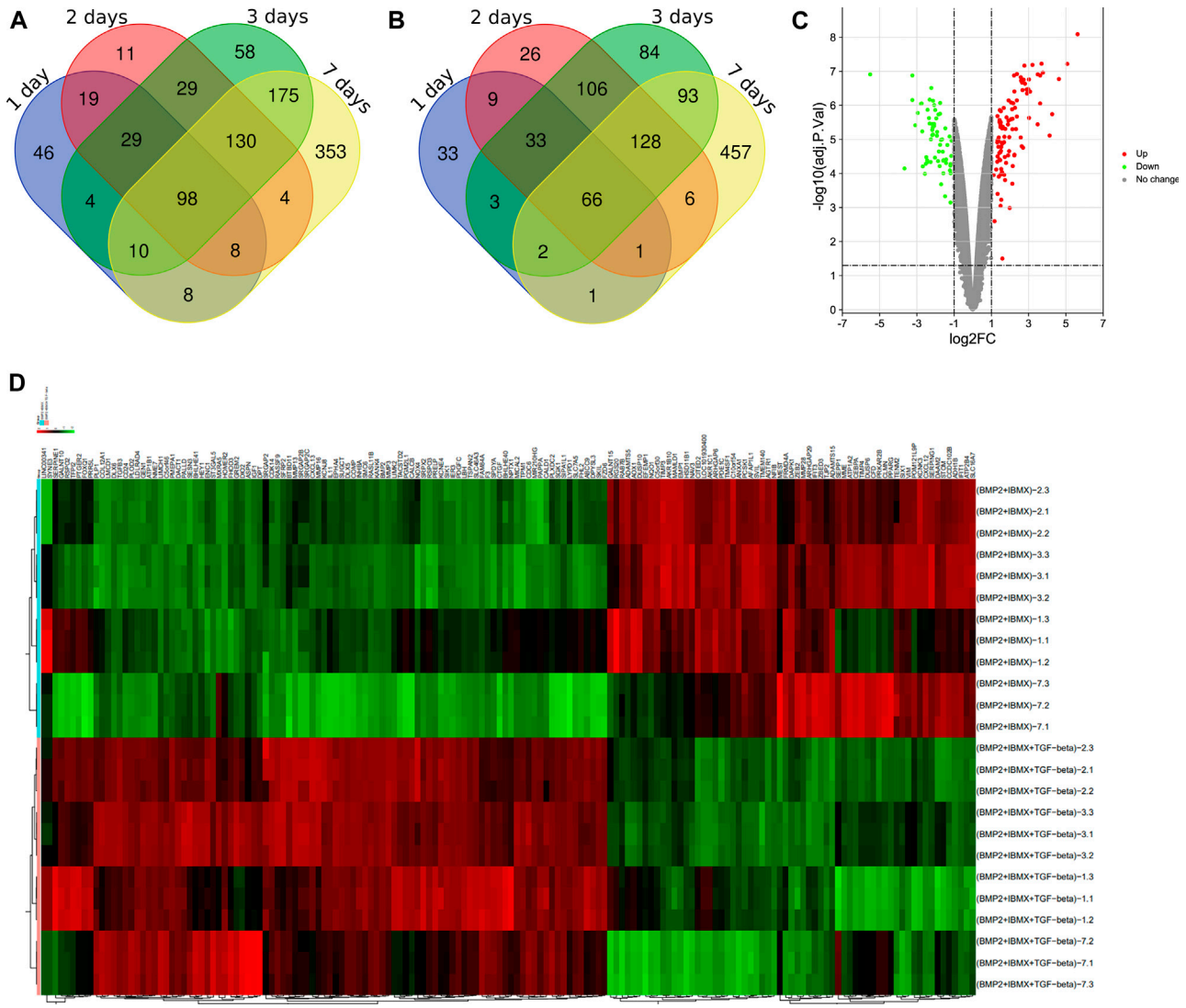

FIGURE 2 | Venn diagrams showing **(A)** the 98 upregulated genes and **(B)** the 66 downregulated genes. A differentially expressed gene (DEG) volcano plot **(C)** and a heatmap **(D)** are shown. Red represents upregulated genes, and green represents downregulated genes ($p < 0.05$, logFC ≥ 1 or logFC ≤ −1).

pathways. Although the *p*-value of "sa05200: Pathways in cancer" was >0.05, it contained a large number of enriched genes.

Protein–Protein Interaction Networks of the Differentially Expressed Genes and Identification of Hub Genes

To systematically analyze the PPIs of DEGs, PPI networks of the upregulated and downregulated genes were constructed using Cytoscape software (**Figures 4A,B**). In the PPI networks of the upregulated genes, the DEGs with the highest connectivity degrees were BMP2, CTGF, IGF1, TGFB3, MMP13, MMP3, SERPINE1, COMP, ASPN, and IL11. Similarly, in the PPI networks of upregulated genes, the DEGs with the highest connectivity degrees were PPARG, TIMP3, ANXA1, ADAMTS5, TIMP4, AGTR1, NQO1, CXCL12, CEBPA, and CFD. The PPI networks of the DEGs from the STRING database were deposited into Cytoscape v3.7.2, and then the cytoHubba plugin from Cytoscape was used to identify hub genes of the PPI networks, and hub genes overlapped by MCC, MNC, and Degree. The top seven upregulated hub genes were CTGF, IGF1, BMP2, MMP13, TGFB3, MMP3, and SERPINE1; and the top seven downregulated hub genes were PPARG, TIMP3, ANXA1, ADAMTS5, AGTR1, CXCL12, and CEBPA (**Figures 4A,B**).

Hub Gene mRNA Expression Levels and Validation

mRNA expression levels of upregulated hub genes involved in osteogenic differentiation were significantly higher in the BIT group than in the BI group. However, the mRNA expression levels of downregulated hub genes involved in adipogenic differentiation were significantly lower in the BIT group than

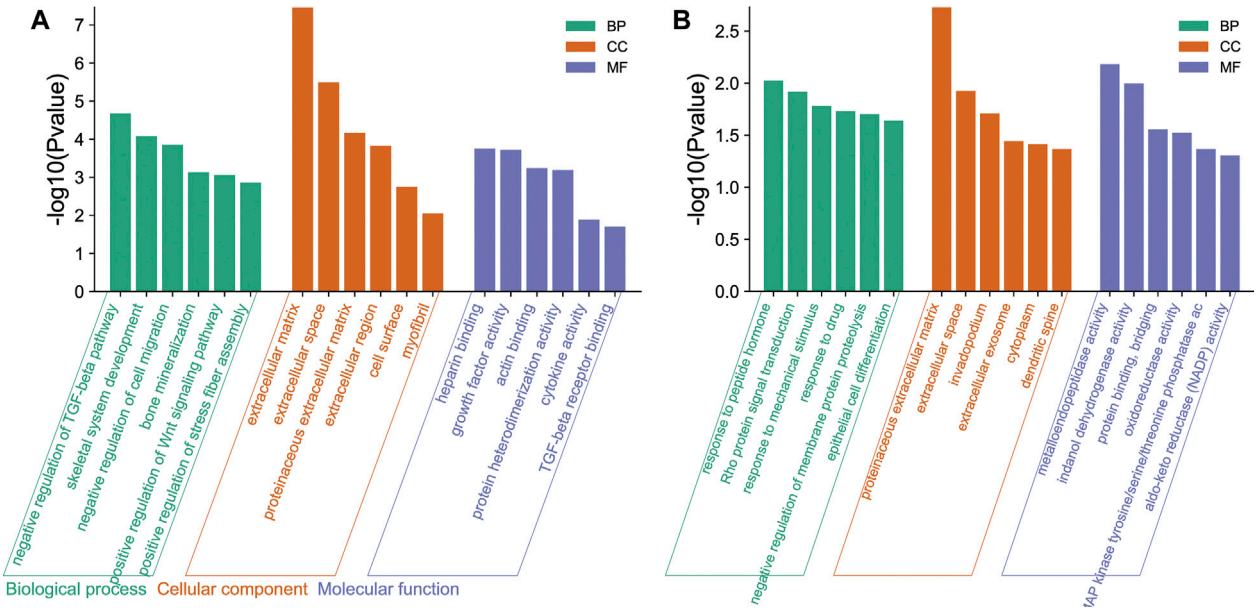

FIGURE 3 | Gene Ontology (GO) functional enrichment of differentially expressed genes (DEGs) in osteogenic and adipogenic differentiation of human bone marrow mesenchymal stem cells (hMSCs). These genes were enriched in various biological process (BP), cellular component (CC), and molecular function (MF) terms. The ordinate is indicated on a –log10 (*p*-value) scale. **(A)** The main enrichment results of the upregulated genes. **(B)** The main enrichment results of the downregulated genes.

TABLE 2 | KEGG pathways enrichment analyses of upregulated DEGs.

Category	Term	Count	*p*-Value	Genes
KEGG_PATHWAY	hsa04550: Signaling pathways regulating pluripotency of stem cells	6	0.0021	BMP2, DLX5, FZD6, IGF1, INHBA, SKIL
KEGG_PATHWAY	hsa04390: Hippo signaling pathway	5	0.0166	BMP2, TGFB3, FZD6, SERPINE1, CTGF
KEGG_PATHWAY	hsa04960: Aldosterone-regulated sodium reabsorption	3	0.0266	IGF1, ATP1B1, SGK1

Note. The three KEGG pathways were selected based on p-*values.*
KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes.

TABLE 3 | KEGG pathways enrichment analyses of downregulated DEGs.

Category	Term	Count	*p*-Value	Genes
KEGG_PATHWAY	hsa00980: Metabolism of xenobiotics by cytochrome P450	3	0.0384	HSD11B1, ADH1B, AKR1C1
KEGG_PATHWAY	hsa05200: Pathways in cancer	5	0.0806	CEBPA, CXCL12, DAPK1, AGTR1, PPARG

Note. The two KEGG pathways were selected based on p-*values. Although the* p-*value of the "sa05200: Pathways in cancer" was >0.05, it contained a large number of enriched genes.*
KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes.

in the BI group. In statistical analyses, mRNA expression levels of all upregulated and downregulated hub genes differed significantly (**Figures 5**, **6**). This indicated that the data were reliable and that these genes were hub genes for TGF-beta-induced upregulated and downregulated genes. These genes can be considered potential targets for TGF-beta-induced osteogenic and adipogenic differentiation of hMSCs.

Construction of MiRNA–mRNA Interaction Networks

The CyTargetLinker plugin from Cytoscape was used to construct miRNA–gene interaction networks for the hub genes of the upregulated and downregulated genes. With respect to upregulated genes, 178 miRNAs were identified using the miRTarBase database, and 178 miRNAs were identified using the TargetScan database. With respect to downregulated genes, 93 miRNAs were identified using the miRTarBase database, and 150 miRNAs were identified using the TargetScan database. After setting an overlap threshold of two for the miRTarBase and TargetScan databases, 36 miRNAs were identified in the upregulated genes, and 17 miRNAs were identified in the downregulated genes. The miRNAs–genes are shown in **Figures 7A–C**. Specifically, 15 miRNAs that coregulate insulin growth factor 1 (IGF1), 10 miRNAs that coregulate SERPINE1, eight miRNAs that coregulate BMP2, six miRNAs that coregulate

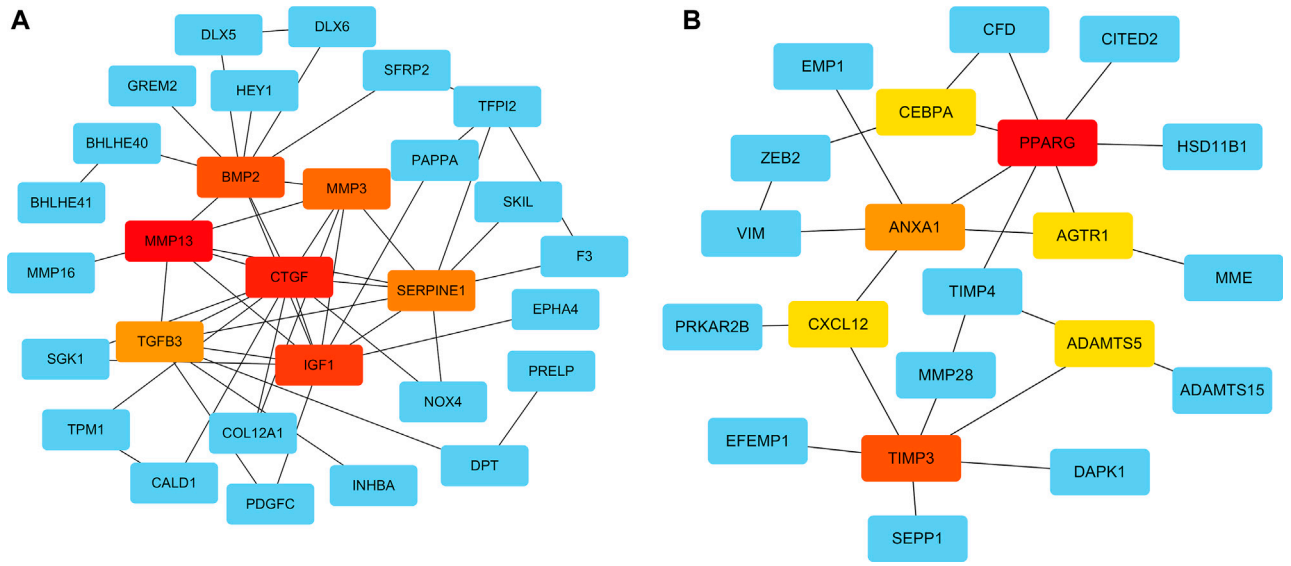

FIGURE 4 | Protein–protein interaction (PPI) networks of the upregulated and downregulated genes were constructed using Cytoscape software. The top seven upregulated **(A)** and downregulated **(B)** hub genes, according to the MCC, MNC, and Degree modules of the cytoHubba were identified.

connective tissue growth factor (CTGF), two miRNAs that coregulate MMP13, seven miRNAs that coregulate ADAMTS5, six miRNAs that coregulate TIMP3, four miRNAs that coregulate PPARG, and two miRNAs that coregulate CXCL12 were identified; six miRNAs (hub miRNAs) that coregulate osteogenic genes and adipogenic genes were also identified (**Table 4**).

DISCUSSION

hMSCs are self-renewing precursor cells that can differentiate into bone, fat, cartilage, and stromal cells of the bone marrow (Frenette et al., 2013). It has been reported that they are ideal seed cells for bone tissue engineering (Fan et al., 2020). Notably, however, the effective cultivation of BMSCs requires a good culture environment and a good *in vitro* culture technique. With increased cell culture time, cell proliferation and stability may be reduced. In the GSE84500 dataset (van Zoelen et al., 2016), adipogenic differentiation of hMSCs increased within 3 days in an optimized medium. Adipogenic differentiation and proliferation entered a plateau phase or began to increase more slowly from 4 to 7 days. Therefore, cells cultured via the GSE84500 dataset are stable and available within 1 week. In order to reduce false-positive results caused by operational error or culture conditions during the cell experiments and to acquire stable genes, the intersection of the DEGs of four time-points was used in the present study. Differential expression was detected at all four time-points (1, 2, 3, and 7 days). This could reduce false-positive results caused by mistakes at a singular time-point.

In the current study, samples were obtained from hMSCs from the mRNA microarray dataset GSE84500 in GEO, undergoing osteogenic and adipogenic differentiation. Through bioinformatics analysis, a total of 164 DEGs were identified, including 98 upregulated genes involved in osteogenic differentiation and 66 downregulated genes involved in adipogenic differentiation. GO enrichment analysis indicated that the upregulated genes were associated with negative regulation of the TGF-beta receptor pathway, skeletal system development, negative regulation of cell migration, bone mineralization, ECM, and extracellular space. Upregulated genes were closely related to bone formation, confirming that osteogenic differentiation of hMSCs could be induced in an optimized microenvironment. Interestingly, the upregulated genes were significantly related to the ECM, which provides a local structural and signaling environment that controls cell proliferation, differentiation, migration, and communication during development (Laczko and Csiszar, 2020). In a previous study, optimized ECM could induce stronger osteogenic effects in mesenchymal stem cells (Freeman et al., 2019). In another recent study, it was reported that ECM mineralization was critical for osteogenesis, and its dysregulation could result in osteoporosis (Hao et al., 2020). The results of the current study are concordant with those previous results. The downregulated genes were

TABLE 4 | Six hub miRNAs from the CyTargetLinker that coregulate five hub genes involved in osteogenic and adipogenic differentiation.

MiRNAs	The upregulated genes	The downregulated genes
hsa-miR-27a-3p	IGF1, MMP13	ADAMTS5, PPARG
hsa-miR-27b-3p	MMP13	ADAMTS5, PPARG
hsa-miR-128-3p	IGF1	ADAMTS5
hsa-miR-1-3p	IGF1	TIMP3
hsa-miR-98-5p	IGF1	ADAMTS5
hsa-miR-130b-3p	IGF1	PPARG

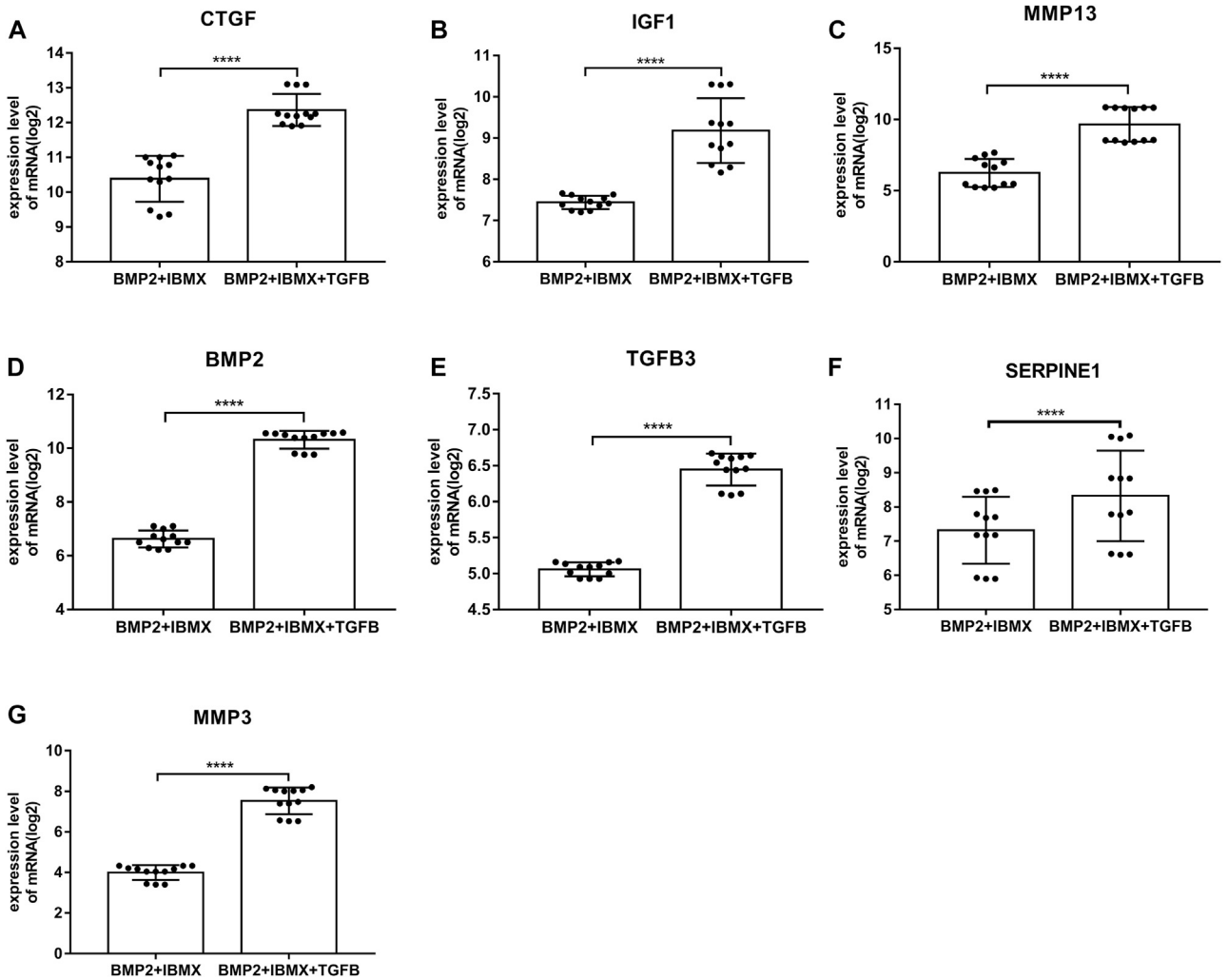

FIGURE 5 | mRNA expression levels of the top seven upregulated hub genes involved in osteogenic differentiation, derived from analysis of 24 samples from four time-points (1, 2, 3, and 7 days; presented on a log2 scale). The data shown are means ± SD. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

involved in the response to peptide hormone, Rho protein signal transduction, responses to mechanical stimuli, proteinaceous ECM, and extracellular space. Peptide hormones such as adiponectin (Kim et al., 2016), parathyroid hormone (Ehrenmann et al., 2019), visfatin (Tsiklauri et al., 2018), and insulin can regulate the metabolism of human tissues and organs and are closely associated with lipid metabolism. Rho GTPases and Rho kinases regulate cell proliferation, migration, and apoptosis by influencing cytoskeletal dynamic stimulation and cell shape (Wang et al., 2017). It has also been shown that Rho GTPase signaling pathways are involved in the regulation of osteoclast activity (Morel et al., 2018). Rho and Rho-related kinase two are inactivated during adipogenesis, which enhances the expression of pro-adipogenic genes, and then induces actin stress fiber loss (Diep et al., 2018). The results of the present study are consistent with those previous reports. In the CC category of GO enrichment, proteinaceous ECM and extracellular space were the most enriched, indicating that intercellular signaling is essential for adipogenic differentiation. In enrichment of KEGG pathway analysis, the main enriched signaling pathways were those regulating pluripotency of stem cells and the Hippo signaling pathway. Studies suggest that the Hippo/YAP1 signaling pathway can promote osteogenic differentiation of mesenchymal stem cells and inhibit their adipogenic differentiation (Zhong et al., 2013; Pan et al., 2018). Li et al. (2021) also reported that the Hippo signaling pathway regulates exosomes from hMSCs to promote osteogenic differentiation and bone formation, preventing osteoporosis. There are currently few reports on the Hippo signaling cascade and TGF-beta in osteogenic differentiation; however, this warrants further research. The downregulated genes indicated that the main enriched component in KEGG analysis was metabolism of xenobiotics by cytochrome P450. In a recent study, repression of cytochrome P450 2b led to obesity (Heintz et al., 2019), and cytochrome P450 2E1 deficiency resulted in reduced adipogenesis (Dang and Yun, 2021).

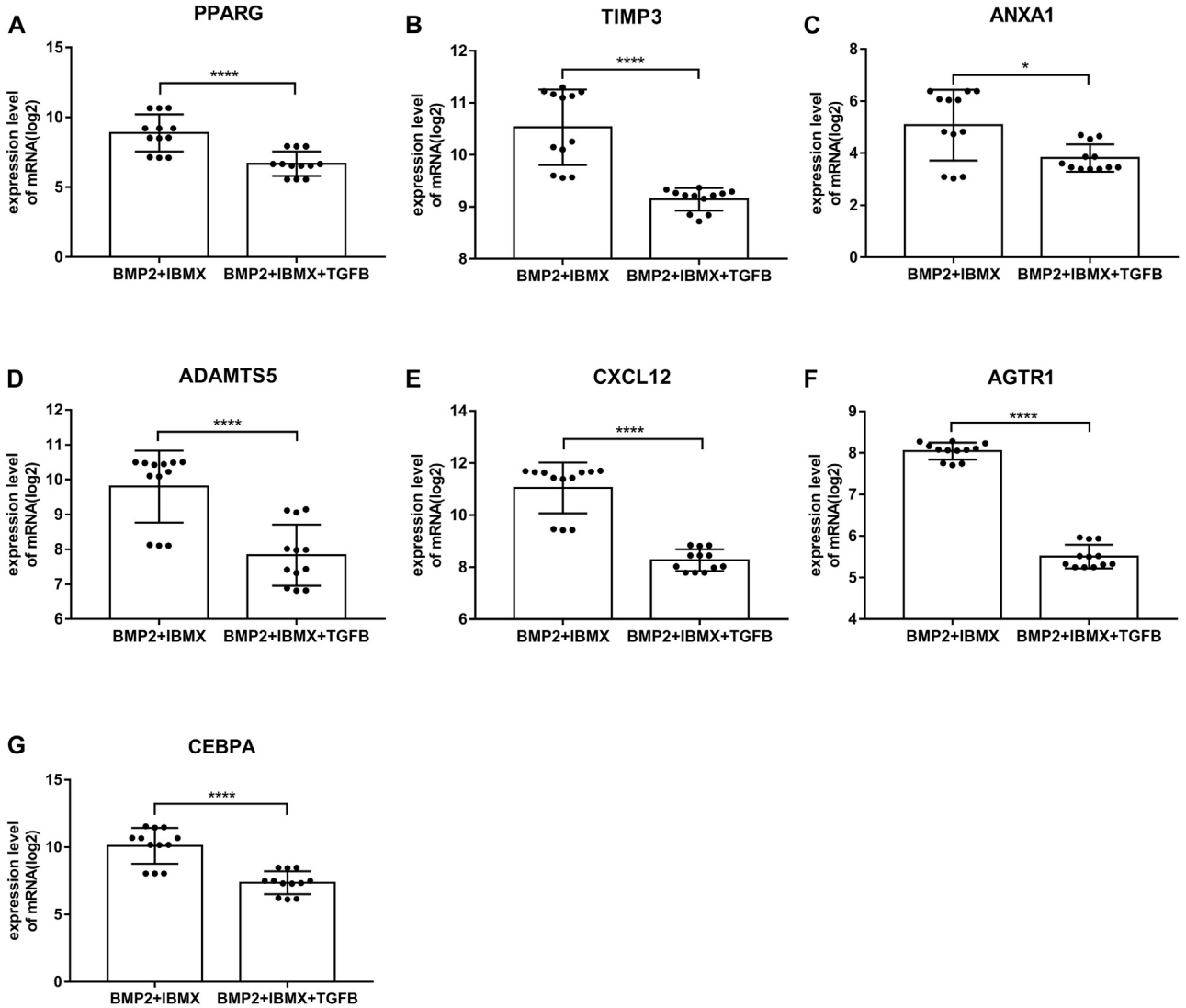

FIGURE 6 | mRNA expression levels of the top seven downregulated hub genes involved in adipogenic differentiation, derived from analysis of 24 samples from four time-points (1, 2, 3, and 7 days; presented on a log2 scale). The data shown are means ± SD. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

On the basis of DEGs, PPI networks of upregulated and downregulated genes were created. The hub genes involved in osteogenic differentiation were CTGF, ICF1, BMP2, MMP13, TGFB3, MMP3, and SERPINE1. The hub genes involved in adipogenic differentiation were PPARG, TIMP3, ANXA1, ADAMTS5, AGTR1, and CXCL12. BMP2 was identified as a master regulator of the differentiation of osteoblasts (Scarfì, 2016), and its overexpression promoted osteogenesis in mesenchymal stems (Cai et al., 2021). Experimental research has suggested that BMP2 is the only growth factor capable of singly inducing bone formation (Cai et al., 2021). CTGF/CCN2 is a matricellular protein that is secreted into the ECM. It is considered a cell adhesion protein, and osteoblasts cultured on a CTGF matrix exhibited enhanced bone nodule formation and matrix mineralization (Hendesi et al., 2015). IGF1 is a multifunctional peptide growth factor that can induce strong proliferation and osteogenic differentiation in BMSCs (Wu et al., 2020b; Feng and Meng, 2021). During osteogenic differentiation, high expression of MMP13 in hMSCs grew on a type I collagen matrix. Additionally, knocking down MMP13 reduced the osteogenic differentiation of hMSCs on a type I collagen matrix (Arai et al., 2021). TGFB3 is a classic growth factor involved in bone generation (Yoon et al., 2018), and its overexpression upregulates alkaline phosphatase activity and induces the osteogenic differentiation of BMSCs (He et al., 2019). It also induces chondrogenesis of hMSCs (Uzieliene et al., 2021). Interactions between SERPINE1 and MMP3 and osteogenic differentiation have rarely been described, however, and warrant future research. Among the downregulated hub genes, peroxisome proliferator-activated receptor-gamma (PPARG) is a critical transcription factor of adipogenesis that is important in the formation of mature adipocytes (Stachecka

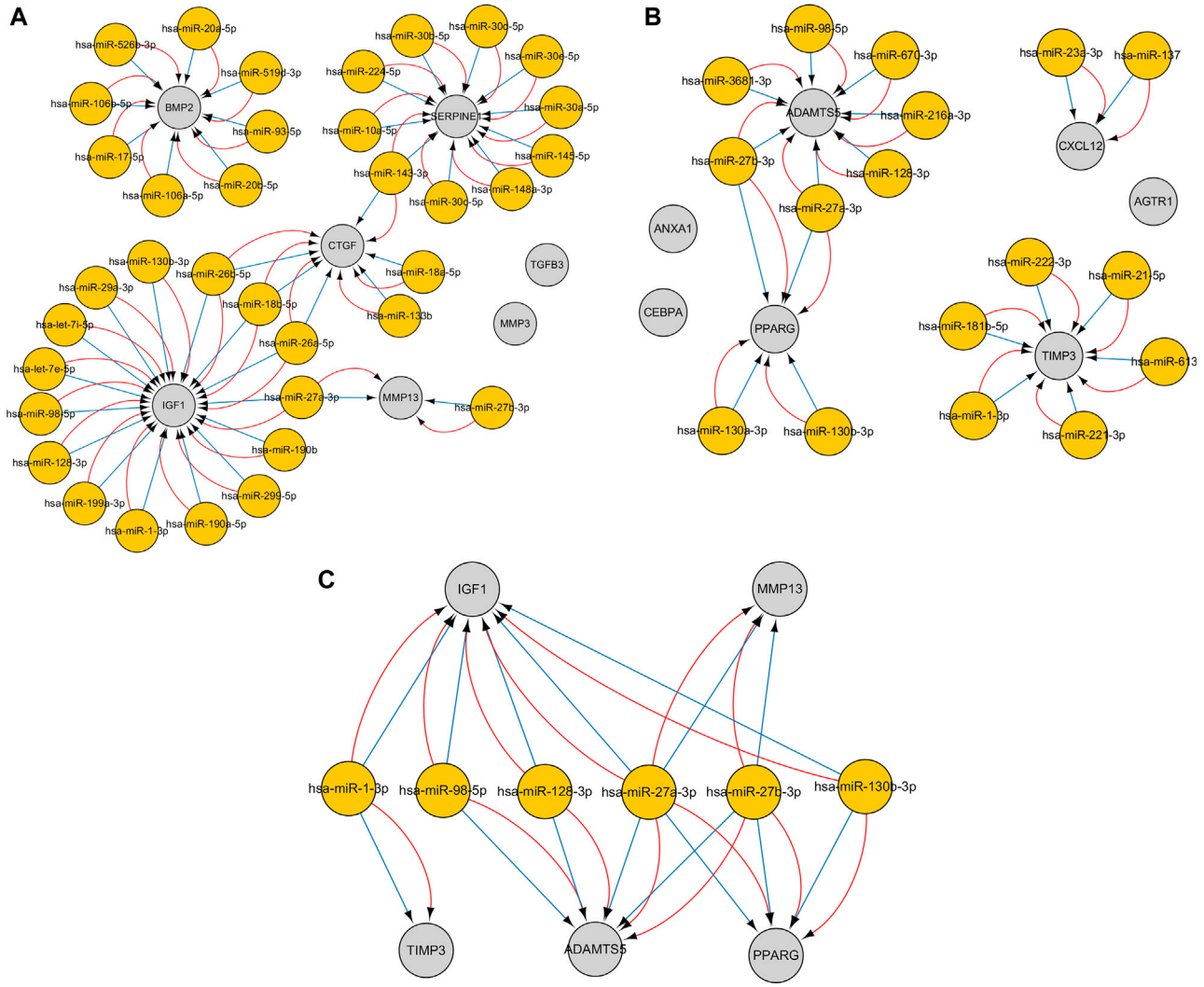

FIGURE 7 | The miRNA–gene interaction networks for the top seven hub genes of the upregulated and downregulated genes were constructed. An overlap threshold was used to display only microRNA–target interactions (MTIs) in two regulatory interaction networks. Genes are indicated by gray circles, and miRNAs are indicated by yellow circles. Blue edges represent MTIs from the TargetScan v7.2 database, and red edges represent MTIs from the MiRTarBase v8.0 database. **(A)** MiRNA–gene pairs of the top seven hub genes of the upregulated genes. **(B)** MiRNA–gene pairs of the top seven hub genes of the downregulated genes. **(C)** Six miRNAs (hub miRNAs) coregulate two osteogenic genes and three adipogenic genes.

et al., 2019). Some studies indicated that PPARG could be used as a new target for weight loss drugs. CEBPA acts as an adipogenic factor and is a key component in adipocyte differentiation (Gao et al., 2015). ADAMTS5 is the major protease that cleaves aggrecan; it reportedly promotes adipogenesis *in vitro* and *in vivo* in an established murine model (Bauters et al., 2016). PPARG, ADAMTS5, TIMP4, ANXA1, AGTR1, and CXCL12 genes are evidently associated with obesity, suggesting that the influence of these genes on obesity may be similar to the influence of fat accumulation in hMSCs. Furthermore, inhibitors of PPARG and ADAMTS5 can block the adipogenic differentiation of hMSCs (van Zoelen et al., 2016). Thus, these genes and corresponding inhibitors could be used as targets for drug development. To further confirm the accuracy of these hub genes, the mRNA expression levels of these hub genes were statistically analyzed. They were significantly higher in the BIT group than in the BI group, whereas the mRNA expression levels of the downregulated hub genes were significantly higher in the BI group than in the BIT group. This was because mesenchymal stem cells tend to differentiate into osteoblasts and inhibit adipogenic differentiation under the regulation of TGF-beta. All hub genes exhibited statistically significant differences. PPARG, ADAMTS5, AGTR1, and CXCL12 expression levels were consistent with a previous report (van Zoelen et al., 2016). Therefore, they are potential therapeutic targets for osteoporosis or obesity.

Integrated miRNA–mRNA regulatory networks of hub genes were constructed to improve understanding of potential molecular relationships between adipogenic differentiation and osteogenic differentiation in osteoporosis. To ensure the

reliability and accuracy of the results, an overlap threshold of two was set for the miRTarBase and TargetScan databases to identify miRNA–gene interactions. Overall, 36 miRNAs were identified in the upregulated hub genes, which were mainly enriched in bone mineralization and the Hippo signaling pathway, whereas 17 miRNAs were identified in the downregulated hub genes, which were mainly enriched in the response to peptide hormone and pathways in cancer. Research has shown that a miRNA can target a number of genes, and a gene can be targeted by various miRNAs (Zhao et al., 2020). In the current study, a single gene was regulated by multiple miRNAs, and these miRNAs were experimentally validated. Interestingly, the results showed that some osteogenic genes and adipogenic genes were regulated by the same miRNA; for example, IGF1, MMP13, PPARG, and ADAMTS5 were regulated by hsa-miR-27a-3p; and ADAMTS5, PPARG, and MMP13 were regulated by hsa-miR-27b-3p. This may be because the miRNAs have the ability to bidirectionally regulate target genes. For example, a miR-149-3p mimic reduced the adipogenic differentiation potential of BMSCs and enhanced their osteogenic differentiation potential (Li et al., 2019). These hub miRNA–mRNA pairs may be therapeutic targets in osteoporosis.

In the current study, an integrated bioinformatics approach and strict screening conditions were used to process datasets. Hub genes were verified using the unpaired t-test. But the study had some limitations. The number of samples in the dataset was small, and larger samples are needed to confirm the study results. The study was based on microarray data obtained *in vitro*, and more *in vitro* and *in vivo* experiments are required to further verify the results. Lastly, the specific regulatory relationship between miRNAs and mRNAs was not further confirmed, and the transformation relationship between adipogenic differentiation and osteogenic differentiation required further confirmation. Nonetheless, we think that the results of the study are valuable and reliable. Identification of the DEGs was derived from the intersection of four time points, which reduced the likelihood of false-positive results. Most of the downregulated hub genes were consistent with van Zoelen et al. (2016). Hub miRNAs were selected from the intersection of two databases, of which miRTarBase is dedicated to collecting MTIs with experimental evidence. These results could provide a reference on osteoporosis or senile obesity, or for bioinformatics research, but more experiments are needed to support the results of the present study.

CONCLUSION

Microarray and bioinformatics approaches were used to identify DEGs involved in adipogenic differentiation and osteogenic differentiation in hMSCs and to identify functions and pathways that the DEGs were involved in. Hub genes of osteogenic differentiation and adipogenic differentiation were identified, and their miRNA–mRNA regulation networks were constructed. The study provides new insight into the osteogenic differentiation and adipogenic differentiation of hMSCs. The hub genes/miRNAs identified may provide a basis for the screening of biomarkers related to osteoporosis or obesity, or for developing new therapies and drugs for osteoporosis or obesity.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the 360 repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GD, YL, and XL conceived and designed the research. GD and XC collected the data, generated the figures based on bioinformatics and online databases, and wrote the manuscript. ZZ and LH analyzed the data and performed literature searches. KW studied the background of the disease. YL reviewed the manuscript. XL supervised the project and reviewed and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Science and Technology Planning Project of Shenzhen (grant numbers JCYJ20180302144355408 and JCYJ20190808100818959), the Administration Project of Traditional Chinese Medicine of Guangdong Province (grant number 20201298), and the Science and Technology Planning Project of Bao'an District, Shenzhen (grant number 2020JD497).

ACKNOWLEDGMENTS

We acknowledge the GEO database for providing the platform and the contributors for uploading their meaningful datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.759596/full#supplementary-material

REFERENCES

Arai, Y., Choi, B., Kim, B. J., Park, S., Park, H., Moon, J. J., et al. (2021). Cryptic Ligand on Collagen Matrix Unveiled by MMP13 Accelerates Bone Tissue Regeneration via MMP13/Integrin α3/RUNX2 Feedback Loop. *Acta Biomater.* 125, 219–230. doi:10.1016/j.actbio.2021.02.042

Atashi, F., Modarressi, A., and Pepper, M. S. (2015). The Role of Reactive Oxygen Species in Mesenchymal Stem Cell Adipogenic and Osteogenic Differentiation: A Review. *Stem Cells Dev.* 24 (10), 1150–1163. doi:10.1089/scd.2014.0484

Bauters, D., Scroyen, I., Deprez-Poulain, R., and Lijnen, H. R. (2016). ADAMTS5 Promotes Murine Adipogenesis and Visceral Adipose Tissue Expansion. *Thromb. Haemost.* 116 (4), 694–704. doi:10.1160/TH16-01-0015

Black, D. M., and Rosen, C. J. (2016). Postmenopausal Osteoporosis. *N. Engl. J. Med.* 374 (3), 254–262. doi:10.1056/NEJMcp1513724

Cai, H., Zou, J., Wang, W., and Yang, A. (2021). BMP2 Induces hMSC Osteogenesis and Matrix Remodeling. *Mol. Med. Rep.* 23 (2), 125. doi:10.3892/mmr.2020.11764

Chang, P. Y., Feldman, D., Stefanick, M. L., McDonnell, D. P., Thompson, B. M., McDonald, J. G., et al. (2019). 27-Hydroxycholesterol, an Endogenous SERM, and Risk of Fracture in Postmenopausal Women: A Nested Case-Cohort Study in the Women's Health Initiative. *J. Bone Miner. Res.* 34 (1), 59–66. doi:10.1002/jbmr.3576

Dang, T. T. H., and Yun, J. W. (2021). Cytochrome P450 2E1 (CYP2E1) Positively Regulates Lipid Catabolism and Induces browning in 3T3-L1 white Adipocytes. *Life Sci.* 278, 119648. doi:10.1016/j.lfs.2021.119648

Diep, D. T. V., Hong, K., Khun, T., Zheng, M., Ul-Haq, A., Jun, H.-S., et al. (2018). Anti-adipogenic Effects of KD025 (SLx-2119), a ROCK2-specific Inhibitor, in 3T3-L1 Cells. *Sci. Rep.* 8 (1), 2477. doi:10.1038/s41598-018-20821-3

Duque, G. (2008). Bone and Fat Connection in Aging Bone. *Curr. Opin. Rheumatol.* 20 (4), 429–434. doi:10.1097/BOR.0b013e3283025e9c

Eastell, R., Rosen, C. J., Black, D. M., Cheung, A. M., Murad, M. H., and Shoback, D. (2019). Pharmacological Management of Osteoporosis in Postmenopausal Women: An Endocrine Society* Clinical Practice Guideline. *J. Clin. Endocrinol. Metab.* 104 (5), 1595–1622. doi:10.1210/jc.2019-00221

Ehrenmann, J., Schöppe, J., Klenk, C., and Plückthun, A. (2019). New Views into Class B GPCRs from the crystal Structure of PTH1R. *FEBS J.* 286 (24), 4852–4860. doi:10.1111/febs.15115

Ensrud, K. E., and Crandall, C. J. (2017). Osteoporosis. *Ann. Intern. Med.* 167 (3), ITC17–C32. doi:10.7326/AITC201708010

Fan, T., Qu, R., Yu, Q., Sun, B., Jiang, X., Yang, Y., et al. (2020). Bioinformatics Analysis of the Biological Changes Involved in the Osteogenic Differentiation of Human Mesenchymal Stem Cells. *J. Cel. Mol. Med.* 24 (14), 7968–7978. doi:10.1111/jcmm.15429

Feng, J., and Meng, Z. (2021). Insulin Growth Factor-1 P-romotes the P-roliferation and O-steogenic D-ifferentiation of B-one M-arrow M-esenchymal S-tem C-ells through the Wnt/β-catenin P-athway. *Exp. Ther. Med.* 22 (2), 891. doi:10.3892/etm.2021.10323

Freeman, F., Browe, D. C., Browe, D., Nulty, J., Von Euw, S., Grayson, W., et al. (2019). Biofabrication of Multiscale Bone Extracellular Matrix Scaffolds for Bone Tissue Engineering. *eCM* 38, 168–187. doi:10.22203/eCM.v038a12

Frenette, P. S., Pinho, S., Lucas, D., and Scheiermann, C. (2013). Mesenchymal Stem Cell: Keystone of the Hematopoietic Stem Cell Niche and a Stepping-Stone for Regenerative Medicine. *Annu. Rev. Immunol.* 31, 285–316. doi:10.1146/annurev-immunol-032712-095919

Gao, Y., Sun, Y., Duan, K., Shi, H., Wang, S., Li, H., et al. (2015). CpG Site DNA Methylation of theCCAAT/enhancer-Binding Protein, Alphapromoter in Chicken Lines Divergently Selected for Fatness. *Anim. Genet.* 46 (4), 410–417. doi:10.1111/age.12326

Haasters, F., Docheva, D., Gassner, C., Popov, C., Böcker, W., Mutschler, W., et al. (2014). Mesenchymal Stem Cells from Osteoporotic Patients Reveal Reduced Migration and Invasion upon Stimulation with BMP-2 or BMP-7. *Biochem. Biophysical Res. Commun.* 452 (1), 118–123. doi:10.1016/j.bbrc.2014.08.055

Halloran, D., Durbano, H. W., and Nohe, A. (2020). Bone Morphogenetic Protein-2 in Development and Bone Homeostasis. *Jdb* 8 (3), 19. doi:10.3390/jdb8030019

Hao, Q., Liu, Z., Lu, L., Zhang, L., and Zuo, L. (2020). Both JNK1 and JNK2 Are Indispensable for Sensitized Extracellular Matrix Mineralization in IKKβ-Deficient Osteoblasts. *Front. Endocrinol.* 11, 13. doi:10.3389/fendo.2020.00013

He, W., Chen, L., Huang, Y., Xu, Z., Xu, W., Ding, N., et al. (2019). Synergistic Effects of Recombinant Lentiviral-Mediated BMP2 and TGF-Beta3 on the Osteogenic Differentiation of Rat Bone Marrow Mesenchymal Stem Cells *In Vitro*. *Cytokine* 120, 1–8. doi:10.1016/j.cyto.2019.03.020

Heintz, M. M., Kumar, R., Rutledge, M. M., and Baldwin, W. S. (2019). Cyp2b-null Male Mice Are Susceptible to Diet-Induced Obesity and Perturbations in Lipid Homeostasis. *J. Nutr. Biochem.* 70, 125–137. doi:10.1016/j.jnutbio.2019.05.004

Hendesi, H., Barbe, M. F., Safadi, F. F., Monroy, M. A., and Popoff, S. N. (2015). Integrin Mediated Adhesion of Osteoblasts to Connective Tissue Growth Factor (CTGF/CCN2) Induces Cytoskeleton Reorganization and Cell differentiationJournal Article; Research Support. *PLoS One* 10 (2), e0115325. doi:10.1371/journal.pone.0115325

Hu, L., Yin, C., Zhao, F., Ali, A., Ma, J., and Qian, A. (2018). Mesenchymal Stem Cells: Cell Fate Decision to Osteoblast or Adipocyte and Application in Osteoporosis Treatment. *Ijms* 19 (2), 360. doi:10.3390/ijms19020360

Kim, H. Y., Bae, E. H., Ma, S. K., Chae, D. W., Choi, K. H., Kim, Y.-S., et al. (2016). Association of Serum Adiponectin Level with Albuminuria in Chronic Kidney Disease Patients. *Clin. Exp. Nephrol.* 20 (3), 443–449. doi:10.1007/s10157-015-1173-4

Laczko, R., and Csiszar, K. (2020). Lysyl Oxidase (LOX): Functional Contributions to Signaling Pathways. *Biomolecules* 10 (8), 1093. doi:10.3390/biom10081093

Li, L., Zhou, X., Zhang, J.-t., Liu, A.-f., Zhang, C., Han, J.-c., et al. (2021). Exosomal miR-186 Derived from BMSCs Promote Osteogenesis through Hippo Signaling Pathway in Postmenopausal Osteoporosis. *J. Orthop. Surg. Res.* 16 (1), 23. doi:10.1186/s13018-020-02160-0

Li, Y., Yang, F., Gao, M., Gong, R., Jin, M., Liu, T., et al. (2019). MiR-149-3p Regulates the Switch between Adipogenic and Osteogenic Differentiation of BMSCs by Targeting FTO. *Mol. Ther. - Nucleic Acids* 17, 590–600. doi:10.1016/j.omtn.2019.06.023

McClung, M. R., O'Donoghue, M. L., Papapoulos, S. E., Bone, H., Langdahl, B., Saag, K. G., et al. (2019). Odanacatib for the Treatment of Postmenopausal Osteoporosis: Results of the LOFT Multicentre, Randomised, Double-Blind, Placebo-Controlled Trial and LOFT Extension Study. *Lancet Diabetes Endocrinol.* 7 (12), 899–911. doi:10.1016/S2213-8587(19)30346-8

Morel, A., Blangy, A., and Vives, V. (2018). Methods to Investigate the Role of Rho GTPases in Osteoclast Function. *Methods Mol. Biol.* 1821, 219–233. doi:10.1007/978-1-4939-8612-5_15

Pan, J.-X., Xiong, L., Zhao, K., Zeng, P., Wang, B., Tang, F.-L., et al. (2018). YAP Promotes Osteogenesis and Suppresses Adipogenic Differentiation by Regulating β-catenin Signaling. *Bone Res.* 6, 18. doi:10.1038/s41413-018-0018-7

Rosen, C. J., and Bouxsein, M. L. (2006). Mechanisms of Disease: Is Osteoporosis the Obesity of Bone? *Nat. Rev. Rheumatol.* 2 (1), 35–43. doi:10.1038/ncprheum0070

Sambrook, P., and Cooper, C. (2006). Osteoporosis. *The Lancet* 367 (9527), 2010–2018. doi:10.1016/S0140-6736(06)68891-0

Scarfì, S. (2016). Use of Bone Morphogenetic Proteins in Mesenchymal Stem Cell Stimulation of Cartilage and Bone Repair. *Wjsc* 8 (1), 1–12. doi:10.4252/wjsc.v8.i1.1

Souza, A. T. P., Freitas, G. P., Lopes, H. B., Totoli, G. G. C., Tarone, A. G., Marostica-Junior, M. R., et al. (2021). Jabuticaba Peel Extract Modulates Adipocyte and Osteoblast Differentiation of MSCs from Healthy and Osteoporotic Rats. *J. Bone Miner. Metab.* 39 (2), 163–173. doi:10.1007/s00774-020-01152-8

Stachecka, J., Nowacka-Woszuk, J., Kolodziejski, P. A., and Szczerbal, I. (2019). The Importance of the Nuclear Positioning of the PPARG Gene for its Expression during Porcine *In Vitro* Adipogenesis. *Chromosome Res.* 27 (3), 271–284. doi:10.1007/s10577-019-09604-2

Tang, S. Y., and Alliston, T. (2013). Regulation of Postnatal Bone Homeostasis by TGFβ. *Bonekey Rep.* 2, 255. doi:10.1038/bonekey.2012.255

Tsiklauri, L., Werner, J., Kampschulte, M., Frommer, K. W., Berninger, L., Irrgang, M., et al. (2018). Visfatin Alters the Cytokine and Matrix-Degrading Enzyme Profile during Osteogenic and Adipogenic MSC Differentiation. *Osteoarthritis and Cartilage* 26 (9), 1225–1235. doi:10.1016/j.joca.2018.06.001

Uzieliene, I., Bagdonas, E., Hoshi, K., Sakamoto, T., Hikita, A., Tachtamisevaite, Z., et al. (2021). Different Phenotypes and Chondrogenic Responses of Human Menstrual Blood and Bone Marrow Mesenchymal Stem Cells to Activin A and TGF-B3. *Stem Cell Res. Ther.* 12 (1), 251. doi:10.1186/s13287-021-02286-w

van Zoelen, E. J., Duarte, I., Hendriks, J. M., and van der Woning, S. P. (2016). Tgfβ-Induced Switch from Adipogenic to Osteogenic Differentiation of Human Mesenchymal Stem Cells: Identification of Drug Targets for Prevention of Fat Cell Differentiation. *Stem Cell Res. Ther.* 7 (1), 123. doi:10.1186/s13287-016-0375-3

Vellucci, R., Mediati, R. D., and Ballerini, G. (2014). Use of Opioids for Treatment of Osteoporotic Pain. *ccmbm* 11 (3), 173–176. doi:10.11138/ccmbm/2014.11.3.173

Wang, T., Kang, W., Du, L., and Ge, S. (2017). Rho-kinase Inhibitor Y-27632 Facilitates the Proliferation, Migration and Pluripotency of Human Periodontal

Ligament Stem Cells. *J. Cell. Mol. Med.* 21 (11), 3100–3112. doi:10.1111/jcmm.13222

Wu, J., Cai, P., Lu, Z., Zhang, Z., He, X., Zhu, B., et al. (2020a). Identification of Potential Specific Biomarkers and Key Signaling Pathways between Osteogenic and Adipogenic Differentiation of hBMSCs for Osteoporosis Therapy. *J. Orthop. Surg. Res.* 15 (1), 437. doi:10.1186/s13018-020-01965-3

Wu, L., Zhang, G., Guo, C., and Pan, Y. (2020b). Intracellular $Ca2^{+}$ Signaling Mediates IGF-1-Induced Osteogenic Differentiation in Bone Marrow Mesenchymal Stem Cells. *Biochem. Biophysical Res. Commun.* 527 (1), 200–206. doi:10.1016/j.bbrc.2020.04.048

Yoon, S.-J., Yoo, Y., Nam, S., Hyun, H., Lee, D.-W., Um, S., et al. (2018). The Cocktail Effect of BMP-2 and TGF-B1 Loaded in Visible Light-Cured Glycol Chitosan Hydrogels for the Enhancement of Bone Formation in a Rat Tibial Defect Model. *Mar. Drugs* 16 (10), 351. doi:10.3390/md16100351

Yu, W., Zhong, L., Yao, L., Wei, Y., Gui, T., Li, Z., et al. (2021). Bone Marrow Adipogenic Lineage Precursors Promote Osteoclastogenesis in Bone Remodeling and Pathologic Bone Loss. *J. Clin. Invest.* 131 (2), e140214. doi:10.1172/JCI140214

Zhao, H., Chang, A., Ling, J., Zhou, W., Ye, H., and Zhuo, X. (2020). Construction and Analysis of miRNA-mRNA Regulatory Networks in the Radioresistance of Nasopharyngeal Carcinoma. *3 Biotech.* 10 (12), 511. doi:10.1007/s13205-020-02504-x

Zhi, F., Ding, Y., Wang, R., Yang, Y., Luo, K., and Hua, F. (2021). Exosomal Hsa_circ_0006859 Is a Potential Biomarker for Postmenopausal Osteoporosis and Enhances Adipogenic versus Osteogenic Differentiation in Human Bone Marrow Mesenchymal Stem Cells by Sponging miR-431-5p. *Stem Cel Res. Ther.* 12 (1), 157. doi:10.1186/s13287-021-02214-y

Zhong, W., Tian, K., Zheng, X., Li, L., Zhang, W., Wang, S., et al. (2013). Mesenchymal Stem Cell and Chondrocyte Fates in a Multishear Microdevice Are Regulated by Yes-Associated Protein. *Stem Cells Dev.* 22 (14), 2083–2093. doi:10.1089/scd.2012.0685

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

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

Copyright © 2021 Du, Cheng, Zhang, Han, Wu, Li and Lin. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 December 2021
doi: 10.3389/fgene.2021.767358

Importance of SNP Dependency Correction and Association Integration for Gene Set Analysis in Genome-Wide Association Studies

Michal Marczyk[1,2], *Agnieszka Macioszek*[1], *Joanna Tobiasz*[1], *Joanna Polanska*[1]* and *Joanna Zyla*[1]

[1]Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland, [2]Yale Cancer Center, Yale School of Medicine, New Haven, CT, United States

A typical genome-wide association study (GWAS) analyzes millions of single-nucleotide polymorphisms (SNPs), several of which are in a region of the same gene. To conduct gene set analysis (GSA), information from SNPs needs to be unified at the gene level. A widely used practice is to use only the most relevant SNP per gene; however, there are other methods of integration that could be applied here. Also, the problem of nonrandom association of alleles at two or more loci is often neglected. Here, we tested the impact of incorporation of different integrations and linkage disequilibrium (LD) correction on the performance of several GSA methods. Matched normal and breast cancer samples from The Cancer Genome Atlas database were used to evaluate the performance of six GSA algorithms: Coincident Extreme Ranks in Numerical Observations (CERNO), Gene Set Enrichment Analysis (GSEA), GSEA-SNP, improved GSEA for GWAS (i-GSEA4GWAS), Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA), and Over-Representation Analysis (ORA). Association of SNPs to phenotype was calculated using modified McNemar's test. Results for SNPs mapped to the same gene were integrated using Fisher and Stouffer methods and compared with the minimum *p*-value method. Four common measures were used to quantify the performance of all combinations of methods. Results of GSA analysis on GWAS were compared to the one performed on gene expression data. Comparing all evaluation metrics across different GSA algorithms, integrations, and LD correction, we highlighted CERNO, and MAGENTA with Stouffer as the most efficient. Applying LD correction increased prioritization and specificity of enrichment outcomes for all tested algorithms. When Fisher or Stouffer were used with LD, sensitivity and reproducibility were also better. Using any integration method was beneficial in comparison with a minimum *p*-value method in specific combinations. The correlation between GSA results from genomic and transcriptomic level was the highest when Stouffer integration was combined with LD correction. We thoroughly evaluated different approaches to GSA in GWAS in terms of performance to guide others to select the most effective combinations. We showed that LD correction and

OPEN ACCESS

Edited by:
Sorin Draghici,
Wayne State University, United States

Reviewed by:
Rostam Abdollahi-Arpanahi,
University of Georgia, United States
Le Li,
Cornell University, United States

***Correspondence:**
Joanna Polanska
joanna.polanska@polsl.pl

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

Received: *30 August 2021*
Accepted: *10 November 2021*
Published: *09 December 2021*

Citation:
Marczyk M, Macioszek A, Tobiasz J, Polanska J and Zyla J (2021) Importance of SNP Dependency Correction and Association Integration for Gene Set Analysis in Genome-Wide Association Studies. Front. Genet. 12:767358. doi: 10.3389/fgene.2021.767358

Stouffer integration could increase the performance of enrichment analysis and encourage the usage of these techniques.

Keywords: gene set analysis, genome-wide association study, statistical integration, single-nucleotide polymorphism, linkage disequilibrium correction

INTRODUCTION

Genome-wide association study (GWAS) is a high-throughput molecular biology technique, which gives insight into understanding the relation of single-nucleotide polymorphism (SNP) frequency and other types of genetic variations with particular traits. In recent years, GWAS reveals plenty of genetic locations related to common diseases, e.g., type 2 diabetes (Billings and Florez, 2010), Alzheimer disease (Marioni et al., 2018), or many types of cancer (Sud et al., 2017). Despite the promising outcomes, the biological functions of many genetic variation loci remain unclear, and the genetic mechanisms of phenotypes are not systematically explained. Yet, the GWAS is still an important tool used to understand the biological mechanisms of different diseases (Wijmenga and Zhernakova, 2018). One of the bioinformatic techniques, which can extend the amount of information from single genetic variations and their impact on the biological systems, is gene set analysis (GSA), and the importance of such a solution has been recently noticed (Wang et al., 2007; Holden et al., 2008; Hirschhorn, 2009; Zhang et al., 2010; Weng et al., 2011; de Leeuw et al., 2015; Mei et al., 2016; Sud et al., 2017; Yoon et al., 2018; Maleki et al., 2020).

The GSA allows summarizing the results of association with phenotype from individual gene level to gene set level, also known as pathway level. Using this concept, it is possible to detect the aggregated impact of multiple genes on phenotype, even when the individual gene has moderate or small effect on the investigated trait. In addition, applying GSA increases understanding of changes observed in complex biological mechanisms under various conditions. Within the last decade of gene set analysis method development, many algorithms were introduced [just to mention a few: GSEA (Subramanian et al., 2005), PADOG (Tarca et al., 2012), SPIA (Tarca et al., 2009) or LEGO (Dong et al., 2016)] and can be classified by their generation (Khatri et al., 2012; Zyla et al., 2017), hypothesis tested (Maciejewski, 2014), or application for a particular omics platform (Das et al., 2020). First, algorithms were created to analyze the gene expression data from microarray experiments, but with rapid advancement in molecular biology techniques, they became widely applied in other omics, resulting in growth of bioinformatic tools, which perform multi-omics gene set analysis (Canzler and Hackermüller, 2020; Kaspi and Ziemann, 2020). Application of GSA techniques to dissimilar omics data is associated with different problems. In the analysis of GWAS results, the key issue is how to transform the observed genetic variation into gene level. One of the most used techniques is to choose the SNP with the strongest association (minimum p-value) to represent a gene (Wang et al., 2007; Zhang et al., 2010). The minimum p-value approach may not be an optimal solution as it favors long genes with many SNPs measured, where obtaining stronger association is more likely compared with shorter ones. Thus, some adjustments were introduced to correct this effect, e.g., adaptive p-value combination of p-values (Yu et al., 2009), selecting representative SNPs for each gene (Weng et al., 2011), or correction of smallest p-value due to some factors, like no. of SNPs per kb, gene size, and linkage disequilibrium units per kb (Segrè et al., 2010). Other aggregation techniques, like Fisher integration, second minimum p-value, or application of Simes' p-value adjusted for the number of SNPs were also proposed (Mei et al., 2016). However, the authors applied those approaches only to the oldest gene set analysis method based on hypergeometric test [over-representation analysis (ORA)]. Also, they performed only basic evaluation, concentrating mostly on detecting target pathways for the analyzed dataset without looking at false positives. Recently, a new method of GSA in GWAS was introduced and compared with other methods by Type 1 error control and statistical power (Yoon et al., 2018; Sun et al., 2019), but without testing different integration methods or SNP dependency correction. Finally, there are solutions where the problem of aggregation from genome to transcriptome level was neglected, e.g., MAGMA (de Leeuw et al., 2015) or GSEA-SNP (Holden et al., 2008).

Even though GSA methods have been used for over a decade in omics data analysis, there still exist many challenges in this research field (Maleki et al., 2020). The knowledge about GSA algorithm efficiency was widely updated in several publications (Mitrea et al., 2013; Tarca et al., 2013; Maleki et al., 2019; Nguyen et al., 2019; Zyla et al., 2019; Geistlinger et al., 2021; Xie et al., 2021). Yet, those studies concentrated on enrichment methods dedicated to gene expression data measured with microarrays or RNA sequencing (RNASeq) technologies, and the overall performance of GSA algorithms in other omics is still not known. In this work, we focused on two major difficulties that occur during applying GSA in GWAS studies. The first goal of the study was to test the impact of aggregation of phenotype association test results from SNP to gene level, which is then transformed to gene set level. For this purpose, three statistical integration techniques were tested in a variety of GSA algorithms. The second goal was to investigate the impact of linkage disequilibrium (LD) control in the process of SNP information aggregation. These two GSA extensions were tested in combination with six gene set analysis methods. Each tested combination of algorithms was evaluated in terms of sensitivity, specificity, prioritization, and reproducibility of gene set analysis. Furthermore, the relation of GSA GWAS results to those obtained on gene expression data was investigated using the same collection of patient samples. Finally, all tested GSA algorithms,

integration techniques, and LD correction were implemented in R package *intGSASNP* (integrative GSA for SNP).

MATERIALS AND METHODS

Data

Data from Affymetrix Genome-Wide Human SNP Array 6.0 platform served for SNP genotyping. Affymetrix SNP 6.0 microarrays include over 906,600 SNPs and over 946,000 probes for copy number variation detection (Affymetrix, 2021). All files are part of The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma collection (Berger et al., 2018) and were downloaded in CEL format from the Genomic Data Commons (GDC) Legacy Archive. Only white female breast cancer patients were selected for the study. For all individuals, data for both primary tumors and solid normal tissues were available.

Subsequently, Illumina paired-end RNA sequencing data were used for the same patients and the same samples (both tumor and normal) as in the case of SNP genotyping. All data files were downloaded from GDC Data Portal as HTSeq-counts. GDC previously preprocessed raw sequencing data according to the bioinformatics pipeline available from GDC Documentation (NCI Genomic Data Commons, 2021).

Consequently, 83 white females were considered. For all of them, RNASeq and SNP microarray results were available for both tumor and normal tissue fresh frozen specimens, which summed up to 166 samples. Hence, all individuals were matched in terms of sample and experiment types. Specimens were collected at five different centers (tissue source sites) participating in TCGA Breast Invasive Carcinoma project. All patients were labeled with a breast cancer subtype as previously described (Marczyk et al., 2019; Marczyk et al., 2020). The summary of breast cancer subtypes, source center, and patient ethnicity is presented in **Supplementary Table S1**.

Single-Nucleotide Polymorphism Data Analysis

For each genotyped SNP, genome location and relation to transcriptomic function were mapped using the ENSEMBL human genome database, v80 (May 2015; http://may2015.archive.ensembl.org; *biomaRt* R package) (Cunningham et al., 2015). During the process of quality control, multiple SNPs were filtered out due to minor allele frequency (MAF; lower than 5% removed) and Hardy–Weinberg equilibrium (HWE; p-value <0.05). Next, only SNPs that are located within the range of 5 kb upstream and 5 kb downstream of the gene were selected. The selected boundary is much narrower in comparison with other studies [e.g., (Segrè et al., 2010; Zhang et al., 2010)], but here, we wanted to reflect the strongest association to possible changes in gene expression. These steps reduced the initial number of SNPs from 905,176 to 240,799. Finally, to compute the association between genotypes and phenotype (breast cancer *vs.* healthy tissue) under genotype genetic model (AA/AB/BB) with the paired design, the multinomial exact test (extension of McNemar's test) was performed with 100,000 Monte Carlo permutations using *rcompanion* R package (Mangiafico, 2016). This method does not allow introducing additional covariates in the analysis. As the collected samples come from white females only, the distributions of other biases between healthy and cancer tissue samples are the same due to the paired design of the experiment. Thus, the calculated model was not adjusted for other covariates.

In most cases, to perform GSA using SNP-level data, a single value per gene is needed. Thus, association results for SNPs within the same transcript need to be integrated into one representative value. Three different techniques for test result integration were applied. Currently, the most common method in GSA GWAS is to take the minimum p-value for SNP i, which falls within the gene g boundaries:

$$p - value_{gene} = \min_{i \in g} \{p - value_{SNP}\} \quad (1)$$

The second integration technique evaluated here was Fisher's probability integration (Fisher, 1925), which calculates the sum of the natural logarithm from k SNP p-values, which fall within the same gene g boundaries:

$$F_{gene} = -2 \sum_{i=1}^{k} \ln(p - value_i) \sim \chi^2_{(2k)} \quad (2)$$

The calculated F statistic per gene, F_{gene}, follows chi^2 distribution with $2*k$ degrees of freedom.

The last statistical integration approach used was the Stouffer method, also known as z-transformation-based integration (Stouffer, 1949). For k SNPs, which fall within the same gene g boundaries, Z_i statistic is first calculated using inverse normal cumulative distribution function (ϕ^{-1}) for each i-th SNP. Then the integrated Z statistic per gene, Z_{gene}, which follows standard normal distribution is calculated.

$$Z_i = \phi^{-1}(p - value_i) \quad (3)$$

$$Z_{gene} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \sim N(0,1) \quad (4)$$

Next, for the integrated p-values, the dependency correction due to LD was applied. The commonly used approach for LD correction requires calculations of r^2 or D′ score. Here, the modification of Dunn–Sidak correction for multiple testing was used instead. As was shown in Saccone et al. (2006), approximately 50% of the SNPs within chromosomes are in high LD; thus, the exponent of Dunn–Sidak was modified as follows:

$$p - value_{corr} = 1 - \left(1 - p - value_{gene}\right)^{\frac{k+1}{2}} \quad (5)$$

where k is the number of SNPs located within gene g. This method of introducing LD correction was proposed by Saccone et al. (2007) and allows for running enrichment analysis even for very limited genotyping data consisting only of two elements: SNP rs number and the result of the association test. Moreover, it was shown that the method is comparable, or slightly better than the regression method of GWAS integration p-value with

correction due to SNPs per kb, gene size, recombination hotspots, linkage disequilibrium units per kb, or genetic distance (Segrè et al., 2010). Each integration approach with or without dependency correction was tested in terms of effectiveness in GSA in GWAS.

Enrichment Algorithms for Single-Nucleotide Polymorphism Data

Several GSA algorithms dedicated to GWAS are based on *p*-value integration to move from SNP to transcriptome level (Das et al., 2020). From this group, the algorithms based on the Gene Set Enrichment Analysis (GSEA) method, mostly used in transcriptomic analysis (Subramanian et al., 2005), were selected.

The basic concept of GSEA is to estimate the enrichment score (ES) by calculating maximum absolute deviation between P_{hit} (normalized metric of genes within gene set *S*) and P_{miss} (normalized metric for genes outside gene set *S*). The ES distribution is calculated for *j*-th gene *g* in gene set *S* at the *i*-th position by modified Smirnov–Kolmogorov statistic using the following formulas:

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|}{N_R} \tag{6}$$

$$N_R = \sum_{g_j \in S} |r_j| \tag{7}$$

$$P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)} \tag{8}$$

Here, as a rank *r*, the negative value of base 10 logarithm of *p*-value was taken [$-\log10(p\text{-value}_{\text{gene}})$]. *N* is the total number of genes, and N_H represents the number of genes in the gene set *S*, and N_R is the sum of ranks of genes within gene set *S*.

The next algorithm used was GSEA-SNP (Holden et al., 2008), which is a simple modification of the standard GSEA approach. In this method, instead of integrating SNP association to transcriptomic level, the *p*-values of all SNPs are taken to calculate rank *r* parameter [$-\log10(p\text{-value}_{\text{SNP}})$]. Moreover, in GSEA-SNP, the SNP label permutation test is performed to assess significance of each gene set, while for GSEA, the gene label permutation is applied. In both GSEA and GSEA-SNP, the ES metric is adjusted for variation in gene set size by dividing the observed *ES* by the mean of permutated ES with the same direction giving normalized enrichment score (NES).

The third algorithm was *i*-GSEA4GWAS (improved GSEA for GWAS) (Zhang et al., 2010). This method has two main modifications compared with the standard GSEA: 1) Instead of gene label permutation, the SNP label permutation is performed, and then integration of *p*-values from SNP association test is executed. 2) The NES is substituted by significance proportion-based enrichment score (SPES). The SPES is multiplication of ES by ratio *k/K*, where *k* is the proportion of significant genes (mapped to 5% of the top SNPs) of the gene set *S*, and *K* is the proportion of significant genes (mapped to 5% of the top SNPs) of the total genes in the study (Zhang et al., 2010). According to the authors, SPES emphasizes the total significance coming from a high proportion of significant genes.

The fourth algorithm was MAGENTA (Meta-Analysis Gene-set Enrichment of variaNT Associations) (Segrè et al., 2010), where gene set significance is estimated as follows: 1) *p*-Values from SNP to gene level were integrated. 2) For each gene set, the number of gene *p*-values within a gene set lower than the cut-off (leading edge fraction) was calculated. The cutoff is a *p*-value of a specific percentile of all gene *p*-values (here set to the 75th percentile and marked as MAGENTA75), 3) to calculate the distribution of leading edge fraction with the permutation approach. In each permutation, the mock gene set is drawn as its leading edge fraction is collected. Finally, to assess the gene set significance, the number of permutation leading edge fractions equal or larger than the observed one for a particular gene set is estimated and divided by the number of permutations. All algorithms described above are modifications of GSEA approach and test *competitive* null hypothesis.

Two other algorithms were added to this list: ORA (over-representation analysis) (Tavazoie et al., 1999) and CERNO (Coincident Extreme Ranks in Numerical Observations) algorithm (Zyla et al., 2019). Both are designed for transcriptome data analysis but can be easily used in GWAS problems. ORA is the first-generation method, which estimates gene set significance via hypergeometric test using information about the number of differentially expressed genes (DEGs) and background genes within and outside the gene set. The CERNO method ranks genes from 1 to *N* (total analyzed genes), where rank 1 is given to the gene with the lowest *p*-value (here *p*-value from integration of SNP association). Next, the given ranks are divided by *N*, and the Fisher probability integration is performed for all genes within the gene set.

RNASeq Data Analysis

Genes that were not represented in SNP data or with no counts within all samples were removed prior to analysis (15,924 genes left). *DESeq2* R package (Love et al., 2014) was used to find genes with different expressions between normal and cancer samples, including the paired nature of the data. Pathway enrichment analysis with the GSEA method was performed using the *fgsea* package in R (Korotkevich et al., 2019) on the same set of pathways used in SNP data analysis. Test statistic from the *DESeq2* package was used as a gene rank value *r* in GSEA to retain information about directionality of expression change on pathway level.

Evaluation of Enrichment Algorithms

The brief evaluation pipeline is presented in **Figure 1**. All described gene set enrichment algorithms were run with three different integration approaches (minimum, Fisher, and Stouffer) and with and without dependency correction for LD. Four metrics were calculated to evaluate the algorithms: sensitivity, specificity, prioritization, and reproducibility (Zyla et al., 2019). Sensitivity represents detection of target gene sets for a particular phenotype. Specifically, gene set *p*-values are collected, and the

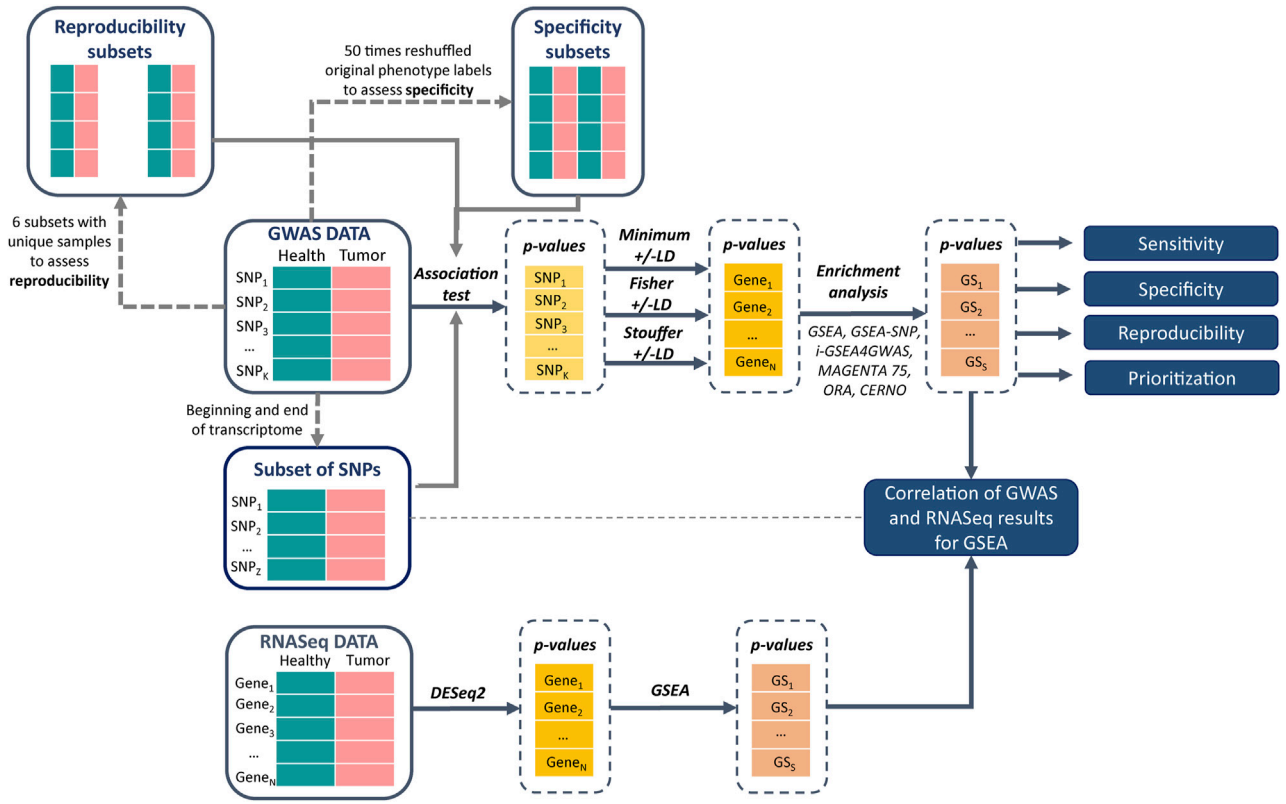

FIGURE 1 | Pipeline of data analysis and gene set analysis (GSA) method evaluation.

proportion of truly alternative hypotheses ($1 - \hat{\pi}$) is calculated with Storey's method (Storey, 2002). Prioritization represents median ranks of target pathways in all analyzed gene sets. Specificity represents deviation of mean false-positive rate (FPR, observed level) from 5% (expected level). Specifically, FPR is the proportion of significant gene sets ($p < 5\%$) among 50 permutations of the original phenotype. Reproducibility is the area under the curve (AUC) from the function of common detected gene sets across five or six data sets of the same phenotype at different cutoffs (Zyla et al., 2019). All used metrics were previously applied in transcriptomic data GSA (Tarca et al., 2013; Zyla et al., 2017; Zyla et al., 2019) and are one of the gold standards in enrichment algorithm evaluation (Geistlinger et al., 2021; Xie et al., 2021).

To test the impact of LD dependency correction, the differences between each metric separately within one enrichment method and integration approach were calculated (e.g., sensitivity of ORA minimum integration with LD correction minus ORA minimum integration without LD correction). Next, the impact of integration was assessed. As the minimum approach is mostly used, we referred its results to the Fisher and Stouffer methods. Again, the difference between performance metrics were calculated, but for different integrations (e.g., sensitivity of ORA Fisher integration with or without LD minus ORA minimum integration with or without LD).

Finally, we investigated similarities and information transition of gene set analysis performed on SNP and RNASeq data. For this purpose, we selected SNPs located at the "5′ UTR and upstream region" (beginning of transcript), as well as the "3′ UTR and downstream" coding region (end of transcript). The results of association test for those SNPs were extracted and aggregated using different integration methods with and without LD correction (the same as previously). Next, only the GSEA algorithm was run, as it has a direct equivalent in transcriptome analysis. The GSEA algorithm for RNASeq can distinguish up- and downregulated pathways; thus, the Spearman rank correlation was calculated for target pathways between GWAS (different SNP locations in transcript) and RNASeq (up-/downregulation). For the results from "5′UTR and upstream" GWAS location and gene set downregulation on RNASeq, the positive correlation is expected as SNPs in this region should block further transcription and translation. Opposite results are expected for "3′ UTR and downstream" where only isoforms of transcript products should be observed (Robert and Pelletier, 2018).

At each step of the evaluation process, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2017) was used as a gene set collection (accessed January 15, 2021). In total, 341 gene sets were analyzed. The 54 target gene sets for breast cancer were selected through the literature search, and

TABLE 1 | Number of significant genes after integration of single-nucleotide polymorphism (SNP) association test results to gene level.

Integration	Minimum		Fisher		Stouffer	
LD correction	No	Yes	No	Yes	No	Yes
# of genes	12,759	10,810	11,401	10,456	7,382	6,948

their detailed description is presented in **Supplementary Table S2**.

Implementation of Gene Set Analysis Algorithms for Single-Nucleotide Polymorphism Data Analysis

The implementation of all evaluated algorithms is provided in the R package *intGSASNP* (integrative GSA for SNP), created for the purpose of this study. This package includes R functions to run selected gene set algorithms (ORA, CERNO, MAGENTA, GSEA, *i*-GSEA4GWAS, and GSEA-SNP) on SNP data. All algorithms were implemented according to the description included in the original manuscripts. *intGSASNP* allows the user to adjust various function parameters depending on the experiment, such as type of integration method (minimum, Fisher, and Stouffer), multiple testing correction method, permutation method (by gene entrez or SNP), number of permutations, incorporation of LD correction, or the number of processing cores required for parallel computing. In addition, an example of a dataset with sample refSNP IDs, entrez IDs, and *p*-values has been provided. Source code and the package documentation are available freely to download on GitHub (https://github.com/ZAEDPolSl/intGSASNP).

RESULTS

At first, results of SNP association tests were transformed to gene level by using minimum, Fisher, and Stouffer methods with and without LD correction. Then these results were used in combination with different GSA algorithms, i.e., GSEA, *i*-GSEA4GWAS, GSEA-SNP, MAGENTA75, ORA, and CERNO, and the four evaluation metrics were established, i.e., sensitivity, specificity, prioritization, and reproducibility (**Figure 1**). Based on those metrics, the impact of integration and correction for LD and the overall performance of tested methods were examined. Detailed results are presented in **Supplementary Figures S1** and **Supplementary Table S3**. The total number of significant pathways is presented in **Supplementary Table S4**.

Single-Nucleotide Integration Results

The number of significant genes ($p < 0.05$) for each method is presented in **Table 1**, while the coverage between approaches is presented in **Supplementary Figure S2**. It can be observed that application of LD correction decreased the number of significant transcripts and more likely reduced false-positive outcomes. Over 50% of genes were common to all integration techniques (54.98% and 55.03% with and without LD correction, respectively). Fisher and minimum approach shared around 30% (32.23% and 29.83% with and without LD correction, respectively) significant genes, which may lead to further similar results of GSA. The association between minimum and Fisher methods is characterized by large correlations (**Supplementary Table S5**). This effect is expected as Fisher integration method is not robust to asymmetrical *p*-values, which result in stronger association assigned to genes, similar to that of the minimum method. Stouffer's technique showed weaker correlation to the minimum and Fisher methods. Also, after using Stouffer, there are not many unique significant genes or genes shared with only one of the other integration methods (**Supplementary Figure S2**). Over 90% of genes (96.56 and 91.81% with and without LD correction, respectively) indicated by the Stouffer method were also significant for the minimum and Fisher methods.

Overall Performance of Gene Set Analysis Methods

Within each evaluation metric, values were first normalized giving the lowest value for the best outcome and the highest for the worst. Next, the sum of all metrics was calculated, and algorithms were ranked from the best to the worst within the study. At last, results were clustered using the k-means approach, where the number of clusters were set by the Silhouette metric (optimal k equals 5). The best performance was obtained for CERNO and MAGENTA75 methods with Stouffer integration regardless of LD correction (**Figure 2A**, cluster 1). The worst outcomes were achieved for *i*-GSEA4GWAS and ORA with Fisher integration (regardless of LD correction) as well as for ORA and MAGENTA75 with minimum integration and LD correction and GSEA-SNP (cluster 5). Original GSEA gave moderate results in comparison with others (cluster 2 or 3). Overall, the results for CERNO and MAGENTA75 were the best (mostly in clusters 1 and 2), while *i*-GSEA4GWAS and ORA were the worst (mostly in clusters 4 and 5).

Next, global similarities of results were investigated by using the UMAP dimensionality reduction technique (McInnes et al., 2018; McInnes and Healy, 2018) on the GSA results for all 341 KEGG pathways (**Figure 2B**). Four major clusters could be distinguished on two first instances of UMAP (**Figure 2B**). *i*-GSEA4GWAS gave similar results regardless of the integration technique as well as incorporation of LD correction. GSEA-SNP, CERNO, MAGENTA75, and GSEA performed on minimum integration and correction for LD are clustered together with *i*-GSEA4GWAS. The middle right cluster includes ORA with Fisher and minimum integration methods regardless of LD correction (color coding of UMAP projection due to integration used is presented in **Supplementary Figure S3**). Moreover, ORA with Stouffer integration gave similar results across all tested pathways to CERNO and MAGENTA75 (with the same integration method; bottom cluster).

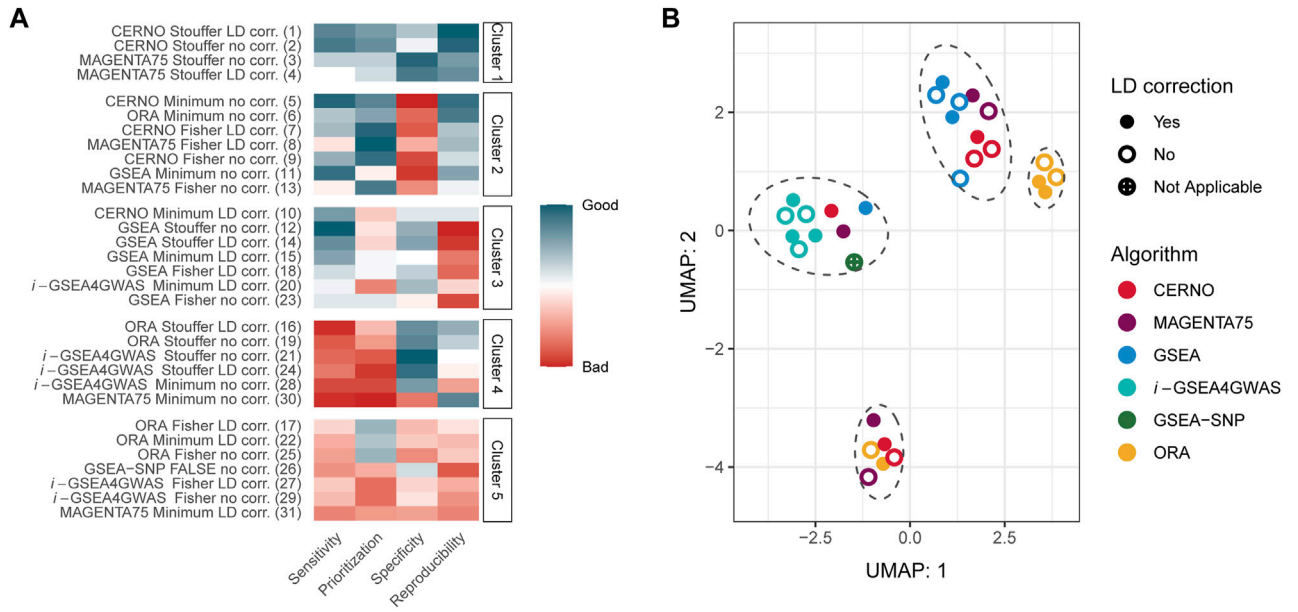

FIGURE 2 | Overall evaluation of GSA algorithms. **(A)** Normalized evaluation metrics together with clustering results. The red color represents poor performance, while the blue color represents good performance. The number in the brackets, next to the algorithm name, illustrates the place in global ranking including all evaluation metrics. **(B)** UMAP projection for results of each algorithm and all 341 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Marked ellipses represent clustering within UMAP projection.

Impact of Linkage Disequilibrium Dependency Correction

To investigate the impact of LD dependency correction, the difference of each evaluation metric within a particular algorithm performed with a specific integration method was calculated (e.g., sensitivity difference between ORA with minimum integration and LD correction, and ORA with minimum integration and without LD correction). Values of these differences are presented in **Supplementary Table S6**. ORA, CERNO, and GSEA showed decreased sensitivity (**Figure 3A**) when LD correction was applied, but the specificity was increased greatly (**Figure 3C**). The LD correction has a positive impact also on prioritization for MAGENTA75 (**Figure 3B**). *i*-GSEA4GWAS with Stouffer and Fisher integration gave similar performance regardless of LD correction usage. However, for minimum integration (default option in original implementation of the algorithm), the LD correction increases the sensitivity of the observed results with only a slight drop of specificity. For the remaining algorithms, when the Stouffer or Fisher integration method is used, the LD correction gave similar or better performance. When minimum integration is applied, the reproducibility decreases with LD correction (**Figure 3D**).

Impact of *p*-Value Integration

As the minimum integration is the most preferred approach, the results of this aggregation technique were compared to those of the Fisher and Stouffer methods separately. For this purpose, the difference between performance metrics was calculated (e.g., sensitivity of ORA minimum integration with LD minus ORA Fisher integration with LD). In the previous paragraph, it was shown that LD correction has a beneficial impact in most of the cases, and it preserves biological insights, so further description will concentrate only on outcomes when dependency correction is applied. CERNO and MAGENTA75 had better results in terms of prioritization, specificity, and reproducibility when the Fisher or Stouffer method was used (**Figure 4**). *i*-GSEA4GWAS showed similar results regardless of the integration method used, with slightly better performance for minimum integration. The ORA algorithm gave similar performance when the minimum or Fisher method was applied, while Stouffer gave better specificity and reproducibility, but decreased sensitivity. Finally, GSEA showed better specificity when both Fisher and Stouffer were used, whereas other evaluation metrics were similar despite the integration techniques used (**Figure 4**).

Within each evaluation metric and obtained differences, the equivalence of mean to zero was tested by one-sample *t*-test. The Stouffer integration method gave significantly better results (p-value = 0.0163) in terms of specificity compared with minimum integration for all tested algorithms (**Figure 4C**). There was no statistically

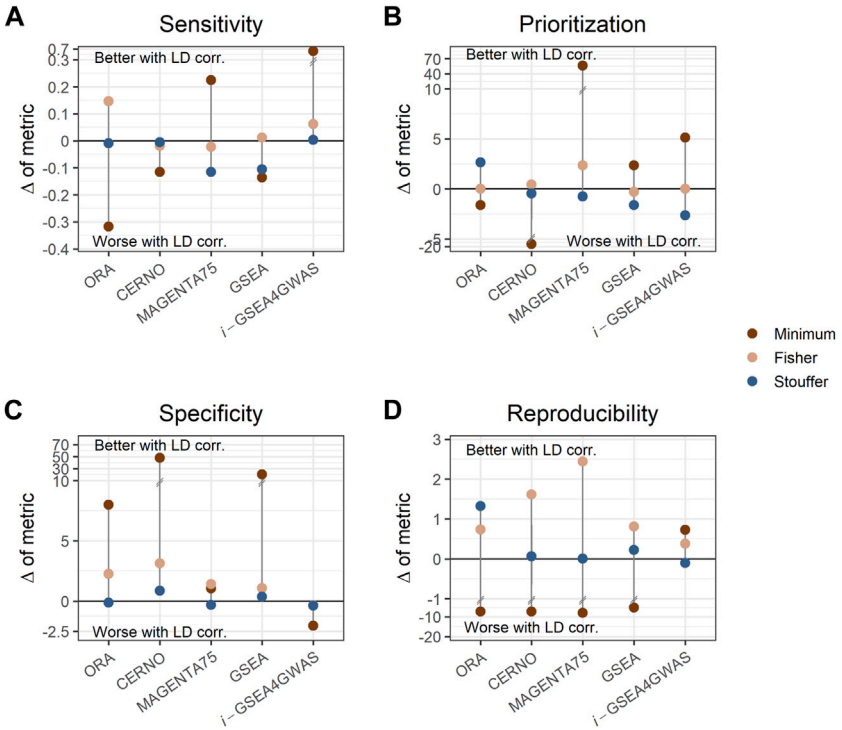

FIGURE 3 | Impact of linkage disequilibrium (LD) correction on evaluation metrics. Each panel shows different metrics, i.e., sensitivity **(A)**, prioritization **(B)**, specificity **(C)**, and reproducibility **(D)**. Each dot represents the difference of metric when LD adjustment is used within a particular algorithm and integration method. Dots above the solid, black line represent better performance when LD correction is applied, while dots below the line represent the opposite. Colors show different integration techniques.

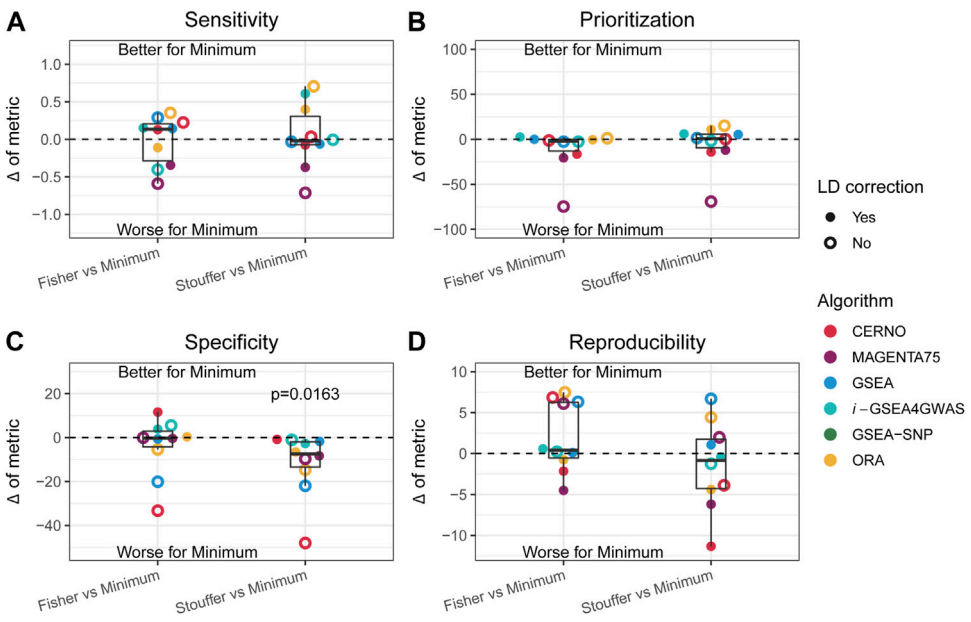

FIGURE 4 | Impact of the Fisher and Stouffer integration technique compared with the minimum approach. **(A–D)** Differences between outcomes for the sensitivity, prioritization, specificity, and reproducibility, respectively. Colors represent different algorithms, and point shape represents whether correction for LD was applied. Dots above the solid, black line represent better performance for the minimum integration approach, while dots below the line represent the opposite.

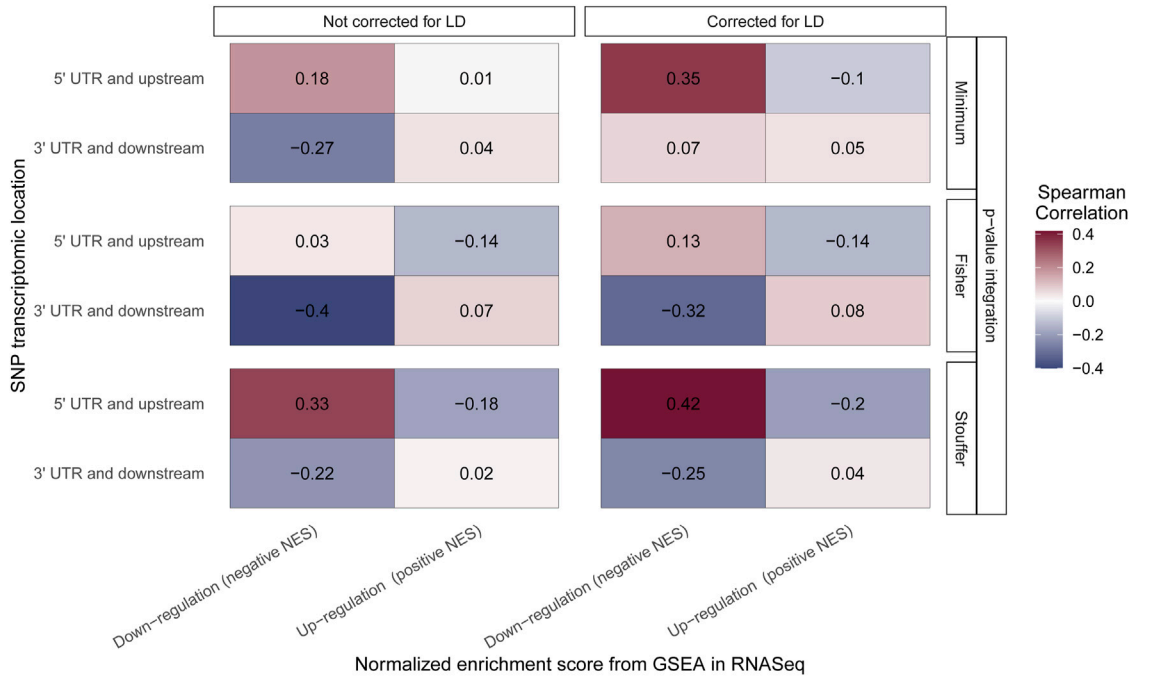

FIGURE 5 | Correlation of the Gene Set Enrichment Analysis (GSEA) algorithm results performed on RNASeq and genomic data. The color of the boxes represents the value of Spearman rank correlation coefficient. The *x*-axis corresponds to the results from the GSEA algorithm performed on RNASeq data with distinction of pathway regulation direction. The *y*-axis corresponds to the results from the GSEA algorithm performed on genomic data with distinction of location of single-nucleotide polymorphisms (SNPs) taken to the aggregation process. All results are grouped by integration technique and LD correction used.

significant difference in other comparisons; however, the variety of effects can be observed for individual algorithms.

Comparison of Gene Set Analysis on Single-Nucleotide and Gene Expression Level

As the enrichment methods were initially designed for transcriptome data analysis, target pathway similarities of outcomes observed for genome and transcriptome were investigated (**Supplementary Figure S4**). The same samples were taken for both omics, and the GSEA algorithm was applied (In RNASeq, it can distinguish up- and downregulated pathways). Moreover, for genomic data, only SNPs from the beginning and the end of transcriptomic regions were selected to catch the regulation directionality. Finally, the Spearman rank correlation coefficient was calculated between $-\log10(p\text{-value}_{\text{GeneSet}})$ from RNASeq and genomic data within pathways up- and downregulated separately.

For Stouffer integration, the highest correlation of downregulated pathways and "5′ UTR and upstream" SNPs is observed, and it increases when LD correction is applied (from 0.33 to 0.42; both medium effect size) (**Figure 5**). Similar results can be observed for the minimum approach, where correlation changes from small effect size when no LD correction is applied to medium effect size with LD correction (from 0.18 to 0.35). For Fisher integration, the small effect size was observed only when LD adjustment was applied (**Figure 5**). When SNPs from "3′ UTR and downstream" (end of transcriptomic region) were analyzed, positive correlation with upregulated pathways was expected, but none of the tested methods showed statistically significant association. Nevertheless, for downregulated pathways, the expected negative correlation is observed for all integration techniques regardless of LD correction.

Comparison of Gene Set Enrichment Analysis and Other Algorithms on Single-Nucleotide Polymorphism Level

Most of the tested algorithms were created by modification of the original GSEA method. Also, we found a correlation between GSEA results on genomic and transcriptomic level. Thus, we wanted to check how the results of GSA for target pathways are correlated between GSEA and other tested enrichment algorithms in GWAS (**Supplementary Figure S5**). The GSEA-SNP does not use integration in the process of enrichment analysis; nevertheless, it showed a small correlation with GSEA only when the Stouffer method is applied. On the other hand, the results of *i*-GSEA4GWAS had negative correlation with GSEA when the Fisher and Stouffer methods were applied. All other methods mostly showed positive correlation with GSEA. The highest correlation was observed for the CERNO and MAGENTA75 algorithms.

DISCUSSION

Incorporation of specific integration methods and LD correction can have significant impact on the performance of gene set analysis in GWAS. Usage of LD correction was beneficial for *i*-GSEA4GWAS, especially when the default minimum integration method was used. Thus, the incorporation of basic SNP dependency correction method, like we did here, or more complex solutions (de Leeuw et al., 2015) is recommended. When Fisher and Stouffer integration were used in the CERNO, ORA, MAGENTA75, and GSEA algorithms, the LD correction was always beneficial, so it should be applied in any case there. Observed decrease in reproducibility after applying LD could be the effect of decreasing the number of significant findings (genes with *p*-values lower than a threshold), which is usual after using correction for multiple testing (like LD correction here). Moreover, the reproducibility experiment was performed on much smaller subsets ($n = 14$ paired samples), which also decreased the power of GSA (Maleki et al., 2019). In *i*-GSEA4GWAS, this effect was not observed due to corrections given by SPES statistic.

Comparing *p*-value integration methods, Stouffer gave the best results in terms of specificity for all tested enrichment methods. Moreover, it gave better or similar results in terms of prioritization and reproducibility for CERNO, MAGENTA, GSEA, and ORA. The Stouffer integration decreases only sensitivity, which is an effect of preserving robustness to asymmetrical *p*-value distribution during the process of integration. Thus, the integrated *p*-value is higher, and some target pathways could not be detected. However, this mechanism prevents *p*-value overestimation that was observed for some GSA methods (Zyla et al., 2019). Also, Stouffer integration gave the highest correlation between GSA analysis results on SNP level and transcriptome level. Thus, this is the method that we recommend most.

SNPs located at the beginning of the gene region have the biggest ability to silence gene expression (Robert and Pelletier, 2018). The GSA outcomes compared between genomic and transcriptomic levels confirmed this effect. Furthermore, results for 5′ UTR and upstream SNPs were negatively correlated with upregulated pathways on the gene level. GSA results for SNPs located at the end of genes were positively correlated with upregulated pathways on the gene level, as it was expected, but the effect was smaller.

All gene set analysis methods, integration approaches, and LD correction method that were tested within the study were implemented in the *intGSASNP* R package and are freely available on GitHub. Therefore, different combinations of methods could be easily tested on any dataset by other researchers. We hope that collecting multiple methods in a single package will help to promote the application of GSA methods in SNP analysis.

In summary, we thoroughly analyzed different methods of gene set analysis in GWAS in terms of performance and its applicability. We showed that LD correction and Stouffer integration could increase the performance of enrichment analysis and encourage the introduction of these techniques into common practice. We believe that this work will guide others to select the most effective combinations of methods.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://www.cancer.gov/tcga.

AUTHOR CONTRIBUTIONS

MM, JP, and JZ conceived the concept of the study and supervised the methodology. AM and JT were responsible for the data acquisition. MM, AM, and JZ were responsible for the data analysis. JT and JZ were responsible for the visualization. AM was responsible for the implementation of the algorithms and R package creation. All authors wrote and approved the final version of the article.

FUNDING

This work was supported by the Silesian University of Technology grant for Support and Development of Research Potential (AM, JP, JZ), the Silesian University of Technology rector's pro-quality grant no. 02/070/RGJ21/0020 (MM), and the European Social Fund grant no. POWR.03.02.00-00-I029 (JT).

ACKNOWLEDGMENTS

We would like to thank Professor Andrzej Polanski for his support regarding TCGA database and GDC Data Portal.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.767358/full#supplementary-material

REFERENCES

Affymetrix (2021). *Genome Wide Human SNP 6.0 Array*. Available at: http://tools.thermofisher.com/content/sfs/brochures/genomewide_snp6_datasheet.pdf (Accessed October 22, 2021).

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 33 (4), 690–e9. e9. Epub 2018/04/07PubMed PMID: 29622464; PubMed Central PMCID: PMCPMC5959730. doi:10.1016/j.ccell.2018.03.014

Billings, L. K., and Florez, J. C. (2010). The Genetics of Type 2 Diabetes: what Have We Learned from GWAS? *Ann. N. Y Acad. Sci.* 1212, 59–77. Epub 2010/11/26PubMed PMID: 21091714; PubMed Central PMCID: PMCPMC3057517. doi:10.1111/j.1749-6632.2010.05838.x

Canzler, S., and Hackermüller, J. (2020). multiGSEA: a GSEA-Based Pathway Enrichment Analysis for Multi-Omics Data. *BMC Bioinformatics* 21 (1), 561. doi:10.1186/s12859-020-03910-x

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2015). Ensembl 2015. *Nucleic Acids Res.* 43, D662–D669. (Database issue)Epub 2014/10/30PubMed PMID: 25352552; PubMed Central PMCID: PMCPMC4383879. doi:10.1093/nar/gku1010

Das, S., McClain, C. J., and Rai, S. N. (2020). Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy* 22 (4), 427. Epub 2020/12/09PubMed PMID: 33286201; PubMed Central PMCID: PMCPMC7516904. doi:10.3390/e22040427

de Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *Plos Comput. Biol.* 11 (4), e1004219. Epub 2015/04/18PubMed PMID: 25885710; PubMed Central PMCID: PMCPMC4401657. doi:10.1371/journal.pcbi.1004219

Dong, X., Hao, Y., Wang, X., and Tian, W. (2016). LEGO: a Novel Method for Gene Set Over-representation Analysis by Incorporating Network-Based Gene Weights. *Sci. Rep.* 6 (1), 18871. doi:10.1038/srep18871

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., et al. (2021). Toward a Gold Standard for Benchmarking Gene Set Enrichment Analysis. *Brief Bioinform* 22 (1), 545–556. Epub 2020/02/07PubMed PMID: 32026945; PubMed Central PMCID: PMCPMC7820859. doi:10.1093/bib/bbz158

Hirschhorn, J. N. (2009). Genomewide Association Studies - Illuminating Biologic Pathways. *N. Engl. J. Med.* 360 (17), 1699–1701. Epub 2009/04/17PubMed PMID: 19369661. doi:10.1056/NEJMp0808934

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: Applying Gene Set Enrichment Analysis to SNP Data from Genome-wide Association Studies. *Bioinformatics* 24 (23), 2784–2785. Epub 2008/10/16PubMed PMID: 18854360. doi:10.1093/bioinformatics/btn516

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 (D1), D353–D361. Epub 2016/12/03PubMed PMID: 27899662; PubMed Central PMCID: PMCPMC5210567. doi:10.1093/nar/gkw1092

Kaspi, A., and Ziemann, M. (2020). Mitch: Multi-Contrast Pathway Enrichment for Multi-Omics and Single-Cell Profiling Data. *BMC Genomics* 21 (1), 447. Epub 2020/07/01PubMed PMID: 32600408; PubMed Central PMCID: PMCPMC7325150. doi:10.1186/s12864-020-06856-9

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *Plos Comput. Biol.* 8 (2), e1002375. Epub 2012/03/03PubMed PMID: 22383865; PubMed Central PMCID: PMCPMC3285573. doi:10.1371/journal.pcbi.1002375

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., and Sergushichev, A. (2019). *Fast Gene Set Enrichment Analysis*, 060012. bioRxiv. doi:10.1101/060012

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

Maciejewski, H. (2014). Gene Set Analysis Methods: Statistical Models and Methodological Differences. *Brief. Bioinform.* 15 (4), 504–518. Epub 2013/02/16PubMed PMID: 23413432; PubMed Central PMCID: PMCPMC4103537. doi:10.1093/bib/bbt002

Maleki, F., Ovens, K., Hogan, D. J., and Kusalik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* 11, 654. Epub 2020/07/23PubMed PMID: 32695141; PubMed Central PMCID: PMCPMC7339292. doi:10.3389/fgene.2020.00654

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2019). Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. *Hum. Genomics* 13 (Suppl. 1), 42. Epub 2019/10/23PubMed PMID: 31639047; PubMed Central PMCID: PMCPMC6805317. doi:10.1186/s40246-019-0226-2

Mangiafico, S. S. (2016)., 125. New Brunswick, NJ, USA, 16–22.Summary and Analysis of Extension Program Evaluation in R*Rutgers Coop. Extension*

Marczyk, M., Jaksik, R., Polanski, A., and Polanska, J. (2019). GaMRed - Adaptive Filtering of High-Throughput Biological Data. *Ieee/acm Trans. Comput. Biol. Bioinf.* 17 (1), 1. doi:10.1109/TCBB.2018.2858825

Marczyk, M., Patwardhan, G. A., Zhao, J., Qu, R., Li, X., Wali, V. B., et al. (2020). Multi-Omics Investigation of Innate Navitoclax Resistance in Triple-Negative Breast Cancer Cells. *Cancers* 12 (9), 2551. PubMed PMID:. doi:10.3390/cancers12092551

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., et al. (2018). GWAS on Family History of Alzheimer's Disease. *Transl Psychiatry* 8 (1), 99. Epub 2018/05/20PubMed PMID: 29777097; PubMed Central PMCID: PMCPMC5959890. doi:10.1038/s41398-018-0150-6

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Joss* 3, 861. doi:10.21105/joss.00861

McInnes, L., and Healy, J. (2018). *Umap: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv preprint arXiv:180203426 (2018).

Mei, H., Li, L., Jiang, F., Simino, J., Griswold, M., Mosley, T., et al. (2016). snpGeneSets: An R Package for Genome-wide Study Annotation. *G3 (Bethesda)* 6 (12), 4087–4095. Epub 2016/11/04PubMed PMID: 27807048; PubMed Central PMCID: PMCPMC5144977. doi:10.1534/g3.116.034694

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., et al. (2013). Methods and Approaches in the Topology-Based Analysis of Biological Pathways. *Front. Physiol.* 4, 278. Epub 2013/10/18PubMed PMID: 24133454; PubMed Central PMCID: PMCPMC3794382. doi:10.3389/fphys.2013.00278

Nci Genomic Data Commons (2021). *Documentation Data*. Available at: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/ (Accessed October 22, 2021).

Nguyen, T.-M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Identifying Significantly Impacted Pathways: a Comprehensive Review and Assessment. *Genome Biol.* 20 (1), 203. Epub 2019/10/11PubMed PMID: 31597578; PubMed Central PMCID: PMCPMC6784345. doi:10.1186/s13059-019-1790-4

Robert, F., and Pelletier, J. (2018). Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. *Front. Genet.* 9, 507. Epub 2018/11/15PubMed PMID: 30425729; PubMed Central PMCID: PMCPMC6218417. doi:10.3389/fgene.2018.00507

Saccone, S. F., Hinrichs, A. L., Saccone, N. L., Chase, G. A., Konvicka, K., Madden, P. A. F., et al. (2007). Cholinergic Nicotinic Receptor Genes Implicated in a Nicotine Dependence Association Study Targeting 348 Candidate Genes with 3713 SNPs. *Hum. Mol. Genet.* 16 (1), 36–49. Epub 2006/12/01PubMed PMID: 17135278; PubMed Central PMCID: PMCPMC2270437. doi:10.1093/hmg/ddl438

Saccone, S. F., Rice, J. P., and Saccone, N. L. (2006). Power-based, Phase-Informed Selection of Single Nucleotide Polymorphisms for Disease Association Screens. *Genet. Epidemiol.* 30 (6), 459–470. Epub 2006/05/11PubMed PMID: 16685721. doi:10.1002/gepi.20159

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., and Altshuler, D. (2010). Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *Plos Genet.* 6 (8), e1001058. PubMed PMID: 20714348; PubMed Central PMCID: PMCPMC2920848. doi:10.1371/journal.pgen.1001058Epub 2010/08/18

Storey, J. D. (2002). A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 64 (3), 479–498. doi:10.1111/1467-9868.00346

Stouffer, S. A. (1949). *The American Soldier: Adjustment during Army Life*. Princeton University Press.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl.*

Acad. Sci. 102 (43), 15545–15550. Epub 2005/10/04PubMed PMID: 16199517; PubMed Central PMCID: PMCPMC1239896. doi:10.1073/pnas.0506580102

Sud, A., Kinnersley, B., and Houlston, R. S. (2017). Genome-wide Association Studies of Cancer: Current Insights and Future Perspectives. *Nat. Rev. Cancer* 17 (11), 692–704. Epub 2017/10/14PubMed PMID: 29026206. doi:10.1038/nrc.2017.82

Sun, R., Hui, S., Bader, G. D., Lin, X., and Kraft, P. (2019). Powerful Gene Set Analysis in GWAS with the Generalized Berk-Jones Statistic. *Plos Genet.* 15 (3), e1007530. Epub 2019/03/16PubMed PMID: 30875371; PubMed Central PMCID: PMCPMC6436759. doi:10.1371/journal.pgen.1007530

Tarca, A. L., Bhatti, G., and Romero, R. (2013). A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS One* 8 (11), e79217. Epub 2013/11/22PubMed PMID: 24260172; PubMed Central PMCID: PMCPMC3829842. doi:10.1371/journal.pone.0079217

Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting Overlapping Genes Improves Gene Set Analysis. *BMC Bioinformatics* 13 (1), 136. doi:10.1186/1471-2105-13-136

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics (Oxford, England)* 25 (1), 75–82. Epub 2008/11/05PubMed PMID: 18990722. doi:10.1093/bioinformatics/btn577

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic Determination of Genetic Network Architecture. *Nat. Genet.* 22 (3), 281–285. Epub 1999/07/03PubMed PMID: 10391217. doi:10.1038/10343

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based Approaches for Analysis of Genomewide Association Studies. *Am. J. Hum. Genet.* 81 (6), 1278–1283. Epub 2007/10/30PubMed PMID: 17966091; PubMed Central PMCID: PMCPMC2276352. doi:10.1086/522374

Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S. G., Yu, Z., et al. (2011). SNP-based Pathway Enrichment Analysis for Genome-wide Association Studies. *BMC Bioinformatics* 12, 99. Epub /04/19PubMed PMID: 21496265; PubMed Central PMCID: PMCPMC3102637. doi:10.1186/1471-2105-12-99

Wijmenga, C., and Zhernakova, A. (2018). The Importance of Cohort Studies in the post-GWAS Era. *Nat. Genet.* 50 (3), 322–328. Epub 2018/03/08PubMed PMID: 29511284. doi:10.1038/s41588-018-0066-3

Xie, C., Jauhari, S., and Mora, A. (2021). Popularity and Performance of Bioinformatics Software: the Case of Gene Set Analysis. *BMC Bioinformatics* 22 (1), 191. Epub 2021/04/17PubMed PMID: 33858350; PubMed Central PMCID: PMCPMC8050894. doi:10.1186/s12859-021-04124-5

Yoon, S., Nguyen, H. C. T., Yoo, Y. J., Kim, J., Baik, B., Kim, S., et al. (2018). Efficient Pathway Enrichment and Network Analysis of GWAS Summary Data Using GSA-SNP2. *Nucleic Acids Res.* 46 (10), e60. Epub 2018/03/22PubMed PMID: 29562348; PubMed Central PMCID: PMCPMC6007455. doi:10.1093/nar/gky175

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., et al. (2009). Pathway Analysis by Adaptive Combination ofP-Values. *Genet. Epidemiol.* 33 (8), 700–709. Epub 2009/04/01PubMed PMID: 19333968; PubMed Central PMCID: PMCPMC2790032. doi:10.1002/gepi.20422

Zhang, K., Cui, S., Chang, S., Zhang, L., Wang, J., and i-Gsea4Gwas (2010). i-GSEA4GWAS: a Web Server for Identification of Pathways/gene Sets Associated with Traits by Applying an Improved Gene Set Enrichment Analysis to Genome-wide Association Study. *Nucleic Acids Res.* 38, W90–W95. (Web Server issue):W90Epub 2010/05/04PubMed PMID: 20435672; PubMed Central PMCID: PMCPMC2896119. doi:10.1093/nar/gkq324

Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S. H. E., Polanska, J., and Weiner, J. (2019). Gene Set Enrichment for Reproducible Science: Comparison of CERNO and Eight Other Algorithms. *Bioinformatics* 35 (24), 5146–5154. Epub 2019/06/06PubMed PMID: 31165139; PubMed Central PMCID: PMCPMC6954644. doi:10.1093/bioinformatics/btz447

Zyla, J., Marczyk, M., Weiner, J., and Polanska, J. (2017). Ranking Metrics in Gene Set Enrichment Analysis: Do They Matter? *BMC Bioinformatics* 18 (1), 256. doi:10.1186/s12859-017-1674-0

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

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

Copyright © 2021 Marczyk, Macioszek, Tobiasz, Polanska and Zyla. 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.

frontiers in Genetics

METHODS
published: 22 December 2021
doi: 10.3389/fgene.2021.783713

PathwayMultiomics: An R Package for Efficient Integrative Analysis of Multi-Omics Datasets With Matched or Un-matched Samples

Gabriel J. Odom[1,2], *Antonio Colaprico*[2], *Tiago C. Silva*[2], *X. Steven Chen*[2,3] *and Lily Wang*[2,3,4,5]*

[1]Department of Biostatistics, Stempel College of Public Health, Florida International University, Miami, FL, United States, [2]Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL, United States, [3]Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Miami, FL, United States, [4]Dr. John T Macdonald Foundation Department of Human Genetics, Miller School of Medicine, University of Miami, Miami, FL, United States, [5]John P. Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL, United States

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Lingling Jin,
University of Saskatchewan, Canada
Yan Yan,
Thompson Rivers University, Canada
Paola Lecca,
Free University of Bozen-Bolzano, Italy

***Correspondence:**
Lily Wang
lily.wang@miami.edu

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

Received: 26 September 2021
Accepted: 07 December 2021
Published: 22 December 2021

Citation:
Odom GJ, Colaprico A, Silva TC, Chen XS and Wang L (2021) PathwayMultiomics: An R Package for Efficient Integrative Analysis of Multi-Omics Datasets With Matched or Un-matched Samples. Front. Genet. 12:783713. doi: 10.3389/fgene.2021.783713

Recent advances in technology have made multi-omics datasets increasingly available to researchers. To leverage the wealth of information in multi-omics data, a number of integrative analysis strategies have been proposed recently. However, effectively extracting biological insights from these large, complex datasets remains challenging. In particular, matched samples with multiple types of omics data measured on each sample are often required for multi-omics analysis tools, which can significantly reduce the sample size. Another challenge is that analysis techniques such as dimension reductions, which extract association signals in high dimensional datasets by estimating a few variables that explain most of the variations in the samples, are typically applied to whole-genome data, which can be computationally demanding. Here we present pathwayMultiomics, a pathway-based approach for integrative analysis of multi-omics data with categorical, continuous, or survival outcome variables. The input of pathwayMultiomics is pathway *p*-values for individual omics data types, which are then integrated using a novel statistic, the MiniMax statistic, to prioritize pathways dysregulated in multiple types of omics datasets. Importantly, pathwayMultiomics is computationally efficient and does not require matched samples in multi-omics data. We performed a comprehensive simulation study to show that pathwayMultiomics significantly outperformed currently available multi-omics tools with improved power and well-controlled false-positive rates. In addition, we also analyzed real multi-omics datasets to show that pathwayMultiomics was able to recover known biology by nominating biologically meaningful pathways in complex diseases such as Alzheimer's disease.

Keywords: pathway analysis, gene set analysis, multi-omics, integrative analysis, R package, Alzheheimer's disease

INTRODUCTION

Recent advances in technology have made multi-omics datasets increasingly available to researchers. For example, The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) have generated comprehensive molecular profiles including genomic, epigenomic, and proteomic expressions on matched samples for many types of human tumors. The underlying hypothesis is that multiple types of molecular profiles (e.g., copy number, DNA methylation, protein) might provide a more coherent and complete signature of the disease process.

To leverage the wealth of information in multi-omics data, a number of integrative analysis strategies have been proposed (Meng et al., 2016; Huang et al., 2017) and compared (Le Cao et al., 2009; Pucher et al., 2019). These methods can be roughly classified into three different categories, characterized by the way they leverage information from the multi-omics datasets. The first group of methods (Parkhomenko et al., 2009; Waaijenborg and Zwinderman, 2009; Witten and Tibshirani, 2009; Lin et al., 2013) analyzes only intersecting (i.e., matched) samples from the multiple omics datasets and only shared genes measured by all types of omics platforms. The second group of methods (Dray and Dufour, 2007; Kaspi and Ziemann, 2020) analyzes only genes shared by multiple types of omics datasets, which may be measured on the same or distinct samples in different omics datasets. The third group of methods (Gao et al., 2004; Kutalik et al., 2008; Zhang et al., 2012; Meng et al., 2014) analyzes matched samples in multi-omics datasets, where each dataset may have the same or distinct genes.

Because of the complexities in multi-omics datasets, effectively extracting biological insights from these datasets remains challenging. A major challenge for multi-omics data analysis is that the samples are often measured on one or a few, but not all, omics data types. Therefore, multi-omics analysis tools that require matched samples (with measurements for all omics data types) as input can significantly limit the sample size when several omics data types are considered. Another challenge is that analysis techniques such as dimension reduction techniques are typically applied to genome-wide data, which can be computationally demanding. Thus, to maximally leverage information from the multi-omics datasets, there is a critical need for developing additional integrative methods that are not restricted to only matched samples and/or shared genes in the input datasets.

Here we present pathwayMultiomics, a pathway-based approach for integrative analysis of multi-omics data. Instead of testing individual genes, pathway analysis tests joint effects of multiple genes belonging to the same biological pathway, such as those defined in the KEGG (Kanehisa et al., 2012) database. Higher power in the pathway-based analysis is achieved by combining weak signals from a number of individual genes in the pathway (Subramanian et al., 2005). The input of pathwayMultiomics is pathway *p*-values for individual omics data types, which are then integrated using a novel statistic, the MiniMax statistic, to prioritize pathways dysregulated in multiple types of omics datasets. Because pathwayMultiomics only requires summary statistics (i.e., pathway *p*-values) as input, it is computationally efficient. In addition, it is also flexible and can be used to analyze multi-omics datasets with categorical, continuous, or survival outcome variables. Importantly, using summary statistics as input allows pathwayMultiomics to maximally leverage information in multi-omics datasets by not restricting to only shared samples and/or genes. Using simulated datasets, we showed that pathwayMultiomics significantly outperforms currently available multi-omics methods with improved power and well-controlled false-positive rates. In addition, we also analyzed multi-omics datasets in Alzheimer's disease to show that pathwayMultiomics was able to recover known biology by nominating biologically meaningful pathways.

MATERIALS AND METHODS

An Overview of pathwayMultiomics Algorithm

Figure 1 illustrates the workflow of the pathwayMultiomics analysis pipeline. We next describe the input datasets, analytical algorithm, and output in detail. The pathwayMultiomics package for R can be accessed from https://github.com/TransBioInfoLab/pathwayMultiomics.

Input Datasets

The input dataset consists of omics datasets for several different molecular traits, such as SNPs, DNA methylation (DNAm), copy number alterations (CNAs), or gene expressions. Of particular interest are dysregulated pathways at multiple molecular levels, for example, those with changes in both DNA methylation and gene expressions. Importantly, pathwayMultiomics is flexible; the samples can be either matched (multiple types of molecular traits are measured on the same set of samples), or un-matched (distinct samples from the same disease are measured with different types of omics technology). Moreover, because the units of analyses for pathwayMultiomics are pathways (i.e., groups of genes participating in the same biological processes), different omics datasets can also include different genes, as long as pathway-level association statistics that relate each type of omics profiles to the phenotype (e.g., pathway *p*-values) can be computed. This flexibility enables pathwayMultiomics to take advantage of different pathway analysis software to model and account for special characteristics in different types of omics datasets. For example, for pathway analysis of DNAm data, the missMethyl method (Phipson et al., 2016), which takes account of the varying number of probes mapped to each gene, could be used. For pathway analysis of gene expression data, pathwayPCA method (Odom et al., 2020), which selects the coherent subset of genes before estimating and testing principal components with phenotypes, could be applied.

MiniMax Statistic

Given pathway *p*-values for each omics data type, pathwayMultiomics next computes the MiniMax statistic. To this end, we first consider all pairs of *p*-values from different

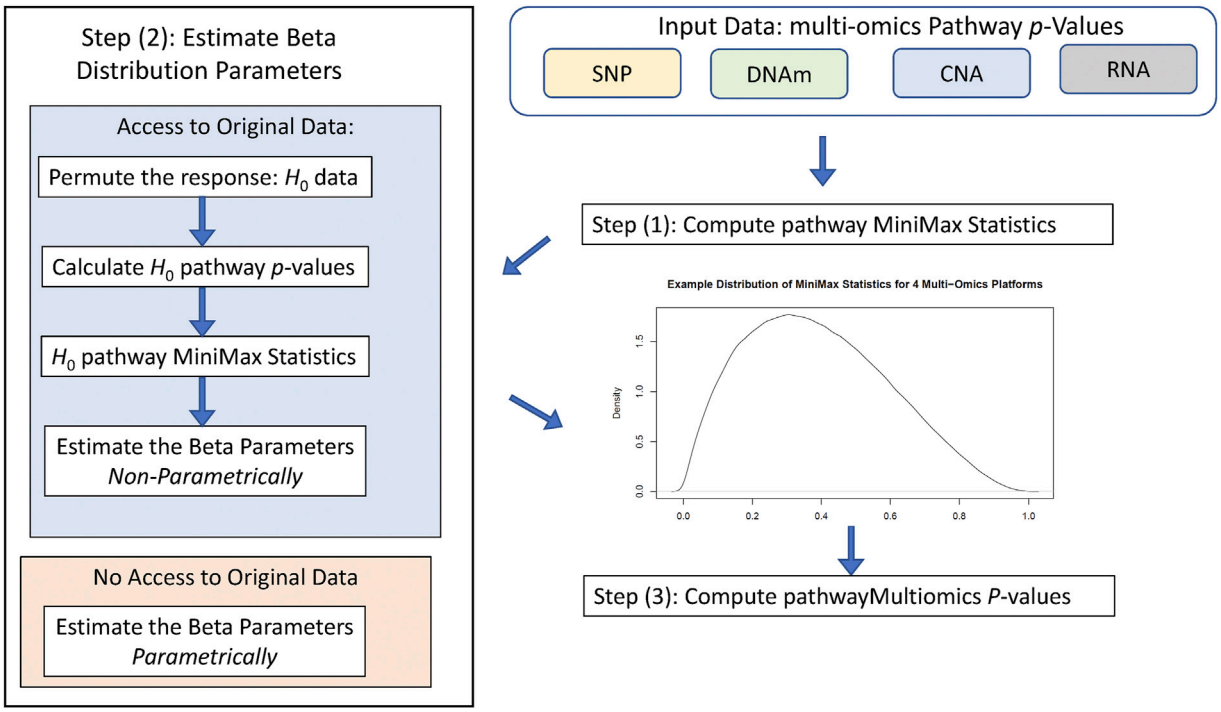

FIGURE 1 | Workflow of pathwayMultiomics analysis.

omics types and take the maximum for each pair of p-values. Next, we take the **mini**mum of all **max**imums computed from the last step. For example, suppose we are interested in an apoptosis pathway for a cancer study, which has p-values of 0.01, 0.03, and 0.05 for copy number variations, gene expressions, and protein assays, respectively. We then have a total of three pairs of p-values (0.01, 0.03), (0.01, 0.05), (0.03, 0.05), with maximums 0.03, 0.05, and 0.05 respectively. The MiniMax statistic is the smallest value of these maximums, which is 0.03. Intuitively, the MiniMax statistic provides a way to identify pathways with differential changes (i.e., small p-values) in *at least two* types of omics data. Note that in this case, the MiniMax statistic is equivalent to taking the second smallest p-value among all p-values; that is, the second-order statistic, $P_{(2)}$, of the pathway p-values. Instead of considering pairs of p-values, the MiniMax statistic can also be computed for triplets or quadruplets of p-values from three, four, or more types of omics data similarly to identify pathways with differential changes (i.e., small p-values) in more than two types of omics data.

Statistical Significance Assessment

To compute p-values for the MiniMax statistic, pathwayMultiomics has two modes: 1) by approximation or 2) by simulation. More specifically, the "approximation" approach is based on the theory that when different types of omics data are independent, the rth order statistic $p_{(r)}$ of the p-values follows a Beta distribution, that is, $P_{(r)} \sim \mathcal{B}(\alpha = r, \beta = G - r + 1)$, where $\mathcal{B}(\cdot, \cdot)$ denotes the Beta distribution and G is the number of different types of omics data (Gentle, 2009; Jones, 2009). Therefore, for integrative analysis that identifies pathways with differential changes in at least two types of omics datasets, the MiniMax statistic is the second-order statistic and has the distribution $P_{(2)} \sim \mathcal{B}(2, 3 - 2 + 1) = \mathcal{B}(2, 2)$ under the null hypotheses. The "approximation" approach is easy to compute and is useful when computational resources are limited or when raw data in different omics data types are not available.

On the other hand, in the "simulation" approach, we simulate the distribution of MiniMax statistics under the null hypothesis, that is, when there is no association between phenotype and the pathway in each type of omics data. More specifically, we generate random phenotype labels for each sample and then re-compute pathway p-values. These resulting p-values are our empirical null p-values. To account for non-independence in the different data types, instead of using the above formula, we estimate values for α and β from the empirical null p-values. In practice, we have found that the more correlated the p-values are across the multi-omics platforms, the smaller $(\hat{\alpha} < 2, \hat{\beta} < G - 1)$ are. The "simulation" approach provides more accurate statistical significance estimation and is recommended when both raw data for different omics and large computational resources are available.

Output

The output of pathwayMultiomics is prioritized pathways with small p-values in multiple omics data types, the MiniMax statistic and significance level for each pathway, and the omics data types that were contributing to the MiniMax statistic. For example, in the apoptosis pathway example we described above, the MiniMax statistic was 0.03, its p-value (using the approximate $\mathcal{B}(2, 2)$

distribution) would be 0.0026, and the omics data that contributed to MiniMax statistic were the copy number variations and gene expression data.

Design of Simulation Studies

We performed a comprehensive simulation study to evaluate and compare the performance of the proposed pathwayMultiomics approach with four alternative methods for prioritizing pathways enriched with concordant but often subtle associations signals. To simulate multi-omics datasets with realistic correlation patterns, we used the TCGA COADREAD dataset (Vasaikar et al., 2018) as our input dataset, which included 614, 222, and 90 samples of copy number alterations (CNAs), gene expression, and proteomics data, respectively. More specifically, the CNA data included gene-level GISTIC2 $\log_2$ ratios for 24,776 genes; gene expression data included normalized counts ($\log_2(x+1)$ transformation) of 6,149 genes generate by the Illumina GenomeAnalyzer platform; and the proteins data include log-ratio normalized protein expression levels of 5,538 genes.

To simulate multi-omics datasets for a collection of pathways, we first created synthetic pathways by performing hierarchical clustering on the 1,710 genes measured by all three types of assays for CNA, gene expression, and protein. More specifically, first, a data matrix with 1,710 genes and 928 samples (from the 623 subjects with at least one type of omics data) was created. Next, within each data type, data for each gene were centered and scaled. Finally, a modified Ward's method (method = "ward.D" in hclust() function) was then used to partition the genes into 50 clusters or 50 synthetic pathways. The number of genes in the resulting pathways ranged from 9 to 74, with an average of 34 genes.

Next, we simulated treated (i.e., true positive) and un-treated (i.e., true negative) pathways. First, we randomly assigned each of the 623 subjects to one of two cancer subtypes: A or B. Next, among the 50 synthetic pathways, we selected five pathways to be our true positive pathways, and treatment effects at different levels (μ = 0.1, 0.2, 0.3, 0.4, 0.5) were added to a subset of genes (p = 20, 40, 60, 80%) within each pathway in each of the multi-omics datasets for samples in subtype A group. This process was then repeated 100 times to create 100 simulated multi-omics datasets, each including 50 pathways, among which 5 pathways are true positive pathways. Overall, we generated datasets for a total of 20 simulation scenarios (5 values for μ × 4 values for p). This benchmark dataset (available at https://zenodo.org/record/5683002#.YZF5SGDMKUk), which was systematically modified from real multi-omics data, can be used for reproducing analyses in this study as well as benchmarking future multi-omics data analysis methods.

To evaluate the false positive rate of each method, we also repeated the same procedures described above, except by setting μ = 0 (i.e., not adding any treatment effect). Multi-omics data was created for a total of 5,000 pathways by generating random sample labels 100 times for the 50 synthetic pathways. The false-positive rate (i.e., test size) for each method was then estimated by the percentage of pathways p-values less than 0.05.

Given the known status of the pathways, we next computed the area under the ROC curve (AUC) for each method. The receiver operating characteristic (ROC) curves is a plot of sensitivity versus 1-specificity as the cutoff for declaring significant pathways is varied. AUC assesses the overall discriminative ability of the methods to determine whether a given pathway is significantly associated with the phenotype (i.e., subtype group of the samples) over all possible significance cutoffs. More specifically, for each of the simulation scenarios, we recorded the rankings of the 50 pathways from most to least extreme (by either a p-value, test statistic, or score returned by a method), constructed ROC curves, and estimated AUC for each method.

Methods Compared in the Simulation Study

We compared pathwayMultiomics with four alternative multi-omics analysis methods: Sparse Multiple Canonical Correlation Analysis (sparse mCCA) (Witten and Tibshirani, 2009), MFA (Dray and Dufour, 2007), iProFun (Song et al., 2019), and mitch (Kaspi and Ziemann, 2020). We chose mCCA to represent multi-omics matrix factorization techniques because it performed best in a recent comparative study of multi-omics analysis methods (Pucher et al., 2019). The last three methods, mitch, iProFun, and MFA were chosen because they were proposed in recent years and can also be applied to un-matched or partially matched datasets (**Table 1**). Note that each of these tools was designed specifically for the analysis of multi-omics data, either matching by samples, genomic features (e.g., gene or probe), or both. In the following, we briefly describe each of the methods compared in our simulation study. In the following, we briefly describe each of the methods compared in our simulation study.

pathwayMultiomics

To compute pathway p-values for single omics data, we used pathwayPCA R package (Odom et al., 2020). PathwayPCA integrates prior biological knowledge to extract Adaptive Elastic-net Sparse PCs (AES-PCs) within each pathway for each omics dataset separately, the first AES-PC with the largest variance was then tested against binary outcome "cancer subtype" using a logistic regression model. The pathway p-values for each type of omics data were then used as input for pathwayMultiomics, to identify pathways dysregulated in more than one omics data type. Because the pathway p-values are calculated for each omics dataset separately, the statistical accuracy and power in pathwayMultiomics analysis will not change as the number of matched samples or shared features decreases.

Sparse Multiple Canonical Correlates Analysis (sCCA)

Sparse Canonical Correlation Analysis (sCCA) is a matrix factorization method that uses penalized multivariate analysis for identifying linear combinations of two groups of variables that are highly correlated. Witten and Tibshirani (2009) (Witten and Tibshirani, 2009) extended sCCA to sparse multiple CCA (mCCA), which can perform integrative analysis of more than two sets of variables measured on the same subjects. In the first step, sparse mCCA finds the set of intersecting (i.e., shared) samples and genes across all multi-omics datasets, i.e., the same set of genes are measured on the same subjects in each of the

TABLE 1 | Methods compared by simulation study. Methods that analyze only matched samples would require multiple types of molecular data (e.g., gene expression and protein) to be generated for the same subject, methods that analyzes only matched genes would require multiple types of molecular data to be generated for the same gene. Summary data refers to resulting statistics such as *p*-values or *t*-statistics from differential expression analysis for genes or pathways. All function calls used default function arguments unless specified.

Method	Matches on	Analyzes only matched samples	Analyzes only matched genes	Can analyze summary data	Implementation R package::function
sCCA	Samples measured by all omics data types	Yes	Yes	No	PMA::MultiCCA.permute() with nperms = 100; and PMA::MultiCCA()
MFA	Features (e.g., genes)	No	Yes	No	ade4::ktab.list.df() and ade4:mfa() with option = "lambda1"
mitch	Features (e.g., genes)	No	Yes	Yes	mitch::mitch_calc() with minsetsize = 5 and priority = "effect"
iProFun	Samples measured on at least two omics data types	No	Yes	No	iProFun::iProFun_permutate() with parameters in package example (pi = rep (0.05, 2); grids = c (seq (0.75, 0.99, 0.01), seq (0.991, 0.999, 0.001), seq (0.9991, 0.9999, 0.0001)); filter = 1; seed = 123).
pathwayMultiomics	Pathways	No	No	Yes	pathwayMultiomics:MiniMax() with parameters orderStat = 2 and method = "parametric"

Abbreviations: sCCA, Sparse Canonical Correlates Analysis; MFA, Multi-Factor Analysis; mitch, multivariate gene set enrichment analysis; iProFun, Integrative Proteogenomic Functional Traits Analysis.

omics datasets. Therefore, the statistical accuracy and power of sparse mCCA to detect multi-omics changes will decrease as the number of shared samples or features decreases because samples or features not shared across all data sets will be discarded. In particular, in the TCGA COADREAD multi-omics datasets, only 71 samples and 1710 genes were measured on all three omics data types (CNA, gene expression, protein). Next, sparse mCCA uses a permutation procedure to determine the thresholds and to extract a single vector of selected genes for each omics data type. The union of these selected genes from each omics data type is then taken as the genes selected by sparse multiple CCA. Finally, a Fisher's Exact Test is used to determine if a pathway is enriched with selected genes. We used mCCA implemented via the MultiCCA() function in the PMA R package (https://cran.r-project.org/web/packages/PMA/index.html), optimal weights and penalties were identified by the MultiCCA.permute() function.

Multi-Factor Analysis (MFA)

The MFA method is also a matrix factorization technique, but it differs from sparse mCCA in that it only requires data to be matched on features rather than samples. For MFA analysis of multi-omics data, the main requirement is that the same set of p genes are measured on all omics data types on potentially different subjects. Therefore, the statistical accuracy and power of MFA to detect multi-omics changes will not be affected by the number of matched samples, but will decrease as the number of shared features decreases, because features not shared across all data sets will be discarded. In the first step, MFA reshapes data by stacking the multi-omics datasets, each with samples as rows and the same p genes as columns. Next, MFA performs a weighted principal components analysis, where the weights from each data set are inversely related to the principal eigenvalue of the data set (a measurement of the overall variability in the dataset). Then, genes are given a score measuring its concordance across the datasets for different omics types, where the distribution of these scores follows $N\ (0,\ p^{-1/2})$ where p is the number of genes measured on all omics data types. Finally, genes with upper-sided p-values < 0.05 are selected, and Fisher's Exact Test is used to identify pathways significantly enriched with selected genes. We implemented the MFA method using the mfa() function in ade4 R package under default settings.

Multi-Contrast Pathway Enrichment Analysis (mitch)

The mitch method is very similar to the proposed MiniMax statistic because it also computes pathway-level enrichment scores from summary statistics rather than using the data itself. There are several steps in the mitch algorithm: first, users identify the set of p genes measured by all G omics data types, and subsets the multi-omics datasets to include only these p genes. Next, for each omics dataset, methods appropriate for each platform (e.g., DESeq2 for RNASeq data) are used to compute gene-wise summary statistics or gene scores (e.g., p-values or t-statistics) that associate each gene with the phenotype. This step produces a $p \times G$ data matrix (i.e., p genes $\times$ G omics data types). Therefore, the statistical accuracy and power of mitch to detect multi-omics changes will not be affected by the number of matched samples, but will decrease as the number of shared features decreases, because features not shared across all data sets will be discarded. Finally, for each pathway, mitch performs a one-way MANOVA to test if gene scores across the G omics data types are different for genes within the pathways compared to background genes. We compared the mitch algorithm, computed using the mitch_calc() routine from the mitch R package with priority = "effect", with two alternative gene-wise summary statistics: the gene-specific t-statistic obtained after fitting a linear model that associated each gene with subtype group effect (labeled as "mitch_tStat" in **Figure 2**), and the gene-specific p-values from the same linear models (labeled as "mitch_pValue"). Note that using the t-statistic accounted for different directions of associations among genes while using the p-value did not.

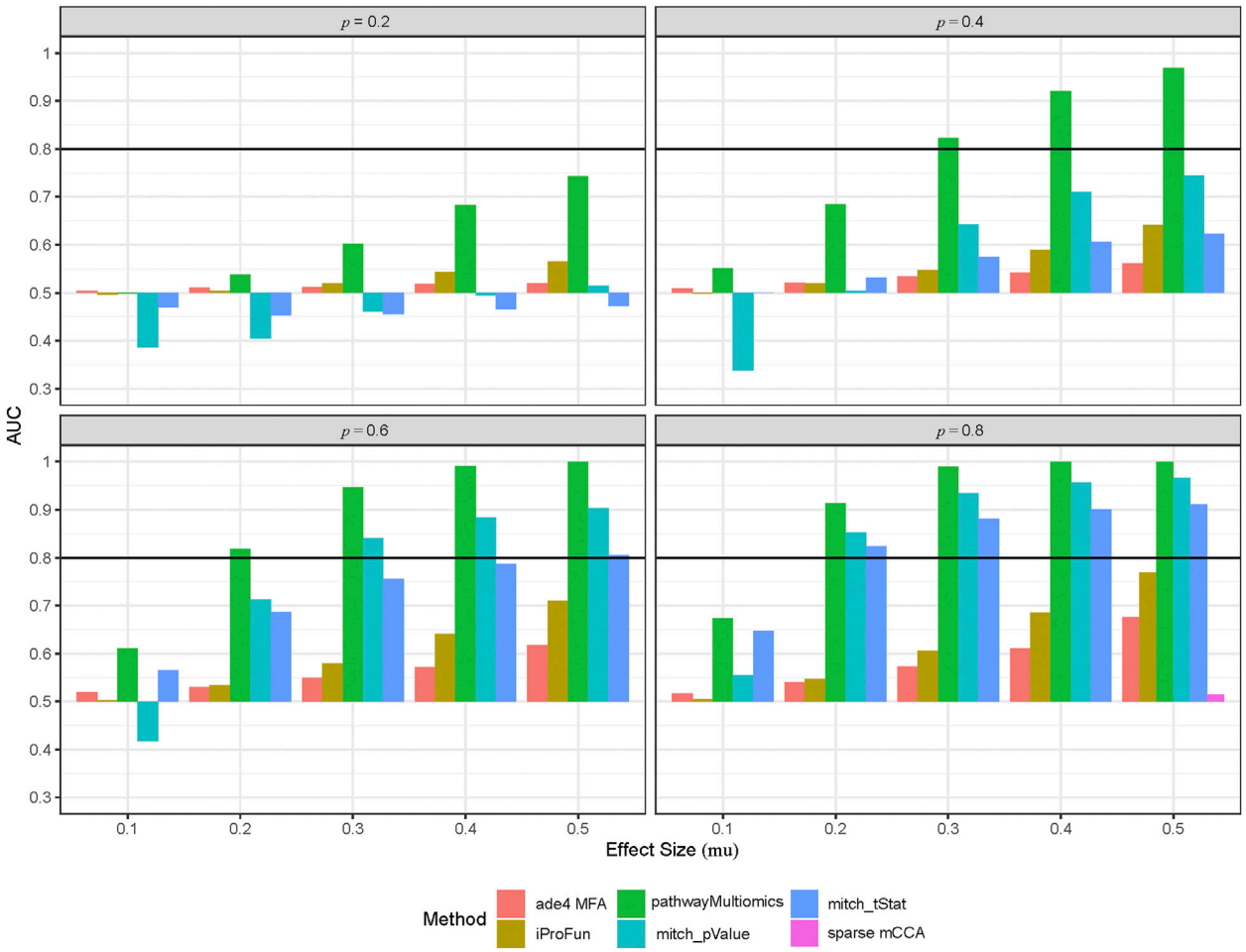

FIGURE 2 | Performance of different multi-omics analysis methods in the simulation study. To simulate multi-omics datasets, we used the TCGA COADREAD datasets (in copy number alterations, gene expressions and proteomics data) as an input, created 50 synthetic pathways by clustering genes measured by all three types of omics data, and then added treatment with different effect sizes (mu) to a proportion (p = 0.2, 0.4, 0.6, 0.8) of the genes. This process was repeated for 100 times to create 100 simulated multi-omics datasets for each simulation scenario (i.e., different combinations of mu and p). Shown are area under ROC curves (AUCs) for each method averaged over 100 simulation datasets at each simulation scenario.

Integrative Screening for Proteogenomic Functional Traits (iProFun)

The iProFun method (Song et al., 2019) aims to detect DNA copy numbers (CNA) and methylation alterations (DNAm) with downstream functional consequences in mRNA expression levels, global protein abundances, or phosphoprotein abundances. In the first step, iProFun fits three linear models, each with a molecular trait (mRNA, global protein, or phosphoprotein) as the outcome, and CNA or DNAm as the predictor, along with additional covariate variables (e.g., age, sex). Next, multiple comparison correction is applied to *p*-values of the predictor (CNA or DNAm) in each of the three linear models, and genes with at least one significant predictor are selected. Finally, Fisher's Exact Test is used to identify pathways enriched with selected genes. Notably, iProFun allows more flexibility in the input dataset and can take advantage of samples not completely measured on all omics types. Specifically, iProFun requires samples to be measured by at least one genomic (e.g., copy number, DNA methylation) trait and at least one transcriptomic (i.e., mRNA) or proteomic (e.g., global, phosphor protein) trait, but it does not require samples to be measured by more than one genomic trait or more than one transcriptomic/proteomic traits. In the simulation study, the number of shared samples analyzed by iProFun were 216 (copy number and RNAseq) and 88 (copy number and proteomics). The statistical accuracy and power of sparse iProFun to detect multi-omics changes will decrease as the number of these shared samples (between copy number and RNAseq, or between copy number and proteomics) decreases, because samples not shared by at least two data sets will be discarded. In our simulation study, we used the iProFun_permutate() function in the iProFun package to independently predict synthetic gene expressions and proteomics data from simulated copy number aberrations.

Default parameter values, as shown in package examples, were used for all functions.

Analysis of Multi-Omics Datasets in Alzheimer's Disease

pathwayMultiomics Analysis

We next applied pathwayMultiomics to analyze a set of multi-omics datasets in Alzheimer's disease. The input of pathwayMultiomics analysis is pathway *p*-values for single omics data. Therefore, we first performed pathway analysis for genetic variants, DNAm, and gene expressions using the mixed model approach (Wang et al., 2011), MissMethyl (Phipson et al., 2016), and fgsea (Korotkevich et al., 2021) methods, which were specifically designed for pathway analyses of these different omics data types.

More specifically, for the analysis of genetic variants, Kunkle et al. (2019) (Kunkle et al., 2019) described a recent large meta-analysis of more than 90,000 individuals to identify genetic variants associated with AD. We downloaded summary statistics for individual variants obtained in this study from https://www.niagads.org/igap-rv-summary-stats-kunkle-p-value-data ("Kunkle_et al._Stage1_results.txt"). Next, we performed GWAS pathway analysis using the mixed model approach (Wang et al., 2011), which tested the combined association signals from a group of variants in the same pathway against the null hypothesis that there is no overall association between SNPs in a pathway and the outcome (i.e., AD status). An empirical null distribution, estimated using the bacon R package (van Iterson et al., 2017), was used to estimate the statistical significance of the pathways.

For the analysis of DNA methylation data, we recently performed a meta-analysis of more than 1,000 prefrontal cortex brain samples (Zhang et al., 2020) to identify epigenetic changes associated with AD Braak stage, a standardized measure of neurofibrillary tangle burden determined at autopsy. Braak scores range from 0 to 6, corresponding to increased severity of the disease (Braak and Braak, 1995). **Supplementary Tables 1, 2** in Zhang et al. (2020) included summary statistics for 3,751 differentially methylated individual CpGs and 119 differentially methylated regions (DMRs) that reached a 5% FDR significance threshold in our meta-analysis. The combined collections of the significant individual CpGs and CpGs located in the DMRs were then used as input for pathway analysis via the MissMethyl R package (Phipson et al., 2016), which performs over-representation analysis by determining if AD Braak-associated CpGs are significantly enriched in a pathway. In particular, MissMethyl models the multiple probes mapped to each gene on the methylation arrays using the Wallenius' noncentral hypergeometric test.

For the analysis of RNASeq data, we analyzed 640 samples of RNAseq data measured on postmortem prefrontal cortex brain samples in the ROSMAP AD study. Normalized FPKM (Fragments Per Kilobase of transcript per Million mapped reads) gene expression values generated by the ROSMAP AD study were downloaded from the AMP-AD Knowledge Portal (Synapse ID: syn3388564). For each gene, we assessed the association between gene expression and Braak stage. More specifically, for each gene, we fitted the linear model log2 (normalized FPKM values +1) ~ Braak stage + ageAtDeath + sex + markers for cell types. The last term, "markers for cell types," included multiple covariate variables to adjust for the multiple types of cells in the brain samples. Specifically, we estimated expression levels of genes that are specific for the five main cell types present in the CNS: ENO2 for neurons, GFAP for astrocytes, CD68 for microglia, OLIG2 for oligodendrocytes, and CD34 for endothelial cells, and included these as variables in the above linear regression model, as was done in a previous large study of AD samples (De Jager et al., 2014). This linear model identifies genes for which gene expressions are associated with AD Braak stage linearly (Zhang et al., 2020). For pathway analysis, we ranked each gene by *p*-values for the Braak stage in the above linear model, which was then used as input for the Fast Gene Set Enrichment Analysis (fgsea) (Korotkevich et al., 2021) software. The fgsea software performs pathway analysis of genome-wide gene expression data by determining if genes within a pathway are enriched on top of the gene list (ranked by gene-wise differential gene expression *p*-values) compared to the rest of the genes.

The pairwise correlations of *p*-values in individual omics data types are very small, at $\rho = 0.0045$ (SNP pathway *p*-values vs. DNAm pathway *p*-values), −0.0263 (SNP pathway *p*-values vs. RNAseq pathway *p*-values), and 0.0432 (DNAm pathway *p*-values vs. RNAseq pathway *p*-values). In pathwayMultiomics, we used the approximation approach, supported by the relatively low pairwise correlations in pathway *p*-values of individual omics data types.

mitch Analysis

The input of mitch R package is summary statistics for genes such as *p*-values for different types of omics data. For the GWAS meta-analysis results described in (Kunkle et al., 2019), we assigned SNPs to a gene if they were located within 5 kb upstream of the first exon or downstream of the last exon (Wang et al., 2011). Next, we represented each gene by the smallest *p*-value if there are multiple SNPs associated with it. To remove selection bias due to different numbers of SNPs associated with each gene (i.e., the smallest *p*-value for a gene with many SNPs is likely to be smaller than the smallest *p*-value for a gene with only a few SNPs), we next fit a generalized additive model using the R package gam: $Y_i \sim f(n.links_i)$ where Y_i is - $\log_{10}$ transformation of the smallest *p*-value for gene *i*, $n.\ links_i$ is the number of SNPs associated with gene *i*, and *f* is a spline function. We assumed gamma distribution for Y_i, as under the null hypothesis of no association, Y_i follows the chi-square distribution (a special case of gamma distribution). The spline model allows us to model linear and nonlinear associations between the number of SNPs mapped to a gene and the strength of significance for the gene as previously described (Zhang et al., 2021). The residuals from this model, which represented -$\log_{10}$ transformation of the *p*-values with gene size effects removed, were then estimated, and used as input for genetic data in mitch.

Similarly, for the analysis of DNA methylation data, we assigned CpGs to genes based on Illumina annotation, represented each gene by the CpG with the smallest *p*-value, and removed the bias due to gene size using the same spline model described above, except *n. links*$_i$ is the number of CpGs associated with gene *i*. The residuals from the spline model were then used as input for DNAm data in mitch.

For the analysis of RNAseq data, we used the R package fgsea (Korotkevich et al., 2021). For each gene, we fit a linear model log2 (normalized FPKM values +1) ~ Braak stage + ageAtDeath + sex + markers for cell types. As described above, the last term, "markers for cell types" included covariate variables (marker gene expressions of ENO2, *GFAP, CD68, OLIG2, CD34*) to adjust for the multiple types of cells in the brain samples. The -log10 transformation of the *p*-values for the Braak stage in the above model was then used as input for RNASeq data in mitch.

All analyses were performed using the R software (version 4.0) and SAS software (version 9.4). We used the venny tool (Oliveros, 2007-2015). To account for multiple comparisons, we computed the false discovery rate using the method of Benjamini and Hochberg (Benjamini Y and Y, 1995). The scripts for the analysis performed in this study can be accessed at https://github.com/TransBioInfoLab/pathwayMultiomics_manuscript_supplement.

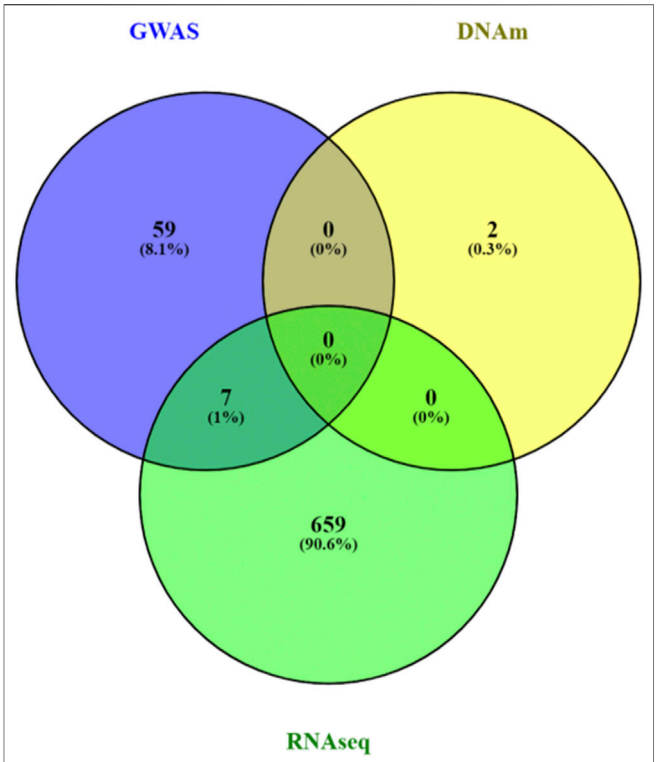

FIGURE 3 | Venn diagram of pathway analyses results for individual omics data types. A total of 666, 2 and 66 significant pathways reached 5% false discovery rate in the analyses of GWAS, DNA methylation (DNAm) and RNASeq data pathway analyses, respectively. Very few pathways (n = 7) were significantly associated with AD in more than one omics data types. The mixed models approach, MissMethyl, and fgsea, which were specifically designed for pathway analyses of genetic variants, DNAm, and gene expression data were used to analyze a total of 2,833 canonical pathways in MsigDB database.

RESULTS

Results of the Simulation Study

As discussed in Methods, pathwayMultiomics has two approaches for computing *p*-values, either by approximation using formula or by simulation. Our results showed the estimated parameters α and β for Beta distribution based on simulation are $\alpha = 1.85$ and $\beta = 1.9$, which are very similar to the theoretical values of $\alpha = 2$ and $\beta = 2$ used in the approximation approach. The results in **Supplementary Table 1** showed that both the simulation and approximation approaches had Type-I error rates close to 5%. Therefore, we next compared AUCs for the pathwayMultiomics method in the approximation approach with the other four methods.

Among all methods, the pathwayMultiomics method performed best with the highest AUCs across all 20 simulation scenarios (**Figure 2**, **Supplementary Table 2**). The second-best performing method is mitch, for which ranking genes by *p*-values performed better than ranking genes by *t*-statistic in most simulation scenarios, except the ones with weak association signals (i.e., effect size = 0.1). The iProFun method also performed well in the simulated pathways that included a high proportion (e.g., 80%) of genes with large association signals (e.g., effect size = 0.5). On the other hand, the sparse mCCA and MFA methods lacked power, probably because these matrix factorization techniques lost information by requiring matched samples or genes across all platforms, and their unsupervised framework also ignored phenotype information. Because sparse mCCA lacked power even in the last simulation scenario with the strongest signal (80% genes in a true positive pathway are treated with an effect size of 0.5), we only included AUC for sparse mCCA in the last simulation scenario.

Case Study: Analysis of Multi-Omics Datasets in Alzheimer's Disease

We next applied the two methods that performed best in our simulation study, pathwayMultiomics and mitch, to analyze a collection of real multi-omics datasets in Alzheimer's disease, which included summary statistics for genetic variants and DNA methylation from two recent large-scale meta-analysis studies (Kunkle et al., 2019; Zhang et al., 2020), as well as a gene expression dataset measured on the prefrontal cortex of brain samples generated by the ROSMAP study (De Jager et al., 2014; De Jager et al., 2018). Note that because we did not have access to raw genotype data included in the meta-analysis, many of the tools that require raw omics data would not be applicable here. In contrast, pathwayMultiomics and mitch can be applied to analyze summary statistics obtained in meta-analyses. For comparison, we also included a third method, the commonly used Venn diagram method, which identifies pathways that are significant in multiple omics data types.

TABLE 2 | Top 10 most significant pathways identified by pathwayMultiomics in the analysis of multiomics Alzheimer's datasets.

Pathway	Size	Single omics *p*-values			Single omics FDRs			pathwayMultiomics			
		SNP	DNAm	RNASeq	SNP	DNAm	RNASeq	MiniMax	*p*-value	FDR	Contributing Omics
PID_PDGFRB_PATHWAY	126	6.99E-01	1.45E-04	1.67E-04	9.99E-01	1.37E-01	3.30E-03	1.67E-04	8.33E-08	2.36E-04	DNAm, RNA
WP_CHEMOKINE_SIGNALING_PATHWAY	155	8.19E-01	3.17E-04	1.94E-05	9.99E-01	1.39E-01	9.00E-04	3.17E-04	3.02E-07	3.28E-04	DNAm, RNA
KEGG_HEMATOPOIETIC_CELL_LINEAGE	80	3.67E-36	3.40E-04	7.61E-01	3.24E-34	1.39E-01	8.11E-01	3.40E-04	3.48E-07	3.28E-04	SNP, DNAm
PID_TCR_PATHWAY	58	4.48E-04	2.75E-02	4.90E-04	2.04E-02	6.43E-01	6.55E-03	4.90E-04	7.20E-07	5.10E-04	SNP, RNA
WP_REGULATION_OF_TOLLLIKE_RECEPTOR_SIGNALING_PATHWAY	128	3.76E-05	3.32E-02	6.55E-04	2.08E-03	6.68E-01	7.70E-03	6.55E-04	1.29E-06	5.69E-04	SNP, RNA
KEGG_CHEMOKINE_SIGNALING_PATHWAY	172	7.90E-01	6.72E-04	2.98E-04	9.99E-01	1.47E-01	4.70E-03	6.72E-04	1.35E-06	5.69E-04	DNAm, RNA
PID_KIT_PATHWAY	52	2.55E-01	6.84E-04	1.10E-04	9.99E-01	1.47E-01	2.69E-03	6.84E-04	1.40E-06	5.69E-04	DNAm, RNA
WP_KIT_RECEPTOR_SIGNALING_PATHWAY	57	3.05E-02	3.68E-04	1.41E-03	5.95E-01	1.39E-01	1.26E-02	1.41E-03	5.94E-06	2.10E-03	DNAm, RNA
PID_CXCR4_PATHWAY	98	4.87E-04	1.50E-03	7.29E-02	2.19E-02	2.36E-01	1.56E-01	1.50E-03	6.76E-06	2.13E-03	SNP, DNAm
REACTOME_TCR_SIGNALING	112	6.06E-52	2.07E-01	2.16E-03	6.11E-50	9.46E-01	1.68E-02	2.16E-03	1.40E-05	3.98E-03	SNP, RNA

We analyzed 2,833 canonical pathways (C2:CP collection) in MSigDB (Subramanian et al., 2005) that included between 3 and 200 genes. Analyzing each omics data type individually, at a 5% false discovery rate (FDR), we identified 66, 2, and 666 pathways associated with AD in SNP, DNAm, and gene expression data, respectively (**Supplementary Table 3–5**). There was little agreement between the FDR-significant pathways identified in different omics datasets (**Figure 3**). A possible reason could be the lack of power in single omics studies for Alzheimer's disease, which has relatively weaker association signals than other complex diseases such as cancers. Among the top pathways, only seven pathways reached 5% FDR in more than one omics data type. These seven pathways, which reached 5% FDR in both GWAS and RNASeq analysis, are MHC Class II antigen presentation, TCR signaling, factors involved in megakaryocyte development and production, Rig I like receptor signaling pathway, DDX58 IFIH1 mediated induction of interferon alpha-beta, and regulation of toll-like receptor signaling pathway, all of which are involved in inflammatory responses, highlighting the importance of immune processes in AD (Cunningham, 2013; Heneka et al., 2015).

At 5% FDR, pathwayMultiomics identified 74 significant pathways (**Supplementary Table 6**). Note that for this analysis example, the MiniMax statistics in pathwayMultiomics is the minimum of all maximums in pairs of *p*-values from individual omics, that is min{ max (SNP pathway *p*-value, DNAm pathway *p*-value), max (SNP pathway *p*-value, RNAseq pathway *p*-value), max (DNAm pathway *p*-value, RNAseq pathway *p*-value) }. For these significant pathways, we next examined which two omics data types contributed to the MiniMax statistics. Among the 74 pathways, the significance of the pathwayMultiomics *p*-value (for MiniMax statistic) was driven by pathway *p*-values for DNAm and RNA in the majority of pathways (n = 40, 54%), followed by pathway *p*-values for SNP and RNA (n = 25, 34%), recapitulating the prominent gene regulatory role of DNAm in AD (Klein et al., 2016). In contrast, pathwayMultiomics *p*-values were driven by *p*-values for SNP and DNAm in only 9 (12%) out of the 74 significant pathways, consistent with the relatively independent contributions of genetic variants and DNA methylations in influencing AD susceptibility (Chibnik et al., 2015; Klein et al., 2016). The majority of the top 10 most significant pathways identified by pathwayMultiomics (**Table 2**) involved signaling pathways activated by the immune system in responses to amyloid-β induced neurotoxicity in AD brains, such as the activation of chemokines (Jorda et al., 2020), toll-like receptors (Landreth and Reed-Geaghan, 2009), T cell receptors (Gate et al., 2020), PDGFR-beta receptors (Liu H. et al., 2018), and CXCR4 receptors (Li and Wang, 2017). Notably, seven out of these top 10 pathways did not reach 5% FDR in more than one type of omics in the analysis of individual omics data types (**Figure 3**), so these pathways would have been missed by the conventional Venn diagram method.

At 5% FDR, mitch identified 237 pathways (**Supplementary Table 7**). The most significant pathway pointed to systemic lupus erythematosus (SLE), an autoimmune disease in which the immune system attacks the body's own tissues. A recent meta-analysis found that patients with SLE have a significantly higher risk for cognitive impairment (Zhao et al., 2018). Other top pathways (**Table 3**)

TABLE 3 | Top 10 most significant pathways identified by the mitch method in the analysis of Alzheimer's disease multi-omics datasets.

Pathway	Size	*p*-value	FDR
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	128	7.49E-19	2.11E-15
REACTOME_SIRT1_NEGATIVELY_REGULATES_RRNA_EXPRESSION	65	3.16E-15	4.46E-12
REACTOME_DNA_METHYLATION	62	1.21E-13	1.14E-10
REACTOME_ACTIVATED_PKN1_ STIMULATES_TRANSCRIPTION_OF_ AR_ANDROGEN_RECEPTOR_REGULATED_ GENES_KLK2_AND_KLK3	64	2.40E-13	1.69E-10
REACTOME_HDACS_DEACETYLATE_HISTONES	91	6.03E-13	3.40E-10
REACTOME_CONDENSATION_OF_PROPHASE_CHROMOSOMES	71	5.28E-12	2.48E-09
REACTOME_HDMS_DEMETHYLATE_HISTONES	45	5.17E-11	2.08E-08
REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	129	4.21E-10	1.48E-07
REACTOME_PRC2_METHYLATES_HISTONES_AND_DNA	70	5.40E-10	1.69E-07
REACTOME_TRANSCRIPTIONAL_REGULATION_OF_GRANULOPOIESIS	88	6.79E-10	1.91E-07

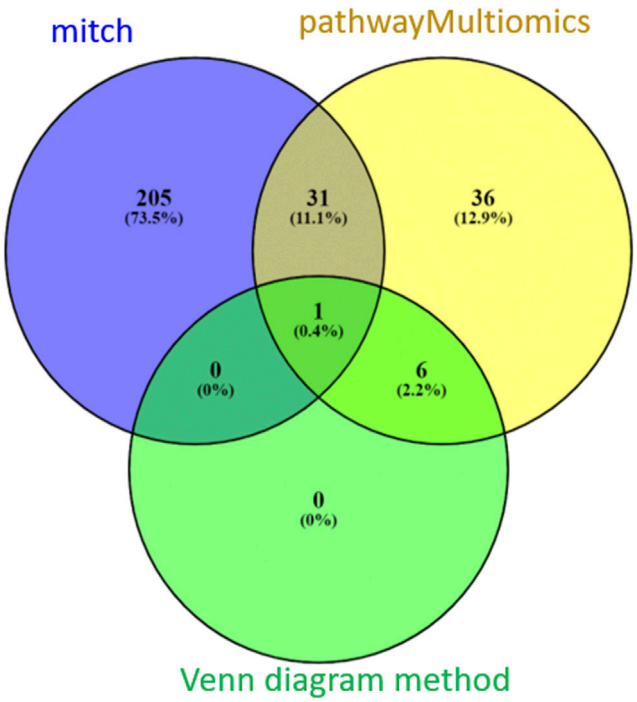

FIGURE 4 | A comparison of FDR significant pathways identified by pathwayMultiomics, mitch, and Venn diagram analyses. At 5% FDR, pathwayMultiomics and mitch identified 74 and 237 pathways, respectively. The Venn diagram method identified 7 pathways with 5% FDR in more than one type of omics data type. There was only modest overlap between the three methods. A total of 32 pathways (11%) were significant in both pathwayMultiomics and mitch methods. PathwayMultiomics identified all the significant pathways using the Venn diagram method. There was no overlap between significant pathways by mitch and Venn diagram, except for one pathway (T cell Receptor pathway), which was identified by all three methods.

highlighted key biological processes regulated by proteins previously shown to be important in AD, such as PRC2 (Zhang et al., 2020), which regulates neuronal lineage specification, proliferation, and differentiation (Liu P.-P. et al., 2018); PKN1, which was shown to have a neuroprotective role (Thauerer et al., 2014); and histone deacetylases (HDACS), which maintains the histone acetylation homeostasis and play important roles in the process of neuronal differentiation, neurite outgrowth and neuroprotection (Shukla and Tekwani, 2020).

Between the three methods (pathwayMultiomics, mitch, and Venn diagram), there was only modest overlap (**Figure 4**). A total of 32 pathways (11%) reached 5% FDR by both pathwayMultiomics and mitch methods. PathwayMultiomics identified all seven significant pathways that were significant in more than one type of omics data type based on the Venn diagram method. There was no overlap between significant pathways by mitch and Venn diagram method, except for one pathway (T cell Receptor pathway), which was identified by all three methods.

DISCUSSION

To identify pathways dysregulated in multiple types of omics datasets, we developed the pathwayMultiomics R package. PathwayMultiomics is flexible and only requires pathway *p*-values for individual omics data types as input, thus making it possible to take advantage of pathway analysis tools that are specially designed for each omics data type. In addition, pathwayMultiomics is computationally efficient, does not require matched samples from multi-omics data, and is applicable in situations when raw omics data are not available, such as when aggregating summary statistics from meta-analyses related to the same disease. PathwayMultiomics is also informative; the individual omics data type that contributed to pathwayMultiomics significance can be used to distinguish pathways with potentially different underlying regulatory mechanisms, such as the pathways for which gene expressions are regulated by DNA methylation versus pathways for which gene expressions are mainly regulated by genetic variants.

We performed a comprehensive simulation study to assess the statistical properties of our method. To emulate correlation patterns in real omics datasets, we generated simulation datasets using real TCGA multi-omics datasets as input. We showed that pathwayMultiomics significantly outperforms currently available multi-omics methods with improved power and well-controlled false-positive rates. A challenge with

analyzing multi-omics datasets is that many of the samples with data recorded for one molecular type did not have matching data from other data types. Therefore, methods that require matched samples across all data types (e.g., mSCCA) would only analyze a subset of the samples, which would result in reduced statistical power. Also, often only a subset of genes is measured by multiple omics platforms. Therefore, methods that require the same set of genes measured on all omics data types (e.g., MFA) may also exclude important biological signals, leading to reduced power. Finally, unsupervised methods (e.g., NMF, sCCA, and iProFun) might also lose power because they do not leverage information in the phenotypes. In contrast, pathwayMultiomics gains power by leveraging information in all samples (including the un-matched samples), and all features (e.g., genes) mapped to the pathways, as well as phenotype information along with multi-omics data.

To further assess the performance of pathwayMultiomics on real datasets, we also compared it with two alternative approaches using the Venn diagram and mitch. When multiple types of omics data are available, a commonly used strategy is to test for marginal associations between each type of omics data with phenotype first, and then use Venn diagram to intersect significant pathways or genes that overlap in different omics data types. Although a good visualization tool, Venn diagrams do not provide prioritization or any statistical assessment for pathways. In addition, it might be overly stringent because when several types of omics data are considered, often few (if any) pathways pass the threshold of statistical significance in all omics data types. In contrast, pathwayMultiomics provides prioritization and statistical assessment for pathways with moderate to strong association signals in multiple omics data types. In our analysis of multi-omics AD datasets, at 5% FDR, pathwayMultiomics identified 67 pathways in addition to the seven FDR-significant pathways in more than one type of omics data as identified by the Venn diagram method. The discrepancy in multi-omics analysis results by pathwayMultiomics and mitch is not unexpected. In addition to the differences in underlying algorithms, an important reason might also be the different hypotheses these methods test. While mitch tests the competitive null hypothesis that the genes in a pathway show the same magnitude of associations with the disease phenotype compared with genes in the rest of the genome, pathwayMultiomics tests the self-contained null hypothesis that the genes in a pathway are not associated with the disease phenotype (Tian et al., 2005). Therefore, mitch and pathwayMultiomics analysis complement each other in the analysis of multi-omics datasets. PathwayMultiomics is available as an R package and can be accessed at https://github.com/TransBioInfoLab/pathwayMultiomics.

CONCLUSIONS

In summary, we have presented the pathwayMultiomics method, which can be used to analyze multi-omics data with any type of outcome variables (e.g., categorical, continuous, or survival phenotypes). We have shown that pathwayMultiomics significantly outperforms currently available multi-omics methods with improved power and well-controlled false-positive rates. In addition, we also analyzed multi-omics datasets in Alzheimer's disease to show that pathwayMultiomics was able to recover known biology, as well as nominate novel biologically meaningful pathways. We expect pathwayMultiomics to be a useful tool for integrative analysis of multiple types of omics data.

DATA AVAILABILITY STATEMENT

The TCGA cancer datasets can be accessed from the LinkedOmics repository http://linkedomics.org/login.php, the Alzheimer's GWAS summary statistics can be accessed from https://www.niagads.org/igap-rv-summary-stats-kunkle-p-value-data (file "Kunkle_et al._Stage1_results.txt"), the ROSMAP RNASeq dataset can be accessed from AMP-AD (accession: syn3388564). The pathwayMultiomics software can be accessed at https://github.com/TransBioInfoLab/pathwayMultiomics The scripts for the analysis performed in this study can be accessed at https://github.com/TransBioInfoLab/pathwayMultiomics_manuscript_supplement The benchmark dataset used in the simulation study is available at https://zenodo.org/record/5683002#.YZF5SGDMKUk.

AUTHOR CONTRIBUTIONS

GO, LW, XC, AC, and TS designed the computational analysis. GO, AC, TS, and LW analysed the data. GO, LW, XC, and AC contributed to the interpretation of the results. GO, LW wrote the paper, and all authors participated in the review and revision of the manuscript. LW conceived the original idea and supervised the project.

FUNDING

This work was supported by the National Institutes of Health (R01CA158472 (XC), R01 CA200987 (XC), P30CA240139 (XC), R01AG061127 (LW), R01AG062634 (LW), and R21AG060459 (LW)). The ROSMAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Bing Zhang for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.783713/full#supplementary-material

REFERENCES

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 (Methodological)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Braak, H., and Braak, E. (1995). Staging of Alzheimer's Disease-Related Neurofibrillary Changes. *Neurobiol. Aging* 16, 271–278. doi:10.1016/0197-4580(95)00021-6

Chibnik, L. B., Yu, L., Eaton, M. L., Srivastava, G., Schneider, J. A., Kellis, M., et al. (2015). Alzheimer's Loci: Epigenetic Associations and Interaction with Genetic Factors. *Ann. Clin. Transl Neurol.* 2, 636–647. doi:10.1002/acn3.201

Cunningham, C. (2013). Microglia and Neurodegeneration: the Role of Systemic Inflammation. *Glia* 61, 71–90. doi:10.1002/glia.22350

De Jager, P. L., Ma, Y., Mccabe, C., Xu, J., Vardarajan, B. N., Felsky, D., et al. (2018). A Multi-Omic Atlas of the Human Frontal Cortex for Aging and Alzheimer's Disease Research. *Sci. Data* 5, 180142. doi:10.1038/sdata.2018.142

De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., et al. (2014). Alzheimer's Disease: Early Alterations in Brain DNA Methylation at ANK1, BIN1, RHBDF2 and Other Loci. *Nat. Neurosci.* 17, 1156–1163. doi:10.1038/nn.3786

Dray, S., and Dufour, A. (2007). The Ade4 Package: Implementing the Duality Diagram for Ecologists. *J. Stat. Softw.* 22, 1–20. doi:10.18637/jss.v022.i04

Gao, F., Foat, B. C., and Bussemaker, H. J. (2004). Defining Transcriptional Networks through Integrative Modeling of mRNA Expression and Transcription Factor Binding Data. *BMC Bioinformatics* 5, 31. doi:10.1186/1471-2105-5-31

Gate, D., Saligrama, N., Leventhal, O., Yang, A. C., Unger, M. S., Middeldorp, J., et al. (2020). Clonally Expanded CD8 T Cells Patrol the Cerebrospinal Fluid in Alzheimer's Disease. *Nature* 577, 399–404. doi:10.1038/s41586-019-1895-7

Gentle, J. E. (2009). *Computational Statistics*. Berlin/Heidelberg, Germany: Springer.

Heneka, M. T., Carson, M. J., Khoury, J. E., Landreth, G. E., Brosseron, F., Feinstein, D. L., et al. (2015). Neuroinflammation in Alzheimer's Disease. *Lancet Neurol.* 14, 388–405. doi:10.1016/s1474-4422(15)70016-5

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

Jones, M. C. (2009). Kumaraswamy's Distribution: A Beta-type Distribution with Some Tractability Advantages. *Stat. Methodol.* 6, 70–81. doi:10.1016/j.stamet.2008.04.001

Jorda, A., Campos-Campos, J., Iradi, A., Aldasoro, M., Aldasoro, C., Vila, J. M., et al. (2020). The Role of Chemokines in Alzheimer's Disease. *Emiddt* 20, 1383–1390. doi:10.2174/1871530320666200131110744

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, D109–D114. doi:10.1093/nar/gkr988

Kaspi, A., and Ziemann, M. (2020). Mitch: Multi-Contrast Pathway Enrichment for Multi-Omics and Single-Cell Profiling Data. *BMC Genomics* 21, 447. doi:10.1186/s12864-020-06856-9

Klein, H.-U., Bennett, D. A., and De Jager, P. L. (2016). The Epigenome in Alzheimer's Disease: Current State and Approaches for a New Path to Gene Discovery and Understanding Disease Mechanism. *Acta Neuropathol.* 132, 503–514. doi:10.1007/s00401-016-1612-7

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., and Sergushichev, A. (2021). Fast Gene Set Enrichment Analysis. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/060012v060013.full.pdf.

Kunkle, B. W., Grenier-Boley, B., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., et al. (2019). Genetic Meta-Analysis of Diagnosed Alzheimer's Disease Identifies New Risk Loci and Implicates Aβ, Tau, Immunity and Lipid Processing. *Nat. Genet.* 51, 414–430. doi:10.1038/s41588-019-0358-2

Kutalik, Z., Beckmann, J. S., and Bergmann, S. (2008). A Modular Approach for Integrative Analysis of Large-Scale Gene-Expression and Drug-Response Data. *Nat. Biotechnol.* 26, 531–539. doi:10.1038/nbt1397

Landreth, G. E., and Reed-Geaghan, E. G. (2009). Toll-like Receptors in Alzheimer's Disease. *Curr. Top. Microbiol. Immunol.* 336, 137–153. doi:10.1007/978-3-642-00549-7_8

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., and Besse, P. (2009). Sparse Canonical Methods for Biological Data Integration: Application to a Cross-Platform Study. *BMC Bioinformatics* 10, 34. doi:10.1186/1471-2105-10-34

Li, H., and Wang, R. (2017). A Focus on CXCR4 in Alzheimer's Disease. *Brain Circ.* 3, 199–203. doi:10.4103/bc.bc_13_17

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., and Wang, Y.-P. (2013). Group Sparse Canonical Correlation Analysis for Genomic Data Integration. *BMC Bioinformatics* 14, 245. doi:10.1186/1471-2105-14-245

Liu, H., Saffi, G. T., Vasefi, M. S., Choi, Y., Kruk, J. S., Ahmed, N., et al. (2018a). Amyloid-β Inhibits PDGFβ Receptor Activation and Prevents PDGF-BBInduced Neuroprotection. *Car* 15, 618–627. doi:10.2174/1567205015666180110110321

Liu, P.-P., Xu, Y.-J., Teng, Z.-Q., and Liu, C.-M. (2018b). Polycomb Repressive Complex 2: Emerging Roles in the Central Nervous System. *Neuroscientist* 24, 208–220. doi:10.1177/1073858417747839

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A Multivariate Approach to the Integration of Multi-Omics Datasets. *BMC Bioinformatics* 15, 162. doi:10.1186/1471-2105-15-162

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension Reduction Techniques for the Integrative Analysis of Multi-Omics Data. *Brief Bioinform* 17, 628–641. doi:10.1093/bib/bbv108

Odom, G. J., Ban, Y., Colaprico, A., Liu, L., Silva, T. C., Sun, X., et al. (2020). PathwayPCA: an R/Bioconductor Package for Pathway Based Integrative Analysis of Multi-Omics Data. *Proteomics* 20, e1900409. doi:10.1002/pmic.201900409

Oliveros, J. C. (2007). Venny. An Interactive Tool for Comparing Lists with Venn's Diagrams. Available at: https://bioinfogp.cnb.csic.es/tools/venny/index.html.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Stat. Appl. Genet. Mol. Biol.* 8–1. doi:10.2202/1544-6115.1406

Phipson, B., Maksimovic, J., and Oshlack, A. (2016). missMethyl: an R Package for Analyzing Data from Illumina's HumanMethylation450 Platform. *Bioinformatics* 32, 286–288. doi:10.1093/bioinformatics/btv560

Pucher, B. M., Zeleznik, O. A., and Thallinger, G. G. (2019). Comparison and Evaluation of Integrative Methods for the Analysis of Multilevel Omics Data: a Study Based on Simulated and Experimental Cancer Data. *Brief Bioinform* 20, 671–681. doi:10.1093/bib/bby027

Shukla, S., and Tekwani, B. L. (2020). Histone Deacetylases Inhibitors in Neurodegenerative Diseases, Neuroprotection and Neuronal Differentiation. *Front. Pharmacol.* 11, 537. doi:10.3389/fphar.2020.00537

Song, X., Ji, J., Gleason, K. J., Yang, F., Martignetti, J. A., Chen, L. S., et al. (2019). Insights into Impact of DNA Copy Number Alteration and Methylation on the Proteogenomic Landscape of Human Ovarian Cancer via a Multi-Omics Integrative Analysis. *Mol. Cell Proteomics* 18, S52–S65. doi:10.1074/mcp.ra118.001220

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Thauerer, B., Zur Nedden, S., and Baier-Bitterlich, G. (2014). Protein Kinase C-Related Kinase (PKN/PRK). Potential Key-Role for PKN1 in Protection of Hypoxic Neurons. *Cn* 12, 213–218. doi:10.2174/1570159x11666131225000518

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering Statistically Significant Pathways in Expression Profiling Studies. *Proc. Natl. Acad. Sci.* 102, 13544–13549. doi:10.1073/pnas.0506577102

Van Iterson, M., Van Zwet, E. W., van Zwet, E. W., and Heijmans, B. T. (2017). Controlling Bias and Inflation in Epigenome- and Transcriptome-wide Association Studies Using the Empirical Null Distribution. *Genome Biol.* 18, 19. doi:10.1186/s13059-016-1131-9

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: Analyzing Multi-Omics Data within and across 32 Cancer Types. *Nucleic Acids Res.* 46, D956–D963. doi:10.1093/nar/gkx1090

Waaijenborg, S., and Zwinderman, A. H. (2009). Sparse Canonical Correlation Analysis for Identifying, Connecting and Completing Gene-Expression Networks. *BMC Bioinformatics* 10, 315. doi:10.1186/1471-2105-10-315

Wang, L., Jia, P., Wolfinger, R. D., Chen, X., Grayson, B. L., Aune, T. M., et al. (2011). An Efficient Hierarchical Generalized Linear Mixed Model for Pathway Analysis of Genome-wide Association Studies. *Bioinformatics* 27, 686–692. doi:10.1093/bioinformatics/btq728

Witten, D. M., and Tibshirani, R. J. (2009). Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Stat. Appl. Genet. Mol. Biol.* 8, Article28. doi:10.2202/1544-6115.1470

Zhang, L., Silva, T. C., Young, J. I., Gomez, L., Schmidt, M. A., Hamilton-Nelson, K. L., et al. (2020). Epigenome-wide Meta-Analysis of DNA Methylation Differences in Prefrontal Cortex Implicates the Immune Processes in Alzheimer's Disease. *Nat. Commun.* 11, 6114. doi:10.1038/s41467-020-19791-w

Zhang, L., Young, J. I., Gomez, L., Silva, T. C., Schmidt, M. A., Cai, J., et al. (2021). Sex-specific DNA Methylation Differences in Alzheimer's Disease Pathology. *Acta Neuropathol. Commun.* 9, 77. doi:10.1186/s40478-021-01177-8

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. *Nucleic Acids Res.* 40, 9379–9391. doi:10.1093/nar/gks725

Zhao, Z., Rocha, N. P., Salem, H., Diniz, B. S., and Teixeira, A. L. (2018). The Association between Systemic Lupus Erythematosus and Dementia A Meta-Analysis. *Dement. Neuropsychol.* 12, 143–151. doi:10.1590/1980-57642018dn12-020006

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

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

Copyright © 2021 Odom, Colaprico, Silva, Chen and Wang. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 January 2022
doi: 10.3389/fgene.2021.778503

Identification of a Ubiquitin Related Genes Signature for Predicting Prognosis of Prostate Cancer

Guoda Song[1,2†], Yucong Zhang[3†], Hao Li[1], Zhuo Liu[1], Wen Song[1], Rui Li[1], Chao Wei[1], Tao Wang[1], Jihong Liu[1] and Xiaming Liu[1]**

[1]Department of Urology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Second Clinical College, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [3]Department of Geriatrics, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Background: Ubiquitin and ubiquitin-like (UB/UBL) conjugations are one of the most important post-translational modifications and involve in the occurrence of cancers. However, the biological function and clinical significance of ubiquitin related genes (URGs) in prostate cancer (PCa) are still unclear.

Methods: The transcriptome data and clinicopathological data were downloaded from The Cancer Genome Atlas (TCGA), which was served as training cohort. The GSE21034 dataset was used to validate. The two datasets were removed batch effects and normalized using the "sva" R package. Univariate Cox, LASSO Cox, and multivariate Cox regression were performed to identify a URGs prognostic signature. Then Kaplan-Meier curve and receiver operating characteristic (ROC) curve analyses were used to evaluate the performance of the URGs signature. Thereafter, a nomogram was constructed and evaluated.

Results: A six-URGs signature was established to predict biochemical recurrence (BCR) of PCa, which included ARIH2, FBXO6, GNB4, HECW2, LZTR1 and RNF185. Kaplan-Meier curve and ROC curve analyses revealed good performance of the prognostic signature in both training cohort and validation cohort. Univariate and multivariate Cox analyses showed the signature was an independent prognostic factor for BCR of PCa in training cohort. Then a nomogram based on the URGs signature and clinicopathological factors was established and showed an accurate prediction for prognosis in PCa.

Conclusion: Our study established a URGs prognostic signature and constructed a nomogram to predict the BCR of PCa. This study could help with individualized treatment and identify PCa patients with high BCR risks.

Keywords: prostate cancer, ubiquitin, prognostic signature, prognosis, bioinformatics

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Feng Gao,
The Sixth Affiliated Hospital of Sun Yat-sen University, China
Karim Farhat,
King Saud University, Saudi Arabia

****Correspondence:***
Jihong Liu
jhliu@tjh.tjmu.edu.cn
Xiaming Liu
xmliu77@hust.edu.cn

[†]These authors have contributed equally to this work and share first authorship

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

Received: *17 September 2021*
Accepted: *30 December 2021*
Published: *17 January 2022*

Citation:
Song G, Zhang Y, Li H, Liu Z, Song W, Li R, Wei C, Wang T, Liu J and Liu X (2022) Identification of a Ubiquitin Related Genes Signature for Predicting Prognosis of Prostate Cancer. Front. Genet. 12:778503. doi: 10.3389/fgene.2021.778503

INTRODUCTION

Prostate cancer (PCa) is one of the most common malignancies worldwide with the third highest cancer-causing deaths following lung cancer and colorectal cancer in American males (Schatten, 2018). The curative therapies for primary tumors are radical prostatectomy or radiation therapy (Mateo et al., 2019). Nearly one third patients would suffer biochemical recurrence (BCR) at 10 years

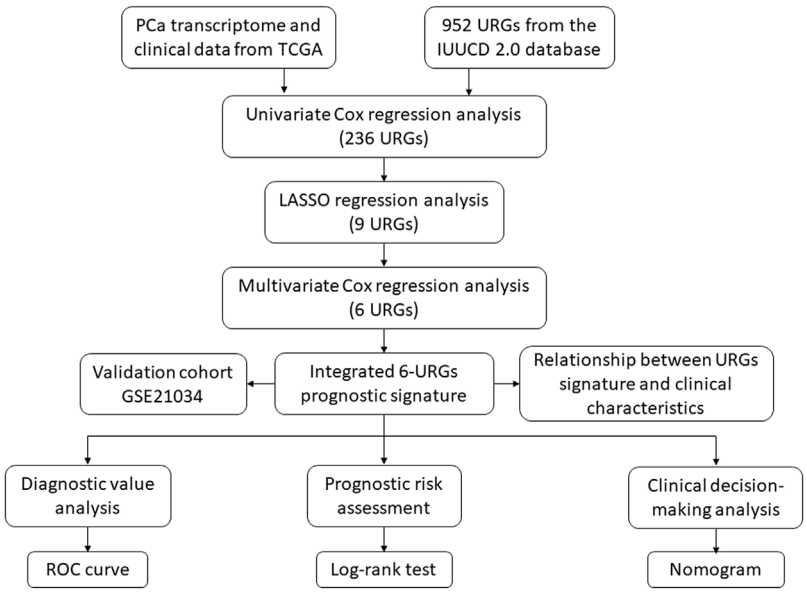

FIGURE 1 | The flowchart of the study procedures. PCa, prostate cancer; TCGA, the Cancer Genome Atlas; URGs, ubiquitin related genes; LASSO, the least absolute shrinkage and selection operator; ROC, receiver operating characteristic.

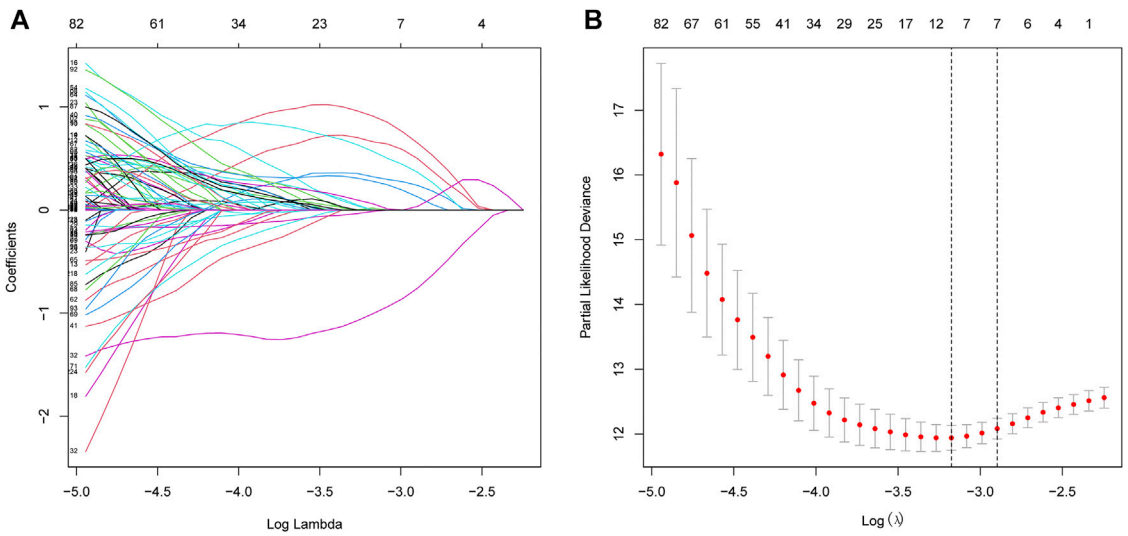

FIGURE 2 | Selection of prognostic URGs by LASSO regression. **(A)** LASSO coefficient profiles of the prognostic genes. **(B)** Parameter selection in LASSO model. URGs, ubiquitin related genes; LASSO, the least absolute shrinkage and selection operator.

after radical prostatectomy without neoadjuvant or adjuvant therapy (Liesenfeld et al., 2017). Without further treatment, the median time from BCR to metastasis and from metastasis to death is 8 and 5 years, respectively (Freedland et al., 2005). Thus, it is important to identify the patients with high risk of BCR.

Ubiquitin and ubiquitin-like (UB/UBL) conjugations are vital post-translational modifications which participate in nearly all biological processes and pathways such as protein degradation and turnover, intercellular signal transduction, cell cycle and DNA damage repair (Swatek and Komander, 2016). Ubiquitin is a highly conserved heat-stable protein with 76 amino acids. The process of ubiquitin conjugation is a successive three-step cascade which is catalyzed by three enzymes including ubiquitin-activating enzymes (E1s), ubiquitin-conjugating enzymes (E2s), and ubiquitin protein ligases (E3s) (Pickart, 2001). However,

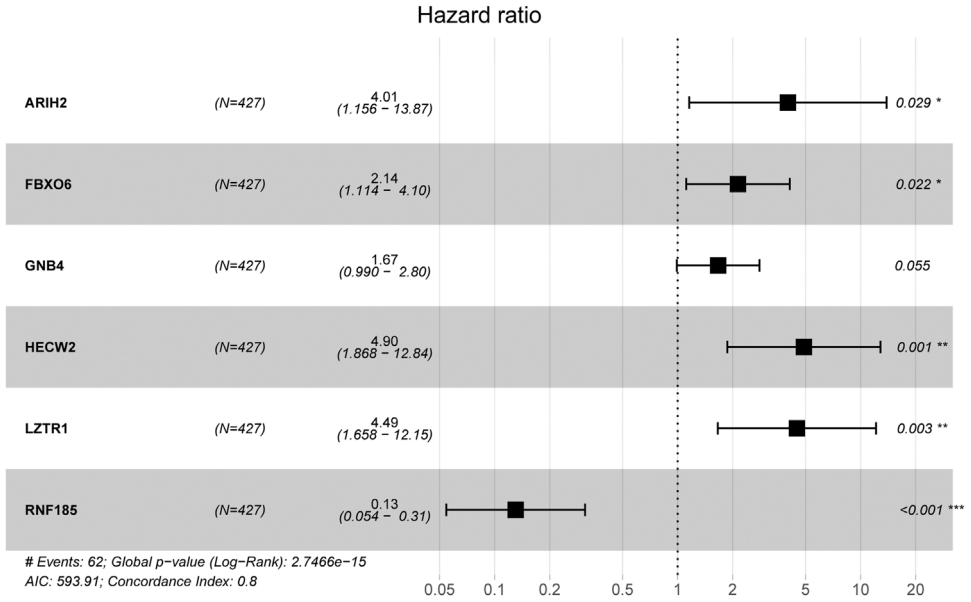

FIGURE 3 | Multivariate Cox regression analysis to identify prognostic genes.

deubiquitinating enzymes (DUBs) remove Ub or UBL moieties from protein and reverse the ubiquitination process (Heride et al., 2014). In addition, the protein containing ubiquitin-binding domain (UBDs) and ubiquitin-like domains (ULDs) also plays an important role in regulating the ubiquitination process (Buchberger, 2002; Heride et al., 2014). Studies have revealed that the dysfunction of protein ubiquitination would be involved in many human pathologies such as tumorigenesis, and neurodegeneration (Popovic et al., 2014). However, no studies have investigated the association between URGs and the prognosis of PCa patients.

In this study, we downloaded transcriptome data and clinicopathological data from the Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases, and performed bioinformatics analysis to identify a URGs signature to predict the prognosis of PCa patients.

MATERIALS AND METHODS

Data Collection and Processing

The transcriptome data and corresponding clinicopathological data, containing 499 tumor and 52 adjacent normal tissues, were downloaded from TCGA (https://portal.gdc.cancer.gov/). The transcriptome data has been background adjusted and normalized with the style of fragments per kilobase million (FPKM) (Mortazavi et al., 2008). Normalized mRNA expression data from GSE21034 dataset, which was served as a validation cohort, were downloaded from GEO database. The genes would be deleted if their expression values were 0 in more than 50% samples. Average expression values were evaluated if genes were duplicated. Then, nine E1s, 43 E2s and 900 E3s were identified from the iUUCD 2.0 database (http://iuucd.biocuckoo.org/). 952 URGs were identified.

Construction and Evaluation of URGs Prognostic Signature for BCR

The expression data of 902 URGs were extracted from processed transcriptome data of TCGA dataset. The batch effect and unwanted variation were removed using "sva" R package between the two datasets. The univariate Cox regression analysis was performed to calculate the association between URGs and BCR-free survival. Next, the least absolute shrinkage and selection operator (LASSO) Cox regression analysis was used to select the valuable prognostic URGs by using the "glmnet" package in R software (Version 4.1.0). Then, a stepwise multivariate Cox regression proportional hazards regression model was established to further select URGs and optimize the model. Finally, a risk score formula was constructed based on the regression coefficients of multivariate Cox analysis and the expression values of corresponding URGs. The risk score formula was listed: *Risk score = (exp Gene1 ×coef Gene1) + (exp Gene2 × coef Gene2) +. . .+ (exp GeneN × coef GeneN)*. Here, exp represents the expression of optimized genes, and coef represents estimated multivariate Cox regression coefficients.

Both median value and optimal cut-off value are widely used for the stratification in survival analysis. Here, we used median risk score of training dataset as the same cut-off value for both training and validation datasets. Kaplan-Meier curve analysis (using "survival" package) and the area (AUC) under receiver operating characteristics (ROC) curve analysis (using the "timeROC" package) were performed to investigate the predictive value of the URGs-based signature. Moreover, GSE21034 dataset was served as a

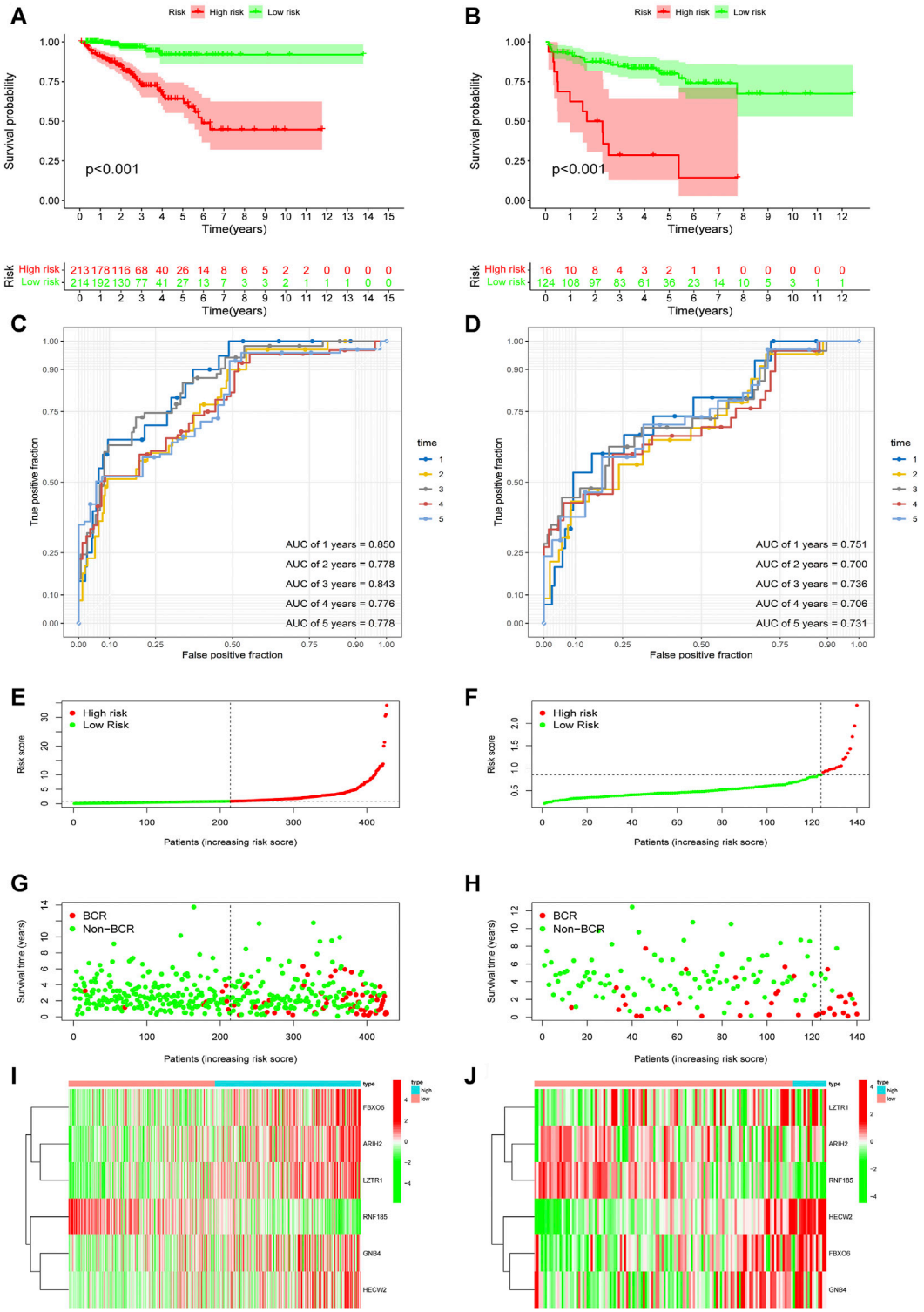

FIGURE 4 | Evaluation and validation of the predictive value of URGs signature in the TCGA dataset and GSE21034 dataset. **(A)** Kaplan-Meier curve analysis between high-risk and low-risk subgroups in the TCGA dataset. The subgroups were stratified by the optimal cut-off value for the risk scores. **(B)** Kaplan-Meier curve analysis between high-risk and low-risk subgroups in the GSE21034 dataset. The subgroups were stratified by the optimal cut-off value for the risk scores. **(C)** The AUCs under ROC for first, second, third, fourth and fifth year BCR predictions based on URGs signature in the TCGA dataset. **(D)** The AUCs under ROC for first, second, third, fourth and fifth year BCR predictions based on URGs signature in GSE21034 dataset. **(E, G, I)** The distribution of survival status and risk score, and heat map of prognostic genes expression in the TCGA dataset. **(F, H, J)** The distribution of survival status and risk score, and heat map of prognostic genes expression in the GSE21034 dataset. URGs, ubiquitin related genes; TCGA, the Cancer Genome Atlas; AUC, area under ROC curve; ROC, receiver operating characteristic; BCR, biochemical recurrence.

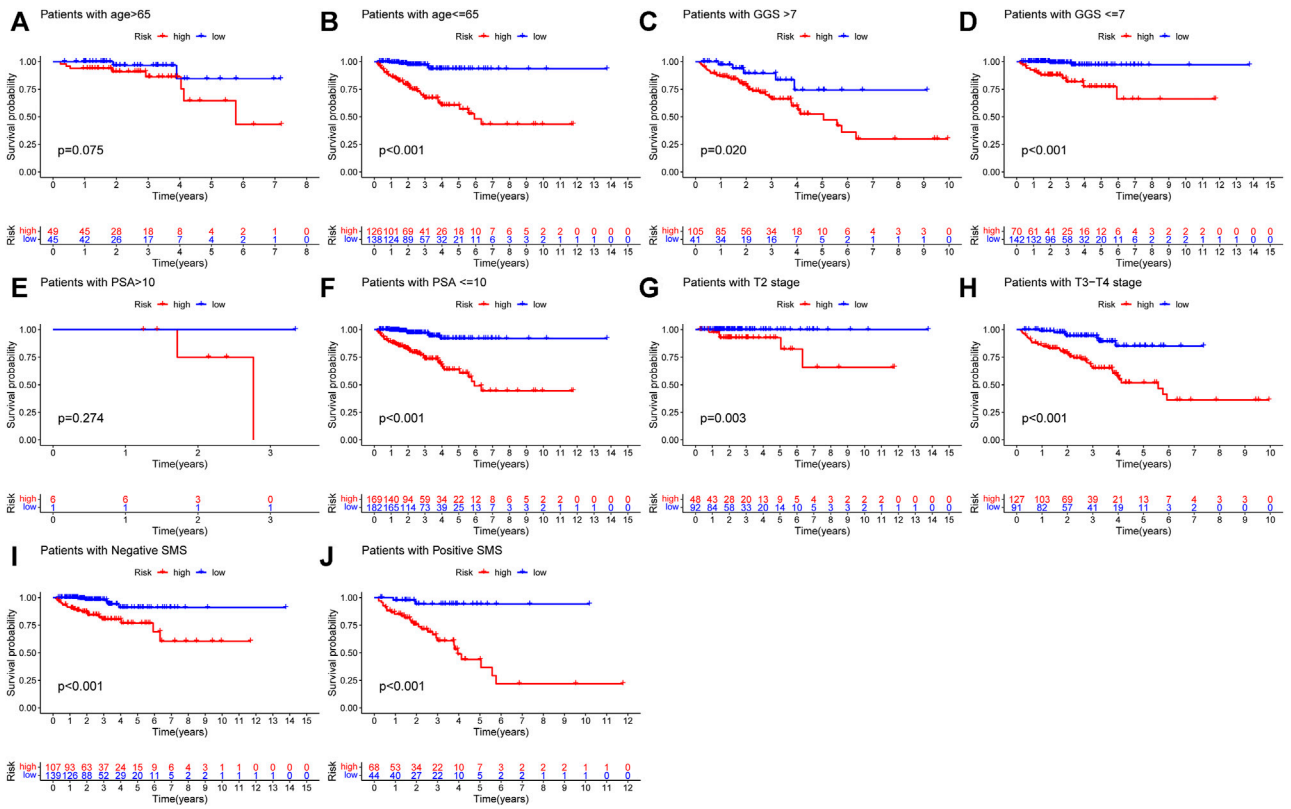

FIGURE 5 | Kaplan-Meier curve analysis of PCa patients in different clinicopathological stratifications in TCGA dataset. **(A)** Age>65. **(B)** Age ≤ 65. **(C)** Gleason score>7. **(D)** Gleason score ≤ 7. **(E)** PSA value >10. **(F)** PSA value <=10. **(G)** Pathological T stage: T2. **(H)** Pathological T stage: T3-T4. **(I)** Negative SMS. **(J)** Positive SMS. PCa, prostate cancer; PSA, prostate specific antigen; SMS, surgical margin status.

validation set to verify the stability and accuracy of URGs signature. p values < 0.05 were considered as statistical significance.

Correlation Between Prognostic Signature and Clinicopathological Parameters

To investigate the clinical significance of the URGs signature, the patients from TCGA were stratified by clinicopathological parameters containing age, pathological T stage (pT), Gleason grade score (GGS), surgical margin status (SMS), and prostate specific antigen (PSA). Kaplan-Meier curve analysis was used to investigate the prognostic value of the signature in different subgroups. In addition, we assessed the URGs signature risk score distribution according to different clinicopathological variables. p values < 0.05 were considered as statistical significance.

Construction and Validation of a Nomogram

Univariate Cox analysis and multivariate Cox analysis were performed to identify independent prognostic parameters combined the URGs signature and clinicopathological parameters in TCGA dataset. Next, a prognostic nomogram was constructed to predict 1-, 3-, and 5-year BCR-free survival for PCa patients. Calibration plots were used to evaluate the reliability of the nomogram. Then, Kaplan-Meier curve analysis, the AUC under ROC curve, and C index (using the "survcomp" package) were performed to evaluate the performance of the nomogram. p values < 0.05 were considered as statistical significance.

RESULTS

Construction and Validation of URGs Prognostic Signature

The procedure of this study is listed in **Figure 1**. BCR information and transcriptome data of 427 patients were obtained from the TCGA database. The univariate Cox regression analysis found that the expression of 236 URGs were significantly correlated with BCR prognosis of PCa patients (p < 0.05; **Supplementary Table S1**). LASSO Cox regression analysis was then applied for further analysis, and nine URGs were identified (**Figure 2**). Subsequently, a six URGs based prognostic signature was established by performing multivariate Cox regression analysis (**Figure 3**). The risk scores were calculated by following formula: *Risk score = (1.3876×ARIH2exp)+(0.7596×FBXO6exp)+(0.5102×GNB4exp)+(1.5888×HECW2exp)+(1.5015×LZTR1exp)+(−2.0379×RNF185exp).*

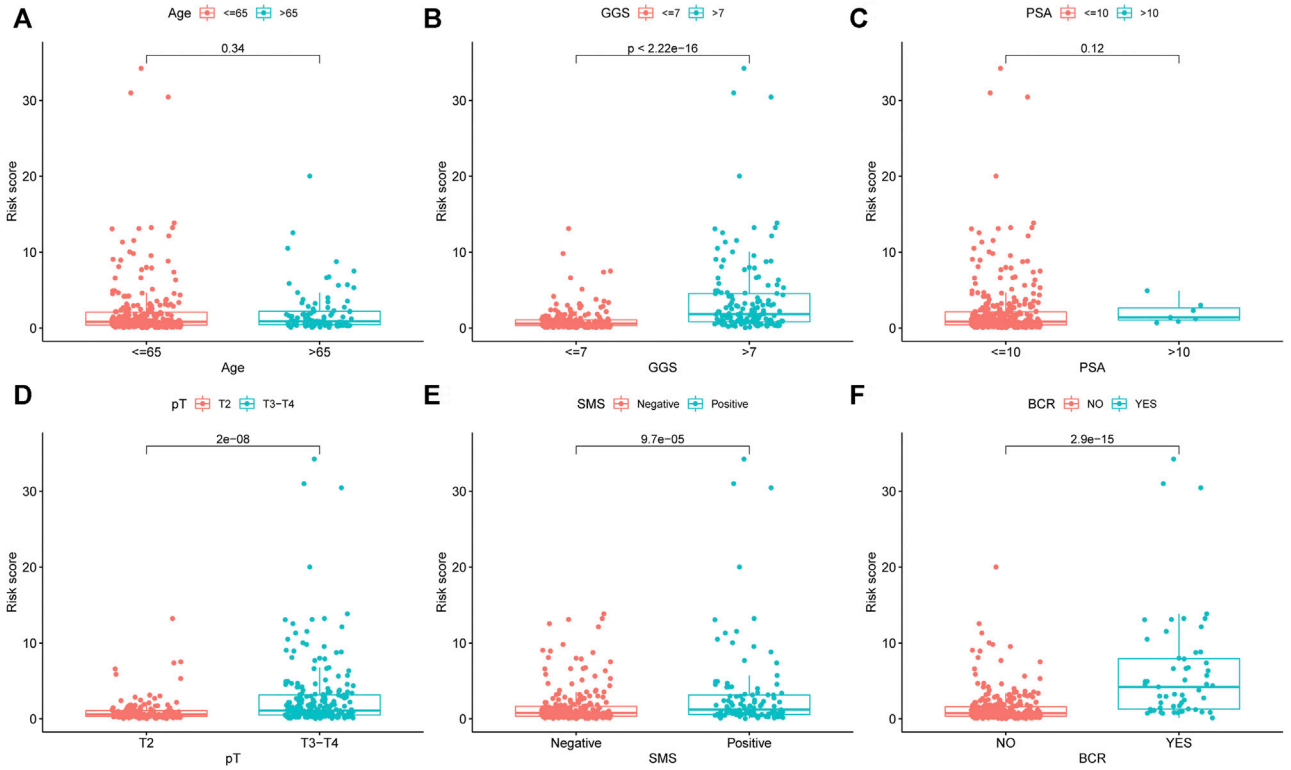

FIGURE 6 | The differential distribution of URGs signature risk scores between different clinicopathological variables and survival status in the TCGA dataset. **(A)** Age. **(B)** GGS. **(C)** PSA. **(D)** pT. **(E)** SMS. **(F)** BCR. URGs: ubiquitin related genes; GGS, Gleason grade score; PSA, prostate specific antigen; pT, pathological T stage; SMS, surgical margin status; BCR, biochemical recurrence.

Based on the URGs signature, patients were divided into high-risk and low-risk subgroups according to the same cut-off value in TCGA dataset and GSE21034 dataset. In the TCGA dataset, Kaplan-Meier curve analysis showed that the patients in the high-risk group had a poorer BCR-free survival prognosis than those in low-risk group (**Figure 4A**). AUC values of different time point were estimated and the results showed that the AUC values were 0.850 at first year, 0.778 at second year, 0.843 at third year, 0.776 at fourth year and 0.778 at fifth year. It indicated that this signature had a good prognostic predictability (**Figure 4C**). Then, GSE21034 dataset was used to verify the performance of the URGs signature. The result of Kaplan-Meier curve analysis also revealed that the BCR-free survival prognosis of patients in high-risk group was poorer than those in low-risk group (**Figure 4B**). The AUC values were 0.751, 0.700, 0.736, 0.706 and 0.731 at first, second, third, fourth and fifth year, respectively (**Figure 4D**). The distribution of risk score, recurrence status and gene expression heat maps were showed in **Figures 4E–J**.

Association Between Prognostic Signature and Clinicopathological Parameters

In order to explore the association between the URGs signature and clinicopathological parameters, we stratified patients in TCGA dataset according to age, GGS, PSA, pT, and SMS. Then, Kaplan-Meier curve analysis was performed and the results showed that high-risk patients had poorer BCR-free survival prognosis compared to low-risk patients in all stratifications except for patients with age >65 and PSA >10 (**Figure 5**).

In addition, we compared the risk score distribution in different clinicopathological stratifications to investigate the association between prognostic signature and the tumor clinical characteristics. The results showed that the patients with higher GGS, higher PSA, higher pT, and positive SMS had higher URGs signature risk scores (**Figures 6B,D,E**). The patients with BCR also had higher risk scores (**Figure 6F**). However, no significant difference existed between the subgroups stratified by age and PSA (**Figures 6A,C**).

Construction and Validation of a Nomogram

First, univariate Cox analysis and multivariate Cox analysis were applied to evaluate the prognostic significance of the URGs signature combined with different clinicopathological parameters in the TCGA dataset (**Figures 7A,B**). Next, a nomogram was constructed to quantitatively predict the prognosis of PCa patients based on clinicopathological parameters and the URGs signature. Age was excluded because of the insignificant prognostic value (p = 0.995) in univariate Cox analysis. Then, PSA, SMS, GGS, and pT and

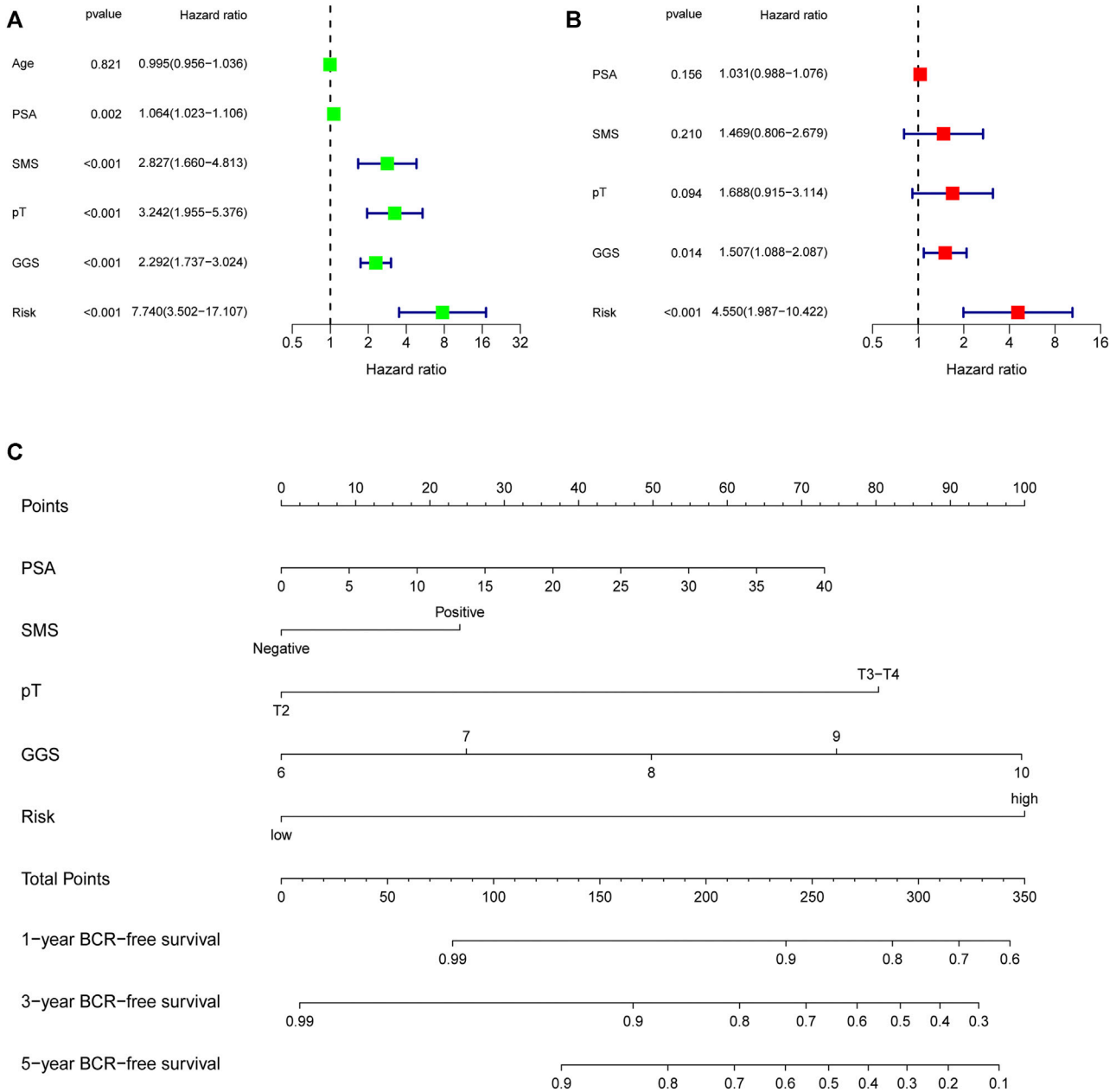

FIGURE 7 | Evaluation of the independent prognostic parameters based on clinicopathological variables and the signature, and construction of a nomogram in the TCGA dataset. **(A)** Univariate Cox analysis and **(B)** multivariate Cox analysis for evaluating independent prognostic factors. **(C)** Nomogram for predicting 1-, 3-, 5-year BCR-free survival of PCa patients. TCGA, the Cancer Genome Atlas; PSA, prostate specific antigen; SMS, surgical margin status; pT, pathological T stage; GGS, Gleason grade score; BCR, biochemical recurrence.

URGs signature were enrolled to constructed the nomogram (**Figure 7C**). The calibration curve showed good consistency between the predicted results and the actual results in both the TCGA dataset and the GSE21034 dataset (**Figures 8A–F**). Using the median risk score of TCGA nomogram model as the cut-off value, the patients were stratified to high-risk and low-risk groups in both the TCGA dataset and GSE21034 dataset. The results of Kaplan-Meier curve analysis showed that patients of high-risk group had poorer BCR-free survival prognosis in both the TCGA dataset and GSE21034 dataset (**Figures 8G,H**). In TCGA dataset, the AUC values for 1-, 2-, 3-, 4-, and 5-year BCR survival were 0.832, 0.797, 0.842, 0.828, and 0.856, respectively (**Figure 8I**) and the C index was 0.810 (95% CI: 0.766–0.853). In the GSE21034 dataset, the AUC values were 0.901, 0.899, 0.902, 0.857, and 0.816

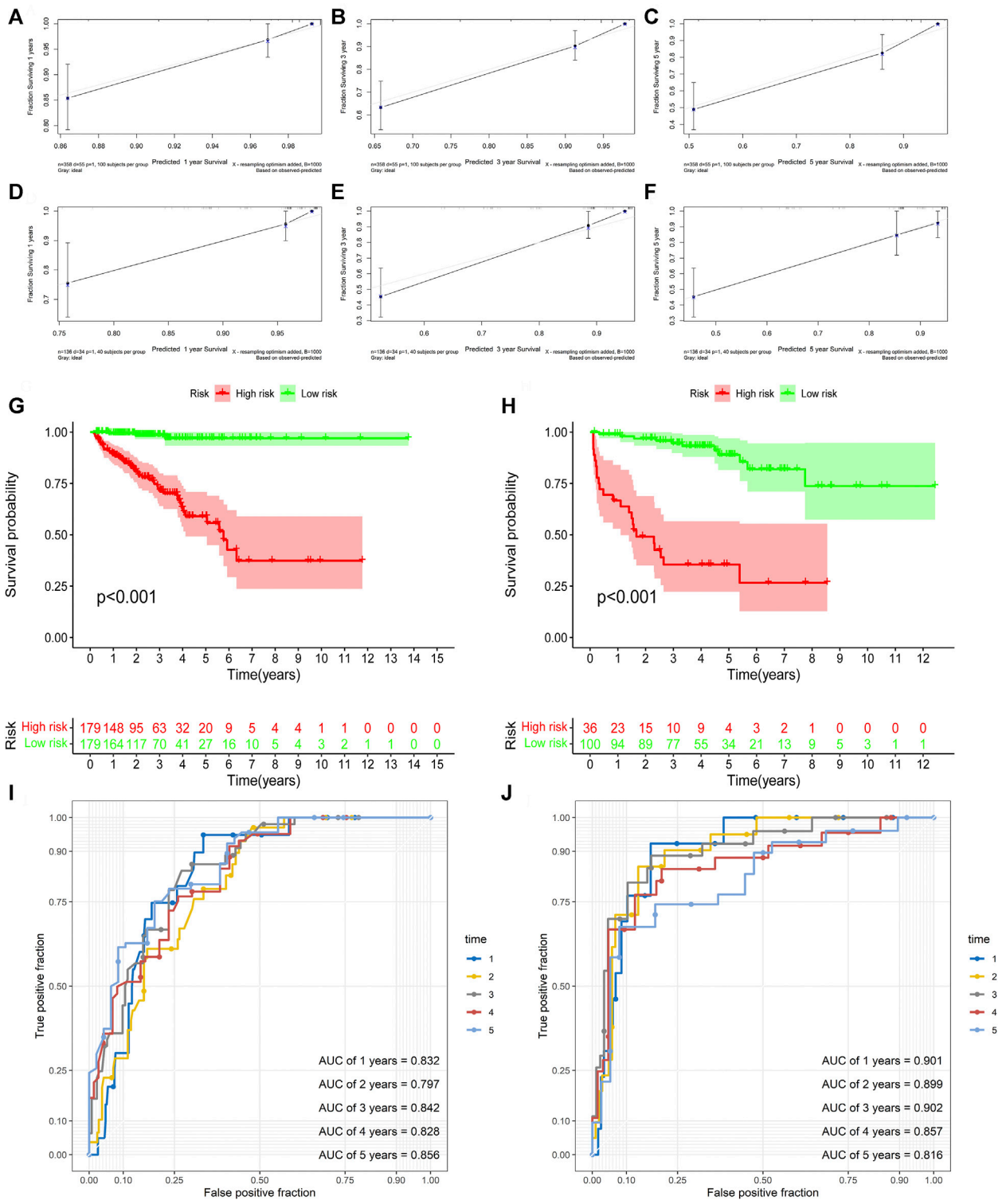

FIGURE 8 | Validation of the nomogram. **(A–C)** the calibration curve of the nomogram for predicting 1-, 3-, 5-year BCR of the PCa patients in the TCGA dataset. **(D–F)** the calibration curve of the nomogram for predicting 1-, 3-, 5-year BCR of the PCa patients in GSE21034 dataset. **(G)** Kaplan-Meier curve analysis based the nomogram in the TCGA dataset. **(H)** Kaplan-Meier curve analysis based the nomogram in GSE21034 dataset. **(I)** ROC curve analysis for predicting 1-, 2-, 3-, 4-, and5-year BCR of the PCa patients in the TCGA dataset. **(J)** ROC curve analysis for predicting 1-, 2-, 3-, 4-, and 5-year BCR of the PCa patients in GSE21034 dataset. BCR, biochemical recurrence; PCa, prostate cancer; TCGA, the Cancer Genome Atlas; AUC, area under ROC curve; ROC, receiver operating characteristic.

at 1-, 2-, 3-, 4, and 5-year, respectively (**Figure 8J**), and C index was 0.854 (95% CI: 0.799–0.910).

DISCUSSION

BCR will occur in a sizeable proportion of patients with localized PCa after radical prostatectomy (Liesenfeld et al., 2017). Early BCR is associated with high risk of recurrence and metastasis of PCa. Therefore, an effective method should be established to early predict BCR to improve the prognosis of PCa patients. It is important to identify the patients with high-risk of BCR.

In our study, we constructed a gene signature based on URGs. These genes included ARIH2, FBXO6, GNB4, HECW2, LZTR1, and RNF185. ARIH2 is a RING-in-between-RING E3 ligase gene, its encoding protein has tumor suppressive function and also involves in the neuronal response to hypoxia (Wang et al., 2020). FBXO6 is a member of the F-box protein family which is characterized by a 40 amino acid motif, F-box. The protein encoded by this gene may participate in the regulation of the cell cycle (Wu et al., 2017). FBXO6 protein can promote the growth and proliferation of gastric cancer cells and normal gastric cells (Zhang et al., 2009). Guanine nucleotide-binding protein subunit beta-4 (GNB4) is an important component of heterotrimeric G protein, which transmits signal from G protein-coupled receptors to downstream pathways. It can participant in regulating various biological behaviors of both normal and tumor cells (Gao et al., 2020). Study has showed that GNB4 promotes the tumor progression and chemoresistance in breast cancer, and the high expression of this gene is associated with worse survival rate of colorectal cancer (Riemann et al., 2009; Wang et al., 2018). HECW2 is a HECT-type E3 ubiquitin ligase belonging to the NEDD4 family. The biological function of HECW2 protein is to regulate ubiquitination and stabilize tumor suppressor p73 (Miyazaki et al., 2003). HECW2 also functions as a mediator of proteasomal degradation of DNA damage checkpoint signaling kinase, ATR, in lamin-misexpressing cells (Lu et al., 2013). LZTR1 encodes Golgi protein belonging to the BTB-Kelch superfamily and may be participated in apoptosis and ubiquitination (Nacak et al., 2006). It is also known as a tumor suppressor and the germline and somatic mutations of this gene are associated with schwannomatosis and glioblastoma (Piotrowski et al., 2014) (Franceschi et al., 2016). RNF185 encodes an E3 ubiquitin ligase. RNF185 protein can impact the degradation of BNIP1 and Dvl2, which induce autophagy and osteogenesis, respectively (Tang et al., 2011; Zhou et al., 2014). In addition, the expression levels of RNF185 are positively associated with the lymph node and distant metastasis in renal cell carcinomas patients (de Martino et al., 2012).

According to the survival and ROC curve analysis of TCGA dataset and GSE21034 dataset, this URGs signature had a good diagnostic ability and could be used to identify the PCa patients with poor prognosis of BCR. In addition, the URGs signature could also predict the BCR-free survival of PCa patients in different clinicopathological stratifications and the signature was significantly associated with advanced clinical stage and pathological grade. Finally, a nomogram was established to provide a straightforward and convenient scoring system and help clinical decision making.

To the best of our knowledge, a prognostic model based on URGs and the associated nomogram in PCa has not been reported yet. This model had a good predictive performance and could help identify the patients with high risk of recurrence and make treatment decision. However, there are also some limitations. First, most of the patients in training and validation dataset were from North America, so it is controversial to apply this model to other ethnicities. Second, the construction and validation of the model was designed by retrospective analysis and prospective clinical study should be conducted to validate the model. Finally, the exact molecular mechanisms and biological functions of the URGs should be further investigated.

CONCLUSION

We systematically analyzed the prognostic value of URGs and constructed a prognostic model in PCa by bioinformatics techniques. This URGs signature was an independent prognostic factor for predicting the BCR-free survival of PCa patients. A nomogram combining clinicopathological parameters and this signature would be useful to identify the patients with high risk of BCR.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Conception and design: GS, YZ, JL, and XL; Data curation and methodology: GS, HL, ZL, and WS; Analysis and interpretation of data: RL, CW, TW, and JL; Writing of the manuscript: GS and YZ; Review of the manuscript: JL and XL; Study supervision: JL and XL.

FUNDING

This study was funded by grants from National Natural Science Foundation of China (Grant Number: 82072838), Tongji Outstanding Young Researcher Funding (Grant number: 2020YQ13), and Huazhong University of Science and Technology (Grant Number: 2019kfyXKJC06).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.778503/full#supplementary-material

REFERENCES

Buchberger, A. (2002). From UBA to UBX: New Words in the Ubiquitin Vocabulary. *Trends Cell Biology* 12 (5), 216–221. doi:10.1016/s0962-8924(02)02269-9

de Martino, M., Klatte, T., Haitel, A., and Marberger, M. (2012). Serum Cell-free DNA in Renal Cell Carcinoma. *Cancer* 118 (1), 82–90. doi:10.1002/cncr.26254

Franceschi, S., Lessi, F., Aretini, P., Mazzanti, C. M., Menicagli, M., La Ferla, M., et al. (2016). Molecular Portrait of a Rare Case of Metastatic Glioblastoma: Somatic and Germline Mutations Using Whole-Exome Sequencing. *Neuro Oncol.* 18 (2), 298–300. doi:10.1093/neuonc/nov314

Freedland, S. J., Humphreys, E. B., Mangold, L. A., Eisenberger, M., Dorey, F. J., Walsh, P. C., et al. (2005). Risk of Prostate Cancer-specific Mortality Following Biochemical Recurrence After Radical Prostatectomy. *Jama* 294 (4), 433–439. doi:10.1001/jama.294.4.433

Gao, J., Pan, H., Zhu, Z., Yu, T., Huang, B., and Zhou, Y. (2020). Guanine Nucleotide-Binding Protein Subunit Beta-4 Promotes Gastric Cancer Progression via Activating Erk1/2. *Acta Biochim. Biophys. Sinica* 52 (9), 975–987. doi:10.1093/abbs/gmaa084

Heride, C., Urbé, S., and Clague, M. J. (2014). Ubiquitin Code Assembly and Disassembly. *Curr. Biol.* 24 (6), R215–R220. doi:10.1016/j.cub.2014.02.002

Liesenfeld, L., Kron, M., Gschwend, J. E., and Herkommer, K. (2017). Prognostic Factors for Biochemical Recurrence More Than 10 Years After Radical Prostatectomy. *J. Urol.* 197 (1), 143–148. doi:10.1016/j.juro.2016.07.004

Lu, L., Hu, S., Wei, R., Qiu, X., Lu, K., Fu, Y., et al. (2013). The HECT Type Ubiquitin Ligase NEDL2 Is Degraded by Anaphase-Promoting Complex/cyclosome (APC/C)-Cdh1, and its Tight Regulation Maintains the Metaphase to Anaphase Transition. *J. Biol. Chem.* 288 (50), 35637–35650. doi:10.1074/jbc.M113.472076

Mateo, J., Fizazi, K., Gillessen, S., Heidenreich, A., Perez-Lopez, R., Oyen, W. J. G., et al. (2019). Managing Nonmetastatic Castration-Resistant Prostate Cancer. *Eur. Urol.* 75 (2), 285–293. doi:10.1016/j.eururo.2018.07.035

Miyazaki, K., Ozaki, T., Kato, C., Hanamoto, T., Fujita, T., Irino, S., et al. (2003). A Novel HECT-type E3 Ubiquitin Ligase, NEDL2, Stabilizes P73 and Enhances its Transcriptional Activity. *Biochem. biophysical Res. Commun.* 308 (1), 106–113. doi:10.1016/s0006-291x(03)01347-0

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nat. Methods* 5 (7), 621–628. doi:10.1038/nmeth.1226

Nacak, T. G., Leptien, K., Fellner, D., Augustin, H. G., and Kroll, J. (2006). The BTB-Kelch Protein LZTR-1 Is a Novel Golgi Protein that Is Degraded upon Induction of Apoptosis. *J. Biol. Chem.* 281 (8), 5065–5071. doi:10.1074/jbc.M509073200

Pickart, C. M. (2001). Mechanisms Underlying Ubiquitination. *Annu. Rev. Biochem.* 70, 503–533. doi:10.1146/annurev.biochem.70.1.503

Piotrowski, A., Xie, J., Liu, Y. F., Poplawski, A. B., Gomes, A. R., Madanecki, P., et al. (2014). Germline Loss-Of-Function Mutations in LZTR1 Predispose to an Inherited Disorder of Multiple Schwannomas. *Nat. Genet.* 46 (2), 182–187. doi:10.1038/ng.2855

Popovic, D., Vucic, D., and Dikic, I. (2014). Ubiquitination in Disease Pathogenesis and Treatment. *Nat. Med.* 20 (11), 1242–1253. doi:10.1038/nm.3739

Riemann, K., Struwe, H., Alakus, H., Obermaier, B., Schmitz, K. J., Schmid, K. W., et al. (2009). Association of GNB4 Intron-1 Haplotypes with Survival in Patients with UICC Stage III and IV Colorectal Carcinoma. *Anticancer Res.* 29 (4), 1271–1274.

Schatten, H. (2018). Brief Overview of Prostate Cancer Statistics, Grading, Diagnosis and Treatment Strategies. *Adv. Exp. Med. Biol.* 1095, 1–14. doi:10.1007/978-3-319-95693-0_1

Swatek, K. N., and Komander, D. (2016). Ubiquitin Modifications. *Cell Res* 26 (4), 399–422. doi:10.1038/cr.2016.39

Tang, F., Wang, B., Li, N., Wu, Y., Jia, J., Suo, T., et al. (2011). RNF185, A Novel Mitochondrial Ubiquitin E3 Ligase, Regulates Autophagy Through Interaction with BNIP1. *PloS one* 6 (9), e24367. doi:10.1371/journal.pone.0024367

Wang, B., Li, D., Rodriguez-Juarez, R., Farfus, A., Storozynsky, Q., Malach, M., et al. (2018). A Suppressive Role of Guanine Nucleotide-Binding Protein Subunit Beta-4 Inhibited by DNA Methylation in the Growth of Anti-estrogen Resistant Breast Cancer Cells. *BMC cancer* 18 (1), 817. doi:10.1186/s12885-018-4711-0

Wang, P., Dai, X., Jiang, W., Li, Y., and Wei, W. (2020). RBR E3 Ubiquitin Ligases in Tumorigenesis. *Semin. Cancer Biol.* 67 (Pt 2), 131–144. doi:10.1016/j.semcancer.2020.05.002

Wu, J., Chen, Z.-P., Shang, A.-Q., Wang, W.-W., Chen, Z.-N., Tao, Y.-J., et al. (2017). Systemic Bioinformatics Analysis of Recurrent Aphthous Stomatitis Gene Expression Profiles. *Oncotarget* 8 (67), 111064–111072. doi:10.18632/oncotarget.22347

Zhang, L., Hou, Y., Wang, M., Wu, B., and Li, N. (2009). A Study on the Functions of Ubiquitin Metabolic System Related Gene FBG2 in Gastric Cancer Cell Line. *J. Exp. Clin. Cancer Res.* 28 (1), 78. doi:10.1186/1756-9966-28-78

Zhou, Y., Shang, H., Zhang, C., Liu, Y., Zhao, Y., Shuang, F., et al. (2014). The E3 Ligase RNF185 Negatively Regulates Osteogenic Differentiation by Targeting Dvl2 for Degradation. *Biochem. biophysical Res. Commun.* 447 (3), 431–436. doi:10.1016/j.bbrc.2014.04.005

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

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

Copyright © 2022 Song, Zhang, Li, Liu, Song, Li, Wei, Wang, Liu and Liu. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 February 2022
doi: 10.3389/fgene.2022.727307

Clinical and Biological Significance of DNA Methylation-Driven Differentially Expressed Genes in Biochemical Recurrence After Radical Prostatectomy

Chao Luo[1†], *Songzhe He*[2†], *Haibo Zhang*[1], *Shuhua He*[1], *Huan Qi*[1*] *and Anyang Wei*[1*]

[1]*Department of Urology, Nanfang Hospital, Southern Medical University, Guangzhou, China,* [2]*Department of Laboratory Medicine, Affiliated Hospital of Guilin Medical University, Guilin, China*

OPEN ACCESS

Edited by:
Renee Menezes, The Netherlands Cancer Institute (NKI), Netherlands

Reviewed by:
Dongrui Zhou, Southeast University, China
Xiaofei Yang, Xi'an Jiaotong University, China
Said El Bouhaddani, University Medical Center Utrecht, Netherlands

***Correspondence:**
Huan Qi
2668241842@qq.com
Anyang Wei
profwei@126.com

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

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

Received: *18 June 2021*
Accepted: *13 January 2022*
Published: *02 February 2022*

Citation:
Luo C, He S, Zhang H, He S, Qi H and Wei A (2022) Clinical and Biological Significance of DNA Methylation-Driven Differentially Expressed Genes in Biochemical Recurrence After Radical Prostatectomy. Front. Genet. 13:727307. doi: 10.3389/fgene.2022.727307

Background: Biochemical recurrence (BCR) after radical prostatectomy indicates poor prognosis in patients with prostate cancer (PCA). DNA methylation (DNAm) is a critical factor in tumorigenesis and has attracted attention as a biomarker for the diagnosis, treatment, and prognosis of PCA. However, the predictive value of DNAm-derived differentially expressed genes (DMGs) in PCA with BCR remains elusive.

Methods: We filtered the methylated genes and the differentially expressed genes (DGEs) for more than 1,000 clinical samples from the TCGA cohort using the chAMP and DESeq2 packages of R language, respectively. Next, we integrated the DNAm beta value and gene expression data with the Mithymix package of R language to obtain the DMGs. Then, 1,000 times Cox LASSO regression with 10-fold cross validation was performed to screen signature DMGs and establish a predictive classifier. Univariate and multivariate cox regressive analyses were used to identify the prognostic factors to build a predictive model, and its performance was measured by receiver operating characteristic, calibration curves, and Harrell's concordance index (C-index). Additionally, a GEO dataset was used to validate the prognostic classifier.

Results: One hundred DMGs were mined using the chAMP and Methymix packages of R language. Of these, seven DMGs (CCK, CD38, CYP27A1, EID3, HABP2, LRRC4, and LY6G6D) were identified to build the prognostic classifier (Classifier) through LASSO analysis. Moreover, univariate and multivariate Cox regression analysis determined that the Classifier and pathological T stage (pathological_T) were independent predictors of BCR (hazard ratio (HR 2.2), (95% CI 1.4–3.5), $p < 0.0012$, and (HR 1.8), (95% CI 1.0–3.2), $p < 0.046$). A nomogram based on the Classifier was constructed, with high prediction accuracy for BCR-free survival in TCGA and GEO datasets. GSEA enrichment analysis showed that the DMGs were mainly enriched in the metabolism pathways.

Conclusion: We identified and validated the nomogram of BCR-free survival for PCA patients, which has the potential to guide treatment decisions for patients at differing risks of BCR. Our study deepens the understanding of DMGs in the pathogenesis of PCA.

Keywords: prostate cancer, biochemical recurrence, classifier (classification tool), DNA methylation-driven genes, biomarker

INTRODUCTION

Prostate cancer (PCA) is a common cancer with the highest prevalence among men worldwide. In 2018, the global incidence of PCA was 29.3 per 100,000 (WHO, 2018). In the United States, it is estimated that more than 30,000 cases of death in men per year are attributable to PCA (Siegel et al., 2019); furthermore, there are 60.3 new cases of PCA per 100,000 and 26.6 deaths per 100,000 individuals in China (Chen et al., 2016). Radical prostatectomy (RP) is considered as an effective therapy for the treatment of localized PCA. However, up to 20–53% of patients experience biochemical recurrence (BCR) after RP (Mottet et al., 2018). BCR is defined as a serum PSA equal to or greater than 0.2 ng/ml on two consecutive occasions after surgery or radiation. However, some studies and guidelines have indicated that PSA cannot be used to predict BCR for each patient with PCA, especially when its value is very low (Eisenberg et al., 2010; Fendler et al., 2019; Wang et al., 2020). Moreover, patients with similar clinical features or PSA levels might have a different clinical endpoint. Among patients with high-risk PCA and clinical stage ≥ T3a, a biopsy Gleason score of 8–10, and/or a serum PSA level >20 ng/ml, approximately 60% had at least 15 years of metastasis-free survival after RP, indicating that not all patients had poor prognosis (Spahn et al., 2010a; Spahn et al., 2010b). The monitoring of BCR was expected to effectively prevent mortality. However, overtreatment owing to misprediction should also be avoided (Artibani et al., 2018).

Epigenetics and PCA have been studied at great length. The evolution of PCA involves a combination of epigenetic and genetic changes, and methylation is an important mechanism. The methylation of KDM1A and CHD1 genes can drive the transcription and translocation of androgen receptors (Metzger et al., 2016). PCA recurrence can lead to many molecular aberrations, including DNA methylation (DNAm), which can be used as biomarkers of PCA prognosis (Fraser et al., 2017). Additionally, the promoter methylation of CRMP4 in biopsied tissue can predict lymph node metastasis of PCA (Gao et al., 2017). As a critical factor in tumorigenesis, DNAm has attracted increasing attention as a biomarker for the diagnosis, treatment, and prognosis of PCA (Wei et al., 2015). CpG islands are rich in cytosine and guanine dinucleotides and are 200 bp to several kilobases in length. To better regulate highly expressed genes, CpG islands are always in close proximity to the promoters of these genes (Nowacka-Zawisza and Wiśnik, 2017). Additionally, CpG islands can modulate cancer proliferation, including that of PCA, *via* the hypomethylation of cytosines at the 5′position in CpG islands within the promoter region of oncogenes. In contrast, hypermethylation of the regulatory (promoter) region of suppressor genes leads to gene silencing (Herman and Baylin, 2003; Baylin and Ohm, 2006). Alterations of tumors at the molecular level always occur before the manifestation of clinicopathological features (Jordan et al., 2017; Devos et al., 2020). However, to date, no reliable BCR biomarkers for PCA have been identified for routine application in clinical practice.

In this study, we established a practical and reliable nomogram based on DNAm-derived differentially expressed gene (DMG) profiling from The Cancer Genome Atlas (TCGA) data to improve risk stratification for patients with PCA. Moreover, we analyzed Gene Expression Omnibus (GEO) datasets to validate the nomogram and related genes and explored the relationship between methylation status and gene expression. Our findings confirm that these DMGs might be potential therapeutic targets in the future.

MATERIALS AND METHODS

Data Collection

TCGA data (gene expression data, methylation data, and associated clinicopathological features) were downloaded from the Genomic Data Commons (GDC) Data Portal of the National Institutes of Health, and TCGA level-3 molecular data and corresponding clinical data were available through the GDC (up to 2020/4/10; **Supplementary Table S1**). The DNAm level was measured with β values ranging from 0 to 1 (the Illumina Infinium Human Methylation 450 platform of the GDC). Furthermore, the inclusion criteria for the discovery cohort (TCGA cohort) were as follows: 1) patients who had undergone RP; 2) patients with associated clinicopathological features, such as BCR time, BCR status, residual tumor data, TNM stage, lymph node number, pathologic Gleason Score, target therapy, radiotherapy, and laterality; and 3) clinical results assessed using BCR time. For the non-BCR samples without BCR time, their last follow-up time was used for further study.

Identification of Differentially Expressed Genes

After downloading raw RNA-sequencing datasets of TCGA prostate adenocarcinoma (PRAD) cohorts (HTSeq-Counts of TCGA-PRAD transcriptome profiling) and deleting the duplicated samples, we extracted DEGs between 474 PCA and 53 nontumorous tissues using the "DESeq2", package (Love et al., 2014). Here, for multiple probes, the average value corresponding to the same gene is taken during the calculation. An absolute logFC >1 and false discovery rate (FDR) < 0.05 were set as the cut-off values. The results were visualized using the R language package "ggplot2."

Filtering and Cleaning of Methylation Data

The DNAm data contained the 499 PCA and 50 nontumorous tissues. The data were filtered using the chAMP package of R language (Phipson et al., 2016) according to the following criteria: 1) filter out probes with a *p*-value greater than 0.01; 2) filter out

probes with a bead count less than 3 in at least 5% of the samples; and 3) filter out probes at non-CpG sites; 4) filter all SNP-related probes (R code: **Supplementary File S1**). In case of multiple CpG sites being annotated by one methylated gene, we could calculate their average value using the "aggregate function"of R language. The CpGs annotation file was obtained from the TCGA dataset.

Identification of DMGs

DMGs were identified by integrating the methylated genes and DEGs with the "MethylMix" package. A new version of MethylMix was developed to automatically preprocess the databases of methylation-driven genes and subsequently analyze their transcriptionally predictive methylation states by applying the MethylMix algorithm (Gevaert, 2015). First, a correlation analysis was performed between the gene expression data of DEGs in the PCA samples and their corresponding methylation data. The target genes with a correlation coefficient < −0.3 and p-value <0.05 were used for subsequent analysis. Second, beta mixture models were used to determine the methylation status of multiple genes. Last, to verify the existence of difference between the PCA samples and the corresponding non-tumor samples, the Wilcoxon rank-sum test was used as the measurement standard. Finally, mixture models and regression analyses of the DMGs were visualized, respectively.

Functional and Pathway Enrichment Analysis of DMGs

Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Set Enrichment Analysis (GSEA) were used to explore the critical pathways associated with DMGs, which were performed using the R packages "org.Hs.eg.db" and "clusterProfiler."

Generation and Validation of the DMG-Based Classifier of BCR-Free Survival

To explore the relationship between the gene expression of DMGs and BCR-free survival, least absolute shrinkage and selector operation (LASSO) regression was performed to identify prognosis-related DMGs and establish a signature. Briefly, LASSO is a method that pushes regression coefficients toward zero *via* the application of an L1 penalty. If the penalty is larger, fewer predictors are selected, and as a result, several variables are diminished. In addition, analysis using the "glmnet" package based on the program with 1,000 iterations of Cox LASSO regression and 10-fold cross-validation led to seed genes being integrated into multiple gene sets. Seed genes with nonzero coefficients were identified as potential prognostic predictors. The linear combination of the regression coefficient (β) multiplied by its mRNA expression level can generate a risk score for candidate genes based on BCR (Tibshirani, 1997; Sauerbrei et al., 2007) as follows:

$$Riskscore = \sum_{i=1}^{k} \beta i Si$$

(k: the number of candidate genes, βi: the coefficient index of candidate genes, and Si: the expression level of candidate genes).

To classify patients into low-, medium-, and high-risk groups, the x-tile (Version 3.6.1) tool was used to determine the cut-off value of the risk score (Camp et al., 2004). Kaplan–Meier (K-M) survival plots and log-rank test were used to estimate BCR-free survival differences. To assess the effectiveness of the Classifier, the area under the curve (AUC) of the time-dependent receiver operating characteristic (ROC) curve was assessed. In this study, the predictive property was evaluated based on the time-dependent ROC curves, which were generated using the "survivalROC" and "rms" R package. In addition, the "ggplot2" R package was used for drawing.

Screening of Prognostic Factors

To identify the meaningful predictive factors of a BCR-free state for PCA patients, univariate Cox regression analysis was performed with the Classifier (risk level) and clinicopathological features of patients. Additionally, multivariate Cox regression with 1000-times bootstrapping was performed using the "survival" package in R to eliminate confounding factors. The hazard ratio and its 95% CI for each variate were obtained. Statistical significance was set at a p-value < 0.05.

Establishment and Validation of the Nomogram

The nomogram was constructed with meaningful predictive factors by multivariate Cox regression analysis. The calibration curves were plotted using the Hosmer–Lemeshow test, which was expected to calibrate the probability of patients with PCA after RP at 1, 3, and 5 years. Furthermore, the identification performance of the nomogram was quantified using Harrell's concordance index (C-index). In total, 1,000 bootstrap resamples were processed for verification to obtain a stable C-index. The C-index ranged from 0.5 (indicative of poor or no predictive ability) to 1.0 (perfect predictive ability). A time-dependent ROC analysis (Heagerty et al., 2000) and area AUC were used to measure the predictive accuracy of the nomogram.

External Validation of the Nomogram

The gene expression dataset (GSE21034), as a validation cohort, was downloaded from the GEO cohort (https://www.ncbi.nlm.nih.gov/geo/). The GSE21034 microarray dataset included gene expression profiles of 140 PCA samples and 29 nontumor samples as well as the related 140 clinicopathological features (Taylor et al., 2010) (GPL5188: Affymetrix Human Exon 1.0 ST Array). As previously mentioned, patients were classified into low-, medium-, and high-risk groups according to the cut-off value of the risk score determined using x-tile. K-M survival plots and log-rank test were used to evaluate the BCR-free survival differences. The time-dependent ROC analysis and AUC were used to measure the predictive accuracy of the nomogram, and the accuracy, sensitivity, and specificity of the model were quantitatively evaluated.

Copy Number Variation, Mutation Features, and GSEA of the Candidate Genes

We collected graphic illustrations for CNVs and the seven-gene mutation profiles of all PCA tissues in the TCGA dataset, searching from the cBioPortal website (http://www.cbioportal.org/). Perl (strawberry-Perl-5.30.2.1) and GSEA 3.0 software (Gene sets database: c2. cp.kegg.v7.2. symbols.gmt) were used to perform GSEA analysis. Differences were considered statistically significant at an FDR <0.05.

Cell Culture and DAC Treatment

The PCA cell line lymph node carcinoma of the prostate (LnCap) was purchased from Jining company (Shanghai, China) and maintained in minimum essential medium (cat no. C11875500BT; Gibco, Grand Island, NY, United States at 37°C and supplemented with 10% fetal bovine serum (cat no. A31608-02, Gibco) in a humidified atmosphere containing 5% CO_2. LnCap cells in culture were treated with 5-aza-2′-deoxycytidine (DAC, Cat No. A3656-5MG; Sigma-Aldrich, St. Louis, MO, United States) for 120 h, and the medium was replaced daily owing to DAC instability. For experiments involving DAC treatment, dimethyl sulfoxide was used as the control. The cells were harvested for extraction of genomic DNA and total RNA for analysis of DNAm and gene expression.

RNA Extraction and Quantitative Reverse-Transcription PCR (qRT-PCR)

RNA extraction and qRT-PCR were performed using AG RNAex Pro Reagent (AG21101, Accurate Biology, Changsha, China). The samples were treated with 20% chloroform, vortexed briefly, and incubated at room temperature for 15 min. The samples were then centrifuged at high speed for 15 min at 4°C after the aqueous phase was transferred to a new tube, and an equal volume of isopropanol was added. Samples were incubated at room temperature for 10 min, followed by centrifugation at high speed for 10 min at 4°C. The pellets were then washed in 95% ethanol, dried, and resuspended in nuclease-free water. cDNA was synthesized using RNAiso plus reagent (Takara, Tokyo, Japan) according to the manufacturer's instructions. qRT-PCR was performed using a LightCycler® 480 II (Roche, Basel, Switzerland) with a SYBR Green PCR kit (Takara Bio). The primer sequences are listed in **Supplementary Table S6**.

Cancer Cell Line Encyclopedia Database

Gene expression of PCA cell lines was obtained from CCLE. We downloaded CCLE from the GEO dataset (Barretina et al., 2012). The gene expression profile GSE36133 (Affymetrix GPL15308 platform, Affymetrix Human Genome U133 plus 2.0 Array) was obtained. The probes were converted into the corresponding gene symbol according to the annotation information of the GPL571 platform. Genes with more than one probe set were averaged using R language.

Statistical Analysis

The gene expression data of the seven DMGs were normalized using the TMM methods implemented in the package "edgeR." The statistical analyses of qRT-PCR data were performed using R language (version 4.0.0) and GraphPad Prism 8.3.0. A p-value < 0.05 was considered statistically significant for two-sided tests.

RESULTS

Identification of DEGs

A flow diagram of the entire process is shown in **Figure 1**. By comparing the mRNA expression between PCA tissues and nontumorous prostate tissues, we identified 3,023 DEGs for further analysis. Among these DEGs, 1,262 were upregulated and 1761 were downregulated (**Supplementary Table S2**).

Identification of DMGs

After identifying 9,574 methylated genes, we evaluated the level of methylation and gene expression level of 1,285 methylated genes from 397 PCA samples and the methylation level of these 1,285 methylation-associated genes from 49 non-tumor samples by integrating the datasets. The MethylMix (Gevaert, 2015) package was used to import these three datasets. Altogether, 100 DMGs were identified (**Supplementary Table S3**). Heatmap was used to show the gene expression (**Figure 2A**) of these 100 DMGs and took seven represented genes in black frames as an example. GO analyses were performed to elucidate the functional properties of the newly identified DMGs, and eight GO terms were obtained (**Figure 2B**), including the organic acid biosynthetic process, benzene-containing compound metabolic process, and cellular modified amino acid metabolic process (p < 0.001). Moreover, pathway analysis using the KEGG revealed that these genes were mainly enriched in glutathione metabolism, drug metabolism -cytochrome P450, platinum drug resistance, and the PPAR signaling pathway (p < 0.05; **Figure 2C**). KEGG pathway analysis revealed that the most abundant pathways were those related to metabolism and drug resistance. GSEA revealed that these genes were enriched in metabolism (**Figure 2D**).

Establishment of a Classifier Related to BCR-Free Survival

These 100 DMGs with 339 PCA samples with BCR time and status were included in LASSO analysis. Of these, *CCK, CD38, CYP27A1, EID3, HABP2, LRRC4,* and *LY6G6D* were recommended as candidate genes (**Figure 3A**). The methylation status of these seven genes was negatively correlated with gene expression (**Figure3B**). Among them, *CCK, CD38, CYP27A1, EID3, LRRC4,* and *LY6G6D* were hypermethylated, whereas *HABP2* was hypomethylated (**Figure 3C**, **Supplementary Figure S3**). Based on the seven genes, a formula for calculating the risk score was generated as follows:

Risk score = −0.066 × CCK mRNA level + (−0.127) × CD38 mRNA level + (−0.0615) × CYP27A1 mRNA level + (−0.833) × EID3 mRNA level + 0.088 × HABP2 mRNA level + 0.473 × LRRC4 mRNA level + (−0.122) × LY6G6D mRNA level.

Please note that the gene expression should be normalized before importing the formula (**Supplementary File S1**).

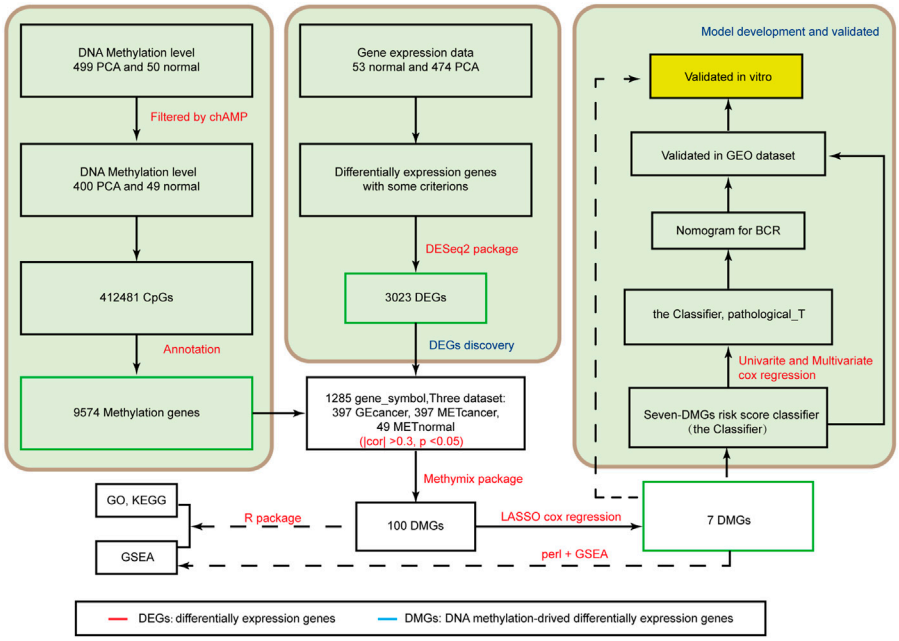

FIGURE 1 | Analysis of the flowchart illustrates the exploration procedure for the PCA prognostic DMGs and establishment of risk score signature.

The range of the risk scores among the 334 patients in the TCGA dataset was between −4.588 and −1.81 (**Supplementary Table S4**). However, by analyzing the risk scores of the Classifier and BCR status, the patients with PCA could be classified into low-, medium-, and high-risk groups with the cut-off value from x-tile. In total, 191 patients with a cut-off value greater than −2.89 were included in the high-risk group, 88 patients with values between −3.33 and −2.89 were included in the medium-risk group, and 55 others were included in the low-risk group. K-M analyses of these three groups demonstrated that patients with lower risk scores had a lesser occurrence of BCR than those with medium-risk scores, which in turn had an even lower occurrence than those with high-risk scores ($p < 0.0001$; **Figure 3D**). The heatmap in **Figure 3E** shows the gene expression of the seven candidate genes based on the risk level. A time-dependent ROC curve was generated to describe the predictive ability of the Classifier, and the AUC values of the Classifier at 1, 3, and 5 years were 0.8243, 0.7878, and 0.7704, respectively (**Figure 3F**). As here, the same data were used to select genes and build the risk score, an association with BCR and the resulting predictive ability were to be expected.

Establishment and Evaluation of the Nomogram for BCR-Free Survival Prediction in PCA

The prognostic classifier (Classifier) and pathological_T were regarded as the key prognostic predictors using univariate and multivariate regression analyses (**Figure 4A**). Furthermore, the relationship between Schoenfeld model residuals and the Classifier was plotted to evaluate the importance of these prediction factors in the combined model. Schoenfeld residuals showed that the combined model satisfied the risk assumption of an equal proportion (**Figure 4B**). A nomogram was established based on the Classifier and pathological_T (**Figure 4C**). Based on the combined model, patients were divided into low-, medium-, and high-risk groups with the risk score from x-tile as the cut-off value (1.21 and 4.09). Patients with the lowest risk scores had the lowest BCR rates and those with the highest risk scores had the highest BCR rates when the K-M survival analysis was applied ($p < 0.0001$; **Figure 4D**). The C index and robust C- index values were 0.802 and 0.810, respectively, which means that the predicted results of the model were nearly consistent with the actual observed results. The calibration curve of the combined model for predicting BCR-free survival at 1, 3, and 5 years revealed favorable forecasting performance (**Figure 4E**). Additionally, the time-dependent ROC curve demonstrated that the AUC of the seven-DMG signature combined with the Classifier and pathological_T was significantly higher than that of the Classifier or Gleason score only at 1, 3, and 5 years (**Figure 4F**), indicating that the sensitivity of the nomogram was considerably better than that of the Classifier or the Gleason score alone. The nomogram offered excellent performance in BCR-free survival predictions, especially with a long term. Taken together, the findings suggest that the nomogram can help physicians provide appropriate recommendations for clinical therapy and follow-up schedules for patients with PCA.

External Validation of the Nomogram

The GEO dataset GSE21034 was subsequently used to verify the newly established nomogram. In total, 140 cases were included in the external study (**Supplementary Table S5**). Based on the Classifier,

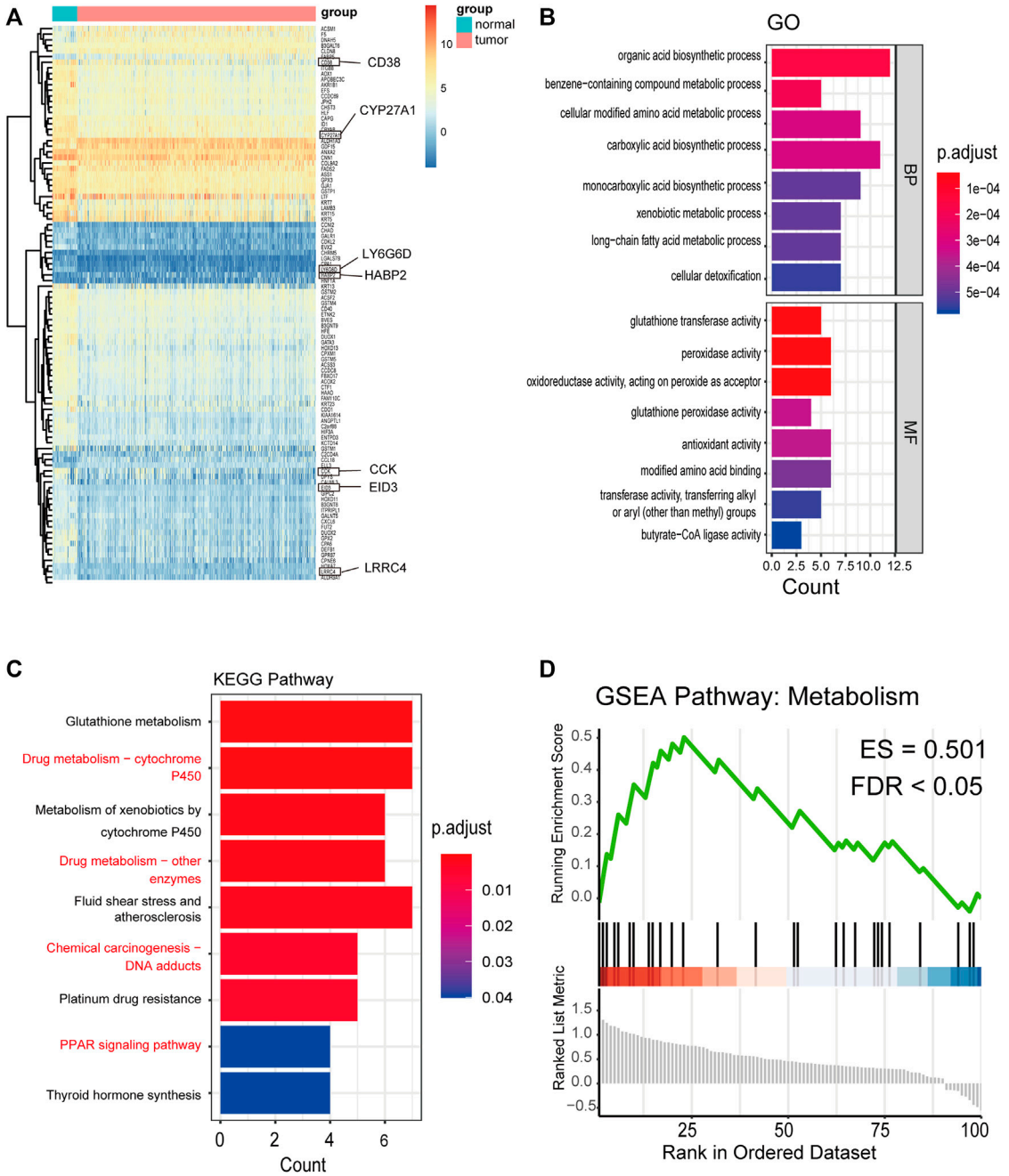

FIGURE 2 | Distribution of the methylation level and gene expression of DNA methylation-driven genes and GO, KEGG, and GSEA pathways of 100 DMGs. **(A)** Distribution of the gene expression of 100 DMGs between PCA and nontumorous prostate tissues (seven representative genes are shown in the black frame). **(B)** GO analysis classified the DEGs into 2 groups (i.e., molecular function and biological process) and significant enriched GO Terms of 100 DMGs based on their functions. **(C)** KEGG pathway analysis. **(D)** GSEA KEGG pathway of 100 DMGs.

patients were divided into low-, medium-, and high-risk groups with the risk score from x-tile as the cut-off value (0.02 and 0.03). Generally, comparing the three cohorts, patients with the lowest risk scores had lowest BCR rates and those with the highest risk scores had the highest BCR rates when the K-M survival analysis was applied ($p = 0.00015$; **Figure 5A**). The AUCs of BCR-free survival at 1-, 3-, and 5-year BCR-free survival were 0.7078, 0.7544, and 0.725, respectively (**Figure 5B**). Based on the combined model of the Classifier and pathologic_T, patients were divided into low-, medium-, and high-risk groups with the risk score from x-tile as the cut-off value (0.52 and 2.12). Comparing the three cohorts, patients with the lowest risk scores had the lowest BCR rates and those with the highest risk score had the highest BCR rates when the K-M survival analysis was applied. ($p < 0.0001$; **Figure 5C**).

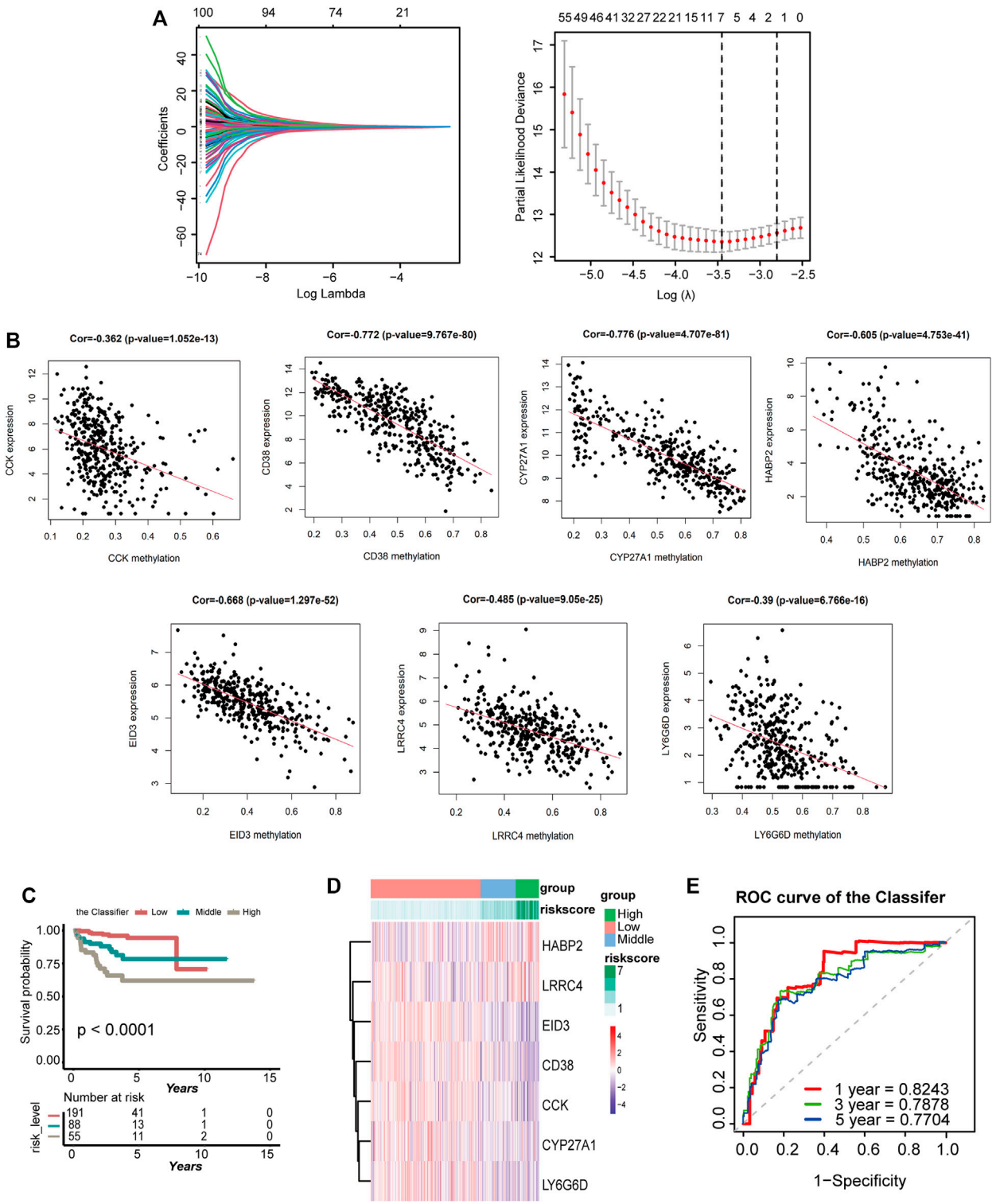

FIGURE 3 | Establishment of the classifier based on seven DMGs in the TCGA cohort. **(A)** LASSO coefficient profiles of the 100 genes in TCGA cohort. A coefficient profile plot was generated against the log(lambda) sequence. Selection of the optimal parameter (lambda) in the LASSO model for TCGA-PRAD. A vertical line is drawn at the optimal value by 1−SE standards and results in seven nonzero coefficients. **(B)** Regression analysis between gene expression and DNA methylation of seven DMGs. **(C)** K-M survival curves compare BCR status among the low-, medium-, and high-expression groups of seven DMGs. **(D)** Heatmap of the seven DMGs expression profiles based on low-, medium-, and high-risk groups. **(E)** Time-dependent ROC for accuracy of BCR-free survival prediction by the seven-DMG signature (the Classifier) among 1, 3, and 5 years in TCGA group.

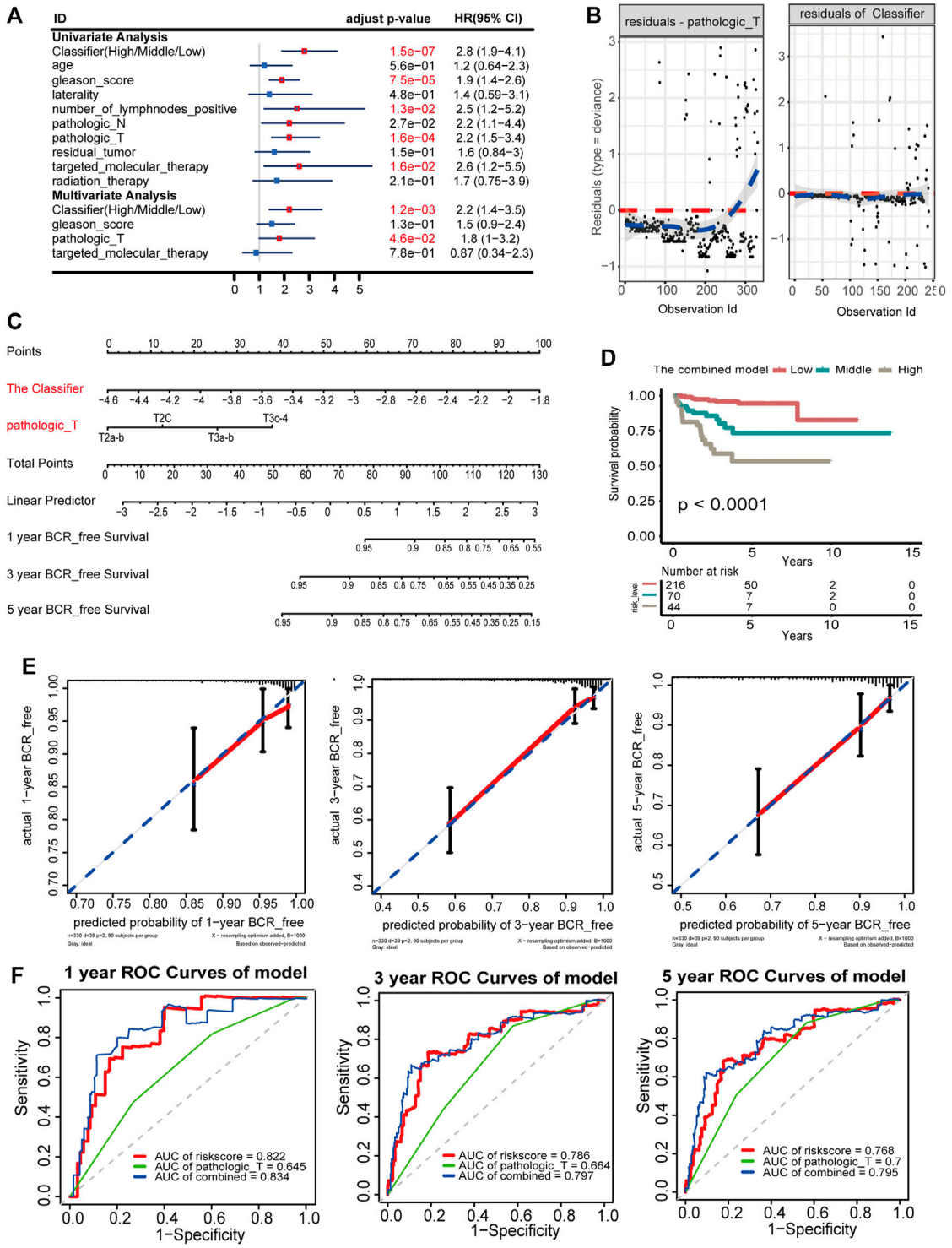

FIGURE 4 | Nomogram to predict 1-, 3-, and 5-year BCR-free survival. The BCR-free survival nomogram was established in the TCGA cohort, incorporating pathological_T and the Classifier. **(A)** Univariate and multivariate analyses of the Classifier, clinical factors, and pathological characteristics with BCR-free survival. The statistical significance is indicated using different colors; red indicates statistical significance, and blue indicates no significance. **(B)** Schoenfeld residual suggested that this model met the equally proportional risk hypothesis. Schoenfeld model residuals versus pathological_T stage and the Classifier were plotted to obtain a preliminary assessment of whether these predictive factors should be incorporated into the model. **(C)** Nomogram to predict the 1-, 3-, and 5-year BCR-free survival of PCA patients. **(D)** K-M survival curves for comparison of BCR-free survival among the low-, medium-, and high-risk groups based on the combined model in the TCGA cohort. **(E)** Calibration curves of 1-, 3-, and 5-year BCR-free survival in the combined model. Blue dotted lines represent the ideal predictive model, and the solid red line represents the observed model. **(F)** Time-dependent ROC for accuracy of BCR-free survival prediction by the combined model among 1, 3, and 5 years.

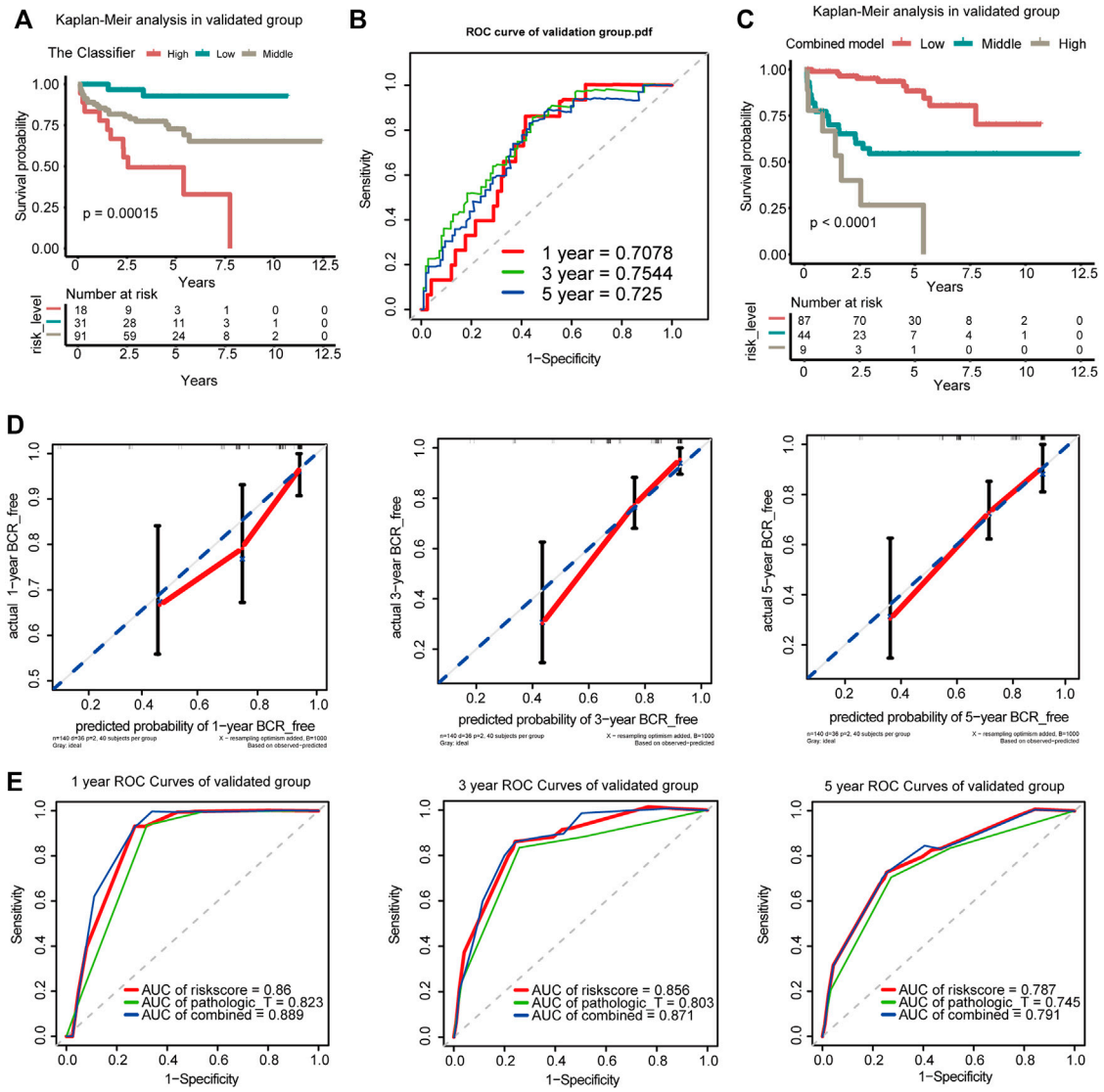

FIGURE 5 | Verification of the Classifier and the combined model in the GEO dataset. **(A)** K-M survival curves for comparison of BCR-free survival among the low-risk, medium-risk, and high-risk score based on the Classifier. **(B)** Time-dependent ROC for accuracy of BCR-free survival prediction by the seven-DMG signature among 1, 3, and 5 years in the validated group. **(C)** K-M survival curves for comparison of BCR-free survival among the low-risk, medium-risk, and high-risk score based on the combined model. **(D)** Calibration curves of 1-, 3-, and 5-year BCR-free survival. Blue dotted lines represent the ideal predictive model, and the solid red line represents the observed model. **(E)** Time-dependent ROC analysis was used to evaluate the accuracy of the BCR-free survival nomograms. The red, blue, and green solid lines represent the combined model, GS, and Classifier, respectively.

The calibration curves for 1-, 3-, and 5-year BCR-free survival status based on the nomogram suggested a significant agreement between the predicted outcomes and those observed in the validation group (**Figure 5D**). The combined model of the Classifier and pathological_T exhibited better predictive ability than either the Classifier or pathological_T alone. The AUCs of 1-, 3-, and 5-year BCR-free survival were 0.889, 0.871, and 0.791, respectively, in our validation group (**Figure 5E**). The 1-, 3-, and 5-year trend in the AUC in the validation cohort was consistent with that in the TCGA cohort, which further illustrates the value of the prediction model for long-term follow-up. Additionally, coincidence analysis of the combined model showed that the C-index was 0.853 and the robust C-index was 0.860.

CNV, Mutation Features, and KEGG Signaling Pathway Based on GSEA

The seven candidate DMGs were affected by methylation, gene amplification, deletion, and mutation. We noted that the rates of genetic alterations among these seven genes were between 0.8 and 1.6% based on the GDC TCGA-PRAD database (**Figure 6A**), indicating that the effect of methylation might promote a change in gene expression.

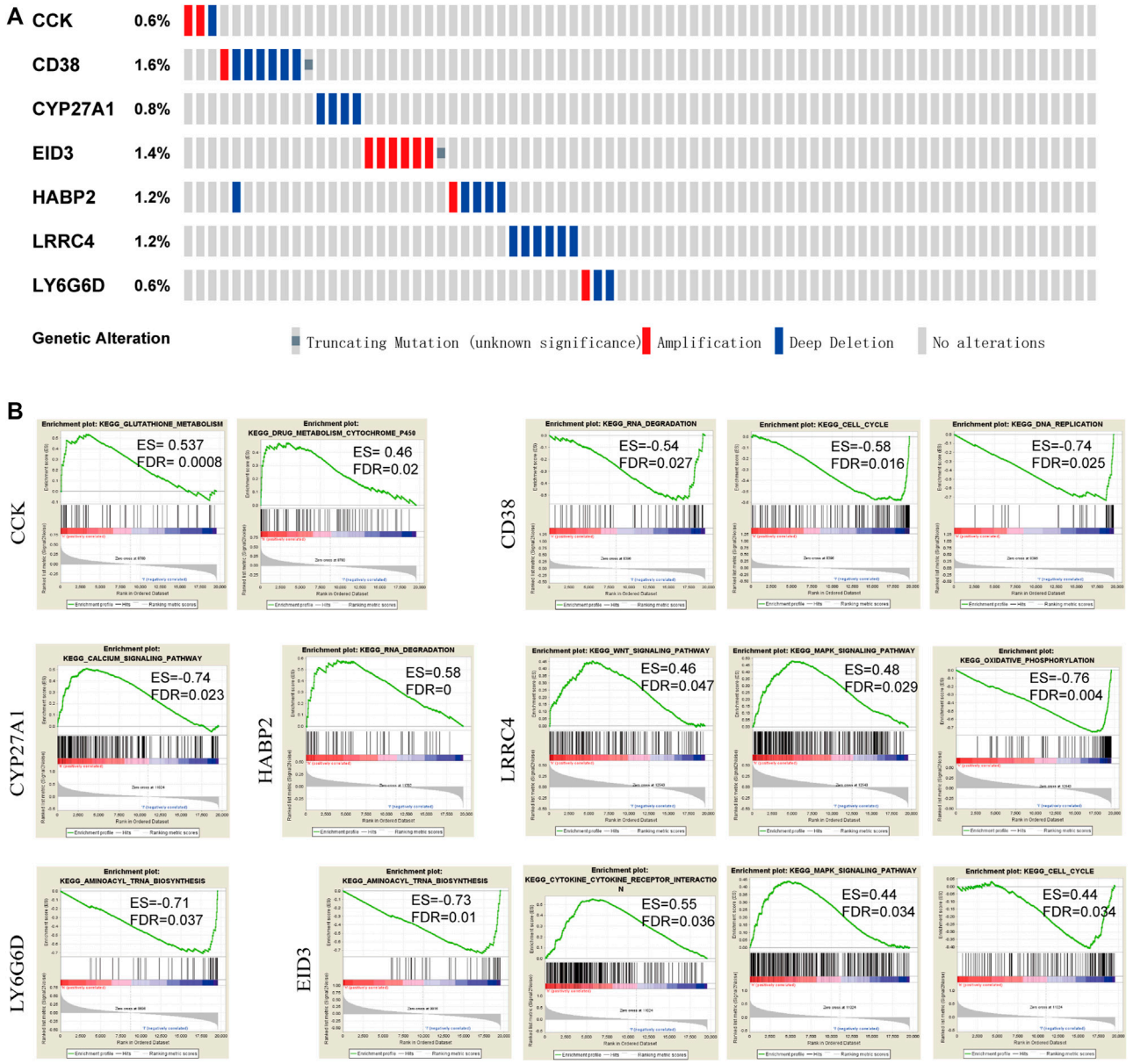

FIGURE 6 | Genetic changes and mutation features of seven DMGs by cBioPortal and pathways by GSEA. **(A)** Genetic alterations of DMGs in PCA samples. Rows and columns represent the genes and tumor samples, respectively. **(B)** KEGG pathways enrichment analysis in each of the seven genes based on GSEA (FDR <0.05).

To determine the potential signaling pathways affecting these seven genes, functional category enrichment analysis was performed to examine their function. *CCK* was mainly related to glutathione metabolism, drug metabolism, and cytochrome P450. *CD38* was mainly associated with RNA degradation, the cell cycle, and DNA replication. *CYP27A1* was mainly associated with the calcium signaling pathway. *HABP2* was mainly associated with RNA degradation. *LRRC4* was mainly associated with the WNT signaling pathway, the MAPK signaling pathway, and oxidative phosphorylation, and *LY6G6D* was mainly associated with aminoacyl tRNA biosynthesis. *EID3* was mainly associated with aminoacyl tRNA biosynthesis, cytokine receptor interaction, the MAPK signaling pathway, and the cell cycle. A NOM q-value (FDR) < 0.05 was set as the threshold value (**Figure 6B**).

Expression of Seven DMGs in DAC-Treated LnCap Cells

As shown in **Figure 3B** and **Supplementary Figure S3**, the methylation levels of *CCK, CD38, CYP27A1, EID3, HABP2, LRRC4,* and *LY6G6D* exhibited the strongest negative correlation with their gene expression, respectively. To confirm this, we analyzed the changes in the expression of the four genes in DAC-treated LNCaP cells to evaluate their functional

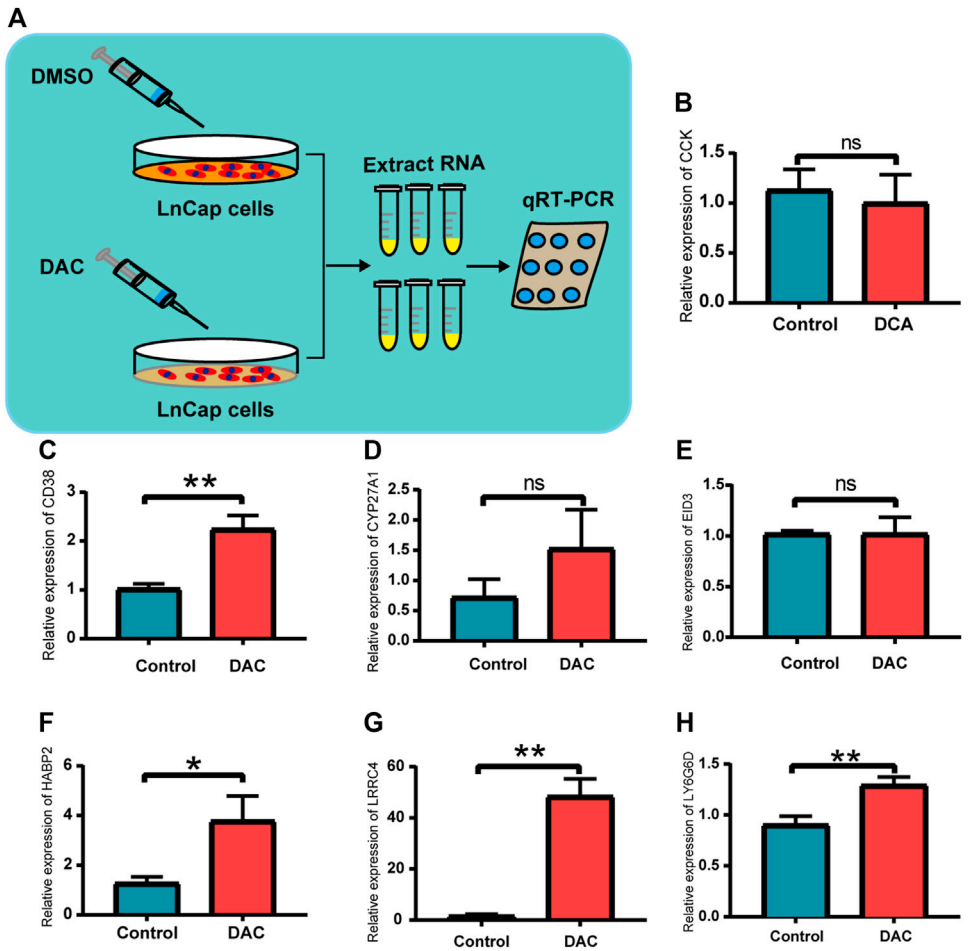

FIGURE 7 | Validation in prostate cancer cells for the seven DMGs. **(A)** Schematic illustration of demethylation of LnCap after DAC treatment. **(B–H)** Relative expression of CCK, CD38, CYP27A1, EID3, HABP2, LRRC4, and LY6G6D between the DAC group and the control; ns: $p > 0.05$, $^{*}p < 0.05$, $^{**}p < 0.01$.

correlation with methylation (**Figure 7A**). Our results indicated that the expression of *CD38, HABP2, LRRC4,* and *LY6G6D* was upregulated in LnCap cells treated with DAC, whereas that of *CCK, CYP27A1,* and *EID3* did not significantly change (**Figures 7B–H**). The results were thus not entirely validated using the LNCaP cells' data.

DISCUSSION

The course of PCA is long after operation, and patients often want to know when the next recurrence will occur so that they can be treated for it as soon as possible (Bianco et al., 2005). Predicting the BCR of PCA is clinically essential, but it is difficult for currently available prediction tools to meet current clinical needs. Thus, considerable effort has been devoted to exploring new technologies to detect early signs of tumors (Wu and Qu, 2015). This study attempted to predict BCR from a new perspective of epigenetic DMGs. We successfully established a predictive model based on seven DMGs to determine low-, medium-, and high-risk groups from the TCGA and GEO datasets. In addition, based on the Classifier, a nomogram was constructed to predict the BCR-free survival rate, which almost unambiguously classified patients into low-, medium-, and high-risk groups and distinguished BCR-free survival with high accuracy, achieving high sensitivity and specificity. Taken together, the findings suggest that the nomogram has the potential to predict BCR in patients with PCA after RP.

Brockman et al. (2015) constructed a nomogram that showed excellent predictive value for BCR, but this nomogram was built based on PSA, which limits its sensitivity at low PSA levels (Fendler et al., 2019). ^{99m}Tc-MIP-1404 PSMA-SPECT/CT was also shown to have high performance for detecting PSMA-positive lesions suggestive of tumor recurrence in patients with PCA BCR and very low serum PSA levels (Schmidkonz et al., 2019). However, it is an invasive examination, which limits its practical application. In contrast, our prediction tool is based mainly on the pathological_T and the Classifier. Most of the specimens collected were postoperative specimens that were not affected by PSA. Thus, the nomogram can use postoperative specimens to detect DMGs to avoid postoperative invasive puncture and unpredictability with low PSA levels.

Regarding the seven DMGs, cholecystokinin, also named *CCK*, as a gastrointestinal hormone, is a chemical messenger that regulates the physiological functions of the intestine and pancreas, including secretion, motility, absorption, and digestion (Thomas et al., 2003). The cholecystokinin hormones affect proliferation by blocking their respective receptors in PCA (Thomas et al., 2003). Moreover, as early as 1997, Jean Claude Reubi studied the role of *CCK-A* and *CCK-B* in some neuroendocrine and reproductive tumors, including PCA (Reubi et al., 1997). In addition, we found that CCK is mainly related to glutathione metabolism and drug metabolism. This suggests that this might be the beginning of a new understanding of *CCK* in neuroendocrine PCA. *CD38* is a glycoprotein that regulates cellular nicotinamide adenine dinucleotide metabolism. One study suggested that the methylation of *CD38* regulates the progression of localized and metastatic PCA. In our study, *CD38* was mainly associated with RNA degradation, the cell cycle, and DNA replication. *CYP27A1* is an enzyme that stimulates the transformation of cholesterol to oxysterol 27-hydroxycholesterol (27-HC). Accumulating evidence suggests that 27-HC acts as an agonist of the estrogen receptor. Moreover, *CYP27A1* is associated with the risk of lethal PCA, another sex hormone–dependent tumor (Shui et al., 2012). The relationship between *CYP27A1* methylation and PCA has not been reported to date. Interestingly, one study showed that the excessive corticosterone-induced downregulation of *CYP27A1* coincides significantly with increased CpG methylation of its promoters (Hu et al., 2017). In this study, *CYP27A1* methylation was associated with the calcium signaling pathway. De-regulation of calcium signals in prostate tumor cells mediates several pathological dysfunctions associated with PCA progression, which plays a relevant role in tumor cell death, proliferation, motility invasion, and tumor metastasis (Ardura et al., 2020).

Furthermore, *HABP2, LRRC4, LY6G6D,* and *EID3* had not been studied in PCA to date. *HABP2* has mostly been studied in thyroid cancer (Zhao et al., 2015; Zhang and Xing, 2016). However, it is expected to be studied in PCA, another endocrine-dependent cancer. As a tumor suppressor gene, inactivation of LRRC4 mediates DNA hypermethylation in central nervous system tumors (Zhang et al., 2008). In our study, *LRRC4* was mainly associated with the WNT signaling pathway, the MAPK signaling pathway, and oxidative phosphorylation, which are the common pathways in the progression of PCA (Schöpf et al., 2016; Murillo-Garzón and Kypta, 2017; Park et al., 2020). *LY6G6D* can lead to the progression of colorectal cancer and colon adenocarcinoma (Sewda et al., 2016; Giordano et al., 2019). However, its relationship with PCA needs further study. High expression of EID3 is an adverse prognostic indicator for patients with colorectal cancer (Munakata et al., 2016). In our study, EID3 was mainly associated with aminoacyl tRNA biosynthesis, cytokine–cytokine receptor interaction, the MAPK signaling pathway, and the cell cycle, which requires further study for validation. In addition, we further studied these seven genes as DMGs in LnCap cells. The changes in the expression of *CD38, HABP2, LRRC4,* and *LY6G6D* after DAC demethylation were statistically significant, whereas *CCK, CYP27A1,* and *EID3* were not statistically significant. We searched the datasets of GSE36,133 and GSE21034 and found that the gene expression of the seven DMGs was also found in other prostate cancer cell lines (**Supplementary Table S7**, **Supplementary Table S8**). Thus, other cell lines might be included for validation in the future.

Notably, this new nomogram excluded the Gleason score, N staging, and the number of positive lymph nodes from being considered as the predictive factors. However, the pathological stage markedly contributes to the predisposition of distant metastasis (Pound et al., 1999). T staging, Gleason score, and PSA can be evaluated to precisely predict the BCR risk stratification of PCA (Eisenberg et al., 2010). Although T staging was included, it had limited impact on the model according to the AUC. The number of lymph node metastases also significantly affected the survival time of patients with PCA. However, Felix Preisser et al. (2020) suggested that there was no significant difference in clinical outcomes in patients with D'Amico high- or intermediate-risk PCA who had or had not undergone pelvic lymph node dissection during radical prostatectomy. Therefore, the therapeutic benefits of pelvic lymph node dissection remain elusive (Preisser et al., 2020). This observation corroborates our finding from the nomogram on early PCA, as it also excludes the influence of the positive lymph nodes. In addition, a positive surgical margin was also an effective predictor of BCR ≥5 years post-surgery (Negishi et al., 2017). However, our prediction model did not take this into account. This may be due to the weakening of the function of this project after the replacement of DNAm biomarkers.

A limitation of this study is that the data obtained were from TCGA and GEO datasets only, and it lacks further validation using third-party clinical data. In addition, family history and ethnic/ethnic background are closely associated with PCA morbidity and affect it significantly. Hence, further investigations are warranted to conclusively establish whether the nomogram is applicable to the Asian population or not. This can be addressed by verifying the function of the Classifier with relevant data. Furthermore, owing to the lack of *in vivo* validation of our data, further evaluation of altered expression of these genes in cancer tissues compared to the normal tissues is required.

CONCLUSION

In this study, we constructed a nomogram based on DMGs that can predict postoperative BCR of PCA with high sensitivity and specificity, which expands our understanding of DMGs in the pathogenesis of PCA. The target genes had high clinical specificity and may function as a molecular marker and a potential therapeutic target for PCA in the future. However, these results were not validated using the data obtained from the LNCaP cells. Further verification using clinical and experimental data is required.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Nanfang Hospital, Southern Medical University.

AUTHOR CONTRIBUTIONS

Conception and design: CL, SH, and HZ. Administrative support: AW and HQ. Statistical analysis and bioinformatics analysis: CL and SH. *In vivo* experiments: CL and SH. Data analysis and interpretation: CL and HZ. Manuscript writing: All authors. Final approval of the manuscript: All authors.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 82060809) and the Natural Science Foundation of Guangdong Province, China (2020A1515010114).

ACKNOWLEDGMENTS

We would like to acknowledge the GEO and TCGA databases for their free use. We thank Yaqian Peng for his assistance in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.727307/full#supplementary-material

REFERENCES

Ardura, J. A., Álvarez-Carrión, L., Gutiérrez-Rojas, I., and Alonso, V. (2020). Role of Calcium Signaling in Prostate Cancer Progression: Effects on Cancer Hallmarks and Bone Metastatic Mechanisms. *Cancers* 12 (5), 1071. doi:10.3390/cancers12051071

Artibani, W., Porcaro, A. B., De Marco, V., Cerruto, M. A., and Siracusano, S. (2018). Management of Biochemical Recurrence after Primary Curative Treatment for Prostate Cancer: A Review. *Urol. Int.* 100 (3), 251–262. doi:10.1159/000481438

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 (7391), 603–607. doi:10.1038/nature11003

Baylin, S. B., and Ohm, J. E. (2006). Epigenetic Gene Silencing in Cancer - a Mechanism for Early Oncogenic Pathway Addiction. *Nat. Rev. Cancer* 6 (2), 107–116. doi:10.1038/nrc1799

Bianco, F. J., Jr., Scardino, P. T., and Eastham, J. A. (2005). Radical Prostatectomy: Long-Term Cancer Control and Recovery of Sexual and Urinary Function ("trifecta"). *Urology* 66 (5 Suppl. l), 83–94. doi:10.1016/j.urology.2005.06.116

Brockman, J. A., Alanee, S., Vickers, A. J., Scardino, P. T., Wood, D. P., Kibel, A. S., et al. (2015). Nomogram Predicting Prostate Cancer-specific Mortality for Men with Biochemical Recurrence after Radical Prostatectomy. *Eur. Urol.* 67 (6), 1160–1167. doi:10.1016/j.eururo.2014.09.019

Camp, R. L., Dolled-Filhart, M., and Rimm, D. L. (2004). X-tile. *Clin. Cancer Res.* 10 (21), 7252–7259. doi:10.1158/1078-0432.Ccr-04-0713

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer Statistics in China, 2015. *CA: A Cancer J. Clinicians* 66 (2), 115–132. doi:10.3322/caac.21338

Devos, G., Witters, M., Moris, L., Van den Broeck, T., Berghen, C., Devlies, W., et al. (2020). Site-specific Relapse Patterns of Patients with Biochemical Recurrence Following Radical Prostatectomy Assessed by 68Ga-PSMA-11 PET/CT or 11C-Choline PET/CT: Impact of Postoperative Treatments. *World J. Urol.* 39, 399–406. doi:10.1007/s00345-020-03220-0

Eisenberg, M. L., Davies, B. J., Cooperberg, M. R., Cowan, J. E., and Carroll, P. R. (2010). Prognostic Implications of an Undetectable Ultrasensitive Prostate-specific Antigen Level after Radical Prostatectomy. *Eur. Urol.* 57 (4), 622–630. doi:10.1016/j.eururo.2009.03.077

Fendler, W. P., Calais, J., Eiber, M., Flavell, R. R., Mishoe, A., Feng, F. Y., et al. (2019). Assessment of 68Ga-PSMA-11 PET Accuracy in Localizing Recurrent Prostate Cancer. *JAMA Oncol.* 5 (6), 856–863. doi:10.1001/jamaoncol.2019.0096

Fraser, M., Sabelnykova, V. Y., Yamaguchi, T. N., Heisler, L. E., Livingstone, J., Huang, V., et al. (2017). Genomic Hallmarks of Localized, Non-indolent Prostate Cancer. *Nature* 541 (7637), 359–364. doi:10.1038/nature20788

Gao, X., Li, L.-Y., Rassler, J., Pang, J., Chen, M.-K., Liu, W.-P., et al. (2017). Prospective Study ofCRMP4Promoter Methylation in Prostate Biopsies as a Predictor for Lymph Node Metastases. *JNCI J. Natl. Cancer Inst.* 109 (6), djw282. doi:10.1093/jnci/djw282

Gevaert, O. (2015). MethylMix: an R Package for Identifying DNA Methylation-Driven Genes. *Bioinformatics* 31 (11), 1839–1841. doi:10.1093/bioinformatics/btv020

Giordano, G., Parcesepe, P., D'Andrea, M. R., Coppola, L., Di Raimo, T., Remo, A., et al. (2019). JAK/Stat5-mediated Subtype-specific Lymphocyte Antigen 6 Complex, Locus G6D (LY6G6D) Expression Drives Mismatch Repair Proficient Colorectal Cancer. *J. Exp. Clin. Cancer Res.* 38 (1), 28. doi:10.1186/s13046-018-1019-5

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* 56 (2), 337–344. doi:10.1111/j.0006-341x.2000.00337.x

Herman, J. G., and Baylin, S. B. (2003). Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N. Engl. J. Med.* 349 (21), 2042–2054. doi:10.1056/NEJMra023075

Hu, Y., Sun, Q., Zong, Y., Liu, J., Idriss, A. A., Omer, N. A., et al. (2017). Prenatal Betaine Exposure Alleviates Corticosterone-Induced Inhibition of CYP27A1 Expression in the Liver of Juvenile Chickens Associated with its Promoter DNA Methylation. *Gen. Comp. Endocrinol.* 246, 241–248. doi:10.1016/j.ygcen.2016.12.014

Jordan, E. J., Kim, H. R., Arcila, M. E., Barron, D., Chakravarty, D., Gao, J., et al. (2017). Prospective Comprehensive Molecular Characterization of Lung Adenocarcinomas for Efficient Patient Matching to Approved and Emerging Therapies. *Cancer Discov.* 7 (6), 596–609. doi:10.1158/2159-8290.Cd-16-1337

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

Metzger, E., Willmann, D., McMillan, J., Forne, I., Metzger, P., Gerhardt, S., et al. (2016). Assembly of Methylated KDM1A and CHD1 Drives Androgen Receptor-dependent Transcription and Translocation. *Nat. Struct. Mol. Biol.* 23 (2), 132–139. doi:10.1038/nsmb.3153

Mottet, N., Briers, E., Van den Broeck, T., Cumberbatch, M. G., De Santis, M., Fanti, S., et al. (2018). EAU Guidelines: Prostate Cancer. Available at: https://uroweb.org/guideline/prostate-cancer/(accessed May 29, 2020).

Munakata, K., Uemura, M., Tanaka, S., Kawai, K., Kitahara, T., Miyo, M., et al. (2016). Cancer Stem-like Properties in Colorectal Cancer Cells with Low Proteasome Activity. *Clin. Cancer Res.* 22 (21), 5277–5286. doi:10.1158/1078-0432.Ccr-15-1945

Murillo-Garzón, V., and Kypta, R. (2017). WNT Signalling in Prostate Cancer. *Nat. Rev. Urol.* 14 (11), 683–696. doi:10.1038/nrurol.2017.144

Negishi, T., Kuroiwa, K., Hori, Y., Tomoda, T., Uchino, H., Tokuda, N., et al. (2017). Predictive Factors of Late Biochemical Recurrence after Radical Prostatectomy. *Jpn. J. Clin. Oncol.* 47 (3), 233–238. doi:10.1093/jjco/hyw181

Nowacka-Zawisza, M., and Wiśnik, E. (2017). DNA Methylation and Histone Modifications as Epigenetic Regulation in Prostate Cancer. *Oncol. Rep.* 38 (5), 2587–2596. doi:10.3892/or.2017.5972

Park, S., Kwon, W., Park, J.-K., Baek, S.-M., Lee, S.-W., Cho, G.-J., et al. (2020). Suppression of Cathepsin a Inhibits Growth, Migration, and Invasion by Inhibiting the P38 MAPK Signaling Pathway in Prostate Cancer. *Arch. Biochem. Biophys.* 688, 108407. doi:10.1016/j.abb.2020.108407

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

Pound, C. R., Partin, A. W., Eisenberger, M. A., Chan, D. W., Pearson, J. D., and Walsh, P. C. (1999). Natural History of Progression after PSA Elevation Following Radical Prostatectomy. *Jama* 281 (17), 1591–1597. doi:10.1001/jama.281.17.1591

Preisser, F., van den Bergh, R. C. N., Gandaglia, G., Ost, P., Surcel, C. I., Sooriakumaran, P., et al. (2020). Effect of Extended Pelvic Lymph Node Dissection on Oncologic Outcomes in Patients with D'Amico Intermediate and High Risk Prostate Cancer Treated with Radical Prostatectomy: A Multi-Institutional Study. *J. Urol.* 203 (2), 338–343. doi:10.1097/ju.0000000000000504

Reubi, J. C., Schaer, J. C., and Waser, B. (1997). Cholecystokinin(CCK)-A and CCK-B/gastrin Receptors in Human Tumors. *Cancer Res.* 57 (7), 1377–1386.

Sauerbrei, W., Royston, P., and Binder, H. (2007). Selection of Important Variables and Determination of Functional Form for Continuous Predictors in Multivariable Model Building. *Statist. Med.* 26 (30), 5512–5528. doi:10.1002/sim.3148

Schmidkonz, C., Goetz, T. I., Kuwert, T., Ritt, P., Prante, O., Bäuerle, T., et al. (2019). PSMA SPECT/CT with 99mTc-MIP-1404 in Biochemical Recurrence of Prostate Cancer: Predictive Factors and Efficacy for the Detection of PSMA-Positive Lesions at Low and Very-Low PSA Levels. *Ann. Nucl. Med.* 33 (12), 891–898. doi:10.1007/s12149-019-01400-6

Schöpf, B., Schäfer, G., Weber, A., Talasz, H., Eder, I. E., Klocker, H., et al. (2016). Oxidative Phosphorylation and Mitochondrial Function Differ between Human Prostate Tissue and Cultured Cells. *Febs j* 283 (11), 2181–2196. doi:10.1111/febs.13733

Sewda, K., Coppola, D., Enkemann, S., Yue, B., Kim, J., Lopez, A. S., et al. (2016). Cell-surface Markers for colon Adenoma and Adenocarcinoma. *Oncotarget* 7 (14), 17773–17789. doi:10.18632/oncotarget.7402

Shui, I. M., Mucci, L. A., Kraft, P., Tamimi, R. M., Lindstrom, S., Penney, K. L., et al. (2012). Vitamin D-Related Genetic Variation, Plasma Vitamin D, and Risk of Lethal Prostate Cancer: a Prospective Nested Case-Control Study. *J. Natl. Cancer Inst.* 104 (9), 690–699. doi:10.1093/jnci/djs189

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer Statistics, 2019. *CA A. Cancer J. Clin.* 69 (1), 7–34. doi:10.3322/caac.21551

Spahn, M., Joniau, S., Gontero, P., Fieuws, S., Marchioro, G., Tombal, B., et al. (2010a). Outcome Predictors of Radical Prostatectomy in Patients with Prostate-specific Antigen Greater Than 20 Ng/ml: a European Multi-Institutional Study of 712 Patients. *Eur. Urol.* 58 (1), 1–7. doi:10.1016/j.eururo.2010.03.001

Spahn, M., Kneitz, S., Scholz, C.-J., Nico, S., Rüdiger, T., Ströbel, P., et al. (2009b). Expression of microRNA-221 Is Progressively Reduced in Aggressive Prostate Cancer and Metastasis and Predicts Clinical Recurrence. *Int. J. Cancer* 127 (2), NA. doi:10.1002/ijc.24715

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., et al. (2010). Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell* 18 (1), 11–22. doi:10.1016/j.ccr.2010.05.026

Thomas, R. P., Hellmich, M. R., Townsend, C. M., Jr., and Evers, B. M. (2003). Role of Gastrointestinal Hormones in the Proliferation of normal and Neoplastic Tissues. *Endocr. Rev.* 24 (5), 571–599. doi:10.1210/er.2002-0028

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statist. Med.* 16 (4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385:aid-sim380>3.0.co;2-3

Wang, J., Ni, J., Beretov, J., Thompson, J., Graham, P., and Li, Y. (2020). Exosomal microRNAs as Liquid Biopsy Biomarkers in Prostate Cancer. *Crit. Rev. Oncology/Hematology* 145, 102860. doi:10.1016/j.critrevonc.2019.102860

Wei, J.-H., Haddad, A., Wu, K.-J., Zhao, H.-W., Kapur, P., Zhang, Z.-L., et al. (2015). A CpG-Methylation-Based Assay to Predict Survival in clear Cell Renal Cell Carcinoma. *Nat. Commun.* 6, 8699. doi:10.1038/ncomms9699

WHO (2018). World Health Organization. Cancer, Available at: http://www.who.int/topics/cancer/en/.

Wu, L., and Qu, X. (2015). Cancer Biomarker Detection: Recent Achievements and Challenges. *Chem. Soc. Rev.* 44 (10), 2963–2997. doi:10.1039/c4cs00370e

Zhang, T., and Xing, M. (2016). HABP2G534E Mutation in Familial Nonmedullary Thyroid Cancer. *JNCI.J* 108 (6), djv415. doi:10.1093/jnci/djv415

Zhang, Z., Li, D., Wu, M., Xiang, B., Wang, L., Zhou, M., et al. (2008). Promoter Hypermethylation-Mediated Inactivation of LRRC4 in Gliomas. *BMC Mol. Biol.* 9, 99. doi:10.1186/1471-2199-9-99

Zhao, X., Li, X., and Zhang, X. (2015). HABP2Mutation and Nonmedullary Thyroid Cancer. *N. Engl. J. Med.* 373 (21), 2084–2087. doi:10.1056/NEJMc1511631

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

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

Copyright © 2022 Luo, He, Zhang, He, Qi and Wei. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 February 2022
doi: 10.3389/fgene.2022.810681

Role of Suprabasin in the Dedifferentiation of Follicular Epithelial Cell-Derived Thyroid Cancer and Identification of Related Immune Markers

*Hao Tan, Lidong Wang and Zhen Liu**

Department of General Surgery, Shengjing Hospital of China Medical University, Shenyang, China

Background: Aberrant regulation of suprabasin (SBSN) is associated with the development of cancer and immune disorders. SBSN influences tumor cell migration, proliferation, angiogenesis, and immune resistance. In this study, we investigated the potential correlation between *SBSN* expression and immune infiltration in thyroid cancer.

Methods: The expression of SBSN in 80 papillary thyroid carcinoma (PTC) specimens was determined using quantitative reverse-transcription polymerase chain reaction, western blotting, and immunohistochemical staining. The expression of SBSN in 9 cases of poorly differentiated thyroid carcinoma (PDTC) and 18 cases of anaplastic thyroid carcinoma (ATC) was evaluated by immunohistochemical staining. Comprehensive bioinformatics analysis of *SBSN* expression was performed using The *Cancer* Genome Atlas and Gene Expression Omnibus datasets, and the relationship of SBSN expression with M2 macrophages and T regulatory cells (Tregs) in ATC and PTC was verified by immunohistochemical staining.

Results: Compared with those in adjacent normal tissues, the expression levels of SBSN mRNA and protein were significantly higher in PTC tissues. SBSN expression level was correlated with that of cervical lymph node metastasis in PTC patients. Immunohistochemical staining results showed statistically significant differences among high-positive expression rates of SBSN in PTC, PDTC, and ATC. Functional enrichment analysis showed that *SBSN* expression was associated with pathways related to cancer, cell signaling, and immune response. Furthermore, analysis of the tumor microenvironment (using CIBERSORT-ABS and xCell algorithms) showed that *SBSN* expression affected immune cell infiltration and the cancer immunity cycle, and immunohistochemistry

OPEN ACCESS

Edited by:
Farhad Maleki, McGill University, Canada

Reviewed by:
Marc Gregory Yu, Joslin Diabetes Center and Harvard Medical School, United States
Yin Detao, First Affiliated Hospital of Zhengzhou University, China

****Correspondence:***
Zhen Liu
liuzhen1973@aliyun.com

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

Received: *07 November 2021*
Accepted: *14 January 2022*
Published: *09 February 2022*

Citation:
Tan H, Wang L and Liu Z (2022) Role of Suprabasin in the Dedifferentiation of Follicular Epithelial Cell-Derived Thyroid Cancer and Identification of Related Immune Markers. Front. Genet. 13:810681. doi: 10.3389/fgene.2022.810681

Abbreviations: ACC, adenoid cystic carcinoma; ATC, anaplastic thyroid carcinoma; CAF, cancer-associated fibroblast; DCs, dendritic cells; ESCC, esophageal squamous cell carcinoma; GSEA, gene set enrichment analysis; HSCs, hematopoietic stem cells; iDCs, immature dendritic cells; MDSCs, marrow-derived suppressor cells; MEP, megakaryocyte–erythroid progenitor; NSCLC, non-small cell lung cancer; PDTC, poorly differentiated thyroid carcinoma; PTC, papillary thyroid carcinomas; qRT-PCR, quantitative reverse-transcription polymerase chain reaction; SBSN, suprabasin; ssGSEA, single-sample gene set enrichment analysis; TAMs, tumor-associated macrophages; TCGA, the *cancer* genome atlas; Th2, T helper cell type 2; TIICs, tumor-infiltrating immune cells; TIP, tumor immunophenotype; TME, tumor microenvironment; Tregs, regulatory T cells.

confirmed a significant increase in M2 macrophage and Treg infiltration in tumor tissues with high-positive SBSN expression.

Conclusion: These findings reveal that SBSN may be involved in thyroid carcinogenesis, tumor dedifferentiation progression, and immunosuppression as an important regulator of tumor immune cell infiltration.

Keywords: suprabasin, thyroid cancer, lymph node metastasis, immune infiltration, tumor immunosuppression, dedifferentiation

INTRODUCTION

Thyroid cancer is one of the most common endocrine tumors and its incidence has increased globally over the last 30 years (La Vecchia et al., 2015). Most thyroid cancers originate from the follicular epithelial cells of the thyroid gland, which secrete iodine-containing thyroid hormones. Follicular epithelium-derived thyroid cancers can be classified as papillary thyroid carcinoma (PTC), follicular thyroid carcinoma, poorly differentiated thyroid carcinoma (PDTC), and anaplastic thyroid carcinoma (ATC) (Dralle et al., 2015). Of these, PTC is the most common pathological type, accounting for 80–90% of all thyroid cancers (Abdullah et al., 2019). Most thyroid cancers exhibit inert biological behavior and have a good prognosis, with a 20-year survival rate of 95% (Gospodarowicz et al., 2001). However, recurrence, metastasis, and resistance to radioiodine therapy in PDTC, ATC, and some invasive PTCs remain the leading causes of death from thyroid cancer (Mazzaferri and Jhiang, 1994; Molinaro et al., 2017; Xu and Ghossein, 2020), with more than 25% of patients with PTC experiencing recurrence during a long-term follow-up (Abdullah et al., 2019). Furthermore, a high rate (up to 85%) of cervical-lymph-node metastasis, which is considered a very high-risk factor for PTC recurrence, has been documented in patients with PTC (Zheng et al., 2019). Currently, some studies suggest that ATC is different from PTC in the early stages of tumor development and that the two tumor types evolve via distinct mechanisms (Capdevila et al., 2018). However, it is also believed that histologically highly differentiated thyroid cancer may dedifferentiate into PDTC or ATC via a multi-step process of genetic and epigenetic alterations (Papp and Asa, 2015) or that ATC can develop from PTC via accumulation of genomic mutations (Landa et al., 2016).

Immunotherapy has long been a focus area in oncology and is effective against non-small cell lung cancer (NSCLC) and kidney cancer (Motzer et al., 2015; Reck et al., 2016). Immune cell infiltration of the tumor microenvironment (TME) has also been associated with survival in many patients with solid tumors (Baxevanis et al., 2019). Infiltrating immune cells may be used as drug targets to improve patient survival (Lote et al., 2015). A previous study showed that the polarization of a higher number of tumor-associated macrophages (TAMs) in a tumor-promoting M2 phenotype implies a poorer survival of patients with ATC (Jung et al., 2015). Fang et al. (2014) found that TAMs purified from human PTC could promote invasiveness of thyroid cancer cell lines by secreting CXCL8. Melillo et al. (2010) found that the density of tumor-associated mast cells was higher in PTC than in normal tissue and correlated with extra-thyroidal tumor infiltration. In PTC, the number of $CD4^+$ T cells correlates with the tumor size, whereas that of Tregs correlates with lymph node metastasis (French et al., 2010). Tregs are enriched in tumor-involved lymph nodes, and their numbers correlate with PTC recurrence (French et al., 2012). Tumor-infiltrating lymphocytes, TAMs, and tumor-infiltrating neutrophils influence the prognosis and efficacy of chemotherapy and immunotherapy (Waniczek et al., 2017; Zhang et al., 2018). In addition, Chen and Mellman (2013) divided the cancer immunity cycle into seven steps, including the release of cancer cell antigens, cancer antigen presentation, initiation and activation of immune cells, transport and infiltration of immune cells into the tumor, and recognition and killing of cancer cells by T cells. Consequently, the cancer immunity cycle has become one of the starting points for cancer immunotherapy research. Therefore, it is imperative to study the TME and identify the distribution and functions of tumor-infiltrating immune cells (TIICs) to find new tumor markers for thyroid cancer.

SBSN was first identified in epithelial tissues (human and murine) and is thought to play a key role in the process of epidermal differentiation (Park et al., 2002). However, in recent years, SBSN has been reported to be aberrantly expressed in certain malignancies, and inhibition of SBSN may lead to the inhibition of cancer cell proliferation, invasion, and metastasis, suggesting that SBSN may be associated with tumor progression. For example, SBSN expression is abnormally regulated in esophageal squamous cell carcinoma (ESCC) (Zhu et al., 2016; Takahashi et al., 2020), salivary adenoid cystic carcinoma (ACC) (Shao et al., 2012), and NSCLC (Glazer et al., 2009). In addition, several studies have shown that SBSN expression in tumors is regulated by several signaling pathways that affect the tumor properties (Alam et al., 2014; Zhu et al., 2016; Takahashi et al., 2020). These results suggest that SBSN may act as an oncogenic factor to promote tumorigenesis and tumor progression. In addition, SBSN plays an important role in the development and progression of immune diseases, such as neuropsychiatric systemic lupus erythematosus (Ichinose et al., 2018) and atopic dermatitis (Aoshima et al., 2019). However, the potential mechanism of *SBSN* as a proto-oncogene in thyroid cancer progression and immunology is not clear.

In this study, we investigated the expression of SBSN in thyroid cancers of follicular epithelial origin with different degrees of differentiation. The relationship of SBSN with clinicopathological features of patients with PTC, as well as the potential involvement of SBSN in cancer immunity, were

also explored. This study suggests an important role for SBSN in thyroid carcinoma and the potential mechanisms by which SBSN may be involved in the processes of thyroid cancer dedifferentiation and immune regulation.

MATERIALS AND METHODS

Data Sources and Pre-Processing

This study used several public datasets, including PTC, PDTC, and ATC datasets. For The *Cancer* Genome Atlas (TCGA) dataset, RNA sequencing (RNA-seq) data and clinical features of patients with thyroid cancer were identified in and extracted from the TCGA portal and validated (Cancer Genome Atlas Research Network, 2014), with a total of 568 samples, including 502 PTCs, 8 metastatic thyroid cancers, and 58 matched normal thyroid samples. High-throughput sequencing fragments per kilobase of transcript per million mapped reads (FPKM) values were further analyzed for all samples after $\log_2$(FPKM+1) transformation. The ATC microarray datasets GSE29265, GSE33630, GSE76039, and GSE65144 were downloaded from the Gene Expression Omnibus database of the National Center for Biotechnology Information by searching for "anaplastic thyroid cancer" and "*Homo sapiens*." The PDTC microarray datasets GSE53157 and GSE76039 were downloaded from the same database by searching for "poorly differentiated thyroid cancer" and "*Homo sapiens*." All microarray data were background adjusted and normalized by removing the batch processing effect using the R package "sva." Probes that did not match the gene symbol in the annotation file were deleted. When more than one probe matched the same gene symbol, the average value was calculated as the final expression value. GSE29265 consisted of 9 ATC tissues and 10 adjacent normal thyroid tissues. GSE33630 consisted of 11 ATC tissues and 45 adjacent normal thyroid tissues. GSE65144 consisted of 12 ATC tissues and 13 adjacent normal thyroid tissues. GSE76039 consisted of 17 PDTC tissues and 20 ATC tissues. GSE53157 consisted of five PDTCs, seven classical PTCs, eight PTC follicular variants, and four follicular thyroid carcinomas.

Immune Infiltration Analysis

The ESTIMATE algorithm can determine the ratio of stromal and immune cells based on the gene expression profile in tumor samples. It has been applied to assess the TME in patients, as well as the stromal score (stromal cell content), immune score (degree of immune cell infiltration), ESTIMATE score (a synthetic marker of the stroma and immunity), and tumor purity, using the R package (Yoshihara et al., 2013). The CIBERSORT-ABS and xCell algorithms were used to estimate the relative proportions of various immune cell types in the TME. For each cell type, xCell was used to analyze the enrichment scores of all samples by integrating a single-sample gene set enrichment analysis (ssGSEA) approach (Aran et al., 2017). CIBERSORT-ABS, an analytical method developed by Newman, uses gene expression data to estimate the abundance ratios of 22 cell types in a mixed cell population at a statistical significance level of $p < 0.05$ (Newman et al., 2015). The reference for the deconvolution of RNA-seq data from patients with cancer was the leukocyte signature matrix (LM22). For cancer immunity cycle analysis, we applied the Tracking Tumor Immunophenotype (TIP) pipeline based on the ssGSEA algorithm (Xu et al., 2018). The TIP scores for the TCGA dataset are available from the online TIP server (http://biocc.hrbmu.edu.cn/TIP/).

Patients and Clinicopathological Data

Eighty specimens from patients with PTC who underwent surgical treatment at the Shengjing Hospital of China Medical University from July 2016 to July 2017 were stored in liquid nitrogen and subjected to quantitative reverse-transcription polymerase chain reaction (qRT-PCR) and western blotting. For immunohistochemical staining, tissues were embedded in paraffin. The selected PTC tissues and paired normal tissues adjacent to cancer were diagnosed by pathology. Normal tissues adjacent to the cancer tissue were collected, from the same patients with PTC, at least 2 cm from the PTC area. Paraffin-embedded tissues from 9 patients with PDTC and 18 patients with ATC who underwent surgical treatment at the Shengjing Hospital of China Medical University from 2010 to 2017 were preserved at the Department of Pathology. All histological sections were reviewed by two specialist pathologists to verify the histological diagnosis. All patients were diagnosed for the first time and did not receive any treatment before surgery. Patients were classified according to the eighth edition of the American Joint Committee on *Cancer* TNM classification system for differentiated thyroid cancer. Clinical information such as patient age, tumor size, and cervical-lymph-node metastasis was retrieved from the clinical files of the patients. All patients provided informed consent for the use of their clinical and pathological data for research purposes, and all tissue specimens and clinical data were collected according to the protocol approved by the Ethics Committee of the Shengjing Hospital, China Medical University (approval number 2014PS47K).

Functional and Pathway Enrichment Analyses

The GeneMANIA (http://www.genemania.org) database was used to construct a gene-gene interaction network for *SBSN*, including genes that are associated with SBSN in terms of physical interactions, co-expression, prediction, co-localization, and genetic interactions. Functional and pathway enrichment analyses of the genes co-expressed with *SBSN* in the cBioPortal database (496 cases from the TCGA Cell 2014 dataset of PTC) were performed using DAVID (https://david.ncifcrf.gov). Genome enrichment analysis was performed using GSEA 3.0 software. The c2.cp.kegg.v6.1.symbols.gmt dataset was downloaded from the Molecular Signatures Database on the GSEA website. Enrichment analysis of $SBSN^{high}$ and $SBSN^{low}$ groups was performed for expression spectrum data and attribute files using the default weighted method. The random classification frequency was set to 1,000.

Immunohistochemical Staining

Formalin-fixed, paraffin-embedded sections (4 μm thick) were prepared and subjected to ethanol gradient dewaxing and endogenous peroxidase blocking. Antigen retrieval was performed by boiling the slides in citrate buffer (pH 6.0) for 7.5 min, followed by cooling to room temperature. The slides were incubated with a polyclonal rabbit anti-human SBSN antibody (1:250; Cat# abx130453; Abbexa, Cambridge, United Kingdom) at 4°C overnight after 30 min of incubation at 37°C with drops of goat blocking serum. Afterward, the slides were rinsed with phosphate-buffered saline (PBS) and incubated with a drop of a horseradish peroxidase-labeled sheep anti-rabbit secondary antibody for 30 min at 37°C. Subsequently, the slides were stained with a 3,3′-diaminobenzidine (Cat# DAB-0031; MXB, Maixin, China) solution for 1–2 min, counterstained with hematoxylin, dehydrated, covered with coverslips, and analyzed by light microscopy. PBS was used instead of the primary antibody in a negative control group. The experimental procedure was carried out according to the SP kit instructions. Five high-magnification fields (×400) were randomly selected from each section under a light microscope and scored by two pathologists. The degree of staining was scored from 0 to 4 (0, none; 1, <10%; 2, 10–50%; 3, 51–80%; and 4, >80%). The staining intensity was scored from 0 to 3 (0, no staining; 1, light yellow; 2, brown-yellow; and 3, brown). To calculate the final score, the two scores were multiplied, and the results were presented as follows: 0–1 point (−), 2–4 points (+), 5–8 points (++), and 9–12 points (+++). We defined −/+ as the low-positive expression group and ++/+++ as the high-positive expression group. To control for errors, the scoring was performed by two independent observers, and a third observer read the films; in cases of disagreement, all three discussed the scores collectively until an agreement was reached.

For M2 macrophages, fields of view with $CD163^+$ M2 macrophages were selected. The number of $CD163^+$ M2 macrophages was counted in five randomized high-magnification fields (×400) per sample, and the mean value was considered as the level of $CD163^+$ M2 macrophages.

For Tregs, fields of view with $Foxp3^+$ Tregs were selected. The number of $Foxp3^+$ Tregs was counted in five randomized high-magnification fields (×400) per sample, and the mean value was considered as the level of $Foxp3^+$ Tregs.

Western Blotting

To extract total proteins, RIPA lysis buffer (Cat#P0013B; Beyotime Biotechnology, Shanghai, China) was added to tissues, and the homogenates were centrifuged at 14,000 × rpm for 45 min at 4°C. Total proteins (40 μg) were separated by SDS-PAGE and transferred to PVDF membranes. The membranes were blocked with 5% bovine serum albumin at room temperature for 2 h and then incubated with primary antibodies at 4°C overnight. Thereafter, the membrane was incubated with a secondary antibody for 2 h at room temperature. The primary antibodies used in this study included a rabbit anti-SBSN polyclonal antibody (1:1,000; Cat# abx130453; Abbexa, Cambridge, United Kingdom) and rabbit polyclonal anti-GAPDH (1:10,000; Cat# 10494-1-AP, Proteintech Group, Inc., Chicago, United States). Peroxidase-labeled goat anti-rabbit or anti-mouse IgG (H + L) (1:2,000; Zhongshan Jinqiao Company, Beijing, China) was used as a secondary antibody. An enhanced chemiluminescence kit (Beyotime Biotechnology) was used for detection. The integrated optical density of each band was measured using Image-Pro Plus 6.0 software (Media Cybernetics Inc., Rockville, MD, United States). The target protein expression level was calculated relative to that of GAPDH, which was used as an internal control.

qRT-PCR

The total RNA was extracted from thyroid tissue specimens using TRIzol reagent (Cat# 9108; Takara, Beijing, China) according to the manufacturer's instructions. After verification of the purity and concentration, RNA was reverse transcribed into cDNA using a cDNA synthesis kit (Cat# RR047A; Takara, Beijing, China). qRT-PCR of the cDNA (2 μl per 20 μl reaction) was performed using the TB Green ® Premix Ex Taq TM II kit (Cat# RR820; Takara, Beijing, China) and a 7,500 Fast instrument. Primer sequences for SBSN were forwad:5′-CATGGCGTTAGTCAGGCTGGAAG-3′ and reverse:5′-CCTCCTTGCTGGCTTGGTTGAC-3′. The primer sequences used for GAPDH were forward:5′-GGAGCGAGATCCCTCCAAAAT-3′ and reverse:5′- GGC TGTTGTCATACTTCTCATGG-3′. The PCR protocol was as follows: 95°C for 2 min, followed by 40 cycles at 95°C for 15 s and 60°C for 30 s. The relative expression level was calculated using the $2^{-\Delta\Delta Ct}$ method using *GAPDH* as a reference gene for normalization.

Statistical Analysis

Statistical analyses were performed using SPSS (20.0) and R (4.0.3) software. The chi-squared test was used to analyze differences in the degree of SBSN immunohistochemical staining among ATC, PDTC, PTC, and adjacent normal thyroid tissues. The relationships between the statistical results of SBSN immunohistochemical staining and the clinicopathological characteristics of PTC were assessed by the chi-squared test and Fisher's exact probability test. A *t*-test was used to assess the relationships between the statistical results of SBSN western blotting and qRT-PCR and the clinicopathological characteristics of PTC. Data are expressed as means ± standard deviation. Box plot analysis was performed using the Wilcoxon rank-sum test; correlation between two variables was calculated using Spearman's rho, and one-way analysis of variance was used for comparison among multiple samples. $p < 0.05$ was considered to be statistically significant.

RESULTS

Expression of SBSN in Thyroid Cancer Tissues

The relative expression level of *SBSN* mRNA was evaluated in the 80 PTC tissues and paired paraneoplastic normal tissues. The qRT-PCR results showed that the relative expression level of

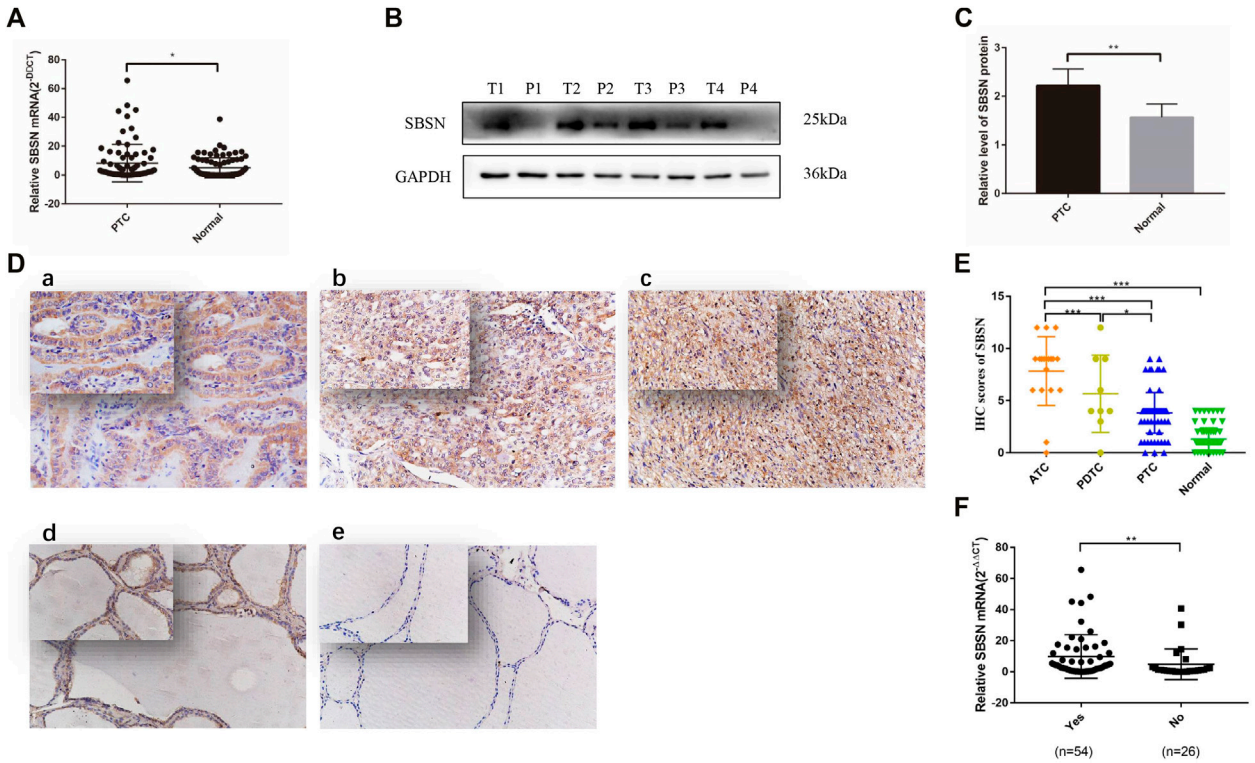

FIGURE 1 | Expression of SBSN in different thyroid tissues. **(A)** Expression of *SBSN* in papillary thyroid carcinoma (PTC) tissues and normal tissues adjacent to cancer was detected by qRT-PCR ($n = 80$ per group). **(B)** Expression of SBSN in PTC and normal tissues adjacent to cancer was detected using western blot ($n = 80$ per group). **(C)** Relative grayscale values of SBSN in PTC tissues and normal tissues adjacent to cancer. **(D)** Expression of SBSN in different thyroid cancer tissue samples (×200, top left ×400). **(a)** Positive expression of SBSN in PTC tissues; **(b)** Positive expression of SBSN in poorly differentiated thyroid carcinoma tissues; **(c)** Positive expression of SBSN in anaplastic thyroid carcinoma tissues; **(d)** Positive expression of SBSN in normal tissues adjacent to cancer; **(e)** Negative expression of SBSN in normal tissues adjacent to cancer. **(E)** Immunohistochemical staining scores of SBSN in various thyroid tissue samples. **(F)** High expression of *SBSN* mRNA levels in PTC tissues was associated with lymph node metastasis in patients. For western blot, GAPDH was used as an internal control. Data are expressed as means ± standard deviation. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 1 | Expression of *SBSN* in different types of thyroid tissue.

Group	*n*	Low		High		High positive rate
		(–)	(+)	(++)	(+++)	(%)
ATC	18	2	0	5	11	88.89[a,b]
PDTC	9	1	4	1	3	44.44[c]
PTC	80	12	58	8	2	12.5[d]
Normal	80	61	19	0	0	0

[a]*ATC, vs. PDTC (*p = 0.023).*
[b]*ATC, vs. PTC (***p < 0.001).*
[c]*PDTC, vs. PTC (*p = 0.044).*
[d]*PTC, vs Normal (***p = 0.001).*

SBSN mRNA was significantly higher in the PTC tissues than that in the paraneoplastic normal tissues (8.228 ± 1.452 and 5.037 ± .783, respectively; $p < 0.05$) (**Figure 1A**). Western blot results showed that the expression level of SBSN in the 80 PTC tissues was significantly higher than that in the paraneoplastic tissues (2.218 ± .343 and 1.563 ± .279, respectively; $p < 0.05$) (**Figures 1B,C**). Immunohistochemical staining showed that SBSN protein was mainly expressed in the cytoplasm (**Figure 1D**). The high-positive expression rate of SBSN was 12.5% (10/80) in the PTC tissues and 0% (0/80) in the adjacent normal tissues ($p = .001$). In addition, the expression of SBSN in the 9 patients with PDTC and 18 patients with ATC was detected by immunohistochemical staining. The results showed that the high-positive expression rate of SBSN in ATC (88.9%, 16/18) was significantly higher ($p < 0.05$) than those in PDTC (44.4%, 4/9) and PTC (12.5%, 10/80). Compared with that in PTC, the high-positive expression rate of SBSN was significantly higher ($p < 0.05$) in the PDTC group (**Table 1**). The immunohistochemical staining scores are shown in **Figure 1E**.

Relationships Between High SBSN Expression Level and Clinicopathological Characteristics of Patients With PTC

We characterized the clinicopathological characteristics, including the expression levels of SBSN, in 80 PTC patients. As shown in **Table 2**, the expression levels of *SBSN* mRNA were significantly different between the groups with and without lymph node metastasis (9.84 ± 1.91 and 4.87 ± 1.93, respectively; $p < 0.05$) (**Figure 1F**). The SBSN protein

TABLE 2 | Relationship between *SBSN* and the clinical pathological characteristics in papillary thyroid carcinoma.

Clinical features	*n*	SBSN expression [cases (%)][a]		*p*	*SBSN* mRNA	*p*	SBSN protein[b]	*p*
		Low positive	High positive					
Gender								
Male	16	13 (81.25)	3 (18.75)	0.673	15.81 ± 5.57	0.104	1.99 ± 0.72	0.338
Female	64	57 (89.06)	7 (10.94)		6.62 ± 1.25		2.27 ± 0.39	
Age(y)								
≥55	23	20 (86.96)	3 (13.04)	0.926	8.59 ± 3.26	0.889	2.86 ± 0.91	0.301
<55	57	50 (87.72)	7 (12.28)		8.11 ± 1.63		2.02 ± 0.35	
Tumor size								
≤2	58	53 (91.38)	5 (8.62)	0.150	7.82 ± 1.75	0.949	1.86 ± 0.36	0.054
>2, ≤4	20	15 (75)	5 (25)		9.69 ± 2.84		2.61 ± 0.76	
>4	2	2 (100)	0 (0)		2.63 ± 0.10		6.71 ± 2.80	
Multifocality								
Yes	43	41 (95.35)	2 (4.65)	0.051	7.56 ± 1.90	0.758	2.22 ± 0.47	0.679
No	37	29 (78.38)	8 (21.62)		8.93 ± 2.23		2.21 ± 0.51	
Bilateral								
No	65	58 (89.23)	7 (10.77)	0.588	7.87 ± 1.70	0.418	2.30 ± 0.42	0.632
Yes	15	12 (80)	3 (20)		9.37 ± 2.82		1.95 ± 0.54	
Extrathyroid invasion								
Yes	8	6 (75)	2 (25)	0.573	8.96 ± 4.29	0.489	3.92 ± 1.88	0.521
No	72	64 (88.89)	8 (11.11)		8.13 ± 1.55		2.00 ± 0.30	
Lymph node metastasis								
Yes	54	44 (81.48)	10 (18.52)	0.047	9.84 ± 1.91	0.010	2.71 ± 0.49	0.042
No	26	26 (100)	0 (0)		4.87 ± 1.93		1.19 ± 0.20	
ACJJ stage								
I-II	74	65 (87.84)	9 (12.16)	0.564	8.24 ± 1.50	0.811	2.25 ± 0.36	0.771
III-IV	6	5 (83.33)	1 (16.67)		7.99 ± 4.69		1.50 ± 0.43	
Tumor grade								
G1	26	23 (88.46)	3 (13.04)	0.857	10.26 ± 3.22	0.486	1.98 ± 0.49	0.731
G2	54	47 (87.04)	7 (12.96)		7.72 ± 1.63		2.28 ± 0.41	

[a]*Immunohistochemical staining.*
[b]*Western blot.*

expression level (2.71 ± .49 vs. 1.19 ± .20) and high-positive expression rate (18.52 vs. 0%) were significantly higher ($p < .05$) in the lymph node metastasis group than in the non-metastasis group, respectively. Other clinicopathological characteristics, such as age, sex, multifocality of PTC, extra-envelope infiltration, tumor size, and clinicopathological stage and grade, were not associated with the relative expression level of SBSN ($p > .05$; **Table 2**). The data suggest that SBSN can influence the malignancy of PTC, thus promoting lymph node metastasis. In addition, we analyzed the relationship between *SBSN* expression and prognosis in 502 PTC patients with complete clinical information in conjunction with the TCGA database and found that *SBSN* expression had no effect on the survival time of patients (**Supplementary Figure S1A**), which may be attributed to the good prognosis of PTC and the low number of deaths during follow-up (16/502).

Biological Interaction Networks of SBSN

We identified the top 20 genes associated with *SBSN* through the GeneMANIA website, which showed SBSN as the central node surrounded by 20 other nodes (**Figure 2A**). The five most relevant genes included PTEN-inducible putative kinase 1 (*PINK1*), tankyrase (*TNKS*), HAUS augmin-like complex subunit 1 (*HAUS1*), DNA polymerase kappa (*POLK*), and glycogen synthase kinase 3β (*GSK3B*), all of which physically interacted with *SBSN*. Further functional analysis showed that these genes were associated with protein hydrolysis, amino acid modification, cell adhesion, and regulation of apoptosis. In addition, using the DAVID tool and the "ggplot2" R package, we performed functional and pathway enrichment analyses of genes in the cBioPortal database that were co-expressed with SBSN to view the biological functions and pathways associated with *SBSN*. Gene Ontology enrichment analysis revealed that *SBSN* co-expressed genes were associated with a variety of processes, including regulation of signaling, regulation of cellular signal transduction and communication, and cellular responses to chemical stimuli. They were also associated with cellular components, including the neuronal fraction and plasma membrane, and molecular functions, such as enzyme binding, activation of nucleic acid binding transcription factors, and binding of some ions (**Figure 2B**). Meanwhile, the enrichment of Kyoto Encyclopedia of Genes and Genomes pathways of *SBSN*-associated genes suggested that *SBSN* was associated with the PI3K/AKT, mitogen-activated protein kinase (MAPK), and other pathways commonly found in tumors. Interestingly, the cytokine receptor interaction pathway, chemokine signaling pathway, and T-cell receptor signaling pathway, which are associated with tumor immunity, were also enriched in *SBSN* co-expressed genes (**Figure 2C**). Similar results were obtained by ssGSEA of the published thyroid dataset (TCGA), wherein samples with high *SBSN* expression levels were associated with levels of cytokines

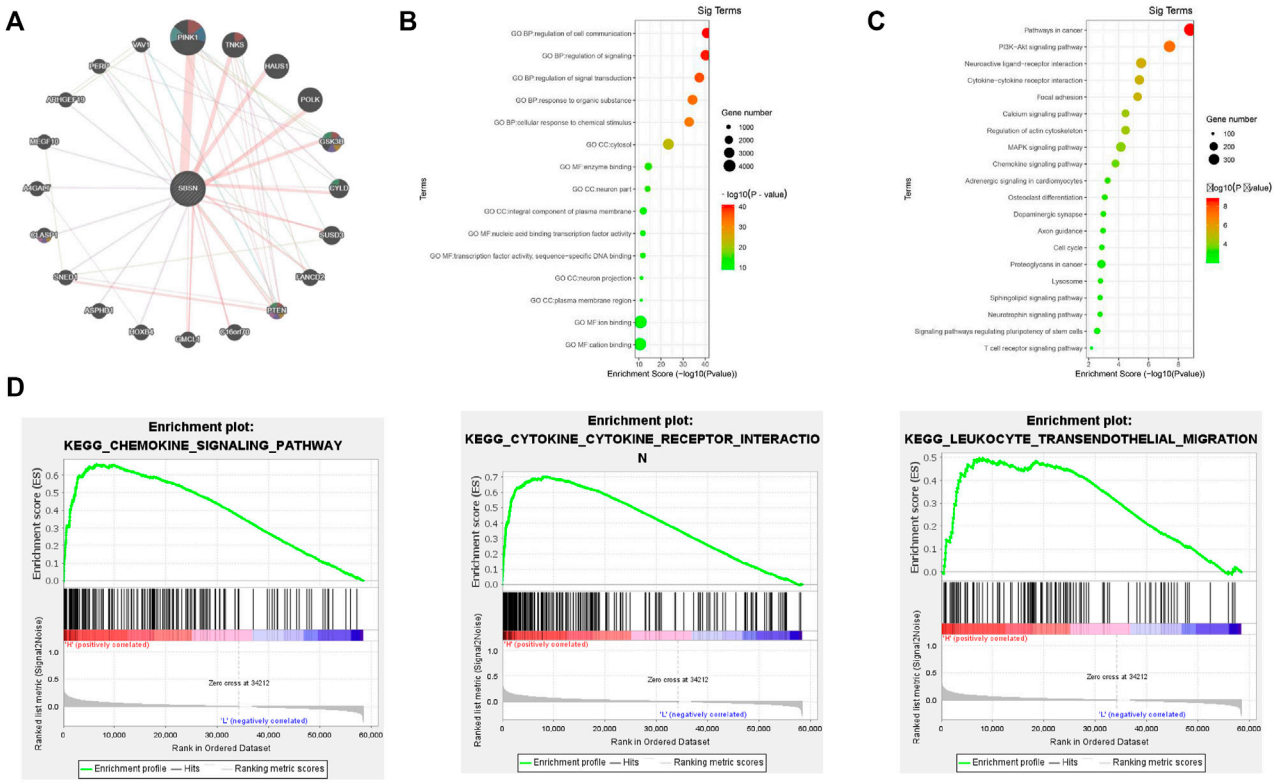

FIGURE 2 | Gene interactions and enrichment analysis of *SBSN*. **(A)** Gene interaction network of the top 20 genes associated with *SBSN* in GeneMANIA. **(B)** Gene ontology enrichment analysis by genes co-expressed with *SBSN* (top five of each type). **(C)** Kyoto Encyclopedia of Genes and Genomes enrichment analysis by genes co-expressed with *SBSN*. **(D)** Gene set enrichment analysis showing *SBSN* in the published thyroid cancer dataset (The *Cancer* Genome Atlas database).

and chemokines and more pronounced leukocyte migration (**Figure 2D**).

Correlation of SBSN Expression Levels With Immune and Stromal Cells in the TME

To assess the distribution of immune and stromal cells in PTC, PDTC, and ATC samples, we analyzed the 502 primary PTC samples from the TCGA dataset; 52 ATC samples from GSE29265, GSE33630, GSE76039, and GSE65144; and 22 PDTC samples from GSE53157 and GSE76039. Based on the median values of *SBSN* expression levels in the TCGA PTC cohort, PDTC joint cohort, and ATC joint cohort, the samples were divided into $SBSN^{high}$ and $SBSN^{low}$ groups. Based on the ESTIMATE algorithm, we found that, in the PTC samples, the immune score was higher in the $SBSN^{high}$ group than in the $SBSN^{low}$ group ($p < 0.01$), while *SBSN* expression level was negatively correlated with tumor purity ($p < 0.01$). There was no significant association between the stromal scores and SBSN expression levels (**Figure 3A**). In the ATC samples, the immune and stromal scores were higher in the $SBSN^{high}$ group than in the $SBSN^{low}$ group ($p < 0.05$), and *SBSN* expression level was negatively correlated with the extent of tumor purity ($p < 0.05$; **Figure 3B**). In the PDTC samples, no significant associations were found between *SBSN* expression levels and the immune scores, stromal scores, or degree of tumor purity (**Supplementary Figure S1B**), which may have been due to the small sample size. In conclusion, these findings suggest that *SBSN* expression may impact immune cells in the PTC TME and immune and stromal cells in the ATC TME.

Relationship Between SBSN Expression Level and Immune Cell Infiltration

To explore the correlation between *SBSN* expression and the immune microenvironment, we inferred the abundances of TIICs using CIBERSORT-ABS and xCell. The association between *SBSN* expression and TIICs in PTC and ATC is shown in **Figure 4**. In the PTC samples, the proportions of TIIC types were significantly different between patients in the $SBSN^{high}$ and $SBSN^{low}$ groups (**Figures 4A,B**). Similarly, the proportions of TIICs in the ATC samples were significantly different between the $SBSN^{high}$ and $SBSN^{low}$ groups (**Figures 4C,D**). In addition, we sought to determine whether the tumor immune microenvironment was different in patients with different *SBSN* expression levels. For the PTC samples, the results of the deconvolution algorithm CIBERSORT-ABS showed that the proportions of effector B cells ($p < 0.05$), resting $CD4^+$ memory T cells ($p < 0.001$), Tregs ($p < 0.01$), activated natural killer (NK) cells ($p < 0.05$), M0 macrophages ($p < 0.01$), M2

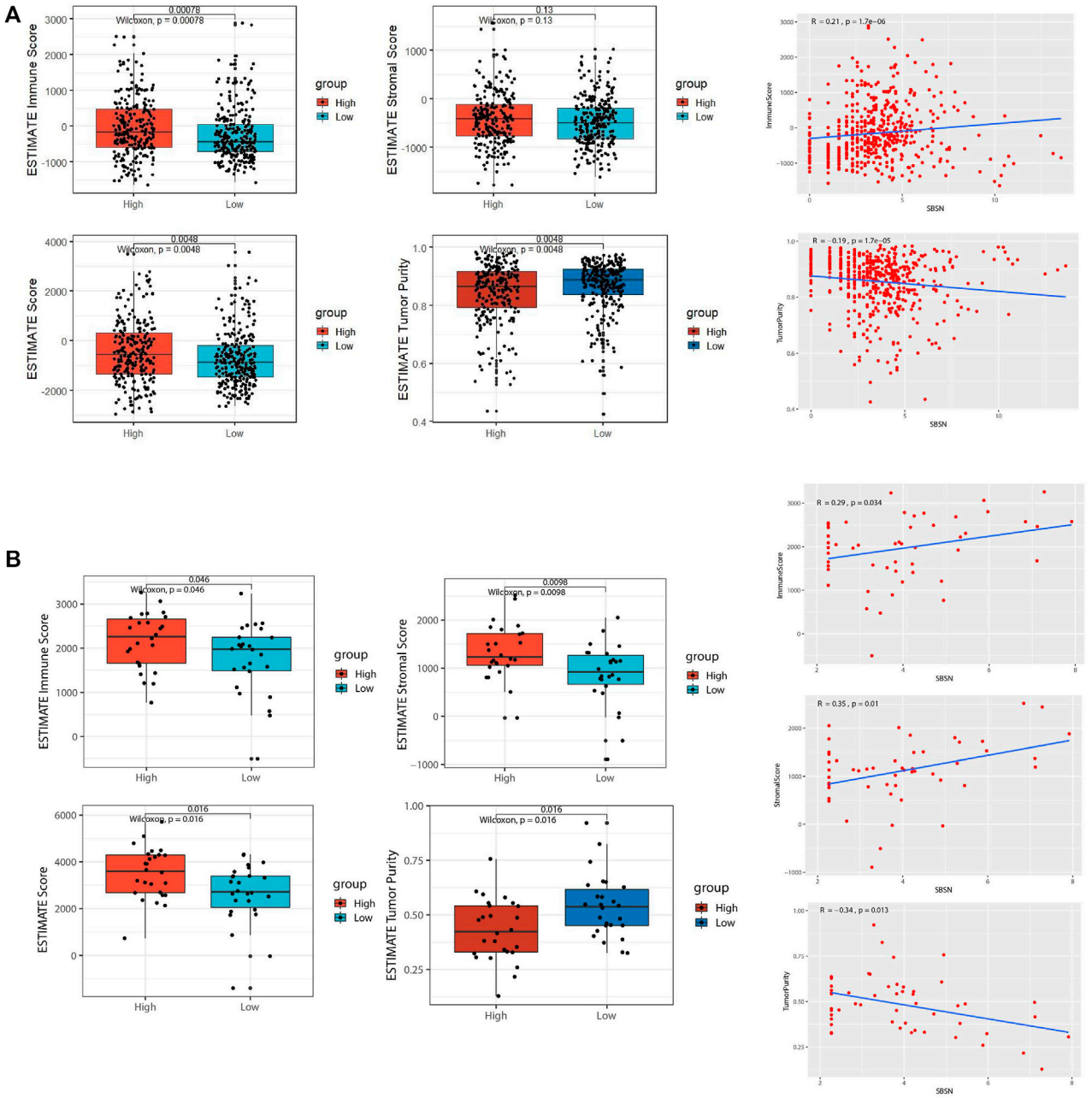

FIGURE 3 | Comparison of tumor microenvironment (TME) scores and *SBSN* expression profiles in follicular epithelial cell-derived thyroid cancer. **(A)** Comparison of the immune score, stromal score, estimate score, and tumor purity in *SBSN*^{high} and *SBSN*^{low} papillary thyroid carcinoma tissue groups as calculated by the ESTIMATE algorithm. **(B)** Comparison of the immune score, stromal score, estimate score, and tumor purity in *SBSN*^{high} and *SBSN*^{low} anaplastic thyroid carcinoma tissue groups as calculated by the ESTIMATE algorithm. Differences between *SBSN*^{high} and *SBSN*^{low} groups were compared using two-sided Wilcoxon rank-sum test.

macrophages ($p < .001$), resting dendritic cells (DCs) ($p < .05$), and activated DCs ($p < .0001$) were significantly elevated in the *SBSN*^{high} group (**Figure 5A**). The results of the xCell algorithm showed that 25 out of 64 noncancerous cell types were correlated and 39 cell types were not correlated with *SBSN* expression (**Supplementary Figure S2**). Among the former, 15 types had higher proportions in the *SBSN*^{high} group, and 10 types had higher proportions in the *SBSN*^{low} group. Numbers of activated DCs ($p < .01$), B cells ($p < .01$), conventional DCs ($p < .01$), DCs ($p < .01$), immature DCs (iDCs) ($p < .01$), macrophages ($p < .05$), M2 macrophages ($p < .05$), monocytes ($p < .01$), and Tregs ($p < .01$) were all significantly elevated in the *SBSN*^{high} group, while those of central memory CD4$^+$ T cells ($p < .01$) and CD8$^+$ naïve T cells ($p < .05$) were significantly elevated in the *SBSN*^{low} group. In addition, some other cell types, such as keratin-forming cells ($p < .01$), epithelial cells ($p < .01$), platelets ($p < .05$), astrocytes

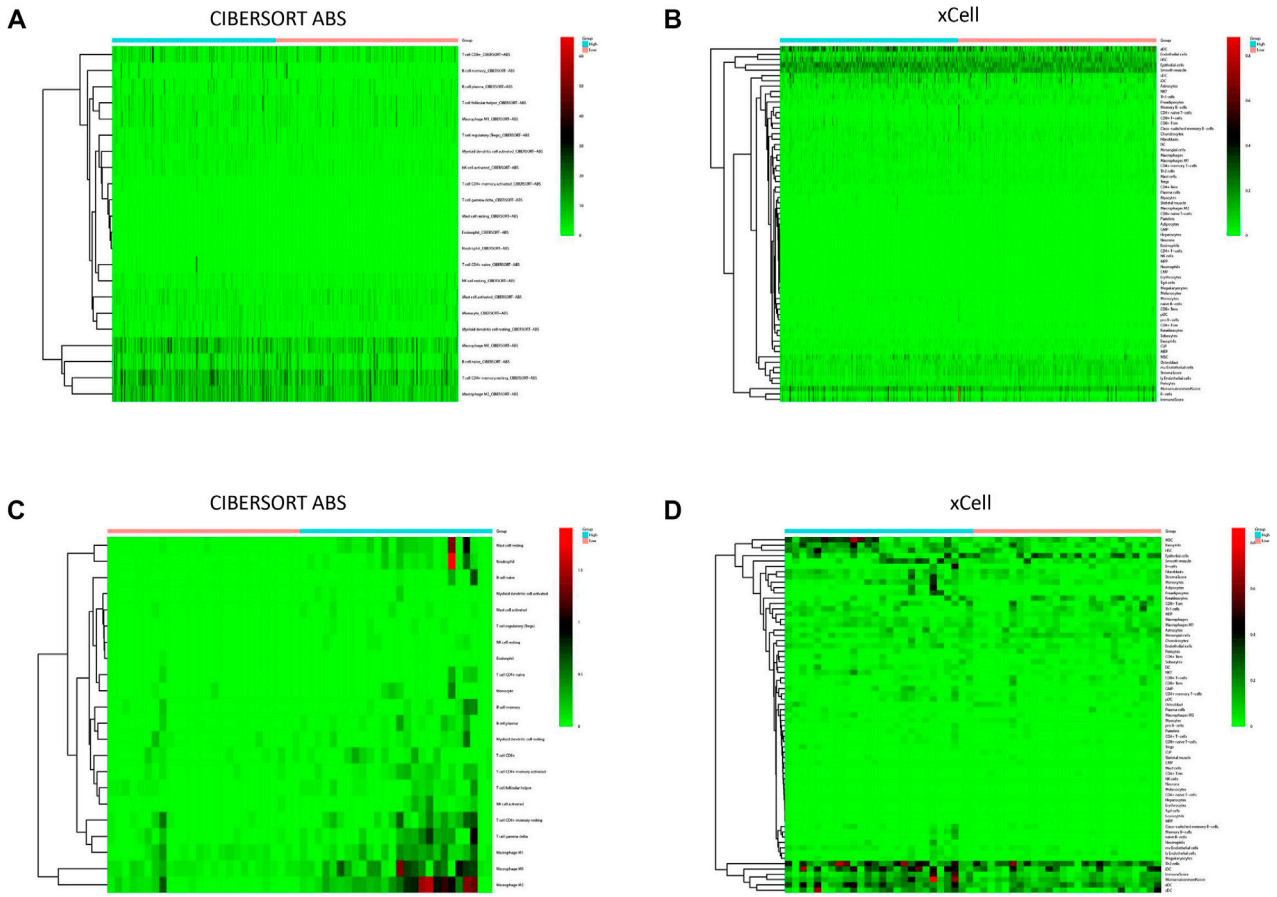

FIGURE 4 | Immune cell infiltration in patients with papillary thyroid carcinoma (PTC) and anaplastic thyroid carcinoma (ATC). **(A)** Heat map of tumor-infiltrating immune cell (TIIC) fractions, quantified by CIBERSORT-ABS, in PTC. **(B)** Heat map of TIIC fractions, quantified by xCell, in PTC. **(C)** Heat map of TIIC fractions, quantified by CIBERSORT-ABS, in ATC. **(D)** Heat map of TIIC fractions, quantified by xCell, in ATC.

($p < .01$), mesangial cells ($p < .01$), and sebocytes ($p < .01$), were more prevalent in the *SBSN*^{high} group; meanwhile, endothelial cells ($p < .01$), ly endothelial cells ($p < .01$), hematopoietic stem cells (HSC) ($p < .01$), megakaryocyte–erythroid progenitor (MEP) ($p < .01$), multinucleated variant endothelial cells ($p < .01$), neurons ($p < .05$), osteoblasts ($p < .01$), and pericytes ($p < .05$) were more predominant in the *SBSN*^{low} group (**Figure 5B**).

Similar to the results of ESTIMATE analysis, the CIBERSORT-ABS results showed no correlation between the immune cell content in PDTC and *SBSN* expression (**Figure 6A**), while xCell analysis yielded higher proportions of only basophils ($p < .05$), hematopoietic stem cells ($p < .05$), myocytes ($p < .01$), smooth muscle cells ($p < .05$), and T helper (Th) cell type 2 (Th2) cells ($p < .05$) in the *SBSN*^{high} group. The *SBSN*^{low} group had a higher proportion of common lymphoid progenitors ($p < .05$) as well as naïve B cells ($p < .05$) (**Figure 6B**).

For the ATC samples, the CIBERSORT-ABS results showed that effector B cells ($p < .001$), resting CD4$^+$ memory T cells ($p < .05$), activated CD4$^+$ memory T cells ($p < .01$), follicular helper T cells ($p < .01$), Tregs ($p < .05$), resting NK cells ($p < .05$), activated NK cells ($p < .05$), macrophages ($p < .01$), M1 macrophages ($p < .01$), M2 macrophages ($p < .0001$), resting mast cells ($p < .01$), and neutrophils ($p < .0001$) were significantly more numerous in the *SBSN*^{high} group (**Figure 7A**). The xCell algorithm showed that 22 out of the 64 noncancerous cell types were correlated and 42 cell types were not correlated with *SBSN* expression (**Supplementary Figure S3**). Among the 22 cell types, 17 had higher proportions in the *SBSN*^{high} group and five in the *SBSN*^{low} group. Proportions of activated DCs ($p < .05$), basophils ($p < .01$), iDCs ($p < .05$), macrophages ($p < .05$), M2 macrophages ($p < .01$), mast cells ($p < .01$), monocytes ($p < .01$), naïve B cells ($p < .05$), and neutrophils ($p < .05$) were significantly elevated in the *SBSN*^{high} group. Some other cell types, such as fibroblasts ($p < .05$), granulocyte–macrophage lineage progenitors ($p < .01$), hematopoietic stem cells ($p < .01$), megakaryocytes ($p < .01$), multinucleated variant endothelial cells ($p < .05$), myocytes ($p < .05$), preadipocytes ($p < .05$), and skeletal muscle cells ($p < .05$) were also more common in the *SBSN*^{high} group. Meanwhile, epithelial cells ($p < .05$), keratin-forming cells ($p < .05$), megakaryocyte–erythroid progenitor cells ($p < .05$), mesangial cells ($p < .05$), and pericytes ($p < .05$) were more highly represented in the *SBSN*^{low} group (**Figure 7B**). Taken together,

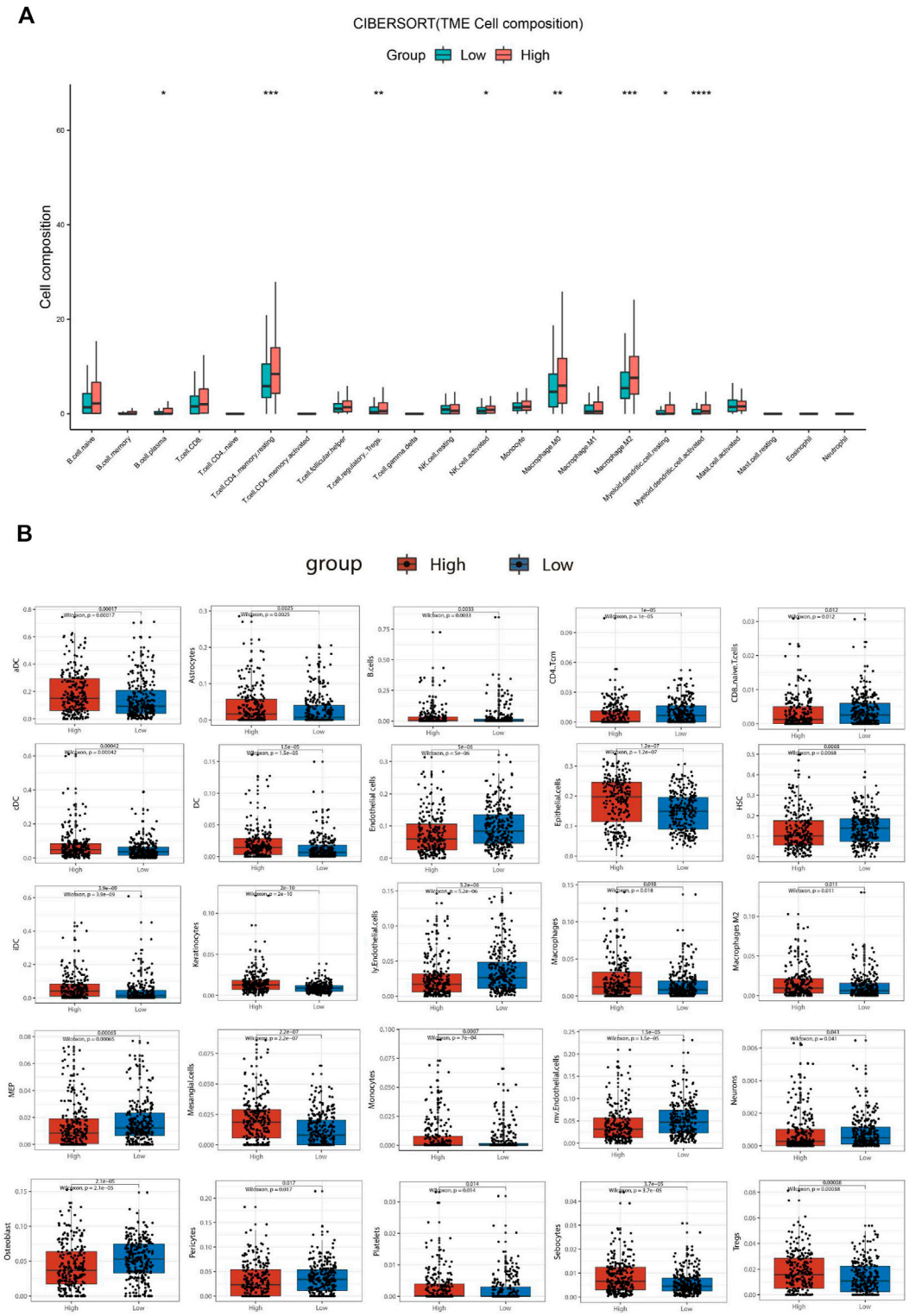

FIGURE 5 | Comparison of different non-cancer cell proportions, quantified by CIBERSORT-ABS algorithm and xCell algorithm, in *SBSN*^{high} and *SBSN*^{low} papillary thyroid carcinoma groups. **(A)** Comparison of different immune cell proportions, quantified by the CIBERSORT-ABS algorithm, in the *SBSN*^{high} and *SBSN*^{low} groups. **(B)** Comparison of different non-cancer cell proportions, quantified by the xCell algorithm, in the *SBSN*^{high} and *SBSN*^{low} groups. Differences between *SBSN*^{high} and *SBSN*^{low} groups were compared using the two-sided Wilcoxon rank-sum test. *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$.

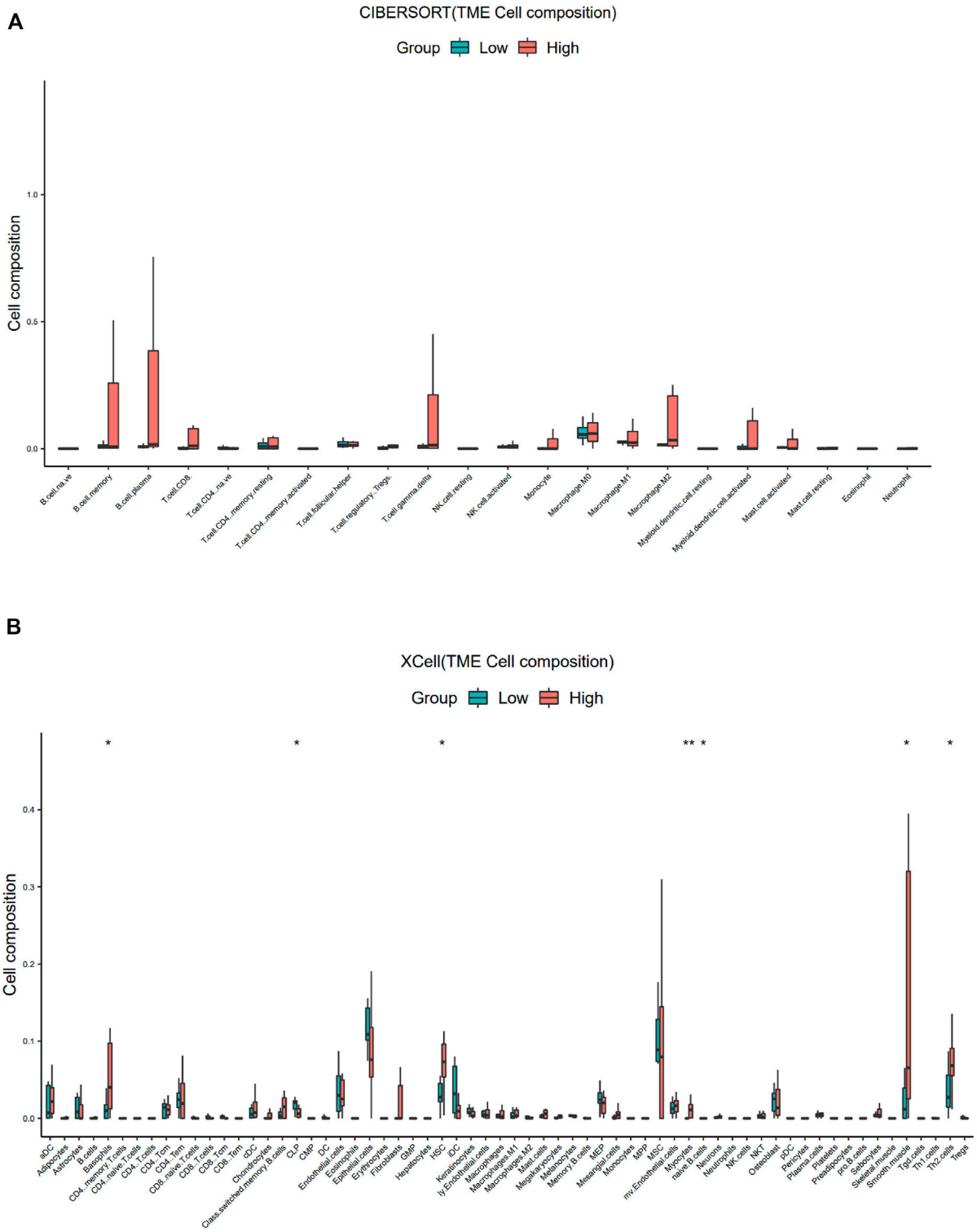

FIGURE 6 | Comparison of different non-cancer cell proportions, quantified by CIBERSORT-ABS algorithm and xCell algorithm, in *SBSN*^{high} and *SBSN*^{low} poorly differentiated thyroid carcinoma groups. **(A)** Comparison of different immune cell proportions, quantified by the CIBERSORT-ABS algorithm, in the *SBSN*^{high} and *SBSN*^{low} groups. **(B)** Comparison of different non-cancer cell proportions, quantified by the xCell algorithm, in the *SBSN*^{high} and *SBSN*^{low} groups. Differences between *SBSN*^{high} and *SBSN*^{low} groups were compared using the two-sided Wilcoxon rank-sum test. *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$.

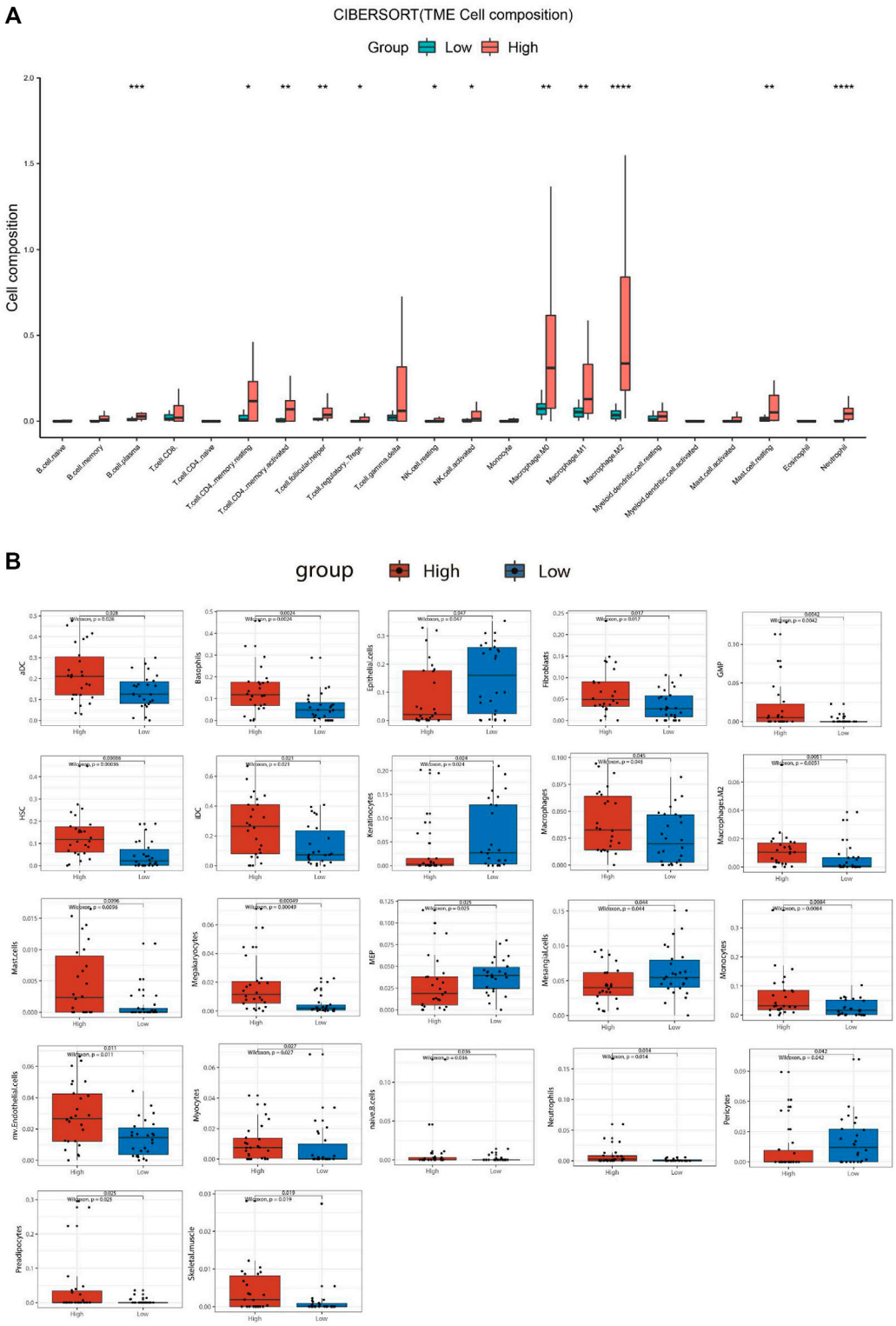

FIGURE 7 | Comparison of different non-cancer cell proportions, quantified by CIBERSORT-ABS algorithm and xCell algorithm, in *SBSN*^{high} and *SBSN*^{low} anaplastic thyroid carcinoma groups. **(A)** Comparison of different immune cell proportions, quantified by the CIBERSORT-ABS algorithm, in the *SBSN*^{high} and *SBSN*^{low} groups. **(B)** Comparison of different non-cancer cell proportions, quantified by the xCell algorithm, in the *SBSN*^{high} and *SBSN*^{low} groups. Differences between *SBSN*^{high} and *SBSN*^{low} groups were compared using the two-sided Wilcoxon rank-sum test. *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$.

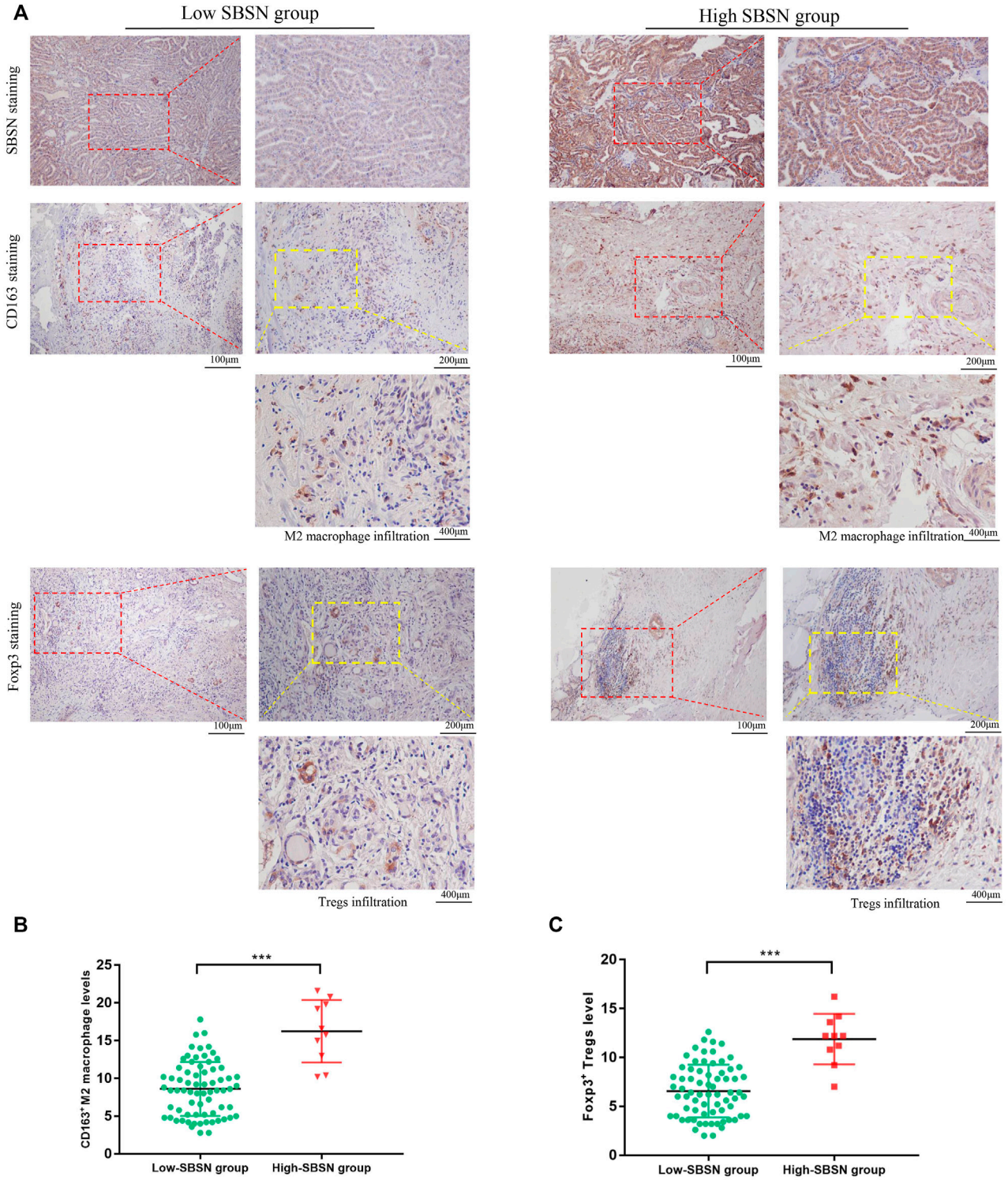

FIGURE 8 | The correlation of SBSN expression with CD163$^+$ M2 macrophages and Foxp3$^+$ T regulatory cells (Tregs) in patients with papillary thyroid carcinoma (PTC) was analyzed by immunohistochemical staining. **(A)** Immunohistochemical staining of SBSN, CD163, and Foxp3 detected M2 macrophage and Treg infiltration in SBSN low-positive expression and SBSN high-positive expression groups. **(B)** Scatter-plot of CD163$^+$ M2 macrophage density in the low- and high-positive SBSN expression PTC groups. **(C)** Scatter-plot of Foxp3$^+$ Treg density in the low- and high-positive SBSN expression PTC groups. Results are presented as means ± standard deviation. ***$p < .001$.

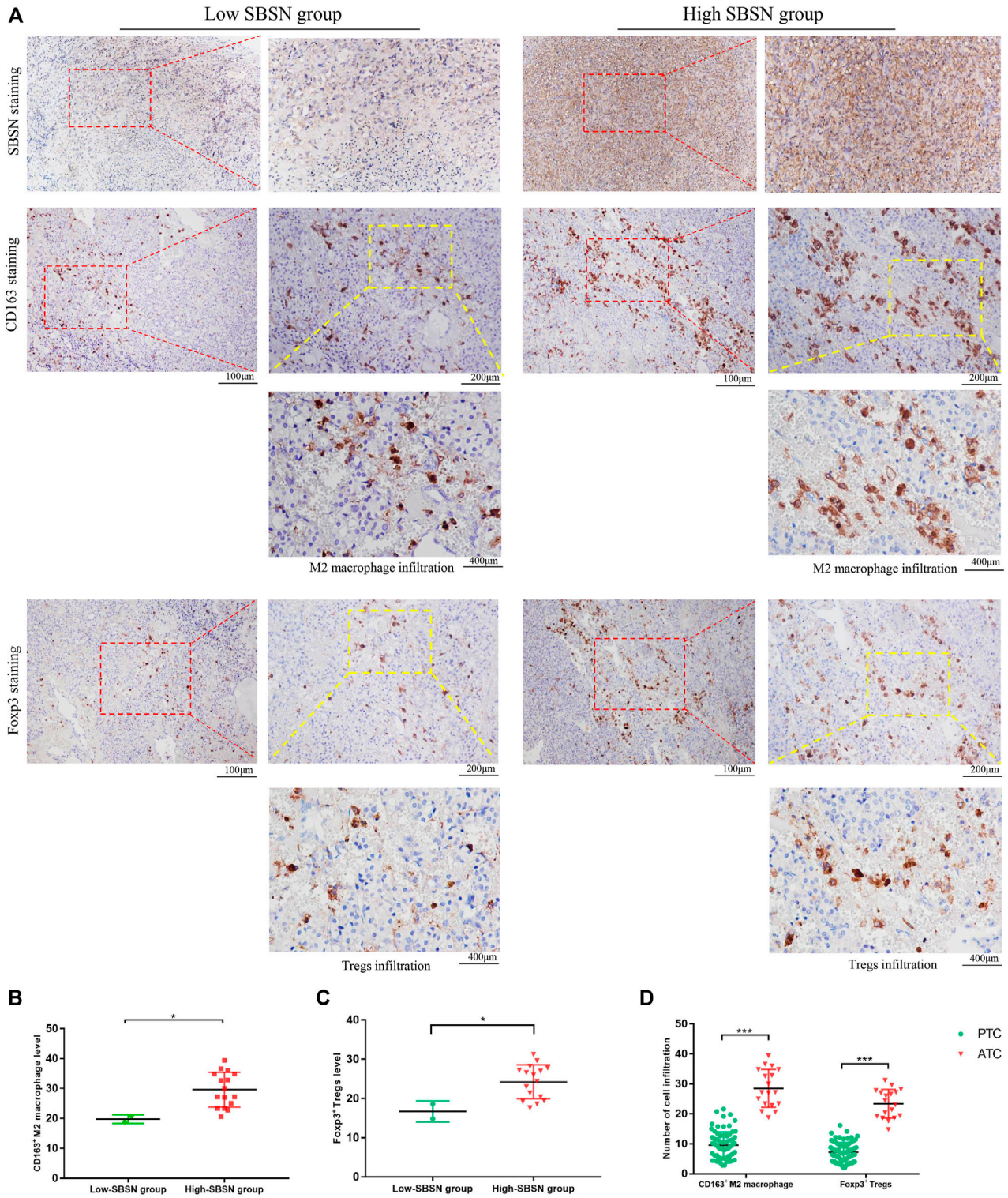

FIGURE 9 | The correlation of SBSN expression with CD163$^+$ M2 macrophages and Foxp3$^+$ T regulatory cells (Tregs) in patients with anaplastic thyroid carcinoma (ATC) was analyzed by immunohistochemical staining. **(A)** Immunohistochemical staining of SBSN, CD163, and Foxp3 detected M2 macrophage and Treg infiltration in SBSN low-positive expression and SBSN high-positive expression groups. **(B)** Scatter-plot of CD163$^+$ M2 macrophage density in the low- and high-positive SBSN expression ATC groups. **(C)** Scatter-plot of Foxp3$^+$ Treg density in the low- and high-positive SBSN expression ATC groups. **(D)** Scatter-plots of CD163$^+$ M2 macrophage and Foxp3$^+$ Treg densities in papillary thyroid carcinoma (PTC) and ATC. Results are presented as means ± standard deviation. $^*p < .05$, $^{***}p < .001$.

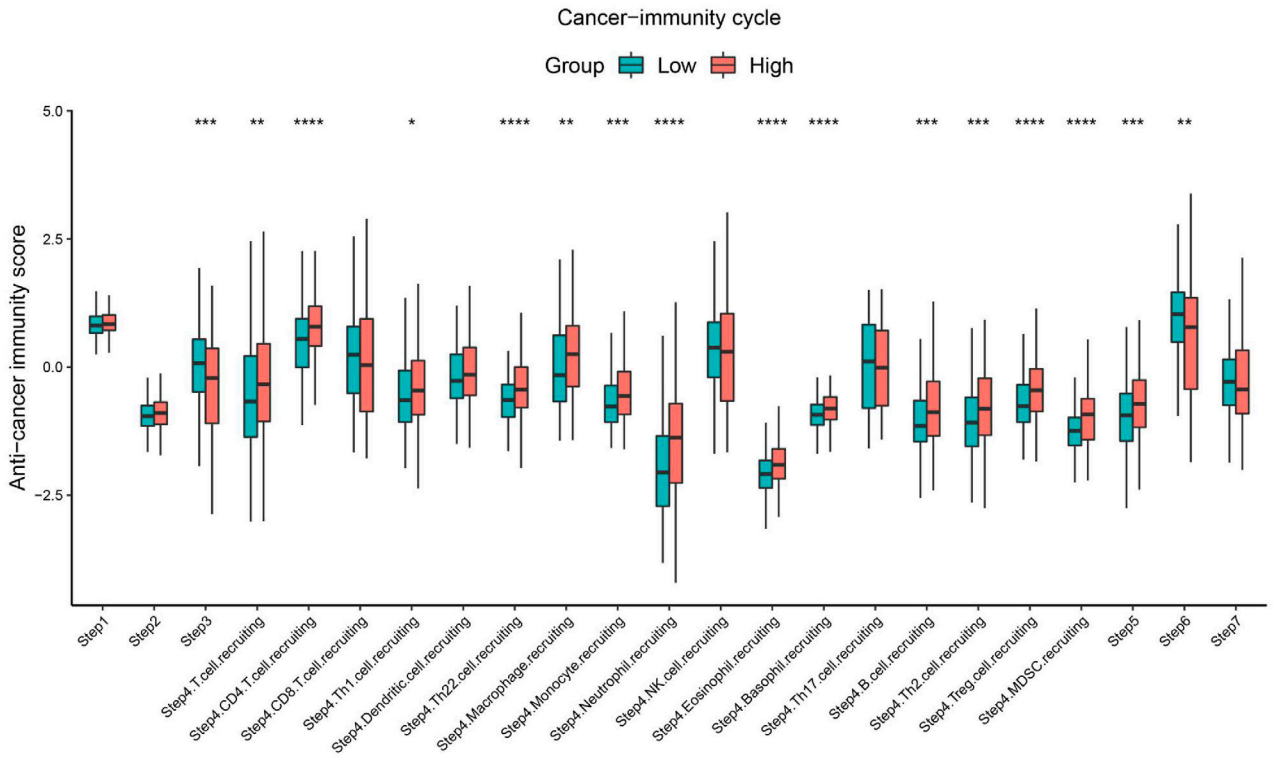

FIGURE 10 | Comparison between the *SBSN*^{high} and *SBSN*^{low} groups in the scores for the cancer immunity cycle in PTC. The thick line represents the median value within each group, and differences between the *SBSN*^{high} and *SBSN*^{low} groups were compared using the two-sided Wilcoxon rank-sum test. *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$.

these results suggest that SBSN can regulate different types of tumor-associated cells in PTC as well as in the ATC TME.

Correlation of SBSN Expression Levels in PTC and ATC Tumor Tissues With the Densities of CD163$^+$ M2 Macrophages and Foxp3$^+$ Tregs

To confirm the results obtained by comprehensive bioinformatics analysis, we performed immunohistochemical staining of tumor tissues from 80 patients with PTC and 18 patients with ATC for the M2 macrophage marker CD163 and the Treg marker Foxp3. CD163+M2 macrophages and Foxp3+Tregs were mainly distributed in the interstitium of tumor tissues and lymphocyte aggregates in PTC (**Figure 8A**) and mainly scattered around cancer cells in ATC (**Figure 9A**). The results showed that, in PTC tissues, the infiltration levels of CD163+M2 macrophages (16.24 ± 4.315) and Foxp3+Tregs (11.88 ± 2.581) were higher in the SBSN high-positive expression group than in the SBSN low-positive expression group (8.606 ± 3.566 and 6.566 ± 2.69, respectively; $p < .01$) (**Figures 8B,C**). In ATC tissues, the infiltration levels of CD163+M2 macrophages (29.6 ± 5.791) and Foxp3+Tregs (24.23 ± 4.318) were higher in the SBSN high-positive expression group than in the SBSN low-positive expression group (19.8 ± 1.414 and 16.7 ± 2.687, respectively; $p < .05$) (**Figures 9B,C**). In addition, the results of staining of CD163+M2 macrophages and Foxp3+Tregs in 80 cases of PTC and 18 of ATC showed that the total infiltration levels of CD163+M2 macrophages (28.51 ± 6.305) and Foxp3+Tregs (23.39 ± 4.775) were higher in ATC than in PTC (9.56 ± 4.417 and 7.23 ± 3.194, respectively; $p < .01$) (**Figure 9D**).

Association of *SBSN* Expression With the Cancer Immunity Cycle

The cancer immunity cycle involves a series of steps that enable the anti-cancer immune response to kill cancer cells effectively: tumor cell release of antigen (step I), tumor antigen presentation (step II), T cell activation (step III), T cell migration to tumor tissue (step IV), tumor tissue T cell infiltration (step V), T cell recognition of tumor cells (step VI), and clearance of tumor cells (step VII) (Chen and Mellman, 2013). The TIP pipeline was used to estimate the activity score of SBSN in the seven-step cancer immunity cycle in PTC samples. The results showed that the *SBSN*^{high} group had higher scores ($p < .05$) than those of the *SBSN*^{low} group in the processes of T cell recruitment and infiltration of recruited T cells into tumor tissue. These included CD4$^+$ T cells, Th1 cells, Th22 cells, macrophages, monocytes, neutrophils, eosinophils, basophils, B cells, Th2 cells, Tregs, and bone

marrow-derived suppressor cells (MDSCs). However, the *SBSN*^{high} group had lower anti-cancer immune scores ($p < .05$) during T-cell activation and T-cell recognition of tumor cells (**Figure 10**). Overall, these results suggest a possible role of *SBSN* in the cancer immunity cycle of PTC.

DISCUSSION

This study aimed to predict the role of SBSN in follicular epithelial cell-derived thyroid cancer using an extensive bioinformatics data mining approach to explore the relationship between SBSN expression levels and the extent of infiltration by different immune cells. To our knowledge, the present study is the first to highlight the relationship between SBSN expression and immune infiltration in thyroid cancer, providing new insights into the role of SBSN in cancer-associated immune regulation and its application as a cancer biomarker.

SBSN was originally identified in epithelial tissues and is thought to be involved in the process of epidermal differentiation under physiological conditions (Park et al., 2002). In recent years, accumulating evidence has indicated that SBSN plays an important role in the development of a variety of tumors. Hypomethylation of the *SBSN* promoter leads to elevated *SBSN* mRNA levels in NSCLC, and aberrant SBSN expression promotes the proliferation of lung squamous cell lines (Glazer et al., 2009). Similarly, in salivary ACC, the CpG island is hypomethylated in *SBSN*, and knockdown of *SBSN* inhibits the proliferation and invasive ability of ACC cell lines (Shao et al., 2012). In ESCC, high SBSN expression level shortens patient survival, and overexpression of SBSN enhances the proliferation and invasive ability of esophageal cancer cells, while *SBSN* knockdown does the opposite (Zhu et al., 2016). In malignant brain tumors, SBSN upregulation is associated with poor prognosis for patients with glioblastoma multiforme (Formolo et al., 2011). Tumor development cannot be separated from angiogenesis, and SBSN is involved in the migration of tumor endothelial cells and angiogenesis through AKT activation (Alam et al., 2014; Takahashi et al., 2020). All of these previous findings suggest that *SBSN* is involved as an oncogene in tumor progression.

In our study, the possible role of SBSN in follicular epithelial cell-derived thyroid cancer and the involved regulatory mechanisms were explored. SBSN was highly expressed in PTC at the protein and mRNA levels. SBSN expression was significantly associated with cervical-lymph-node metastasis in PTC ($p < .05$). These results suggest that SBSN expression may be an indicator to assess the aggressiveness of PTC. Importantly, SBSN expression levels increased with decreasing extent of differentiation and increasing rates of malignancy in follicular epithelial cell-derived thyroid cancer; high-positive expression levels were most pronounced in ATC. This finding indicates that SBSN may be involved in regulating the malignancy of tumors and that its expression level may be related to the degree of tumor malignancy, accurately reflecting the progression of tumor malignancy.

We used two types of enrichment analysis to further elucidate the potential mechanisms of *SBSN*'s involvement in follicular-derived thyroid cancer. *SBSN* and its co-expressed genes were associated with a variety of biological processes, cellular components, and molecular functions, especially signaling, regulation, and responses to stimuli. *SBSN* was also found to be associated with several cancer signaling pathways, such as the PI3K/AKT and MAPK pathways; this result is consistent with those of previous studies (Alam et al., 2014; Takahashi et al., 2020). Interestingly, SBSN expression was also significantly enriched in several cancer immunity-related pathways, such as the chemokine pathway, the cytokine and its receptor interaction pathway, and the T-cell receptor signaling pathway, which are usually involved in tumorigenic processes (Liu T. et al., 2012; Li and Rudensky, 2016). These results suggest that *SBSN* may play complex roles in multiple biological processes.

A deeper analysis of the complexity within the TME may help identify patient populations with the potential to respond to current immune checkpoint therapies and may contribute to the identification of new adjuvant therapeutic targets (Lin et al., 2019). Previously, several studies have reported that elevated immune scores are associated with poor prognosis in patients with different cancers, such as renal cell carcinoma and osteosarcoma (Xu et al., 2019; Zhang et al., 2020), while stromal cells are also thought to play an important role in tumor growth, disease progression, and drug resistance (Denton et al., 2018). Therefore, we explored the relationship between TME scores and SBSN expression using the ESTIMATE algorithm. Our study is the first to demonstrate that, in PTC, SBSN expression level was positively correlated with immune scores and negatively correlated with tumor purity, while no significant association was observed between stromal scores and SBSN expression levels. In ATC, SBSN expression level was significantly positively correlated with immune and stromal scores and negatively correlated with tumor purity. These results suggest that SBSN may contribute to increasing numbers of immune and stromal cells in the TME, which diminishes tumor purity. Indeed, the elevated immune and stromal scores in the TME may indicate either protection or a poor prognosis for the host depending on the type of immune cells that infiltrate the tumor and the specific roles they play in tumor development (Cunha et al., 2014).

The types and relative proportions of TIICs in the TME may correlate with the clinical prognosis of patients (Orhan et al., 2020). Based on the functional enrichment analysis of SBSN and the correlation of its expression with immune and stromal cells in the TME, we further explored the relevance of SBSN expression to immune cell infiltration. Large amounts of RNA-seq data have enabled algorithms that employ the deconvolution principle or gene markers to map the TME of samples. Since there is no gold standard for inferring immune infiltration from RNA-seq data, we selected two algorithms, based on the deconvolution principle and the gene marker principle, to infer the composition of TIICs and confirmed their correlation by immunohistochemical staining. The proportion of M2 macrophages was higher in the *SBSN*^{high} group in both the PTC and ATC samples. The proportion of M2 macrophages in the *SBSN*^{high} group was much

higher than that in the *SBSN*^{low} group and higher than that of M1 macrophages in the *SBSN*^{high} group in the ATC samples, suggesting that macrophages may be more polarized to the M2 phenotype in ATC. In most tumors, M2 macrophages are present as immunosuppressive cells, which can release growth factors to promote tumor development (Ho et al., 2016). More importantly, M2 macrophages tend to promote neoangiogenesis as well as stromal activation and remodeling (Afik et al., 2016), thus positively influencing cancer progression and negatively affecting patient prognosis (Tiainen et al., 2015). PTC tissues have been shown to have high levels of immunity, especially due to M2 macrophages, and compared with early PTC, advanced PTC exhibits a higher degree of immune infiltration and a higher proportion of M2 macrophages, which accelerate tumor cell migration (Zhang et al., 2021), producing a pro-cancer effect and exacerbating immune escape (Xie et al., 2020). Therefore, a high expression level of SBSN in patients with PTC and ATC may accelerate tumor progression by stimulating the polarization of M2 macrophages.

Furthermore, both CIBERSORT-ABS and xCell results showed that SBSN expression was positively correlated with the percentages of multiple DCs, including iDCs, and Tregs in PTC samples. Similarly, in ATC tissues, the percentage of iDCs was positively correlated with SBSN expression. The numbers of Tregs and DCs have been shown to increase in PTC tissues (Yu et al., 2013). Usually, tumor-infiltrating DCs exhibit an immature phenotype, resulting in altered antigen presentation (Tran Janco et al., 2015). iDCs do not effectively induce T- and NK-cell-mediated immune responses and can even suppress immune responses by producing suppressive cytokines, such as IL-10 and TGF-β (Scouten and Francis, 2006). In addition, Tregs, as well as TAMs, alter the normal replication of endothelial cells by creating a hypoxic environment in the tumor tissue and can achieve immunosuppressive and escape effects by inhibiting the antigen presentation by DCs and the activation of $CD8^+$ T cells in the tumor (Facciabene et al., 2017; Jang et al., 2017). The proportion of Tregs in PTC also correlates with lymph node metastasis and extrathyroidal expansion (French et al., 2010), which is consistent with our results. These findings suggest that patients with PTC and ATC with high SBSN expression levels may benefit from targeting SBSN to reduce the proportions of Tregs and iDCs.

In addition, SBSN expression levels in ATC samples were positively correlated with the proportions of mast cells and neutrophils, unlike those in PTC. The role of neutrophils in cancer is still controversial, but in recent years, thyroid cancer cells have been shown to be able to recruit neutrophils by releasing CXCL8/IL-8, improve their own survival by releasing granulocyte colony-stimulating factor, enhancing the pro-inflammatory response of neutrophils, and upregulate the expression of pro-oncogenic factors (Galdiero et al., 2018). Regarding mast cells in tumors, a previous study showed that the presence of mast cells in ATC correlates with tumor aggressiveness and that mast cells induce epithelial-mesenchymal transition in thyroid cancer cell lines, mainly by activating CXCL8 in the AKT/SLUG pathway (Visciano et al., 2015). These results suggest that SBSN may be involved in the regulation of multiple infiltrating immune cells in the process of tumor development and contribute to tumor progression.

Interestingly, as suggested by the ESTIMATE algorithm, our exploration in PDTC yielded different results from those for ATC and PTC. CIBERSORT-ABS and xCell results showed a low level of immune cell infiltration in PDTC, and most immune cells and stromal cells did not correlate with *SBSN* expression. This is partly due to the small sample size of PDTC compared with those of PTC and ATC. However, we found similar results in the study by Giannini et al. (2019) who found that PDTC showed poor or absent immune cell infiltration compared with that in ATC and PTC, that the degree of immune cell infiltration in PDTC was even lower than that in normal thyroid tissue, and that, in most cases, PDTC appeared as non–T-cell-inflamed "cold" tumors.

Furthermore, our results showed that, in the xCell analysis of PTC, compared with the *SBSN*^{high} group, the *SBSN*^{low} group had a higher proportion of central memory $CD4^+$ T cells as well as $CD8^+$ naïve T cells. Central memory $CD4^+$ T cells were reported to have stronger anti-tumor capacity compared with that of effector memory $CD4^+$ T cells (Klebanoff et al., 2005). They have a strong self-renewal and replication ability and not only survive for a long time *in vivo* but also can be efficiently expanded *in vitro* to ensure the number of T cells returned for infusion, which can play a long-term anti-tumor role (Zhou et al., 2005; Berger et al., 2008). Naïve $CD8^+$ T cells will differentiate into many effector cells when encountering antigens such as tumor cells, and these effector cells migrate to the corresponding sites to produce antitumor effects (Brummelman et al., 2018). The decrease in these cellular components coupled with an increase in *SBSN* expression indicates that the *in vivo* anti-tumor capacity is weakened and that tumor cells have a greater chance to develop immune escape ability.

Our results also demonstrated that some blood cells showed differential expression in the *SBSN*^{high} and *SBSN*^{low} groups. HSCs were able to promote tumor growth and progression in the solid TME (Hassan and Seno, 2020). However, the correlation between *SBSN* expression and HSCs in PTC and ATC produced opposite results, which may be attributed to the difference in the degree of malignancy of the two subtypes. Compared with PTC, ATC usually exhibits strong invasive and metastatic abilities, which are dependent on the supply of hematopoietic cells, reflecting the pro-carcinogenic role of *SBSN* in ATC. We also observed that the content of MEPs in ATC and PTC decreased with the expression of *SBSN*. Usually, in solid tumors or leukemia, MEPs can be transformed into erythroblast-like cells, erythrocytes, or megakaryocytes, thus promoting tumor progression (Wickrema and Crispino, 2007; Han et al., 2018). This implies that elevated *SBSN* expression may lead to increased conversion of MEPs into the aforementioned cells and may promote tumor progression.

In addition, studies on the dedifferentiation of thyroid cancer have gained interest in recent years. A previous study showed that immune scores are significantly negatively correlated with thyroid cancer differentiation scores (Na and Choi, 2018) and that the least differentiated ATC usually has higher stromal and immune scores than those for highly differentiated PTC (Cunha et al., 2021), which suggests that the immune microenvironment

may be involved in the process of thyroid cancer dedifferentiation. In our study, the xCell results showed that SBSN expression in ATC was positively correlated with fibroblast content in tumors, while no such relationship was found in PTC. Wen et al. (2021) found that cancer-associated fibroblast (CAF) content was significantly higher in ATC than in PTC or normal thyroid tissue. The content of CAFs was positively correlated with dedifferentiation, aggressiveness, and poor prognosis of thyroid cancer, which suggests that SBSN may promote dedifferentiation processes and contribute to poor outcomes of thyroid cancer by regulating the content of CAFs. In addition, M2 macrophages can activate the Wnt/β-catenin pathway by secreting Wnt1 and Wnt3a, participating in the dedifferentiation, migration, and proliferation of invasive thyroid cancer cells (Lv et al., 2021). Our study indicated that SBSN expression level was positively correlated with the numbers of M2 macrophages in both PTC and ATC; however, compared with that in PTC, the proportion of M2 macrophages in the $SBSN^{high}$ ATC group was significantly higher than that in the $SBSN^{low}$ group. These findings suggest that SBSN may promote the dedifferentiation and aggressiveness of thyroid cancer cells by regulating the content of M2 macrophages, especially in ATC.

Regarding the cancer immunity cycle, patients with PTC with high SBSN expression levels scored higher in T-cell migration to tumor tissue (step IV) and tumor tissue T-cell infiltration (step V). However, according to the results of the cancer immunity cycle, most of the recruited and infiltrated immune cells, such as TAMs, neutrophils, and Tregs, have immunosuppressive effects on the tumor. There are also some helper T cells, such as Th2 cells, whose secreted cytokines can mediate the polarization of macrophages into the M2 phenotype (Shapouri-Moghaddam et al., 2018), and the increase in Th22 cell number has been reported to be associated with the progression of gastric cancer (Liu Z. et al., 2012). In addition, the recruitment of monocytes and MDSCs is positively correlated with SBSN expression in the cancer immunity cycle. Monocytes have been shown to be direct precursors of HSC-derived macrophages, and upon recruitment to tumor tissue, they can differentiate into TAMs and support tumorigenesis, local progression, and distant metastasis (Richards et al., 2013). MDSCs are rare in healthy subjects, but their numbers are elevated in patients with cancer, in whom they show a potent immunosuppressive potential and are associated with a poor prognosis (Marvel and Gabrilovich, 2015). Furthermore, patients with PTC with high SBSN expression levels showed lower scores in processes such as T cell activation and T cell recognition of tumor cells, which might be due to higher levels of iDCs and Tregs in their tissues, as DCs with an immature phenotype are usually unable to fully activate T cells (Yu et al., 2013) and Tregs also suppress the activation of $CD8^+$ T cells (Jang et al., 2017). In addition, iDCs have abnormally altered antigen-presenting functions (Tran Janco et al., 2015), and Tregs in tumors also suppress antigen-presenting functions of DCs, leading to impaired T-cell recognition (Jang et al., 2017). These results imply that SBSN may suppress T-cell activation and inhibit T-cell recognition by affecting the levels of DCs and Tregs and may contribute to the immune escape of tumor cells.

There are several limitations to the current study. First, more thyroid cancer tissue samples are needed to validate the relationships between SBSN expression and immune and stromal cells in the TME and further explore the correlation between SBSN expression and immune cell infiltration. Second, there is no canonical method to analyze infiltrating immune cells in the TME, and we used two different methods that are based on different principles. Thus, additional studies are needed to elucidate the mechanism of SBSN's effects on immune cell infiltration in thyroid cancer. Third, RNA-seq-based algorithms may not be sufficiently accurate; overcoming this limitation requires *in vivo* models to explore the underlying biological mechanisms of SBSN's effects and its interaction with tumor immunity in thyroid cancer.

CONCLUSION

In conclusion, our study showed that SBSN is highly expressed in follicular epithelial cell-derived thyroid cancer. The expression level of SBSN increases with decreasing extent of cancer cell differentiation and is associated with lymph node metastasis in patients with PTC, which can be used as a potential biomarker for follicular epithelial cell-derived thyroid cancer. In addition, SBSN can influence the cancer immunity cycle and promote thyroid cancer dedifferentiation by regulating the level of immune cell infiltration in the TME. This suggests that *SBSN* may be a therapeutic target whose inhibition may promote anti-tumor immune response. Thus, a comprehensive understanding of the relationship between SBSN expression and immune infiltration may provide new insights into the immunotherapy of thyroid cancer.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics Committee of the Shengjing Hospital, China Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HT and ZL conceived and designed the experiments. HT collected the clinical data. HT, LW, and ZL analyzed and interpreted the data of the experiments. HT performed the experiments. The first draft was written by HT. LW participated in language editing. All authors read and approved the final manuscript.

FUNDING

This work was supported by the 345 Talent Project of Shengjing Hospital of China Medical University.

ACKNOWLEDGMENTS

The authors would like to thank the investigators who contributed to this study. We thank all team members of TCGA, GSE29265, GSE33630, GSE76039, GSE65144, GSE53157, and GSE76039 projects for providing public-access data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.810681/full#supplementary-material

Supplementary Figure S1 | (A) Survival curves of 502 patients with papillary thyroid carcinoma in the Cancer Genome Atlas (TCGA) database. **(B)** Comparison between *SBSN*^{high} and *SBSN*^{low} poorly differentiated thyroid carcinoma groups in the immune score, stromal score, estimate score, and tumor purity, calculated by ESTIMATE algorithm.

Supplementary Figure S2 | Non-cancerous cells that do not correlate with SBSN expression as derived by the xCell algorithm in papillary thyroid carcinoma.

Supplementary Figure S3 | Non-cancerous cells that do not correlate with *SBSN* expression as derived by the xCell algorithm in anaplastic thyroid carcinoma.

REFERENCES

Abdullah, M. I., Junit, S. M., Ng, K. L., Jayapalan, J. J., Karikalan, B., and Hashim, O. H. (2019). Papillary Thyroid Cancer: Genetic Alterations and Molecular Biomarker Investigations. *Int. J. Med. Sci.* 16 (3), 450–460. doi:10.7150/ijms.29935

Afik, R., Zigmond, E., Vugman, M., Klepfish, M., Shimshoni, E., Pasmanik-Chor, M., et al. (2016). Tumor Macrophages Are Pivotal Constructors of Tumor Collagenous Matrix. *J. Exp. Med.* 213 (11), 2315–2331. doi:10.1084/jem.20151193

Alam, M. T., Nagao-Kitamoto, H., Ohga, N., Akiyama, K., Maishi, N., Kawamoto, T., et al. (2014). Suprabasin as a Novel Tumor Endothelial Cell Marker. *Cancer Sci.* 105 (12), 1533–1540. doi:10.1111/cas.12549

Aoshima, M., Phadungsaksawasdi, P., Nakazawa, S., Iwasaki, M., Sakabe, J.-i., Umayahara, T., et al. (2019). Decreased Expression of Suprabasin Induces Aberrant Differentiation and Apoptosis of Epidermal Keratinocytes: Possible Role for Atopic Dermatitis. *J. Dermatol. Sci.* 95 (3), 107–112. doi:10.1016/j.jdermsci.2019.07.009

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18 (1), 220. doi:10.1186/s13059-017-1349-1

Baxevanis, C. N., Sofopoulos, M., Fortis, S. P., and Perez, S. A. (2019). The Role of Immune Infiltrates as Prognostic Biomarkers in Patients with Breast Cancer. *Cancer Immunol. Immunother.* 68 (10), 1671–1680. doi:10.1007/s00262-019-02327-7

Berger, C., Jensen, M. C., Lansdorp, P. M., Gough, M., Elliott, C., and Riddell, S. R. (2008). Adoptive Transfer of Effector CD8+ T Cells Derived from central Memory Cells Establishes Persistent T Cell Memory in Primates. *J. Clin. Invest.* 118 (1), 294–305. doi:10.1172/JCI32103

Brummelman, J., Pilipow, K., and Lugli, E. (2018). The Single-Cell Phenotypic Identity of Human CD8+ and CD4+ T Cells. *Int. Rev. Cel Mol Biol.* 341, 63–124. doi:10.1016/bs.ircmb.2018.05.007

Cancer Genome Atlas Research Network (2014). Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell* 159 (3), 676–690. doi:10.1016/j.cell.2014.09.050

Capdevila, J., Mayor, R., Mancuso, F. M., Iglesias, C., Caratù, G., Matos, I., et al. (2018). Early Evolutionary Divergence between Papillary and Anaplastic Thyroid Cancers. *Ann. Oncol.* 29 (6), 1454–1460. doi:10.1093/annonc/mdy123

Chen, D. S., and Mellman, I. (2013). Oncology Meets Immunology: The Cancer-Immunity Cycle. *Immunity* 39 (1), 1–10. doi:10.1016/j.immuni.2013.07.012

Cunha, L. L., Domingues, G. A. B., Morari, E. C., Soares, F. A., Vassallo, J., and Ward, L. S. (2021). The Immune Landscape of the Microenvironment of Thyroid Cancer Is Closely Related to Differentiation Status. *Cancer Cel Int.* 21 (1), 387. doi:10.1186/s12935-021-02084-7

Cunha, L. L., Marcello, M. A., and Ward, L. S. (2014). The Role of the Inflammatory Microenvironment in Thyroid Carcinogenesis. *Endocr. Relat. Cancer* 21 (3), R85–R103. doi:10.1530/ERC-13-0431

Denton, A. E., Roberts, E. W., and Fearon, D. T. (2018). Stromal Cells in the Tumor Microenvironment. *Adv. Exp. Med. Biol.* 1060, 99–114. doi:10.1007/978-3-319-78127-3_6

Dralle, H., Machens, A., Basa, J., Fatourechi, V., Franceschi, S., Hay, I. D., et al. (2015). Follicular Cell-Derived Thyroid Cancer. *Nat. Rev. Dis. Primers* 1, 15077. doi:10.1038/nrdp.2015.77

Facciabene, A., De Sanctis, F., Pierini, S., Reis, E. S., Balint, K., Facciponte, J., et al. (2017). Local Endothelial Complement Activation Reverses Endothelial Quiescence, Enabling T-Cell Homing, and Tumor Control during T-Cell Immunotherapy. *Oncoimmunology* 6 (9), e1326442. doi:10.1080/2162402X.2017.1326442

Fang, W., Ye, L., Shen, L., Cai, J., Huang, F., Wei, Q., et al. (2014). Tumor-associated Macrophages Promote the Metastatic Potential of Thyroid Papillary Cancer by Releasing CXCL8. *Carcinogenesis* 35 (8), 1780–1787. doi:10.1093/carcin/bgu060

Formolo, C. A., Williams, R., Gordish-Dressman, H., MacDonald, T. J., Lee, N. H., and Hathout, Y. (2011). Secretome Signature of Invasive Glioblastoma Multiforme. *J. Proteome Res.* 10 (7), 3149–3159. doi:10.1021/pr200210w

French, J. D., Kotnis, G. R., Said, S., Raeburn, C. D., McIntyre, R. C., Klopper, J. P., et al. (2012). Programmed Death-1+ T Cells and Regulatory T Cells Are Enriched in Tumor-Involved Lymph Nodes and Associated with Aggressive Features in Papillary Thyroid Cancer. *J. Clin. Endocrinol. Metab.* 97 (6), E934–E943. doi:10.1210/jc.2011-3428

French, J. D., Weber, Z. J., Fretwell, D. L., Said, S., Klopper, J. P., and Haugen, B. R. (2010). Tumor-associated Lymphocytes and Increased Foxp3+ Regulatory T Cell Frequency Correlate with More Aggressive Papillary Thyroid Cancer. *J. Clin. Endocrinol. Metab.* 95 (5), 2325–2333. doi:10.1210/jc.2009-2564

Galdiero, M. R., Varricchi, G., Loffredo, S., Bellevicine, C., Lansione, T., Ferrara, A. L., et al. (2018). Potential Involvement of Neutrophils in Human Thyroid Cancer. *PLoS ONE* 13, e0199740. doi:10.1371/journal.pone.0199740

Giannini, R., Moretti, S., Ugolini, C., Macerola, E., Menicali, E., Nucci, N., et al. (2019). Immune Profiling of Thyroid Carcinomas Suggests the Existence of Two Major Phenotypes: An ATC-like and a PDTC-like. *J. Clin. Endocrinol. Metab.* 104 (8), 3557–3575. doi:10.1210/jc.2018-01167

Glazer, C. A., Smith, I. M., Ochs, M. F., Begum, S., Westra, W., Chang, S. S., et al. (2009). Integrative Discovery of Epigenetically Derepressed Cancer Testis Antigens in NSCLC. *PLoS One* 4 (12), e8189. doi:10.1371/journal.pone.0008189

Gospodarowicz, M., Mackillop, W., O'Sullivan, B., Sobin, L., Henson, D., Hutter, R. V., et al. (2001). Prognostic Factors in Clinical Decision Making: the Future. *Cancer* 91 (8 Suppl. l), 1688–1695. doi:10.1002/1097-0142(20010415)91:8+<1688:aid-cncr1184>3.0.co;2-7

Han, Y., Liu, Q., Hou, J., Gu, Y., Zhang, Y., Chen, Z., et al. (2018). Tumor-Induced Generation of Splenic Erythroblast-like Ter-Cells Promotes Tumor Progression. *Cell* 173 (3), 634–648. doi:10.1016/j.cell.2018.02.061

Hassan, G., and Seno, M. (2020). Blood and Cancer: Cancer Stem Cells as Origin of Hematopoietic Cells in Solid Tumor Microenvironments. *Cells* 9 (5), 1293. doi:10.3390/cells9051293

Ho, V. W., Hofs, E., Elisia, I., Lam, V., Hsu, B. E., Lai, J., et al. (2016). All Trans Retinoic Acid, Transforming Growth Factor β and Prostaglandin E2 in Mouse

Plasma Synergize with Basophil-Secreted Interleukin-4 to M2 Polarize Murine Macrophages. *PLoS One* 11 (12), e0168072. doi:10.1371/journal.pone.0168072

Ichinose, K., Ohyama, K., Furukawa, K., Higuchi, O., Mukaino, A., Satoh, K., et al. (2018). Novel Anti-suprabasin Antibodies May Contribute to the Pathogenesis of Neuropsychiatric Systemic Lupus Erythematosus. *Clin. Immunol.* 193, 123–130. doi:10.1016/j.clim.2017.11.006

Jang, J.-E., Hajdu, C. H., Liot, C., Miller, G., Dustin, M. L., and Bar-Sagi, D. (2017). Crosstalk between Regulatory T Cells and Tumor-Associated Dendritic Cells Negates Anti-tumor Immunity in Pancreatic Cancer. *Cel Rep.* 20 (3), 558–571. doi:10.1016/j.celrep.2017.06.062

Jung, K. Y., Cho, S. W., Kim, Y. A., Kim, D., Oh, B.-C., Park, D. J., et al. (2015). Cancers with Higher Density of Tumor-Associated Macrophages Were Associated with Poor Survival Rates. *J. Pathol. Transl Med.* 49 (4), 318–324. doi:10.4132/jptm.2015.06.01

Klebanoff, C. A., Gattinoni, L., Torabi-Parizi, P., Kerstann, K., Cardones, A. R., Finkelstein, S. E., et al. (2005). Central Memory Self/tumor-Reactive CD8+ T Cells Confer superior Antitumor Immunity Compared with Effector Memory T Cells. *Proc. Natl. Acad. Sci.* 102 (27), 9571–9576. doi:10.1073/pnas.0503726102

La Vecchia, C., Malvezzi, M., Bosetti, C., Garavello, W., Bertuccio, P., Levi, F., et al. (2015). Thyroid Cancer Mortality and Incidence: a Global Overview. *Int. J. Cancer* 136 (9), 2187–2195. doi:10.1002/ijc.29251

Landa, I., Ibrahimpasic, T., Boucai, L., Sinha, R., Knauf, J. A., Shah, R. H., et al. (2016). Genomic and Transcriptomic Hallmarks of Poorly Differentiated and Anaplastic Thyroid Cancers. *J. Clin. Invest.* 126 (3), 1052–1066. doi:10.1172/JCI85271

Li, M. O., and Rudensky, A. Y. (2016). T Cell Receptor Signalling in the Control of Regulatory T Cell Differentiation and Function. *Nat. Rev. Immunol.* 16 (4), 220–233. doi:10.1038/nri.2016.26

Lin, P., Guo, Y.-n., Shi, L., Li, X.-j., Yang, H., He, Y., et al. (2019). Development of a Prognostic index Based on an Immunogenomic Landscape Analysis of Papillary Thyroid Cancer. *Aging* 11 (2), 480–500. doi:10.18632/aging.101754

Liu, T., Peng, L., Yu, P., Zhao, Y., Shi, Y., Mao, X., et al. (2012a). Increased Circulating Th22 and Th17 Cells Are Associated with Tumor Progression and Patient Survival in Human Gastric Cancer. *J. Clin. Immunol.* 32 (6), 1332–1339. doi:10.1007/s10875-012-9718-8

Liu, Z., Sun, D.-X., Teng, X.-Y., Xu, W.-X., Meng, X.-P., and Wang, B.-S. (2012b). Expression of Stromal Cell-Derived Factor 1 and CXCR7 in Papillary Thyroid Carcinoma. *Endocr. Pathol.* 23 (4), 247–253. doi:10.1007/s12022-012-9223-x

Lote, H., Cafferkey, C., and Chau, I. (2015). PD-1 and PD-L1 Blockade in Gastrointestinal Malignancies. *Cancer Treat. Rev.* 41 (10), 893–903. doi:10.1016/j.ctrv.2015.09.004

Lv, J., Feng, Z. P., Chen, F. K., Liu, C., Jia, L., Liu, P. J., et al. (2021). M2-like Tumor-associated Macrophages-secreted Wnt1 and Wnt3a Promotes Dedifferentiation and Metastasis via Activating β-catenin Pathway in Thyroid Cancer. *Mol. Carcinogenesis* 60 (1), 25–37. doi:10.1002/mc.23268

Marvel, D., and Gabrilovich, D. I. (2015). Myeloid-derived Suppressor Cells in the Tumor Microenvironment: Expect the Unexpected. *J. Clin. Invest.* 125 (9), 3356–3364. doi:10.1172/JCI80005

Mazzaferri, E. L., and Jhiang, S. M. (1994). Long-term Impact of Initial Surgical and Medical Therapy on Papillary and Follicular Thyroid Cancer. *Am. J. Med.* 97 (5), 418–428. doi:10.1016/0002-9343(94)90321-2

Melillo, R. M., Guarino, V., Avilla, E., Galdiero, M. R., Liotti, F., Prevete, N., et al. (2010). Mast Cells Have a Protumorigenic Role in Human Thyroid Cancer. *Oncogene* 29 (47), 6203–6215. doi:10.1038/onc.2010.348

Molinaro, E., Romei, C., Biagini, A., Sabini, E., Agate, L., Mazzeo, S., et al. (2017). Anaplastic Thyroid Carcinoma: from Clinicopathology to Genetics and Advanced Therapies. *Nat. Rev. Endocrinol.* 13 (11), 644–660. doi:10.1038/nrendo.2017.76

Motzer, R. J., Escudier, B., McDermott, D. F., George, S., Hammers, H. J., Srinivas, S., et al. (2015). Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* 373 (19), 1803–1813. doi:10.1056/NEJMoa1510665

Na, K. J., and Choi, H. (2018). Immune Landscape of Papillary Thyroid Cancer and Immunotherapeutic Implications. *Endocr. Relat. Cancer* 25 (5), 523–531. doi:10.1530/ERC-17-0532

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Orhan, A., Vogelsang, R. P., Andersen, M. B., Madsen, M. T., Hölmich, E. R., Raskov, H., et al. (2020). The Prognostic Value of Tumour-Infiltrating Lymphocytes in Pancreatic Cancer: a Systematic Review and Meta-Analysis. *Eur. J. Cancer* 132, 71–84. doi:10.1016/j.ejca.2020.03.013

Papp, S., and Asa, S. L. (2015). When Thyroid Carcinoma Goes Bad: a Morphological and Molecular Analysis. *Head Neck Pathol.* 9 (1), 16–23. doi:10.1007/s12105-015-0619-z

Park, G. T., Lim, S. E., Jang, S.-I., and Morasso, M. I. (2002). Suprabasin, a Novel Epidermal Differentiation Marker and Potential Cornified Envelope Precursor. *J. Biol. Chem.* 277 (47), 45195–45202. doi:10.1074/jbc.m205380200

Reck, M., Rodríguez-Abreu, D., Robinson, A. G., Hui, R., Csőszi, T., Fülöp, A., et al. (2016). Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-small-cell Lung Cancer. *N. Engl. J. Med.* 375 (19), 1823–1833. doi:10.1056/NEJMoa1606774

Richards, D. M., Hettinger, J., and Feuerer, M. (2013). Monocytes and Macrophages in Cancer: Development and Functions. *Cancer Microenvironment* 6 (2), 179–191. doi:10.1007/s12307-012-0123-x

Scouten, W. T., and Francis, G. L. (2006). Thyroid Cancer and the Immune System: a Model for Effective Immune Surveillance. *Expert Rev. Endocrinol. Metab.* 1 (3), 353–366. doi:10.1586/17446651.1.3.353

Shao, C., Tan, M., Bishop, J. A., Liu, J., Bai, W., Gaykalova, D. A., et al. (2012). Suprabasin Is Hypomethylated and Associated with Metastasis in Salivary Adenoid Cystic Carcinoma. *PLoS One* 7 (11), e48582. doi:10.1371/journal.pone.0048582

Shapouri-Moghaddam, A., Mohammadian, S., Vazini, H., Taghadosi, M., Esmaeili, S. A., Mardani, F., et al. (2018). Macrophage Plasticity, Polarization, and Function in Health and Disease. *J. Cel Physiol.* 233 (9), 6425–6440. doi:10.1002/jcp.26429

Takahashi, K., Asano, N., Imatani, A., Kondo, Y., Saito, M., Takeuchi, A., et al. (2020). Sox2 Induces Tumorigenesis and Angiogenesis of Early-Stage Esophageal Squamous Cell Carcinoma through Secretion of Suprabasin. *Carcinogenesis* 41 (11), 1543–1552. doi:10.1093/carcin/bgaa014

Tiainen, S., Tumelius, R., Rilla, K., Hämäläinen, K., Tammi, M., Tammi, R., et al. (2015). High Numbers of Macrophages, Especially M2-like (CD163-Positive), Correlate with Hyaluronan Accumulation and Poor Outcome in Breast Cancer. *Histopathology* 66 (6), 873–883. doi:10.1111/his.12607

Tran Janco, J. M., Lamichhane, P., Karyampudi, L., and Knutson, K. L. (2015). Tumor-infiltrating Dendritic Cells in Cancer Pathogenesis. *J.I.* 194 (7), 2985–2991. doi:10.4049/jimmunol.1403134

Visciano, C., Liotti, F., Prevete, N., Cali', G., Franco, R., Collina, F., et al. (2015). Mast Cells Induce Epithelial-To-Mesenchymal Transition and Stem Cell Features in Human Thyroid Cancer Cells through an IL-8-Akt-Slug Pathway. *Oncogene* 34 (40), 5175–5186. doi:10.1038/onc.2014.441

Waniczek, D., Lorenc, Z., Śnietura, M., Wesecki, M., Kopec, A., and Muc-Wierzgoń, M. (2017). Tumor-Associated Macrophages and Regulatory T Cells Infiltration and the Clinical Outcome in Colorectal Cancer. *Arch. Immunol. Ther. Exp.* 65 (5), 445–454. doi:10.1007/s00005-017-0463-9

Wen, S., Qu, N., Ma, B., Wang, X., Luo, Y., Xu, W., et al. (2021). Cancer-Associated Fibroblasts Positively Correlate with Dedifferentiation and Aggressiveness of Thyroid Cancer. *Ott* 14, 1205–1217. doi:10.2147/OTT.S294725

Wickrema, A., and Crispino, J. D. (2007). Erythroid and Megakaryocytic Transformation. *Oncogene* 26 (47), 6803–6815. doi:10.1038/sj.onc.1210763

Xiangqian, Z., Chen, P., Ming, G., Jingtai, Z., Xiukun, H., Jingzhu, Z., et al. (2019). Risk Factors for Cervical Lymph Node Metastasis in Papillary Thyroid Microcarcinoma: a Study of 1,587 Patients. *Cancer Biol. Med.* 16 (1), 121–130. doi:10.20892/j.issn.2095-3941.2018.0125

Xie, Z., Li, X., He, Y., Wu, S., Wang, S., Sun, J., et al. (2020). Immune Cell Confrontation in the Papillary Thyroid Carcinoma Microenvironment. *Front. Endocrinol.* 11, 570604. doi:10.3389/fendo.2020.570604

Xu, B., and Ghossein, R. (2020). Poorly Differentiated Thyroid Carcinoma. *Semin. Diagn. Pathol.* 37 (5), 243–247. doi:10.1053/j.semdp.2020.03.003

Xu, L., Deng, C., Pang, B., Zhang, X., Liu, W., Liao, G., et al. (2018). TIP: A Web Server for Resolving Tumor Immunophenotype Profiling. *Cancer Res.* 78 (23), 6575–6580. doi:10.1158/0008-5472.CAN-18-0689

Xu, W.-H., Xu, Y., Wang, J., Wan, F.-N., Wang, H.-K., Cao, D.-L., et al. (2019). Prognostic Value and Immune Infiltration of Novel Signatures in clear Cell Renal Cell Carcinoma Microenvironment. *Aging* 11 (17), 6999–7020. doi:10.18632/aging.102233

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Yu, H., Huang, X., Liu, X., Jin, H., Zhang, G. e., Zhang, Q., et al. (2013). Regulatory T Cells and Plasmacytoid Dendritic Cells Contribute to the Immune Escape of Papillary Thyroid Cancer Coexisting with Multinodular Non-toxic Goiter. *Endocrine* 44 (1), 172–181. doi:10.1007/s12020-012-9853-2

Zhang, C., Gu, X., Pan, M., Yuan, Q., and Cheng, H. (2021). Senescent Thyroid Tumor Cells Promote Their Migration by Inducing the Polarization of M2-like Macrophages. *Clin. Transl Oncol.* 23 (6), 1253–1261. doi:10.1007/s12094-020-02516-2

Zhang, C., Zheng, J.-H., Lin, Z.-H., Lv, H.-Y., Ye, Z.-M., Chen, Y.-P., et al. (2020). Profiles of Immune Cell Infiltration and Immune-Related Genes in the Tumor Microenvironment of Osteosarcoma. *Aging* 12 (4), 3486–3501. doi:10.18632/aging.102824

Zhang, H., Liu, H., Shen, Z., Lin, C., Wang, X., Qin, J., et al. (2018). Tumor-infiltrating Neutrophils Is Prognostic and Predictive for Postoperative Adjuvant Chemotherapy Benefit in Patients with Gastric Cancer. *Ann. Surg.* 267 (2), 311–318. doi:10.1097/SLA.0000000000002058

Zhou, J., Dudley, M. E., Rosenberg, S. A., and Robbins, P. F. (2005). Persistence of Multiple Tumor-specific T-Cell Clones Is Associated with Complete Tumor Regression in a Melanoma Patient Receiving Adoptive Cell Transfer Therapy. *J. Immunother.* 28 (1), 53–62. doi:10.1097/00002371-200501000-00007

Zhu, J., Wu, G., Li, Q., Gong, H., Song, J., Cao, L., et al. (2016). Overexpression of Suprabasin Is Associated with Proliferation and Tumorigenicity of Esophageal Squamous Cell Carcinoma. *Sci. Rep.* 6, 21549. doi:10.1038/srep21549

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

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

Copyright © 2022 Tan, Wang and Liu. 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.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 12 April 2022
doi: 10.3389/fgene.2022.767455

miRModuleNet: Detecting miRNA-mRNA Regulatory Modules

Malik Yousef[1]*†, *Gokhan Goy*[2,3]† *and Burcu Bakir-Gungor*[2]

[1]Department of Information Systems, Zefat Academic College, Zefat, Israel, [2]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey, [3]The Scientific and Technological Research Council of Turkey, Ankara, Turkey

Increasing evidence that microRNAs (miRNAs) play a key role in carcinogenesis has revealed the need for elucidating the mechanisms of miRNA regulation and the roles of miRNAs in gene-regulatory networks. A better understanding of the interactions between miRNAs and their mRNA targets will provide a better understanding of the complex biological processes that occur during carcinogenesis. Increased efforts to reveal these interactions have led to the development of a variety of tools to detect and understand these interactions. We have recently described a machine learning approach miRcorrNet, based on grouping and scoring (ranking) groups of genes, where each group is associated with a miRNA and the group members are genes with expression patterns that are correlated with this specific miRNA. The miRcorrNet tool requires two types of -omics data, miRNA and mRNA expression profiles, as an input file. In this study we describe miRModuleNet, which groups mRNA (genes) that are correlated with each miRNA to form a *star shape*, which we identify as a miRNA-mRNA regulatory module. A scoring procedure is then applied to each module to further assess their contribution in terms of classification. An important output of miRModuleNet is that it provides a hierarchical list of significant miRNA-mRNA regulatory modules. miRModuleNet was further validated on external datasets for their disease associations, and functional enrichment analysis was also performed. The application of miRModuleNet aids the identification of functional relationships between significant biomarkers and reveals essential pathways involved in cancer pathogenesis. The miRModuleNet tool and all other supplementary files are available at https://github.com/malikyousef/miRModuleNet/

Keywords: gene expression, multi omics, machine learning, integrative "omics", feature selection

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Flavia Figueira Aburjaile,
Federal University of Minas Gerais, Brazil
Wenyu Zhang,
Max Planck Institute for Evolutionary Biology, Germany

***Correspondence:**
Malik Yousef
malik.yousef@gmail.com

†These authors have contributed equally to this work

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

Received: *30 August 2021*
Accepted: *24 March 2022*
Published: *12 April 2022*

Citation:
Yousef M, Goy G and Bakir-Gungor B (2022) miRModuleNet: Detecting miRNA-mRNA Regulatory Modules. Front. Genet. 13:767455. doi: 10.3389/fgene.2022.767455

1 INTRODUCTION

The World Health Organization (WHO) reported in 2019 that cancer is the leading cause of death in three out of four countries in the world (Sung et al., 2021). Approximately 19.3 million people were diagnosed with cancer in 2020 and 10 million people lost their lives due to cancer. Lifestyles, environmental, demographic and cultural factors all contribute to these problematic statistics. If these statistics are to change, it is important to better understand the complex molecular processes that lead to cancer development and progression as precisely as possible. This information is critical to both traditional drug development approaches and for personalized medicine approaches (Schmidt, 2014).

miRNAs are non-coding RNAs, roughly 22–25 nucleotides in length (Bartel, 2004; Allmer and Yousef, 2016; Allmer and Yousef, 2022) and are present in animals and plants, as well as in humans.

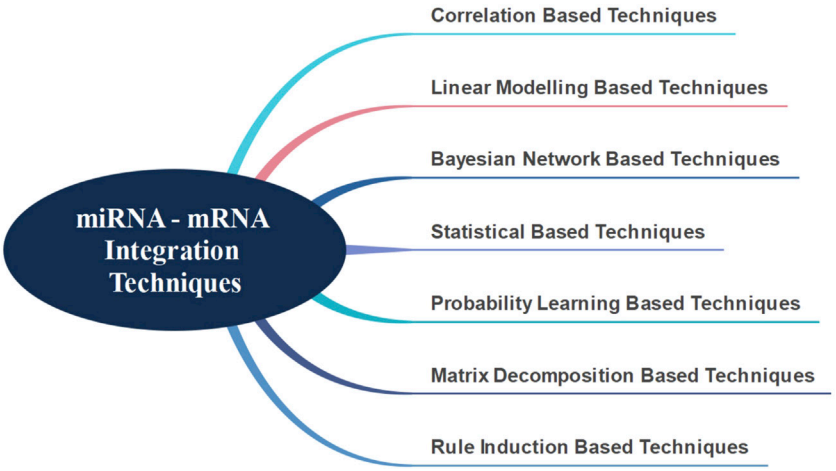

FIGURE 1 | miRNA - mRNA integration techniques.

The observations that miRNAs with similar sequences are detected in all living things further support the idea that miRNAs perform critical biological functions (Cai et al., 2009). miRNAs are known to be responsible for the regulation of approximately 60% of protein coding genes (Friedman et al., 2009) and cellular processes including cell proliferation, apoptosis and necrosis (Keller et al., 2011). miRNAs can affect gene expression by binding to the seed regions of mRNAs (Ivey and Srivastava, 2015; Yousef et al., 2018) and, in general, repress the expression of their target mRNAs via physically interacting with them. In other words, miRNAs tend to have a negative correlation with mRNAs. The elucidation of the relationships between miRNAs and mRNAs is important in order to understand the mechanisms of complex diseases such as cancer (Pencheva and Tavazoie, 2013; Yousef et al., 2014). A better understanding of the associations between miRNAs and the mRNAs can reveal important information on normal and aberrant gene regulation and cell biology.

There are presently seven major techniques in literature for the integration of miRNA-mRNAs, as shown in **Figure 1** (Masud Karim et al., 2016). In general, the correlation-based techniques primarily start by identifying differentially expressed mRNAs and miRNAs. Using various correlation metrics, mRNA-miRNA pairs are identified and the integration is achieved through these pairs (Feng et al., 2018; Li et al., 2018; Liu et al., 2018; Yang et al., 2019; Yao et al., 2019). Hailu et al. (2021) have used Spearman's correlation and attempted to identify target genes and signaling pathways associated with pediatric dilated cardiomyopathy by integrating miRNA and mRNA data.

Correlation-based techniques have the following disadvantages. These techniques assume that one miRNA affects only one mRNA, an assumption that is not entirely true (Huang et al., 2007). Linear modeling based techniques have been developed in order to overcome this assumption. Huang et al. (2007) suggested modeling mRNA expressions as linear combinations of miRNAs to address this problem and applied the Bayesian algorithm to discover hidden miRNA targets. They also used a different distribution technique, integrating sequence information with their previous study. Stingo et al. (2010) proposed a comparable approach. However, they did not consider the effect of different tissues and suggested that miRNAs had a different promoter effect on each mRNA (Le and Bar-Joseph, 2013) attempted to find the mRNA modules that affect the functionality of miRNAs, using interaction, expression and sequence information; and a regression-based solution. They claimed that by using this method, they could identify relevant modules in a more robust and accurate way.

Another approach used for the integration of miRNA and mRNA interactions is the Bayesian network technique. Liu et al. (2009) performed an integrated analysis using differentially expressed miRNAs and mRNAs through Bayesian network technique. Due to the large amount of biological data available, (Madadjim, 2021) emphasized the necessity of producing a scalable solution and suggested that the Bayesian network-based machine learning model could be a valid solution.

All events that take place in a living system happen within a specific biological organization. In other words, the events that occur at the molecular level are not random. This understanding has motivated the development of statistical solutions for miRNA and mRNA integration (Jayaswal et al., 2011). Along this line, (Hecker et al., 2013) evaluated different miRNA-mRNA expression data using statistical approaches, without any other prior knowledge; and developed a method to distinguish different tissues. Using a similar approach, (Nersisyan et al., 2021) developed a new tool to generate miRNA-gene-TF networks.

Another method that generates miRNA-mRNA groups is the probability learning based technique. In this approach, the interaction probabilities of known miRNA-mRNA pairs are estimated (Joung et al., 2007). However, in order for this operation to be performed robustly and effectively, more than one source of information is needed. The Non-Negative Matrix Factorization technique is another important method. This method accomplishes the integration process by successfully

TABLE 1 | Details of the datasets utilized in miRModuleNet.

TCGA data	Abbreviation	Control	Case	PMID
Bladder urothelial carcinoma	BLCA	405	19	24476821
Breast invasive carcinoma	BRCA	760	87	31878981
Kidney chromophobe	KICH	66	25	25155756
Kidney renal papillary cell carcinoma	KIRP	290	32	28780132
Kidney renal clear cell carcinoma	KIRC	255	71	23792563
Lung adenocarcinoma	LUAD	449	20	25079552
Lung squamous cell carcinoma	LUSC	342	38	22960745
Prostate adenocarcinoma	PRAD	493	52	26544944
Stomach adenocarcinoma	STAD	370	35	25079317
Papillary thyroid carcinoma	THCA	504	59	25,417,114
Uterine corpus endometrial carcinoma	UCEC	174	23	23636398

Control and case columns denote the number of samples. Column PMID refers to Pubmed ID of the related publication, where further information about the dataset can be found.

separating different information sources (Zhang et al., 2011) was able to successfully integrate information obtained from different sources and generate significant miRNA-mRNA groups. Additional approaches use rule induction-based techniques based on information theory. Generally, as in the other techniques, data obtained from more than one data source needs to be integrated (Tran et al., 2008) used a rule induction-based technique to find miRNA-mRNA groups while (Lavrac et al., 2004) used the CN2-SD system as the rule generation system to identify miRNA-mRNA groups.

With the advancements in technology we now have access to data which describes different levels of molecular regulation from the same individual. These rich and complicated data sets require the development of novel techniques to integrate and understand this data. All the tools that we have surveyed above are based on statistical approaches. To the best of our knowledge, there are only two available tools that can adequately address the classification problem using integrated miRNA-mRNA groups. These bioinformatics tools are maTE (Yousef et al., 2019) and miRcorrNet (Yousef et al., 2021b). The main difference between these two tools is the miRNA-mRNA grouping methodology. While maTE adopts a biological grouping methodology, miRcorrNet tool uses correlation information in order to generate the groups. These two tools not only solve the classification problem, but also provide a score for each group, where the score reflects the contribution of each group to classification.

In this study, we present a novel bioinformatics tool named miRModuleNet. miRModuleNet differs from our two previous approaches in that miRNA-mRNA integration has been developed using statistical information. In this paper, we have comparatively evaluated these three different grouping methodologies and showed the superiority of miRModuleNet against state of the art methods.

2 MATERIALS AND METHODS

2.1 Datasets

In this study, miRNA and mRNA expression profiles which have been obtained from the same individuals have been used. Due to the aforementioned reasons, in this study we focused on cancer. In this context, 11 different cancer datasets were downloaded from The Cancer Genome Atlas (TCGA) data portal (Tomczak et al., 2015). The details of these datasets are presented in **Table 1**.

2.2 The G-S-M Approach

miRModuleNet was developed based on the generic approach named G-S-M. This generic approach was adopted by different tools such as SVM RCE, SVM-RCE-R (Yousef et al., 2007; Yousef et al., 2021a), maTE (Yousef et al., 2019), CogNet (Yousef et al., 2021d), miRcorrNet (Yousef et al., 2021b), and Integrating Gene Ontology Based Grouping and Ranking (Yousef et al., 2021c). Recently, these tools and their competitors were reviewed in (Yousef et al., 2020).

As illustrated in **Figure 2**, the algorithm mainly consists of 3 components (shown as circles):

1. G Component: Detect groups of genes
2. S Component: Score the groups.
3. M Component: Creates the model by training a classifier (Random Forest)

In the first component G, bioD is a biological database, or another prior biological knowledge that will be used to create the *groups* that contain the genes from the mRNA (gene expression) data. This operation is represented as the G component whose output is the set of groups. Group names are the names of the biological entity such as miRNA names, where a group of genes may be targeted by that miRNA, a KEGG pathway name, or a disease phenotype name. Note that, in most of the cases each group has an important biological meaning. The resulting set of groups is indicated in the Groups box in **Figure 2**.

Assume that we have n samples and k genes in our dataset C. The C data is split into two parts as C_{train} and C_{test}, where the C_{train} is used to score the groups and to train the classifier in the M component. The C_{test} is used for testing and reporting the performance.

Let m = *size*(Gr) be the number of groups generated by the G component and let Gr be the collection of all the groups as Gr = [bioD_grp$_f$, where $f = 1,..,size(Gr)$]. From now on, we will refer to one group of Gr as *bioD_grp*.

In Component S, each *bioD_grp* in Gr is scored, as shown in **Figure 2**. In order to perform this task, we generate *size* (Gr)

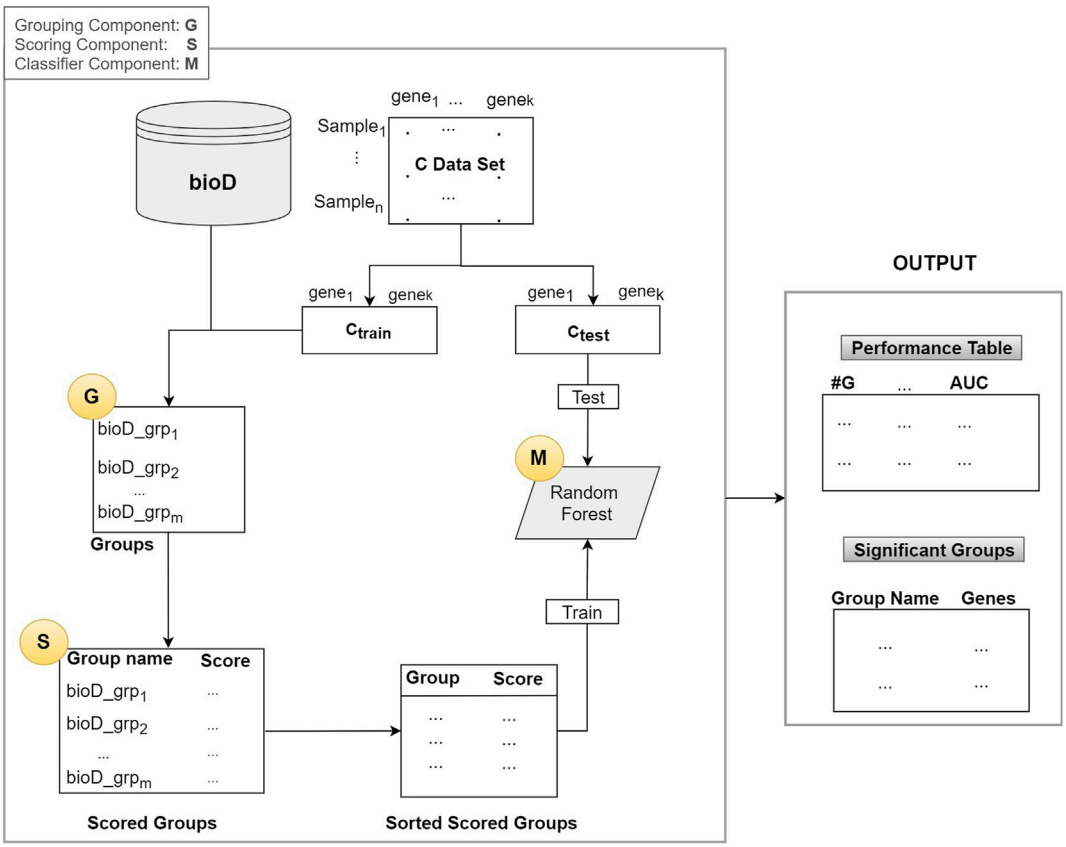

FIGURE 2 | The general integrative approach that is based on grouping and scoring/ranking.

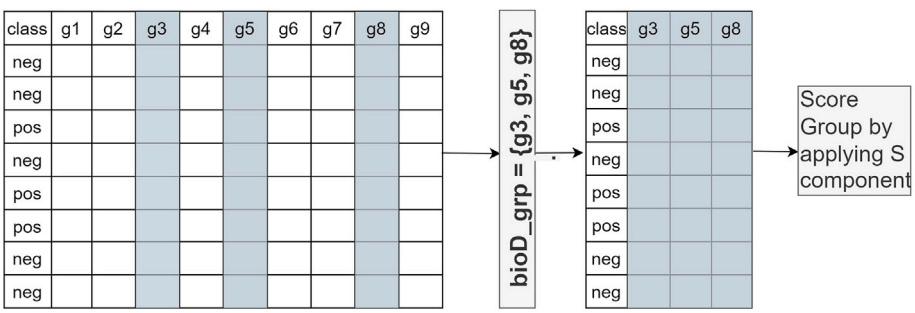

FIGURE 3 | An example of how to create sub_data based on a group of genes and then this sub-data is subject to the scoring component S. g_i refers to $gene_i$.

different sub_data sets which are the sub matrices of the gene expression matrix C_{train} (illustrated in **Figure 2**). Each sub_data set includes the columns from the original data matrix C_{train}, corresponding to the genes in *bioD_grp*. In other words, each sub_data set contains only the gene expression values of specific genes included in that group and associated class labels. We will refer to each sub_data as $C_{train}sub_f$, where f = 1,..,size(Gr) that contains genes that belong to the group of *bioD_grp*. **Figure 3** is an example of how to create sub_data based on a group of mRNAs and then this sub-data is subject to a procedure for scoring those groups.

Let S (sub_data) be the k-fold cross validation procedure that computes and returns some performance measurements such as accuracy, specificity, sensitivity and Area Under the ROC Curve (AUC). We used AUC as the major performance metric to score for the sub_data. Next, we score all the groups using the S function which produces scores for groups, named as grp_scores and grp_scores = [($bioD\_grp_f$, $score_f$) f = *1,..,size*(G)]. Then we sort

TABLE 2 | A sample output of scoring component when applied on THCA data, downloaded from TCGA.

Group	Accuracy	Sensitivity	Specificity	FM	Precision	Cohen's kappa
hsa-miR-101-3p	0.89	0.82	0.92	0.85	0.88	0.73
hsa-miR-200c-3p	0.95	0.92	0.97	0.92	0.94	0.89
hsa-miR-508-3p	0.98	0.93	1.00	0.96	1.00	0.94
hsa-miR-629-5p	0.99	0.97	1.00	0.98	0.99	0.97

Each miRNA ID represents a group, which is generated by the Grouping Component G. Groups are sorted according to the accuracy metric.

TABLE 3 | Pseudocode of component M, which calculates the performance.

For f = 1 to top_f
genes_set = $\cup_{j=1}^{f}$ {bioD_grp_sorted$_j$}
X_train = sub_set of C_{train} that includes the genes from the genes_set
X_test = sub_set of C_{test} that includes the genes from the genes_set
RF_Model < - train Random Forest (X_train)
Performances = test RF_Model (X_test)

this list based on score and obtain grp_scores_sorted = [(bioD_grp_sorted$_f$, score_sorted$_f$) f = 1,..,size(G)]. **Table 2** presents an example output of this S component. In **Table 2**, microRNAs are shown as the group name since in this example miRNAs are used within the G component to group a set of genes targeted by that miRNA.

The last component is the M component, which creates the model by training a classifier. In order to build the Random Forest (RF) model and report the cumulative performance of the algorithm, we implement the procedure presented in **Table 3**. Here, top_f specifies the number of top groups defined by the user.

In **Table 3**, RF_Model is the model created by training Random Forest on the X_train data set. This model will be used to test on the X_test. In **Table 3**, grp_bioD_sorted$_f$ is one of the groups of Gr (for example, of miRNA, KEGG, GO databases). The Performance Table in **Figure 2** describes the cumulative performance of the G-S-M approach, where *#G* is the number of genes in the cumulative group. The output of this step is the Performance Table shown in the right hand side of **Figure 2**.

2.3 miRModuleNet

miRModuleNet tool is developed as a specific application of our G-S-M approach on the -omics data integration problem including miRNA and mRNA expression profiles. Hence, miRModuleNet makes use of the above-mentioned G-S-M approach with further additions. Before utilizing the G-S-M method, miRModuleNet includes some preprocessing steps as explained in detail below. The main idea behind miRModuleNet is illustrated in **Figure 4**. Initially, both miRNA and mRNA expression datasets are split into training and testing parts. Following the general idea presented in **Figure 2**, the training part is used to create the groups, define the features in each group and to build the model, while the testing part is only considered in the evaluation step.

In the 1st step of miRModuleNet, both miRNA and mRNA expression profiles are cleaned by removing the columns containing the missing data. For miRNA-seq profiles, raw read counts were normalized to reads per million mapped reads (RPM). For mRNA-seq profiles, the raw read counts were normalized to Reads Per Kilobase Million Mapped Reads (RPKM). Subsequently, whole data at different ranges were normalized using z-score normalization. Second step identifies statistically important miRNAs and mRNAs that were to be used in the following steps. In the 3rd step, using statistically significant miRNAs and mRNAs, differentially expressed miRNAs and mRNAs are detected using the edgeR package (Robinson et al., 2010). In step 4, the mutual information matrix is generated in order to determine the miRNAs and mRNAs that will be used to form the miRNA-mRNA groups. Instead of considering each pair in this matrix, we only select the pairs that exceeded a certain threshold. We experiment with the values of 0.15, 0.25, and 0.5 as the Mutual Information (MI) threshold and present data identifying the value of 0.25 as the optimal threshold value. This value is used in the later steps of miRModuleNet. The 5th step corresponds to the grouping component in the general approach. In this step the miRNA-mRNA regulatory groups i.e., modules are generated according to the **Algorithm 1**. Here, I(x,y) denotes the mutual information between two variables x and y. I(x,y) = H(x)−H(y|x), where H(y) and H(y|x) are the entropy of y and the conditional entropy of y given x. The strategy for obtaining miRNA-mRNA regulatory modules is explained in the following section.

Algorithm 1. Generate the "Star shaped" module that contains single miRNA and multiple mRNAs.

1) Let C= {gene1, gene2, ...genek } be the profiles of the mRNAs from data Dgenes
2) Let Str ←∅ be the "Star" group for the miRNA
3) Compute I_i = I(gene$_i$, miRNA) of each mRNA gene$_i$ in C.
4) Let gene*= max$_i$ {I_i}, Select the gene with the highest value of mutual information

2.4 Generating the miRNA-mRNA Regulatory Modules/Groups

In order to detect the miRNA-mRNA regulatory modules, we have used the RFCM^3 approach suggested by (Paul and Madhumita, 2020). The RFCM^3 considers two types of -omics data, the miRNA and mRNA expression profiles from the same samples. Here, we will use the terms module and groups interchangeably. miRNA-mRNA modules consist of a miRNA and its related mRNA genes. As illustrated in the Step 5 of **Figure 4**, we have generated the module called the star shaped module, where it contains a single miRNA and multiple mRNAs/genes. As suggested by (Paul and Madhumita, 2020), mRNAs for such modules are selected in such a way that they are

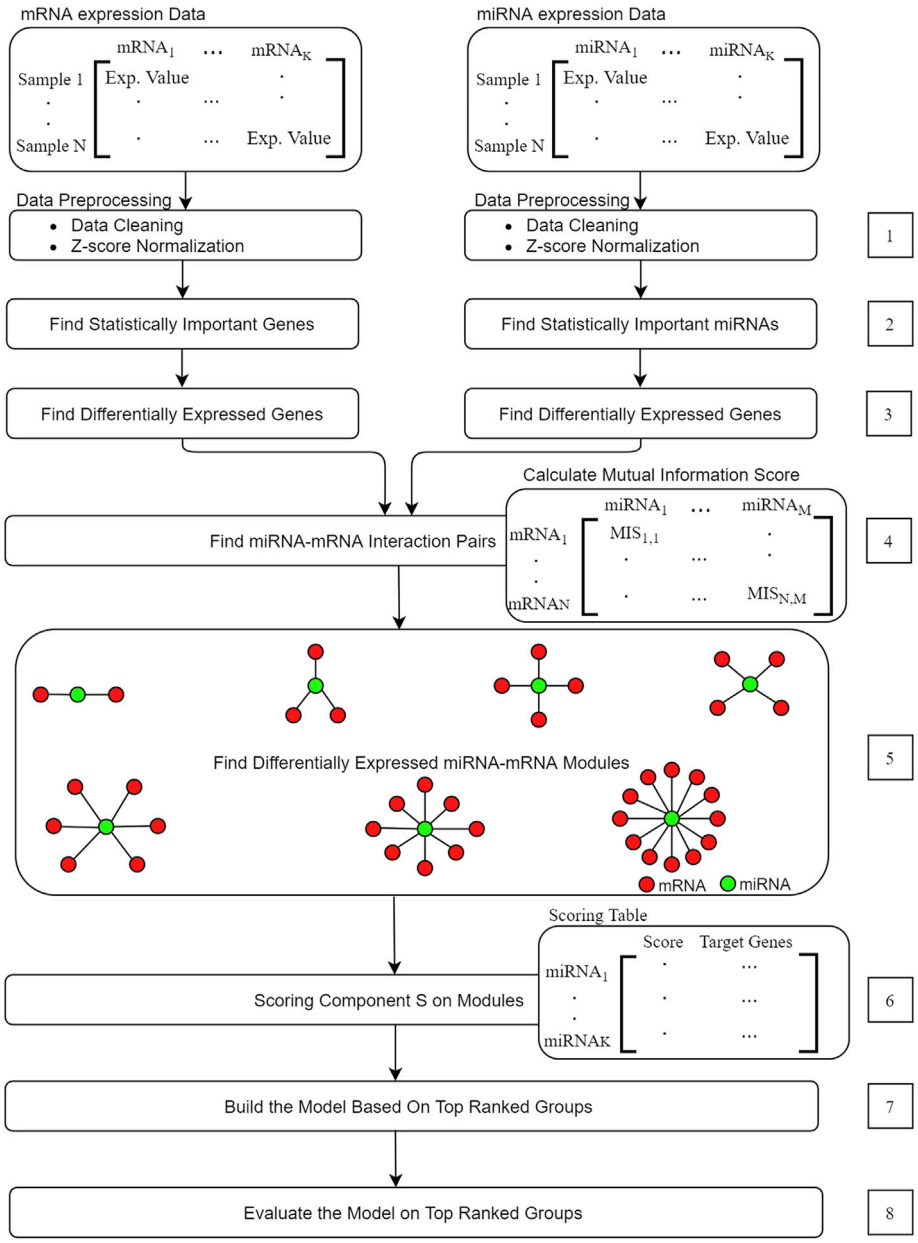

FIGURE 4 | miRModuleNet flowchart.

simultaneously and functionally similar to the corresponding miRNA.

In creating these groups, we first identify the miRNA-mRNA pair with the highest score. As shown in green in Step 5 of **Figure 4**, we detect the center of the star (the miRNA that serves as the group name). The mRNA in this pair is the starting point for the addition of other mRNAs forming the star shape. The relationship of the miRNA to other mRNAs is determined by looking at the Mutual Information matrix. For mRNAs to be included in the group, the mutual information score between them and the relevant mRNA must exceed the threshold set by the end user and this relationship is then considered to be potentially important.

The 6th step corresponds to the scoring component S in the general approach. In this step, the classification power of each group is evaluated by calculating the scores, which indicate how powerful a group is in terms of distinguishing the two classes (case/control). At the end of this step a Scoring Table is produced containing the miRNA in rows and the score of the corresponding mRNA group in the columns. In the 7th step, a machine model is trained using the top ranked groups. In other words, the machine learning model which uses Random Forest is trained via only

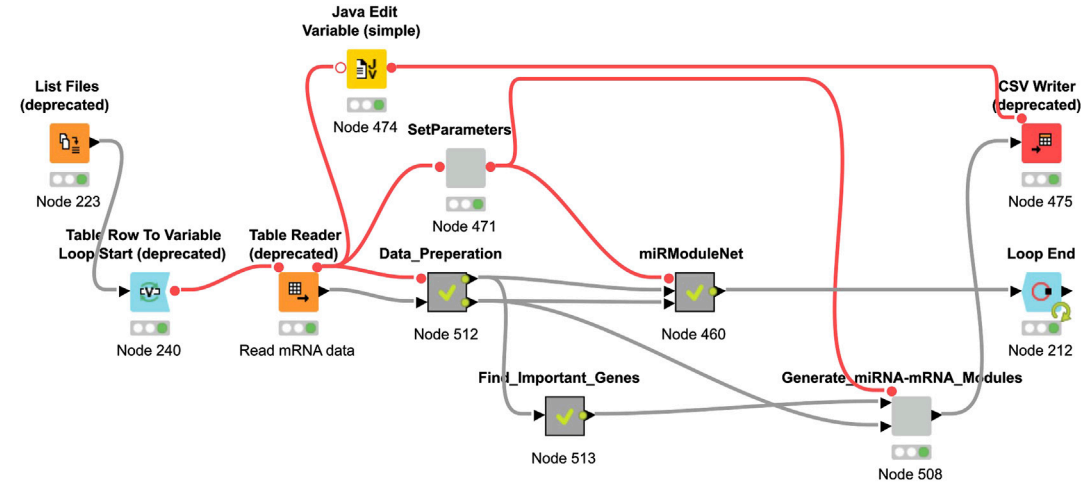

FIGURE 5 | miRModuleNet workflow in KNIME.

TABLE 4 | An example performance table of miRModuleNet for top ranked 10 modules for BLCA dataset.

#Groups	#Genes	Accuracy	Sensitivity	Specificity	AUC
10	1422.96	0.92	0.89	0.94	0.98
9	1254.76	0.92	0.88	0.93	0.98
8	1110.82	0.91	0.87	0.93	0.97
7	962.83	0.91	0.88	0.93	0.97
6	799.7	0.92	0.88	0.94	0.97
5	628.14	0.92	0.87	0.94	0.97
4	489.59	0.91	0.87	0.93	0.98
3	331.02	0.90	0.85	0.93	0.97
2	205.08	0.90	0.84	0.93	0.97
1	79.25	0.89	0.82	0.92	0.95

considering top *f* groups. This means that miRModuleNet is using all of the genes within top *f* groups in a unified manner. The default value of *f* is set as 10 and miRModuleNet generates 10 different machine learning models where each model is trained using a different number of groups from 1 to 10. The user can also change the value of the *f*. Classification strategy is explained more in detail in the following section. Then the last step is the evaluation step that uses the test part.

2.5 Classification Approach

In this study, the Random Forest algorithm (Breiman, 2001), which is a supervised machine learning algorithm, was used to solve the classification problem. This algorithm consists of two stages. In the first stage, a forest is created by producing a large number of decision trees. In the second stage, the classification process is carried out through the feedback obtained from these trees. As an advantage of this use, a model with better generalization can be produced. On the one hand, a more robust solution is obtained, on the other hand, overfitting is potentially prevented.

While generating the model, 100 fold Monte Carlo Cross Validation (MCCV) was used in the learning phase (Xu and Liang, 2001). In order to evaluate the performance, miRModuleNet repeats the process 100 times. In each iteration, 90% of the data is selected for training and the remaining 10% is selected for testing. In addition, an under sampling method was used to solve the imbalanced class problem encountered while training the model. This method aims to provide the desired rate of data distribution by randomly eliminating samples from the class with too many samples. Hence, miRModuleNet randomly selects samples with a ratio of 1:2 for under-sampling. Under-sampling was performed in every iteration of cross validation. In each iteration, our approach generates lists of miRNA modules/groups and their associated genes that are slightly different. Hence, there is a need to apply a prioritization approach on those lists. As utilized in miRcorrNet, we have used rank aggregation methods. In this respect, we have embedded the RobustRankAggreg R package, developed by (Kolde et al., 2012) into miRModuleNet workflow. The RobustRankAggreg assigns a *p*-Value to each element in the aggregated list, which describes how good each element/entity was ranked compared to the expected value.

2.6 Implementation of miRModuleNet

The KNIME Analytics platform is used for the implementation of miRModuleNet (Berthold et al., 2008). The KNIME environment is easy to use, it is an open source platform and it can be used for a wide variety of operations and for a wide variety of data types. In the KNIME environment, all operations work based on workflows. miRModuleNet's workflow is shown in **Figure 5**.

As it can be seen in **Figure 5**, KNIME workflows consist of nodes, where each of these nodes perform a specific task. For example, using the List Files node, the directory where the data is located is specified. By using the Table Reader node, it is ensured that the data is imported into the KNIME environment. By using the Data Preparation metanode, above-mentioned preprocessing operations are performed. miRModuleNet metanode is the node of the main algorithm. In addition to these, within the SetParameters node, two critical parameters of the workflow

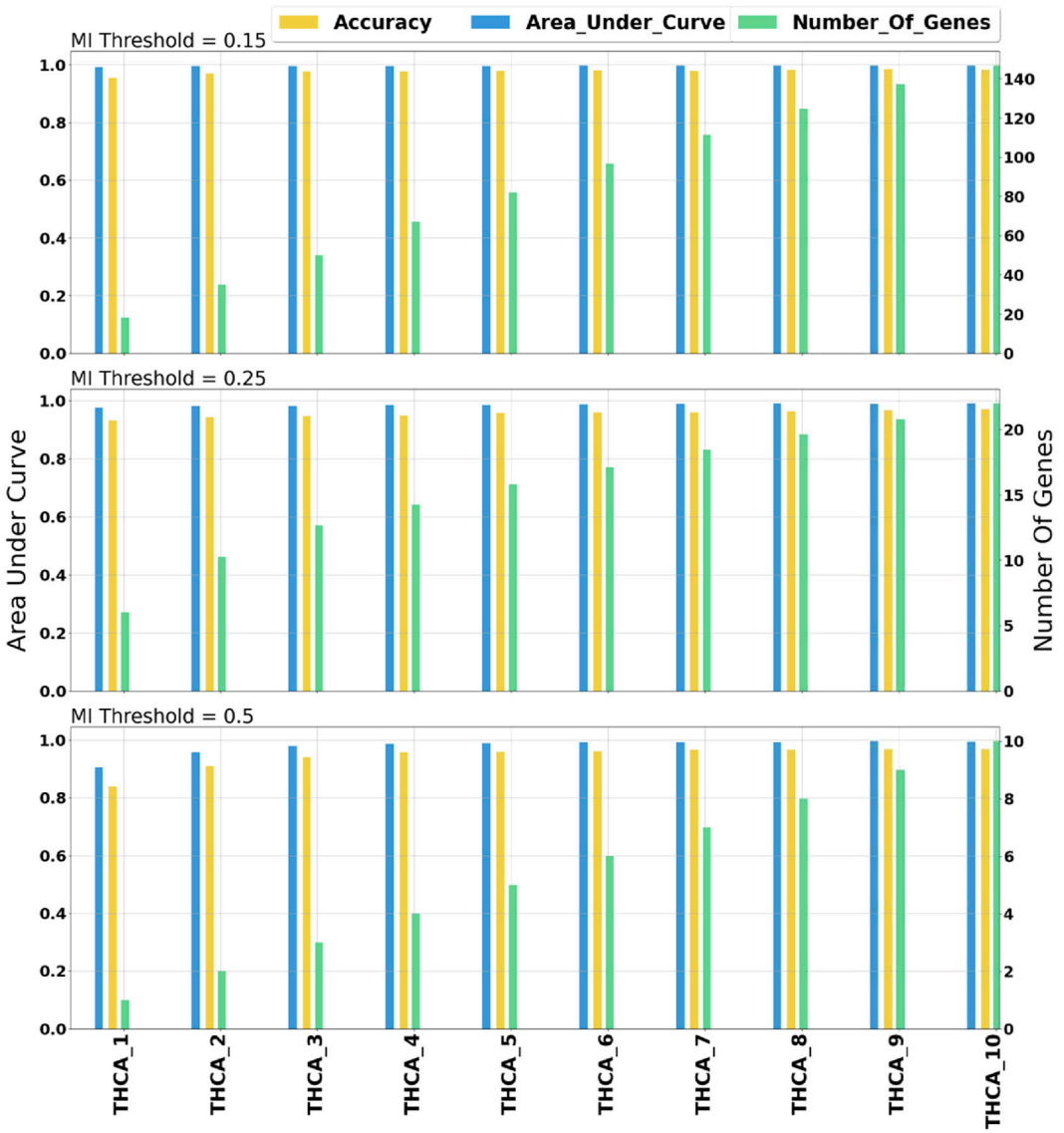

FIGURE 6 | Comprehensive evaluation of different mutual information threshold values. The numbers following the underscore values correspond to the number of groups.

can be entered by the end user. These parameters are the number of iterations and the mutual information threshold.

Results are obtained after running the KNIME workflow, which is shown in **Figure 5**. One of these results is the comparison of the performances of the machine learning models depending on the k (number of top groups) parameter. An example of this comparison is shown in **Table 4**. **Table 4** presents an example performance table of miRModuleNet for top ranked 10 modules for BLCA data. The last row presents the performance of the top ranked module/group (#Groups = 1). In other words, an accuracy of 89% is obtained using 79.25 genes on average. The row of #Groups = 2 presents the performance metrics obtained for the top 2 groups where the genes of the top ranked group and the second highest scoring group are aggregated together. That is to say that miRModuleNet reports the performance results for top 10 groups cumulatively.

3 RESULTS

3.1 Performance Evaluation Metrics

The performance of machine learning models can be evaluated through several quantitative metrics. In this respect, statistical metrics such as Accuracy, Sensitivity, Specificity and Precision could be calculated by constructing the confusion matrix. For the problems involving imbalanced data, it is essential to prove the consistency of the results. In this regard, Area Under the Curve

TABLE 5 | Performance results of miRModuleNet over the top-ranked group.

Disease	#Genes	ACC	SEN	SPE	FM	AUC	Precision
BLCA	79	0.89	0.82	0.92	0.85	0.95	0.88
BRCA	22	0.95	0.92	0.97	0.92	0.98	0.94
KICH	40	0.98	0.93	1.00	0.96	0.99	1.00
KIRC	64	0.99	0.97	1.00	0.98	0.99	0.99
KIRP	41	1.00	0.99	1.00	0.99	1.00	1.00
LUAD	4	0.94	0.90	0.96	0.90	0.98	0.93
LUSC	12	0.98	0.99	0.98	0.98	1.00	0.97
PRAD	5	0.86	0.76	0.91	0.77	0.92	0.82
STAD	115	0.90	0.81	0.95	0.85	0.97	0.92
THCA	6	0.93	0.90	0.95	0.90	0.98	0.92
UCEC	33	0.94	0.89	0.96	0.89	0.99	0.94

ACC stands for Accuracy, SEN stands for Sensitivity, SPE stands for Specificity, FM stands for F-Measure, AUC stands for Area Under the ROC curve.

(AUC) metric is reported as an accurate metric in terms of evaluating the performance results in such problems (Hand, 2004).

3.2 Performance Results

3.2.1 Optimization of Mutual Information Threshold

miRModuleNet tool uses (MI) to detect the relationships between miRNAs and mRNAs. In order to identify the optimal value of the MI threshold, we experimented with three different values (0.15, 0.25, 0.5). As stated above, we selected 0.25 as the optimal threshold. In our comparison, the AUC value versus the number of genes is taken into account. Such a comparison on THCA data is demonstrated in **Figure 6**. As illustrated in **Figure 6**, when the MI threshold value was set to 0.15, the AUC value was in the range of 0.98–0.99, and the number of genes increased from 18 to 146 as the number of groups (star shaped modules) increased. Using the MI threshold value as 0.25, AUC values in the range of 0.97–0.99 were obtained, and the number of genes increased from 6 to 22. When the MI threshold value was set to 0.5, the AUC value was in the range of 0.92–0.99, and the number of genes increased from 1 to 10. Such comparisons were made for all cancer types. As a result of these comparative evaluations, we have decided to set the MI threshold as 0.25.

In this study, we have tested miRModuleNet using 11 different cancer datasets presented in **Table 1**. Our machine learning models generate the most important group as an output; and the performance evaluation metrics were obtained by using the identified most important group. As presented in **Table 5**, the average number of selected genes for the most important groups was 38.27 for 11 tested cancer types. Likewise, the average of obtained AUC values using the top group was 0.98. All performance results reported in this study were obtained by calculating the mean of the 100-fold Monte Carlo Cross Validation (MCCV).

In addition, in terms of performance, miRModuleNet has been compared with other existing tools i.e., SVM-RFE, maTE and miRcorrNet. These tools differ in terms of the data they use and the way they produce results. While miRcorrNet and miRModuleNet both use miRNA and mRNA expression profiles, SVM-RFE and maTE tools use only mRNA data. In addition, while miRcorrNet, miRcorrNet and maTE give the results on group level, the SVM-RFE tool gives the results directly at the gene level. In other words, miRcorrNet, maTE and miRModuleNet tools give their results by building a Random Forest model over the top 1 to 10 cumulative groups of genes. On the other hand, SVM-RFE tool gives its results using different levels of genes, i.e., 1, 2, 4, 6, 8, 10, 20, 40, 60, 80, 100, 125, 250, 500 and 1,000 genes. In order to make a fair comparison of the existing methods involving different approaches, it has become necessary to determine benchmarks at both the group level and the gene level. The comparison level for miRcorrNet, miRModuleNet and maTE, which produced results at the group level, was determined as two according to the number of genes criterion. When these three tools used two as the group level, the lowest number of genes was found to be 7.48, and the highest used number of genes was found to be 141.26. Therefore, it was decided to use gene levels 8 and 125 to be able to include the SVM-RFE tool in the comparison. In **Table 6**, the performance evaluation of all these tools are presented. The calculated performance metrics are number of genes, accuracy, sensitivity, specificity, F-Measure, AUC and Precision. **Table 6** indicates that miRModuleNet achieved a similar performance by using nearly half of the genes compared to another newly developed tool called miRcorrNet. Although there are no serious differences in results, the increase in the AUC metric is considered to be very important and noteworthy. Additionally, the close performances of the tools show that the developed tool miRModuleNet is a consistent and robust tool.

3.3 Functional Enrichment Analysis Results

In order to better understand the disrupted mechanisms of the disease at the molecular level, functional enrichment analysis was carried out. Hence, we investigated whether the obtained results have biological meaning. For this purpose, GeneCodis (Tabas-Madrid et al., 2012) and DAVID (Huang et al., 2009a; Huang et al., 2009b), which have been widely used in literature, are utilized. For each disease, all enriched KEGG pathways were found separately. Overrepresented KEGG pathways of our

TABLE 6 | Comparative evaluation of existing tools using 11 cancer datasets.

Method	#Genes	Accuracy	Sensitivity	Specificity	AUC	SD
miRModuleNet	78.31	0.96	0.91	0.98	0.99	0.04 ± 0.02
miRcorrNet	141.26	0.96	0.94	0.97	0.98	0.05 ± 0.05
maTE	7.48	0.96	0.94	0.96	0.98	0.034 ± 0.02
SVM-RFE	8	0.84	0.85	0.85	0.91	0.07 ± 0.04
SVM-RFE	125	0.96	0.97	0.95	0.98	0.05 ± 0.03

AUC column refers to the area under the curve values. All the presented values are average values over 100 MCCV for the level of top 2 groups for miRModuleNet, maTE and miRcorrNet; 8 and 125 genes for SVM-RFE. Standard Deviation (SD) values are given for AUC.

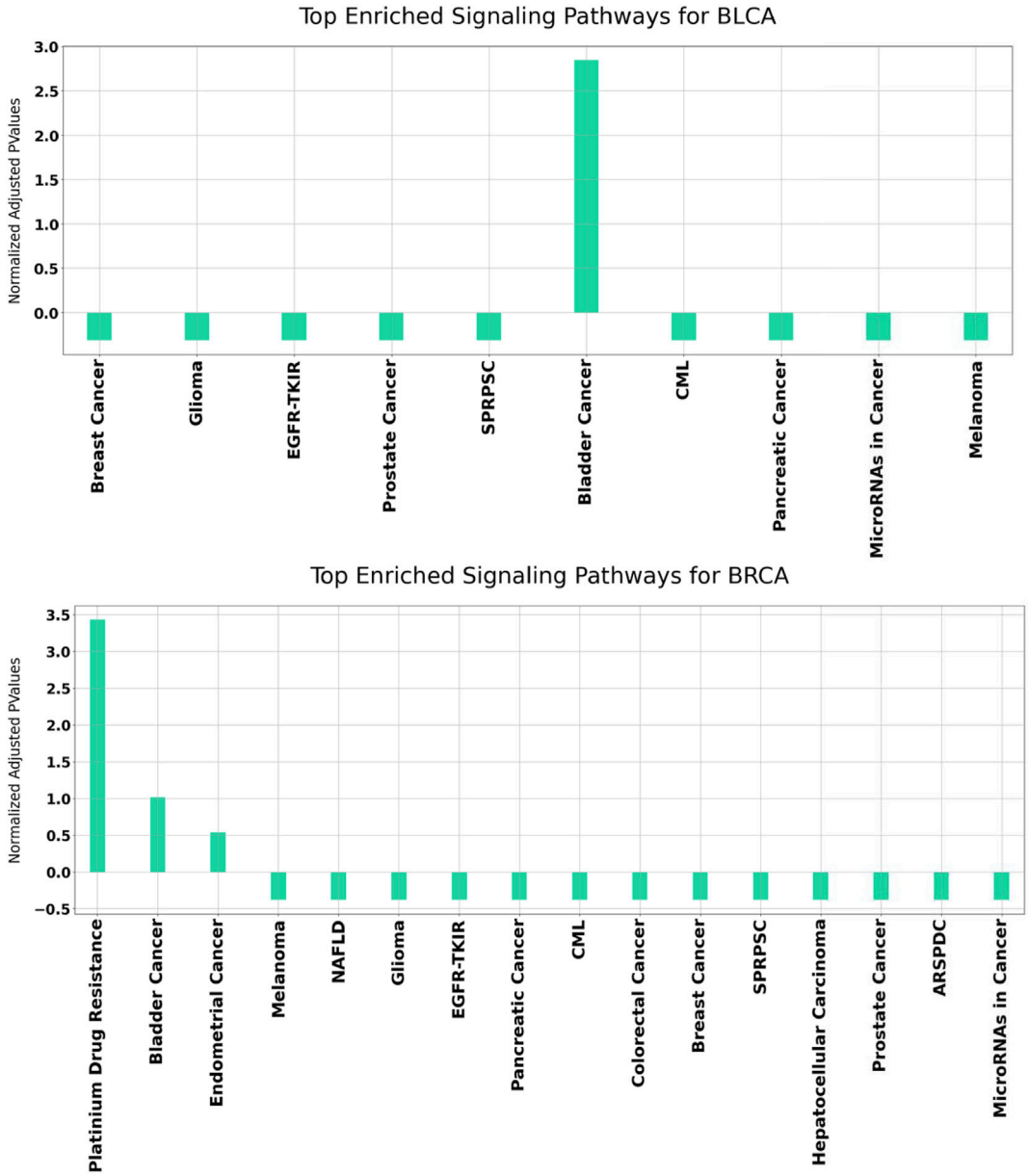

FIGURE 7 | Functional enrichment results for BLCA and BRCA using GeneCodis. The *p* Values of the enriched KEGG pathways refer to the normalized values using mean normalization. SPRPSC stands for Signaling Pathways Regulating Pluripotency of Stem Cells, EGFR-TKIR stands for EGFR Tyrosine Kinase Inhibitor Resistance, CML stands for Chronic Myeloid Leukemia, NAFLD stands for Non-Alcoholic fatty Liver Disease, ARSPDC stands for AGE-RAGE Signaling Pathway in Diabetic Complications.

identified set of genes in BLCA and BRCA datasets are presented in **Figure 7**.

It can be observed from **Figure 7** that for both BLCA and BRCA, the overrepresented pathways are directly related to the specific cancer types. We also felt that it was important to determine the pathways affecting different cancers and, we carried out additional procedures to better understand the molecular level relational networks of cancer. Using DAVID, we found that 55 pathways were commonly enriched in all the cancers tested. For these 55 pathways, a pathway - pathway interaction network was generated using the method that was developed in (Goy et al., 2019). A pathway network was obtained by examining the commonalities among the genes of the overrepresented pathways. Kappa statistics were used as distance metric. In order to construct a pathway - pathway interaction network, 3,025 pairwise relationships were analyzed for 55 commonly overrepresented pathways for 11 cancer types. To be able to find biologically relevant pairs, we used a Kappa score threshold. In this way, we aimed to keep only the interaction pairs, which are considered to be statistically important in terms

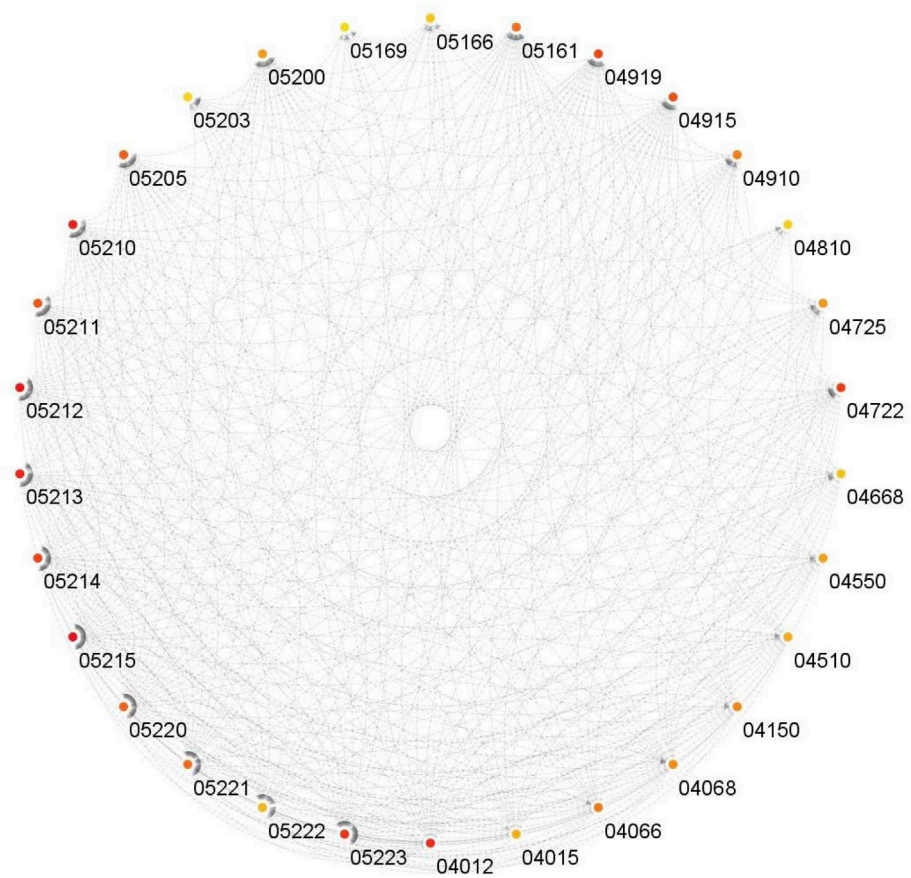

FIGURE 8 | Interactions of the commonly overrepresented pathways in all datasets.

TABLE 7 | Performance results on the external validation data.

Experiments using different gene levels (1–5–30–50)	Sensitivity	Specificity	Accuracy	F-measure
Random 1 gene	0.43	0.58	0.51	0.44
Top 1 gene of mirModuleNet	0.84	0.88	0.87	0.85
Random 5 genes	0.46	0.61	0.55	0.48
Top 5 genes of mirModuleNet	0.94	0.81	0.88	0.87
Random 30 genes	0.57	0.91	0.76	0.68
Top 30 genes of mirModuleNet	0.94	0.92	0.93	0.92
Random 50 genes	0.76	0.94	0.86	0.83
Top 50 genes of mirModuleNet	0.94	0.97	0.95	0.94

In all experiments, the model is trained on TCGA- LUSC data and tested on external data, which is LUSC_E.

of understanding the mechanisms of diseases at the molecular level. When this threshold was set as 0.15, the number of pathway pairs decreased to 403. The cytoHubba plugin (Chin et al., 2014) in the Cytoscape (Shannon et al., 2003) was used to detect the most important nodes in this pathway-pathway interaction network and Matthews Correlation Coefficient (MCC) values of each node (pathway) were calculated. We observed that 30 of the 55 pathways had very high MCC scores (between E^{14} and E^{30}). The constructed pathway-pathway interaction network is presented in **Figure 8**.

3.4 Validation of miRModuleNet's Results Using External Data

In order to check the robustness and reliability of miRModuleNet, an external dataset was considered. In this

TABLE 8 | Biological validation of the identified miRNAs for LUSC data by miRModuleNet, against five disease databases, i.e., dbDEMC, miRcancer, miR2Disease, PhenomiR, HMDD.

miRNA	Score (*p*-value)	Source(s)
hsa-miR-181a-5p	4.83E-58	dbDEMC, miRcancer, PhenomiR
hsa-miR-126-5p	2.79E-57	dbDEMC, miRcancer, miR2Disease, PhenomiR, HMDD
hsa-miR-140-3p	5.9E-55	dbDEMC, miRcancer, miR2Disease, PhenomiR, HMDD
hsa-miR-708-5p	5.9E-55	dbDEMC, miRcancer
hsa-miR-195-5p	5.9E-55	dbDEMC, miRcancer, miR2Disease, PhenomiR, HMDD
hsa-miR-30d-5p	7.76E-53	dbDEMC, miRcancer, PhenomiR,HMDD
hsa-miR-30a-5p	7,76E-53	dbDEMC, miRcancer, miR2Disease, PhenomiR, HMDD

TABLE 9 | Summary of the comparison against the databases of miRNA–disease associations.

Disease	Number of miRNA-disease associations identified by miRModuleNet	Number of databases containing the specific miRNA—disease association				
		1	2	3	4	5
BLCA	62	21	17	9	6	2
BRCA	51	4	15	19	11	—
KICH	61	34	15	—	—	—
KIRC	46	27	9	5	—	—
KIRP	87	44	19	4	—	—
LUAD	91	11	26	31	15	8
LUSC	54	2	6	10	15	20
PRAD	53	9	11	14	13	4
STAD	35	8	14	6	4	2
THCA	55	28	9	8	2	4
UCEC	87	46	20	—	—	—

The numbers in the table indicate the number of identified miRNA–disease associations included in 1, 2, 3, 4, or 5 different databases.

context, the GSE40419 dataset (Seo et al., 2012) was downloaded from the Gene Expression Omnibus database (Barrett & Edgar, 2006). The GSE40419 dataset was derived from 87 lung carcinoma cases and 77 normal people not having the disease. In this study, we refer to this dataset as LUSC_E. In our validation experiments, while the TCGA LUSC data is used as a train set, the LUSC_E dataset is used as a test set. To this end, we have used another KNIME workflow, which is developed for this type of tests. This workflow has also been added as a supplementary material.

Testing was carried out as follows. All genes for specific diseases in the train data and significant genes obtained by miRModuleNet are kept in separate files. To make a fair comparison, the number of random and significant genes was determined as 1, 5, 30, and 50. Subsequently, using the test KNIME workflow, the results were obtained both using these random genes and using the significant genes found by miRModuleNet. While the accuracy obtained using only 1 random gene was 51%, the accuracy reached 87% when the most important 1 gene found by miRModuleNet was used. Likewise, when comparing 50 genes, accuracy increased by approximately 11% with miRModuleNet, and reached 95%. Summary of these results are shown in **Table 7**. It can be concluded from **Table 7** that miRModuleNet is robust, reliable and noteworthy. Moreover, the performance for the training data (TCGA LUSC) is also presented as a supplementary file.

4 DISCUSSIONS

4.1 Biological Interpretation of the Results

In bioinformatics problems, the biological value that the tool is providing is as important as the comparative performance evaluation with existing tools. In this section, we explore those features and provide a biological validation of our tool.

4.2 Validation of miRModuleNet's Results on miRNA-Disease Association Databases

miRModuleNet produces multiple files as an output. One of these output files is the list of significant miRNA groups that are predicted to have a relationship with the disease and the genes targeted by these miRNAs. In the output file, these miRNAs are sorted according to their *p*-Values, which are assigned by the RobustRankAggreg method. In order to show the biological relevance of our findings, we refer to the miRNA - Disease association databases that are widely used in the literature. These databases are HMDD (Huang et al., 2019), miR2Disease (Jiang et al., 2009), miRcancer (Xie et al., 2013), dbDEMC (Yang et al., 2010) and PhenomiR (Ruepp et al., 2010). For each disease, miRNAs which were scored high in miRModuleNet and have *p*-Value less than 0.05 were checked in these databases to see whether there was a known relationship with the disease under study. **Table 8** presents the comparison of the miRNAs identified for Lung squamous cell

carcinoma (LUSC) against these five databases. This table displays the identified miRNA, its *p*-Value and the databases in which the miRNA is known to be associated with the relevant disease. For 11 different cancer datasets, a total of 682 miRNAs were found to be important by miRModuleNet. Among these selected miRNAs, approximately 34% of them were found in only one database, 23% were present in 2 databases, 15% in 3 databases, 10% in 4 databases, and 6% in 5 databases and 75 of the identified miRNAs were not listed in any of the databases. The details are presented in **Table 9**.

It is very difficult to develop a sound machine learning model for diseases such as cancer, which have complex molecular mechanisms. In order to overcome this challenge, it is crucial to integrate different types of -omics data. Hence, effective machine learning models that provide reliable results need to be developed. To this end, in this study we aimed to develop a robust machine learning model that can classify the samples as cancer patients and controls via integrating miRNA and mRNA expression profiles. A variety of studies have been reported that use either mRNA or miRNA data alone or in combined fashion. Some studies are only presented as methods and others as publicly available tools. However, most of the existing tools are limited in use and, to the best of our knowledge, are web based and R based. MMIA (Nam et al., 2009), MAGIA (Sales et al., 2010), miRConnX (Huang et al., 2011) originally offered as web servers and are currently not available. anamiR (Wang et al., 2019) and miRComb (Vila-Casadesús et al., 2016) which are offered as R packages, cannot be used with the latest versions of R.

In comparison, the miRModuleNet has a user-friendly structure and is evaluated on 11 different cancer datasets. In addition, although we focused on a biological problem in miRModuleNet, the same approach can be adapted to any classification problem including two dimensional data. This is not the case with most of the models listed above. miRModuleNet KNIME workflow generates different output files. These outputs provide information about identified mRNAs, miRNAs and their groupings. The mRNAs, miRNAs and mRNA-miRNA groups that were considered to be potentially important were identified and all results were validated using the following two methods. The first is a literature based validation of the miRNA - disease relationships that were predicted by the miRModuleNet using five widely used databases, i.e., dbDEMC, miRcancer, miR2Disease, PhenomiR, HMDD. The second method is validation using an independent external dataset that was not included in training. Such experiments evaluate whether the generated model can be utilized on a totally independent cohort. Our findings using four different levels (1, 5, 30 and 50 genes) imply that miRModuleNet maintains good performance metrics when applied to new independent data.

5 CONCLUSION

Exploring the biological functions of differentially expressed genes through the integration of different types of -omics data such as miRNA and mRNA expression profiles remains an important research topic. However, the problems associated with how to best assess the repression effect on target genes using integrated miRNA/mRNA expression profiles are not fully resolved. To address this problem, we have proposed a novel tool, miRModuleNet, which conducts a machine learning-based integration of two-omics datasets to detect miRNA-mRNA modules that are most significant to the classification task. The tool detects the miRNA/mRNA groups, which are later subjected to Rank procedure. The strength of miRModuleNet is that the identified set of genes that are represented in groups are guaranteed to distinguish two classes (cases vs. controls) and may serve as a biomarker for the specific disease under investigation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.767455/full#supplementary-material

REFERENCES

Allmer, J., and Yousef, M. (2022). "miRNomics: microRNA Biology and Computational Analysis," in *Methods in Molecular Biology* (Totowa, NJ, US: Humana Press).

Allmer, J., and Yousef, M. (2016). Computational miRNomics. *J. Integr. Bioinformatics* 13, 1–2. doi:10.1515/jib-2016-302

Barrett, T., and Edgar, R. (2006). [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis. *Methods Enzymol.* 411, 352–369. doi:10.1016/S0076-6879(06)11019-8

Bartel, D. P. (2004). MicroRNAs. *Cell* 116, 281–297. doi:10.1016/S0092-8674(04)00045-5

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2008). "KNIME: The Konstanz Information Miner," in Data Analysis, Machine Learning And Applications. *Studies in Classification, Data Analysis, and Knowledge Organization*. Editors C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Berlin, Heidelberg: Springer), 319–326. doi:10.1007/978-3-540-78246-9_38

Breiman, L. (2001). Random forests.. *Machine learning 45.1*, 5–32.

Cai, Y., Yu, X., Hu, S., and Yu, J. (2009). A Brief Review on the Mechanisms of miRNA Regulation. *Genomics, Proteomics & Bioinformatics* 7, 147–154. doi:10.1016/S1672-0229(08)60044-3

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., and Lin, C.-Y. (2014). cytoHubba: Identifying Hub Objects and Sub-networks from Complex Interactome. *BMC Syst. Biol.* 8, S11. doi:10.1186/1752-0509-8-S4-S11

Feng, Y., Xing, Y., Liu, Z., Yang, G., Niu, X., and Gao, D. (2018). Integrated Analysis of microRNA and mRNA Expression Profiles in Rats with Selenium Deficiency and Identification of Associated miRNA-mRNA Network. *Sci. Rep.* 8, 6601. doi:10.1038/s41598-018-24826-w

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most Mammalian mRNAs Are Conserved Targets of microRNAs. *Genome Res.* 19, 92–105. doi:10.1101/gr.082701.108

Goy, G., Yazici, M. U., and Bakir-Gungor, B. (2019). "A New Method to Identify Affected Pathway Subnetworks and Clusters in Colon Cancer," in 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 11-15 Sept. 2019 (Piscataway, NJ, US: IEEE), 671–675. doi:10.1109/UBMK.2019.8907141

Hailu, F. T., Karimpour-Fard, A., Toni, L. S., Bristow, M. R., Miyamoto, S. D., Stauffer, B. L., et al. (2021). Integrated Analysis of miRNA-mRNA Interaction in Pediatric Dilated Cardiomyopathy. *Pediatr. Res.* 2021, 1–11. doi:10.1038/s41390-021-01548-w

Hand, D. J. (2004). A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems. *Machine Learn.* 2004, 171–186.

Hecker, N., Stephan, C., Mollenkopf, H.-J., Jung, K., Preissner, R., and Meyer, H.-A. (2013). A New Algorithm for Integrated Analysis of miRNA-mRNA Interactions Based on Individual Classification Reveals Insights into Bladder Cancer. *PLoS ONE* 8, e64543. doi:10.1371/journal.pone.0064543

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). 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. (2009a). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Huang, G. T., Athanassiou, C., and Benos, P. V. (2011). mirConnX: Condition-specific mRNA-microRNA Network Integrator. *Nucleic Acids Res.* 39, W416–W423. doi:10.1093/nar/gkr276

Huang, J. C., Morris, Q. D., and Frey, B. J. (2007). Bayesian Inference of MicroRNA Targets from Sequence and Expression Data. *J. Comput. Biol.* 14, 550–563. doi:10.1089/cmb.2007.R002

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: a Database for Experimentally Supported Human microRNA-Disease Associations. *Nucleic Acids Res.* 47, D1013–D1017. doi:10.1093/nar/gky1010

Ivey, K. N., and Srivastava, D. (2015). microRNAs as Developmental Regulators. *Cold Spring Harb Perspect. Biol.* 7, a008144. doi:10.1101/cshperspect.a008144

Jayaswal, V., Lutherborrow, M., Ma, D. D., and Yang, Y. H. (2011). Identification of microRNA-mRNA Modules Using Microarray Data. *BMC Genomics* 12, 138. doi:10.1186/1471-2164-12-138

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a Manually Curated Database for microRNA Deregulation in Human Disease. *Nucleic Acids Res.* 37, D98–D104. doi:10.1093/nar/gkn714

Joung, J.-G., Hwang, K.-B., Nam, J.-W., Kim, S.-J., and Zhang, B.-T. (2007). Discovery of microRNA mRNA Modules via Population-Based Probabilistic Learning. *Bioinformatics* 23, 1141–1147. doi:10.1093/bioinformatics/btm045

Keller, A., Leidinger, P., Bauer, A., ElSharawy, A., Haas, J., Backes, C., et al. (2011). Toward the Blood-Borne miRNome of Human Diseases. *Nat. Methods* 8, 841–843. doi:10.1038/nmeth.1682

Lavrac, N., Kavsek, B., Flach, P., and Todorovski, L. (2004). Subgroup Discovery with CN2-SD. *J. Mach. Learn. Res.* 5, 153–188.

Le, H.-S., and Bar-Joseph, Z. (2013). Integrating Sequence, Expression and Interaction Data to Determine Condition-specific miRNA Regulation. *Bioinformatics* 29, i89–i97. doi:10.1093/bioinformatics/btt231

Li, L., Peng, M., Xue, W., Fan, Z., Wang, T., Lian, J., et al. (2018). Integrated Analysis of Dysregulated Long Non-coding RNAs/microRNAs/mRNAs in Metastasis of Lung Adenocarcinoma. *J. Transl Med.* 16. doi:10.1186/s12967-018-1732-z

Liu, B., Li, J., Tsykin, A., Liu, L., Gaur, A. B., and Goodall, G. J. (2009). Exploring Complex miRNA-mRNA Interactions with Bayesian Networks by Splitting-Averaging Strategy. *BMC Bioinformatics* 10, 408. doi:10.1186/1471-2105-10-408

Liu, Y., Zhang, J., Xu, Q., Kang, X., Wang, K., Wu, K., et al. (2018). Integrated miRNA-mRNA Analysis Reveals Regulatory Pathways Underlying the Curly Fleece Trait in Chinese Tan Sheep. *BMC Genomics* 19, 360. doi:10.1186/s12864-018-4736-4

Madadjim, R. (2021). *Using an Integrative Machine Learning Approach to Study microRNA Regulation Networks in Pancreatic Cancer Progression*. Lincoln, NE, US: University of Nebraska–Lincoln.

Masud Karim, S. M., Liu, L., Le, T. D., and Li, J. (2016). Identification of miRNA-mRNA Regulatory Modules by Exploring Collective Group Relationships. *BMC Genomics* 17, 7. doi:10.1186/s12864-015-2300-z

Nam, S., Li, M., Choi, K., Balch, C., Kim, S., and Nephew, K. P. (2009). MicroRNA and mRNA Integrated Analysis (MMIA): a Web Tool for Examining Biological Functions of microRNA Expression. *Nucleic Acids Res.* 37, W356–W362. doi:10.1093/nar/gkp294

Nersisyan, S., Galatenko, A., Galatenko, V., Shkurnikov, M., and Tonevitsky, A. (2021). miRGTF-Net: Integrative miRNA-Gene-TF Network Analysis Reveals Key Drivers of Breast Cancer Recurrence. *PLOS ONE* 16, e0249424. doi:10.1371/journal.pone.0249424

Paul, S., and Madhumita (2020). RFCM^3: Computational Method for Identification of miRNA-mRNA Regulatory Modules in Cervical Cancer.. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17 (5), 1729–1740. doi:10.1109/TCBB.2019.2910851

Pencheva, N., and Tavazoie, S. F. (2013). Control of Metastatic Progression by microRNA Regulatory Networks. *Nat. Cel Biol* 15, 546–554. doi:10.1038/ncb2769

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Ruepp, A., Kowarsch, A., Schmidl, D., Bruggenthin, F., Brauner, B., Dunger, I., et al. (2010). PhenomiR: a Knowledgebase for microRNA Expression in Diseases and Biological Processes. *Genome Biol.* 11, R6. doi:10.1186/gb-2010-11-1-r6

Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S., and Romualdi, C. (2010). MAGIA, a Web-Based Tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res.* 38, W352–W359. doi:10.1093/nar/gkq423

Schmidt, M. F. (2014). Drug Target miRNAs: Chances and Challenges. *Trends Biotechnol.* 32, 578–585. doi:10.1016/j.tibtech.2014.09.002

Seo, J.-S., Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, J. K., Bleazard, T., et al. (2012). The Transcriptional Landscape and Mutational Profile of Lung Adenocarcinoma. *Genome Res.* 22, 2109–2119. doi:10.1101/gr.145144.112

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A Bayesian Graphical Modeling Approach to Microrna Regulatory Network Inference. *Ann. Appl. Stat.* 4, 2024–2048. doi:10.1214/10-AOAS360

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012). GeneCodis3: a Non-redundant and Modular Enrichment Analysis Tool for Functional Genomics. *Nucleic Acids Res.* 40, W478–W483. doi:10.1093/nar/gks402

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review the Cancer Genome Atlas (TCGA): an Immeasurable Source of Knowledge. *wo* 1A, 68–77. doi:10.5114/wo.2014.47136

Tran, D. H., Satou, K., and Ho, T. B. (2008). Finding microRNA Regulatory Modules in Human Genome Using Rule Induction. *BMC Bioinformatics* 9, S5. doi:10.1186/1471-2105-9-S12-S5

Vila-Casadesús, M., Gironella, M., and Lozano, J. J. (2016). MiRComb: An R Package to Analyse miRNA-mRNA Interactions. Examples across Five Digestive Cancers. *PLOS ONE* 11, e0151127. doi:10.1371/journal.pone.0151127

Wang, T.-T., Lee, C.-Y., Lai, L.-C., Tsai, M.-H., Lu, T.-P., and Chuang, E. Y. (2019). anamiR: Integrated Analysis of MicroRNA and Gene Expression Profiling. *BMC Bioinformatics* 20, 239. doi:10.1186/s12859-019-2870-x

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-Cancer Association Database Constructed by Text Mining on Literature. *Bioinformatics* 29, 638–644. doi:10.1093/bioinformatics/btt014

Xu, Q.-S., and Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems 56.1*, 1–11. doi:10.1093/bioinformatics/btt014

Yang, L., Li, L., Ma, J., Yang, S., Zou, C., and Yu, X. (2019). miRNA and mRNA Integration Network Construction Reveals Novel Key Regulators in Left-Sided

and Right-Sided Colon Adenocarcinoma. *Biomed. Res. Int.* 2019, 1–9. doi:10.1155/2019/7149296

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a Database of Differentially Expressed miRNAs in Human Cancers. *BMC Genomics* 11, S5. doi:10.1186/1471-2164-11-S4-S5

Yao, Y., Jiang, C., Wang, F., Yan, H., Long, D., Zhao, J., et al. (2019). Integrative Analysis of miRNA and mRNA Expression Profiles Associated with Human Atrial Aging. *Front. Physiol.* 10, 1226. doi:10.3389/fphys.2019.01226

Yousef, M., Abdallah, L., and Allmer, J. (2019). maTE: Discovering Expressed Interactions between microRNAs and Their Targets. *Bioinformatics* 35, 4020–4028. doi:10.1093/bioinformatics/btz204

Yousef, M., Bakir-Gungor, B., Jabeer, A., Goy, G., Qureshi, R., and C. Showe, L. (2021a). Recursive Cluster Elimination Based Rank Function (SVM-RCE-R) Implemented in KNIME. *F1000Res* 9, 1255. doi:10.12688/f1000research.26880.2

Yousef, M., Goy, G., Mitra, R., Eischen, C. M., Jabeer, A., and Bakir-Gungor, B. (2021b). miRcorrNet: Machine Learning-Based Integration of miRNA and mRNA Expression Profiles, Combined with Feature Grouping and Ranking. *PeerJ* 9, e11458. doi:10.7717/peerj.11458

Yousef, M., Jung, S., Showe, L. C., and Showe, M. K. (2007). Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. *BMC Bioinformatics* 8, 144. doi:10.1186/1471-2105-8-144

Yousef, M., Kumar, A., and Bakir-Gungor, B. (2020). Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. *Entropy* 23, 2. doi:10.3390/e23010002

Yousef, M., Levy, D., and Allmer, J. (2018). "Species Categorization via MicroRNAs - Based on 3'UTR Target Sites Using Sequence Features," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4: BIOINFORMATICS* (Setúbal, Portugal: SciTePress), 112–118. doi:10.5220/0006593301120118

Yousef, M., Sayıcı, A., and Bakir-Gungor, B. (2021c). "Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis," in Database And Expert Systems Applications - DEXA 2021 Workshops. *Communications in Computer and Information Science*. Editors G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, et al. (Berlin, Heidelberg: Springer International Publishing), 205205–214214. doi:10.1007/978-3-030-87101-7_20

Yousef, M., Trinh, H., and Allmer, J. (2014). Intersection of MicroRNA and Gene Regulatory Networks and Their Implication in Cancer. *Cpb* 15, 445–454. doi:10.2174/13892010156661405191208 55

Yousef, M., Ülgen, E., and Uğur Sezerman, O. (2021d). CogNet: Classification of Gene Expression Data Based on Ranked Active-Subnetwork-Oriented KEGG Pathway Enrichment Analysis. *PeerJ Computer Sci.* 7, e336. doi:10.7717/peerj-cs.336

Zhang, S., Li, Q., Liu, J., and Zhou, X. J. (2011). A Novel Computational Framework for Simultaneous Integration of Multiple Types of Genomic Data to Identify microRNA-Gene Regulatory Modules. *Bioinformatics* 27, i401–i409. doi:10.1093/bioinformatics/btr206

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

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

Copyright © 2022 Yousef, Goy and Bakir-Gungor. 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.

frontiers | Frontiers in Genetics

BRIEF RESEARCH REPORT
published: 12 April 2022
doi: 10.3389/fgene.2022.820361

PAGER Web APP: An Interactive, Online Gene Set and Network Interpretation Tool for Functional Genomics

Zongliang Yue[1], *Radomir Slominski*[1,2], *Samuel Bharti*[1] *and Jake Y. Chen*[1]*

[1]Informatics Institute in the School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States, [2]Graduate Biomedical Sciences Program, The University of Alabama at Birmingham, Birmingham, AL, United States

Functional genomics studies have helped researchers annotate differentially expressed gene lists, extract gene expression signatures, and identify biological pathways from omics profiling experiments conducted on biological samples. The current geneset, network, and pathway analysis (GNPA) web servers, e.g., DAVID, EnrichR, WebGestaltR, or PAGER, do not allow automated integrative functional genomic downstream analysis. In this study, we developed a new web-based interactive application, "PAGER Web APP", which supports online R scripting of integrative GNPA. In a case study of melanoma drug resistance, we showed that the new PAGER Web APP enabled us to discover highly relevant pathways and network modules, leading to novel biological insights. We also compared PAGER Web APP's pathway analysis results retrieved among PAGER, EnrichR, and WebGestaltR to show its advantages in integrative GNPA. The interactive online web APP is publicly accessible from the link, https://aimed-lab.shinyapps.io/PAGERwebapp/.

Keywords: PAGER, melanoma, functional genomics, geneset analysis, network visualization and analysis, PAGER Web APP, GNPA

OPEN ACCESS

Edited by:
Sorin Draghici,
Wayne State University, United States

Reviewed by:
Bhanwar Lal Puniya,
University of Nebraska-Lincoln, United States
Parul Gupta,
Oregon State University, United States

***Correspondence:**
Jake Y. Chen
jakechen@uab.edu

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

Received: *22 November 2021*
Accepted: *17 March 2022*
Published: *12 April 2022*

Citation:
Yue Z, Slominski R, Bharti S and Chen JY (2022) PAGER Web APP: An Interactive, Online Gene Set and Network Interpretation Tool for Functional Genomics. Front. Genet. 13:820361. doi: 10.3389/fgene.2022.820361

INTRODUCTION

Functional genomics analysis is widely performed to characterize genes and intergenic regulatory regions in the genome that contribute to different biological processes (Yang et al., 2020; Angeloni et al., 2021). Essentially, functional genomics provides a way to reveal the molecules' coordination in mechanisms due to a specific phenotype (Raamsdonk et al., 2001; Rahaman et al., 2015). By tracking the molecular activities in the specific biological conditions, we could identify those driver and passenger genes working in a model linking genotype to phenotype. Numerous studies have shown that the molecules working in pathways could help in disease diagnosis (Zhang and Chen, 2010; Drier et al., 2013; Livshits et al., 2015; Bock and Ortea, 2020; Pian et al., 2021), cancer subtyping (Zhang and Chen, 2013; Mallavarapu et al., 2020; Lafferty et al., 2021), and personalized medicine (Chen et al., 2007; Hamburg and Collins, 2010; Raghavan et al., 2017). Additionally, multi-omics analysis provides a complex map linking transcriptomics, proteomics, and metabolomics (Subramanian et al., 2020; Andrieux and Chakraborty, 2021). In multi-omics studies, the challenges for functional genomics are the coverage of contents, the rendering of the complex network-based models, and the easy-to-use

software with advanced features. Therefore integrative geneset, network, and pathway analysis (GNPA) have emerged in the past decade to lessen the burden of multi-omics data analysis users (Wu et al., 2014). Pathway analysis, especially topology-based approaches that exploit all the knowledge about how genes and proteins interact in a pathway, have been developed to discover the mechanical changes through pathway-level scoring and pathway significance assessment (Draghici et al., 2007; Mitrea et al., 2013; Nguyen et al., 2018). To better understand the impact of perturbations or genetic modifications in a system-level, System-level PAThway Impact AnaLysis using map (SPATIAL), Signaling Pathway Impact Analysis - Global Perturbation Factor (SPIA-GPF), and SPATIAL-GPF have been introduced (Bokanizad et al., 2016).

During the last decade, several GNPA web servers have been developed (Subramanian et al., 2005; Khatri et al., 2012), including DAVID (Jiao et al., 2012), EnrichR (Kuleshov et al., 2016), WebGestalt (Liao et al., 2019), and pathways, annotated gene lists and gene signatures electronic repository (PAGER) (Yue et al., 2018). The highlights of those webservers are interactive and comprehensive data coverage. The first version of the DAVID tool was published in 2003 (Dennis et al., 2003), and it is one of the earliest geneset enrichment analysis webservers. The most updated version of DAVID implements many advanced features such as gene ranking, which gives a quick focus on the most likely important candidate genes, gene with annotation in each single view, and gene extension to make functional inferences (Jiao et al., 2012). EnrichR was initially developed in 2013, and its merits come from comprehensive data coverage and interactive visualization panel (Chen et al., 2013). EnichR provides 190 libraries and adds Appyter to visualize EnrichR results in different styles (Kuleshov et al., 2016). WebGestalt was introduced in 2005 (Zhang et al., 2005), and it highlights the visualization of gene ontology hierarchy structure and pathway view of wikiPathway. WebGestaltR implemented with R language in the recent updates (Liao et al., 2019).

PAGER was initially conceived in 2014 (Harini et al., 2008) and subsequently developed in 2015 (Yue et al., 2015) with a standardized concept called "PAGs" (Pathways, Annotated gene lists, and Gene signatures) that integrates different levels of gene-sets. PAGER highlights the measurement of biological relevance using normalized Cohesion Coefficient (nCoCo) and advances the network interpretation of functional genomics results in several aspects. Additionally, PAGER introduced the computational strategies in generating m-type (co-membership) or r-type (regulatory) PAG-to-PAG relationships. PAGER also provides gene prioritization in each PAG. For the intra-PAG network construction, PAGER adopts the protein-protein interactions from the HAPPI database (Chen et al., 2017), a comprehensive and high-quality map of Human annotated and predicted protein interactions, and gene regulations validated *in vitro* experiment. Hence, PAGER enables gene prioritization using the network topology in each PAG (Yue et al., 2018). All four web servers support API (Application Programming Interface) services.

In this study, we developed the PAGER Web APP, an interactive online application to perform the gene set enrichment analysis and network interpretation of the functional genomics result. PAGER Web APP provides preprocessed RNA-seq data from UALCAN-processed TCGA data (Chandrashekar et al., 2017) and a melanoma drug resistant-sensitive case study (Snyder et al., 2014) from cBioPortal (Gao et al., 2013). We illustrated how the PAGER Web APP enhances the potential to discover biological insights using network-based computational strategy by comparing the enriched pathways from the three leading web servers using their application programming interfaces (APIs). We performed three additional case studies, multiple sclerosis (MS), colonic mucosa in Crohn's disease (CD), and ulcerative colitis (UC) study, to compare the three web server performances and further validate the pathways using PubMed co-citations. We intend for PAGER Web APP to become a popular application for researchers interested in integrative GNPA.

METHODS

Workflow and User Interface

We developed a four-step procedure in performing the functional genomics analysis in PAGER Web APP for Human genomics results (**Figure 1**). Firstly, users need to either load Demo data or upload their data. In the Demo data, PAGER Web APP provides a melanoma dataset, a multiple sclerosis dataset, a Crohn's disease dataset, an ulcerative colitis dataset and 16 cancer types collected from UALCAN TCGA data (Chandrashekar et al., 2017). If users need to upload the data, we ask users to provide a tab-delimited.txt format file, check the log2 fold change column and *p*-value column, and click on the "proceed" button. Secondly, PAGER Web APP will generate a volcano plot using the gene's log2 fold changes and colors the over-expressed candidate genes red and under-expressed candidate genes blue using the default threshold p-value ≤ 0.05 and absolute log2 fold change ≥ 1. PAGER Web APP allows users to adjust the log2foldchange and negative log2-based p-value to optimize the candidate gene list. Users need to click on the proceed button to the next step. Thirdly, PAGER Web APP will perform the gene-set enrichment analysis with the pathway type geneset sources (P-type PAGs) in default. Users can add or remove the source name in the source multiple-choice field. PAGER Web APP also allows users to change the minimum number of overlapped genes, similarity score, and "-log2p-value" cutoff. The similarity score is based on the combination of overlap coefficient and Jaccard index using the methods described previously (Huang et al., 2012). In the table of enriched genesets results, users can use the column "PAGER link" to navigate to the web-hosted PAGER entries of the given PAG, including the metadata, gene members, and gene networks. PAGER Web APP offers two additional leading gene set enrichment analysis tools (EnrichR and WebGestaltR) using the API service. We didn't include DAVID due to the API failure. Lastly, PAGER Web APP summarizes the similarity of the terms and displays a Venn diagram of the overlapped terms.

FIGURE 1 | The PAGER Web APP data analysis workflow. The workflow consists of four steps with visualization panels to help biologists quickly understand the results.

PAGER Web APP also provides the corresponding tables to deliver similar terms with similarity scores by comparing the three tools. All the tables and plots are downloadable.

Term to Term Distance-Based Similarity of Terms Enriched From the Three Tools

The term similarity is generated based on a string metric using the Stringdist library (https://cran.r-project.org/web/packages/stringdist/index.html). We clean up the terms or names by removing irrelevant content, such as species, identifier, etc., and making all the terms lower case. We also remove the redundant terms enriched from different data sources, such as "MAPK signaling pathway" may come from KEGG and wikiPathway at the same time. Then we apply the string similarity using optimal string alignment (OSA) distance (Boytsov, 2011) to generate the similarity matrix between two sets of terms, set A and set B.

Assume there are two terms regarded as two strings a and b, the restricted distance is defined as $d_{a,b}(i,j)$ in a recursive calculation, the i is the prefix of string a, and the j is the prefix of string b.

$$d_{a,b}(i,j) = \min \begin{cases} 0 \ if \ i = j = 0 \\ d_{a,b}(i-1,j) + 1 \ if \ i > 0 \ (deletion) \\ d_{a,b}(i,j-1) + 1 \ if \ j > 0 \ (insertion) \\ d_{a,b}(i-1,j-1) \ if \ a_i = b_j \ (match) \\ d_{a,b}(i-1,j-1) + 1 \ if \ a_i \neq b_j \ (substitution) \\ d_{a,b}(i-2,j-2) + 1 \ if \ i,j > 1 \ and \ a_i = b_{j-1} \ and \ a_{i-1} = b_j \ (transposition) \end{cases}$$

The string similarity is calculated by:

$$1 - \frac{d_{a,b}(i,j)}{\max(|a|,|b|)}$$

where $|a|$ represents the length of string a, and $|b|$ represents the length of string b.

After generating the similarity matrix between the two lists of terms, we check each row (a term from the set A) and take the highest score with the term as the most similar term. Therefore, we generate a list of pairwise term-to-term similarities. Finally, we use the default or customized similarity cutoff to filter low similar term-to-term pairs.

Apply Louvain Clustering in m-Type PAG-To-PAG Networks to Identify PAG Communities

We apply the Louvain clustering function in the igraph library in R (https://cran.r-project.org/web/packages/igraph/index.html) to find the community structure in m-type PAG-to-PAG networks. The Louvain clustering is based on the modularity in a scale between -0.5 (non-modular clustering) to 1 (fully modular clustering) described in the paper (Blondel et al., 2008).

Extract the Critical Concepts From Pathways and Show Them in Word-Clouds

We create bag-of-words from the space-separated PAG names to present the frequently appearing words in each PAG for any enriched PAG set. We create word corpus, remove the potential punctuation such as comma, colon, etc., make all the words lower case, remove both irrelevant words and common words, "pathway," "signaling," "human," "homo," "sapiens," "has," "or," and "and". Finally, we apply wordcloud2 function in the wordcloud2 library in R (https://cran.r-project.org/web/packages/wordcloud2/index.html) for the visualization.

Implementing the Software

The PAGER Web Application user interface is designed using bs4Dash (https://cran.r-project.org/web/packages/bs4Dash/index.html) package in R. The application is supported by R Shiny (https://shiny.rstudio.com/) framework. In addition to data processing and statistical analysis, GNPA Analysis are implemented using PAGER API, EnrichR API and WebGestaltR API. Graphing libraries like Plotly (https://plotly.com/r/), igraph in R (https://igraph.org/r/), ggplot2, wordcloud2, and VennDiagram have been used.

Prepare the Melanoma Drug Resistant-Sensitive Data From cBioPortal

We downloaded the melanoma dataset of 64 patient samples from cBioPortal, and it was initially published in a paper in the New England Journal of Medicine (Snyder et al., 2014). We identified a cohort from the patients who are in the metastasis stage (m1c) with Neuroblastoma RAS Viral Oncogene Homolog (NRAS gene) or/and v-Raf murine sarcoma viral oncogene homolog B (BRAF gene) mutations. Hence, we obtained three drug response patients, two drug weakly response patients and seven non-response patients. We applied the DEseq2 library in R (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) to generate the differentially expressed genes that compared drug-resistant patients to drug-sensitive patients. The output file is stored in the PAGER Web APP as a demo.

Prepare the Multiple Sclerosis Data From the EMBL-EBI Database

We loaded the differentially expressed gene table, "extdata/E-GEOD-21942.topTable.RData", preprocessed in the ROntoTools library (Ansari et al., 2016). This dataset contains a genome-wide array expression study in peripheral blood mononuclear cells (PBMC) from 12 multiple sclerosis (MS) patients and 15 controls (Kemppinen et al., 2011). We selected differentially expressed genes using adjusted p-value ≤ 0.01 (2,864 genes) and saved their fold changes as input of ROntoTools. We set the adjusted p-value ≤ 0.01 and the absolute logFC >0.5 to get the 1,470 candidate genes as the input of PAGER, EnrichR and WebGestaltR.

Prepare the Colonic Mucosa Data From the EMBL-EBI Database

We downloaded the transcription profiling by array of RNA from inflamed and non-inflamed colonic mucosa (E-MTAB-2967). In Crohn's disease, there are 15 inflamed colonic mucosa and 15 controls. In ulcerative colitis, there are 14 inflamed colonic mucosa and 14 controls. We performed the normalization and linear regression using the limma library in R (https://bioconductor.org/packages/release/bioc/html/limma.html). We set the cutoffs of adjusted p-value ≤ 0.05 and the absolute logFC >0.5 to get the 518 candidate genes in Crohn's disease and the 528 candidate genes in the ulcerative colitis study.

Validation of Pathways Using the Co-citations in PubMed Literature

To demonstrate the significance of the keywords in pathways related to a disease, we applied a co-citation enrichment analysis using the hypergeometric test and odds ratio. We applied the NCBI e-utils application programming interface (API) that implements semantic searches of PubMed abstracts to report biomedical literature citations (Sayers, 2008). We implied that the likelihood of observing articles co-mentioning disease names and the keywords from pathways is statistically higher than random using the PubMed score (Yue et al., 2019a). In this study, the background citations using the word "disease" denoted as N, the citations of the specific disease using the word "melanoma" represented as K, the citations of the keywords from a pathway denoted as n, and the joint citations of "melanoma" and the keywords from a pathway represented as k. We performed the co-citation enrichment analysis to generate *PubMed score* using the formula:

$$PubMed\ score = -\ln\left(\sum_{t=k}^{\min(n,K)} \frac{\binom{K}{t}\binom{N-K}{n-t}}{\binom{N}{n}}\right)$$

We calculated the odds ratio based on the formula $\frac{k/(K-k)}{(n-k)/(N-K-n+k)}$. We also manually checked the contents and subsequently confirmed them using the PubTator annotation API (Wei et al., 2019; Wei et al., 2013), i.e., https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/pubtator?pmids=[PMID]. We took a sample list of PubMed IDs from each retrieved entry. To remove biases and further confirm the mentioned keywords, we applied the analysis described in the previously developed tool called biomedical entity expansion ranking and exploration (BEERE) (Yue et al., 2019b) to extract those semantic relationships that co-mention "melanoma" and the pathways' keywords.

To evaluate how well the method can identify "correct" pathways, we introduced a new hybrid validation technique. It involves first defining the ground truth and subsequently developing a statistical model to assess the significance of results retrieved using a receiver operating characteristics (ROC) curve and the area-under-the-curve (AUC) value. The hybrid technique also includes performing a literature co-citation-based assessment. We constructed the ground truth using ROntoTools, the best performing method reported in the review paper (Nguyen et al., 2019), in three steps. Firstly, we took the candidate genes from the differential expression analysis using the adjusted p-value cutoff 0.05 and the absolute gene's log fold-changes larger than or equal to 0.5. Secondly, we performed pathway enrichment analysis using the ROntoTools. Thirdly, we defined the "true" data set as the significantly enriched pathways with adjusted combined p-values ≤ 0.05 (combined p-values were generated by the function "comb.pv.func" (Kemppinen et al., 2011) in ROntoTools) and the "false" data set as the retrieved pathways with adjusted combined p-values > 0.05 but with at least one gene overlapping with the input gene list.

TABLE 1 | A comparison of data coverage and features among PAGER, EnrichR, and WebGestalt web servers.

Webserver		PAGER	EnrichR	Webgestalt
Data coverage (Human)	Unique library	35	89	22
	Metadata	Yes	Partial	Partial
	Gene prioritization	Yes	No	No
Geneset intra-network	Interactions	Yes	No	Partial
	Regulations	Yes	No	Partial
Geneset inter-network	m-type (co-membership)	Yes	Partial	Partial
	r-type (regulatory)	Yes	No	No
Additional feature	Term searching	Yes	Yes	No
	API	Yes	Yes	Yes

In the literature co-citation validation, we developed a t-test based statistical model on evaluating how significant the *p*-value ranked pathways can be supported by the PubMed scores. Particularly, we ranked the PAGs based on adjusted *p*-values, and compared the top n% PAGs' PubMed scores to the bottom (100-*n*)% PAGs' PubMed scores for each method, where n ranges from 10 to 90 with a step increment of 10. And then, we reported their average *p*-values, respectively. The smaller *p*-values are, the better performance the methods have.

RESULTS

Comparison of Data Coverage and Features Among the Three Web Servers

Compared to EnrichR and WebGestalt, PAGER progresses the network interpretation of functional genomics results. Although there are 35 unique geneset libraries reported in most updated PAGER, which are less than EnrichR, each of PAG in PAGER contains metadata other than EnrichR and WebGestalt, including PAG-type (pathways, annotated gene lists and gene signatures), PAG descriptions, source link, publication reference, curator, and nCoCo score (described in PAGER 2.0). In addition, PAGER provides geneset intra-network views, including the protein-protein interaction network and gene-gene regulation network members in each geneset, while WebGestalt reports pathway maps in wikiPathway source only. For the geneset's inter-network, WebGestalt inherits the Gene Ontology (GO) hieratical structure from the GO consortium. We extent the relationship concepts by introducing m-type (co-membership) PAG-to-PAG relationships and r-type (regulatory) PAG-to-PAG relationships described in PAGER. The m-type PAG-to-PAG relationships represent co-memberships between two PAGs, which reveals signaling cross-talk between PAGs that share signaling components within signal transduction pathways in response to external stimuli. The r-type PAG-to-PAG relationships represent the PAG causal ordering inferred from gene-to-gene regulations by adapting our method previously described in PAGER (Yue et al., 2018). The PAGER Web APP fulfills all the additional features in **Table 1**, such as term searching and API service.

Melanoma Drug Resistant-Sensitive Patients Enriched Pathway Case Study in Demo

To better identify the cohorts in melanoma cancer to improve the treatment, functional genomics has been applied to the next-generation sequencing data for an in-depth understanding of the molecular mechanisms in the drug resistance cases. We collected the transcriptomes from the cBioPortal database in this study. In the result, we found 164 P-type PAGs (pathways) to be significantly enriched.

In the 164 P-type PAGs, they are two PAGs that are derived from more than one data source, i.e., "PI3K-Akt signaling pathway" and "Bladder cancer", each of which is simultaneously recorded in both "WikiPathway_2021" and "KEGG_2021_HUMAN" data sources. Compared to the results from EnrichR and WebGestaltR, PAGER had the greatest number of enriched pathways, which is 164, EnrichR has 98, and WebGestaltR has 52 (**Figure 2A**). PAGER also had the greatest number of unique pathways, which is 101 (48%). We found 33 (16%) overlapped pathways among the three tools. In addition, 23 (11%) pathways were shared between PAGER and EnrichR, 5 (2.4%) pathways were shared between PAGER and WebGestaltR, and 6 (2.8%) pathways were shared between EnrichR and WebgestaltR.

In the 164 P-type PAGs reported by PAGER, there were 4 major sources and 1 minor source (**Figure 2B**). 50 (30.5%) are from wikiPathway, 46 (28%) are from Reactome, 37 (22.6%) are from Protein Lounge, 28 (17.1%) are from KEGG, and 3 (1.83%) are from Spike. We showed the top-30 enriched P-type PAGs colored by the sources in the horizontal bar-plot in **Figure 2C**, and the details of the enriched PAGs are in **Supplementary Table S1**.

Critical Terms Extraction From the Louvain Clustered PAGs in the m-Type PAG-To-PAG Networks

The 164 P-type PAGs form a densely connected m-type PAG-to-PAG network (2,749 m-type PAG-to-PAG relationships) with an average degree of 18. After the community detection using Louvain clustering, we found 5 PAG clusters in the m-type

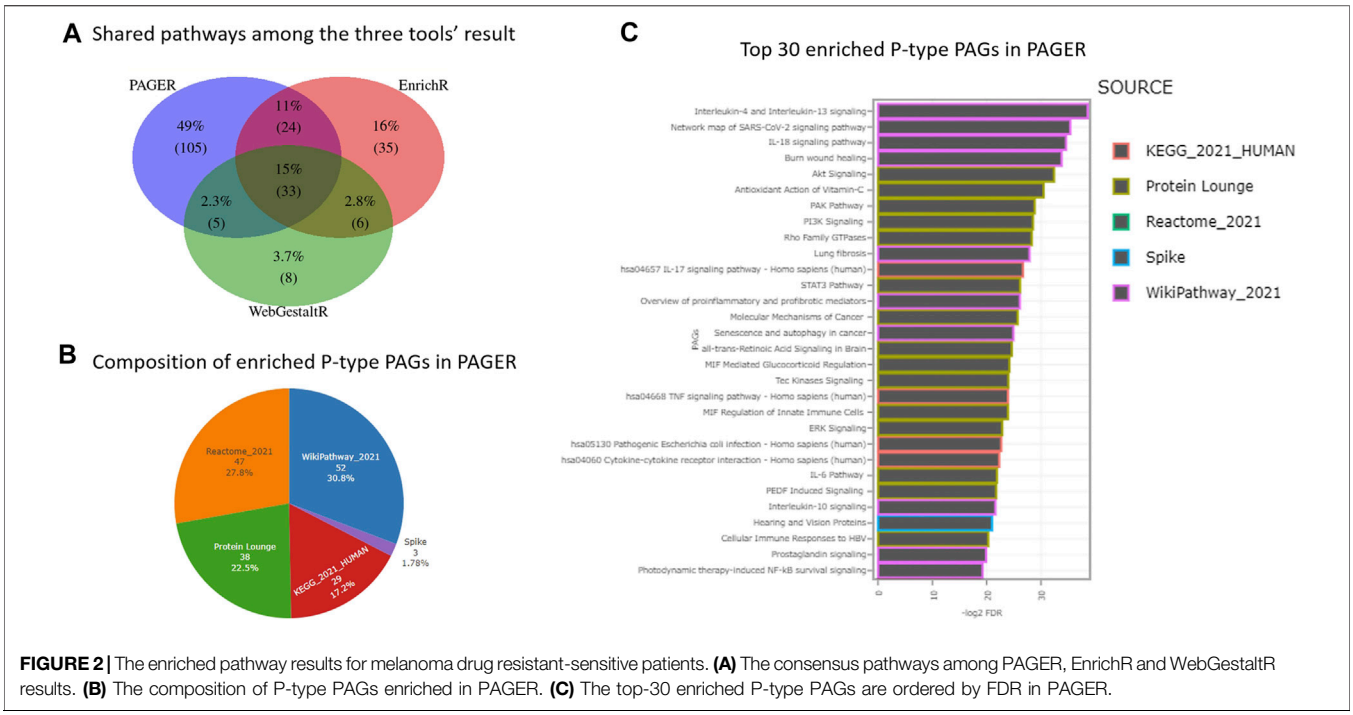

FIGURE 2 | The enriched pathway results for melanoma drug resistant-sensitive patients. **(A)** The consensus pathways among PAGER, EnrichR and WebGestaltR results. **(B)** The composition of P-type PAGs enriched in PAGER. **(C)** The top-30 enriched P-type PAGs are ordered by FDR in PAGER.

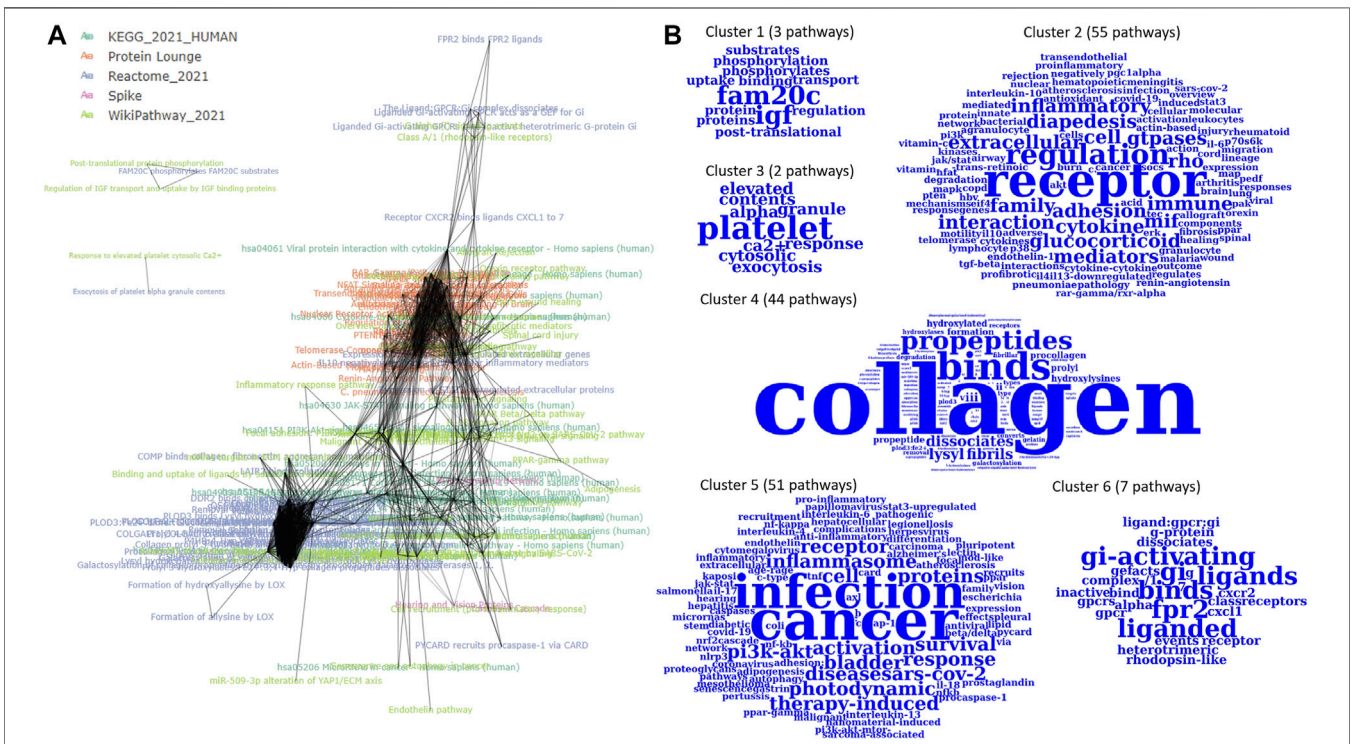

FIGURE 3 | The m-type PAG-to-PAG network of enriched P-type PAGs in PAGER for melanoma drug resistant-sensitive patients. **(A)** The m-type PAG-to-PAG network overview and **(B)** The extracted word clouds from the Louvain clusters in the network.

PAG-to-PAG network (**Figure 3**). The extracted concepts reveal the general pathway functions in the clusters. **Cluster 1** consists of 3 pathways with represented terms "FAM20C" protein (Golgi-associated secretory pathway kinase), "IGF" protein (insulin-like growth factor). **Cluster 2** has 7 pathways related to the Gi-activating and ligand-receptor bindings. **Cluster 3** is

TABLE 2 | The 33 consensus pathways among PAGER, EnrichR, and WebGestaltR results with PubMed literature support. W vs. P represents the term similarities between WebGestaltR and PAGER results. P vs. E represents the term similarities between PAGER and EnrichR results. W vs. E represents the term similarities between WebGestaltR and EnrichR. ***k*** represents the citations of "melanoma" and the keywords from a pathway. ***OR*** represents the odds ratio. Score represents the ***PubMed score***. **PMID** represents one PubMed ID example from each entry. BEERE validation represents the semantic relationships retrieved. 1 stands for Yes, and 0 stands for No. All these abbreviations are applied to **Table 3** and **Table 4**.

Term	W vs. P (%)	P vs. E (%)	W vs. E (%)	Keywords	k	OR	Score	PMID	BEERE validation
Photodynamic therapy-induced ap-1 survival signaling.	100	100	100	Photodynamic therapy	1,076	1.150	1.20E+01	31378787	1
mir-509-3p alteration of yap1/ecm axis	100	100	100	mir-509-3p	3	2.376	1.94E+00	33968718	1
Transcriptional misregulation in cancer	100	100	100	Transcriptional misregulation in cancer	10	1.261	1.27E+00	32079144	1
Photodynamic therapy-induced nf-kb survival signaling	100	100	100	Photodynamic, nf-kb	2	1.261	7.40E-01	16524427	1
Apoptosis-related network due to altered notch3 in ovarian cancer	100	100	100	Notch3 ovarian cancer	2	1.154	6.49E-01	28165469	1
Senescence and autophagy in cancer	100	100	100	Senescence and autophagy in cancer	35	0.722	1.97E-02	12789281	1
Focal adhesion: pi3k-akt-mtor-signaling pathway	96	96	100	pi3k-akt-mtor-signaling pathway	36	0.699	1.04E-02	31370278	0
Cytokine-cytokine receptor interaction	100	100	100	Cytokine-cytokine receptor	14	0.442	1.72E-04	34824546	1
il-18 signaling pathway	100	100	100	il-18 pathway	25	0.482	1.54E-05	31731729	1
Mirna targets in ecm and membrane receptors	100	100	100	mirna membrane receptors	2	0.104	1.22E-07	34680340	0
c-type lectin receptor signaling pathway	100	100	100	c-type lectin receptor signaling pathway	28	0.360	4.03E-11	29497419	1
Nod-like receptor signaling pathway	100	100	100	Nod-like receptor signaling pathway	50	0.394	5.13E-15	34747716	0
il-17 signaling pathway	100	100	100	il-17 pathway	20	0.215	2.80E-20	30079767	1
Protein digestion and absorption	100	100	100	Protein digestion and absorption	5	0.062	4.10E-29	30900145	0
Assembly of collagen fibrils and other multimeric structures	100	100	100	Collagen assembly	13	0.126	2.30E-29	29216889	1
Bladder cancer	100	100	100	Bladder cancer	1815	0.708	6.34E-53	35059301	1
Class a/1 (rhodopsin-like receptors)	100	100	100	Adenosine a1 receptor	10	0.056	8.79E-63	8463264	1
Legionellosis	100	100	100	Legionellosis	1	0.006	7.10E-77	17870669	0
Prostaglandin synthesis and regulation	100	100	100	Prostaglandin synthesis and regulation	75	0.162	5.42E-110	3149408	1
Response to elevated platelet cytosolic ca2+	100	100	100	Platelet, calcium	44	0.105	1.95E-120	32562975	1
Hepatitis c and hepatocellular carcinoma	100	100	100	Hepatitis c and hepatocellular carcinoma	20	0.054	4.73E-127	31538700	0
Interleukin-6 family signaling	100	100	100	il-6 signaling pathway	170	0.237	2.42E-131	22713796	1
tnf signaling pathway	100	100	100	tnf signaling pathway	246	0.260	1.65E-159	30591049	1
Inflammatory response pathway	100	100	100	Inflammatory response pathway	123	0.173	2.30E-161	32517213	1
Amoebiasis	100	100	100	Amoebiasis	5	0.012	2.14E-177	31173190	0
Pertussis	100	100	100	Pertussis	83	0.083	1.46E-303	23737697	1
Cytokines and inflammatory response	100	100	100	Cytokines, inflammatory response	559	0.212	0.00E+00	31176707	0
Lung fibrosis	100	100	100	Lung fibrosis	118	0.057	0.00E+00	31249780	1
Malaria	100	100	100	Malaria	137	0.044	0.00E+00	14657217	1
Micrornas in cancer	100	100	100	Micrornas	1,211	0.342	0.00E+00	28118616	1
Rheumatoid arthritis	100	100	100	Rheumatoid arthritis	357	0.074	0.00E+00	27307502	0
salmonella infection	100	100	100	*salmonella*	168	0.057	0.00E+00	11773163	0
Spinal cord injury	100	100	100	Spinal cord injury	82	0.034	0.00E+00	30008656	0

formed by 43 pathways related to collagen formation and binding events. **Cluster 4** has 37 pathways related to inflammasome responses in cancer or infection. **Cluster 5** contains 66 pathways with the regulation of several inflammatory and cytokine responses through the receptor interactions. Hence, PAGER Web APP enables screening for the critical terms and quickly identifying the specific molecular mechanism communities in the m-type PAG-to-PAG network.

TABLE 3 | The 23 consensus pathways between PAGER, EnrichR results with PubMed literature support.

Term	P vs. E (%)	Keywords	k	OR	Score	PMID	BEERE validation
axl signaling pathway	86	axl signaling	45	1.533	5.30E+00	31871265	0
g alpha (i) signaling events	97	g protein alpha signaling events	15	0.699	6.33E-02	33588787	1
Vitamin d receptor pathway	100	Vitamin d receptor pathway	23	0.734	5.49E-02	28218743	0
Age-rage signaling pathway in diabetic complications	100	Age-rage signaling pathway, diabetes	1	0.200	7.12E-03	25909054	0
Activation of nlrp3 inflammasome by sars-cov-2	100	Viral protein interaction, cytokine receptor	154	0.580	8.48E-14	26920710	0
Viral protein interaction with cytokine and cytokine receptor	100	nlrp3 inflammasome	14	0.225	6.84E-14	33649199	1
pi3k-akt signaling pathway	100	pi3k-akt signaling pathway	475	0.693	2.33E-17	22453015	0
Jak-stat signaling pathway	100	Jak-stat signaling pathway	103	0.478	1.39E-17	32194688	0
Kaposi sarcoma-associated herpesvirus infection	100	Kaposi sarcoma-associated herpesvirus infection	12	0.085	2.15E-45	16443048	0
Proteoglycans in cancer	100	Proteoglycans	766	0.554	8.83E-72	31140988	0
Hematopoietic cell lineage	100	Hematopoietic cell lineage	63	0.130	3.82E-128	26391013	0
Adipogenesis	100	Adipogenesis	25	0.060	1.48E-139	27216185	0
nf-kappa b signaling pathway	100	nf-kappa b signaling	432	0.341	1.46E-158	22433222	1
Glucocorticoid receptor pathway	100	Glucocorticoid receptor	131	0.168	1.91E-179	31911848	1
Gastrin signaling pathway	100	Gastrin	23	0.034	3.80E-246	1,6242076	1
Allograft rejection	100	Allograft rejection	76	0.081	2.85E-288	26951628	0
Human papillomavirus infection	100	Papillomavirus	405	0.237	6.17E-308	10767787	1
Selenium micronutrient network	100	Selenium	129	0.112	2.82e-318	23470450	1
Nanomaterial-induced inflammasome activation	100	Nanotechnology	563	0.160	0.00E+00	28303522	0
Covid-19 adverse outcome pathway	100	Covid-19	289	0.043	0.00E+00	32734626	0
Pathogenic *Escherichia coli* infection	100	*Escherichia coli*	431	0.033	0.00E+00	34912719	0
Lipid and atherosclerosis	100	Lipid, atherosclerosis	21	0.012	0.00E+00	29903879	1
Human cytomegalovirus infection	100	Cytomegalovirus	195	0.125	0.00E+00	15922119	1

TABLE 4 | The 5 consensus pathways between PAGER, WebGestaltR results with PubMed literature support.

Term	W vs. P (%)	Keywords	k	OR	Score	PMID	BEERE validation
Binding and uptake of ligands by scavenger receptors	100	Ligands, scavenger receptors	12	0.398	6.77E-05	31244937	0
Interleukin-4 and interleukin-13 signaling	100	Interleukin-4, interleukin-13	9	0.114	5.68E-24	23972995	1
Collagen chain trimerization	100	Collagen chain	153	0.257	4.14E-102	21853302	0
Interleukin-10 signaling	100	Interleukin-10	349	0.332	1.05E-136	7852279	1
Post-translational protein phosphorylation	100	Protein phosphorylation	2,850	0.301	0	17973544	0

Validation of the Enriched Pathways Using Literature Support in the Melanoma Drug Resistant-Sensitive Study

We found all of them relevant to melanoma cancer for the 33 consensus pathways among PAGER, EnrichR, and WebGestaltR results. We listed the results using pathway name, the pairwise term similarities, keywords used for co-citation retrieval, number of co-citations, odds ratio, PubMed score, one of PubMed IDs, and BEERE validation (in **Table 2**; **Supplementary Table S2**). All the pathways were determined to be related to Melanoma with PubMed literature support. For the 23 consensus pathways between PAGER and EnrichR results, we found that all of them have at least one literature support (**Table 3**). We showed all the BEERE-identified semantic relationships in **Supplementary Table S3**. We found that all the 5 consensus pathways between PAGER and WebGestaltR results to be supported by PubMed literature citations, and we ranked them based on the PubMed score (**Table 4**; **Supplementary Table S4**). Each of the six consensus pathways between EnrichR and WebGestaltR also had at least one literature citation support (**Supplementary Table S5**).

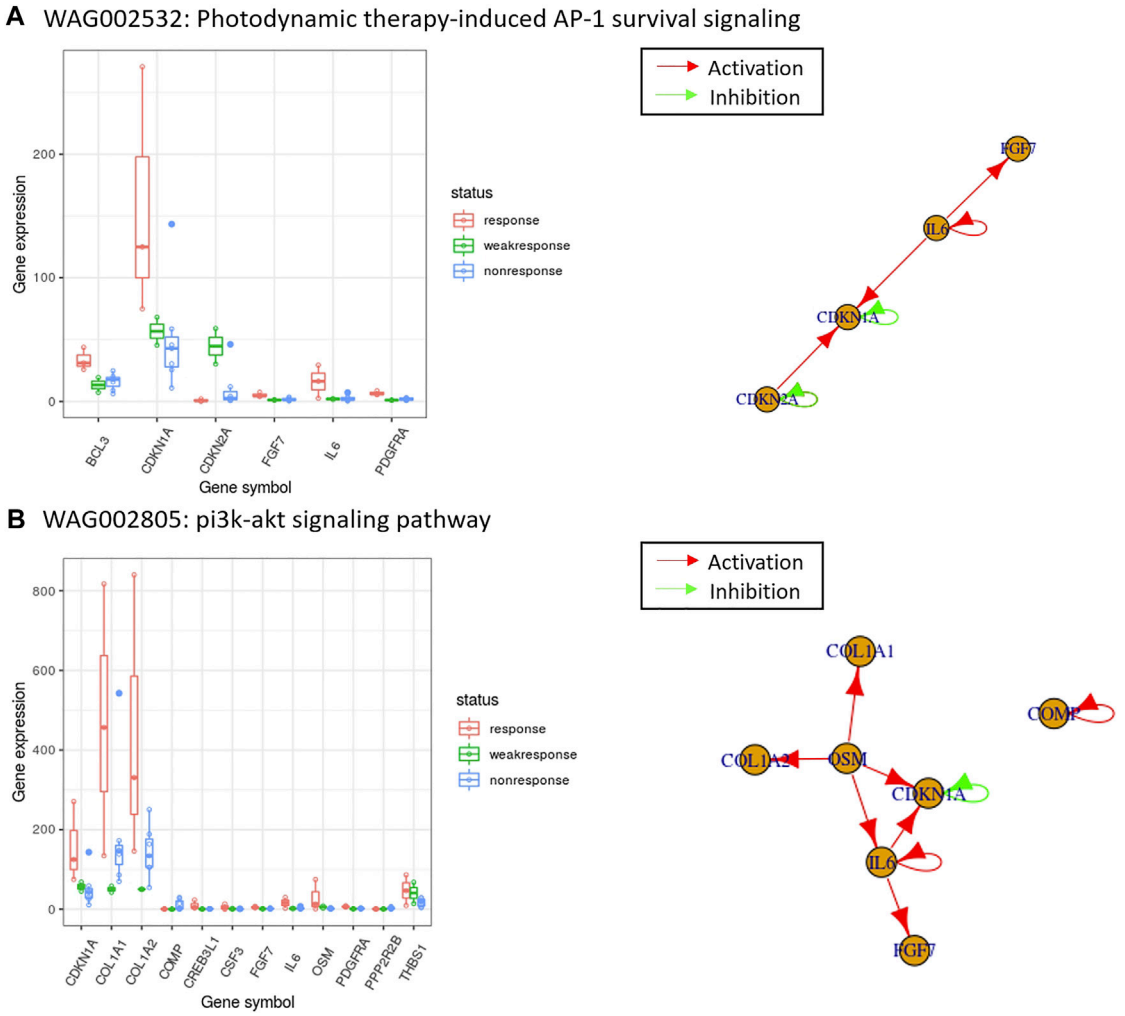

FIGURE 4 | The top-ranked enriched pathways using the PubMed score and the expression of those overlapped genes with gene regulatory networks for melanoma drug resistant-sensitive patients. In the box plots, the *x*-axis are the overlapped genes between differentially expressed gene candidates and pathway gene members, and the *y*-axis are the gene expression values. In the gene regulatory networks, a red arrow indicates the direction of activation, and a green arrow indicates the direction of inhibition. WAG002532 and WAG002805 are the PAG IDs of the enriched pathways shown in **(A)** The pathway with the highest PubMed score in **Table 2**, and **(B)** one of the PubMed literature validated pathways in **Table 3**. The details of pathways shown can be retrieved online from: http://discovery.informatics.uab.edu/PAGER/index.php/geneset/view/[PAG ID].

TABLE 5 | The performance of the three tools. ***k*** represents the citations of "melanoma" and the keywords from a pathway. ***OR*** represents the odds ratio. ***Score*** represents the ***PubMed score***.

Tool	Precision		
	***k* > 0**	**OR > 0.1**	**score > 10e-5**
PAGER	0.95	0.75	0.30
EnrichR	0.89	0.65	0.29
WebGestalt	0.99	0.70	0.24

We ranked the pathways based on the *PubMed score* (Yue et al., 2019a). As reported the highest PubMed score in **Table 2**, the photodynamic therapy has been frequently reported for melanoma treatment in recent years (Shivashankarappa and Sanjay, 2019; Turkoglu et al., 2019; Abramova et al., 2021; Yordi et al., 2021). We observed that in the overlapped genes between differentially expressed gene candidates and the photodynamic therapy-induced ap-1 survival signaling's gene

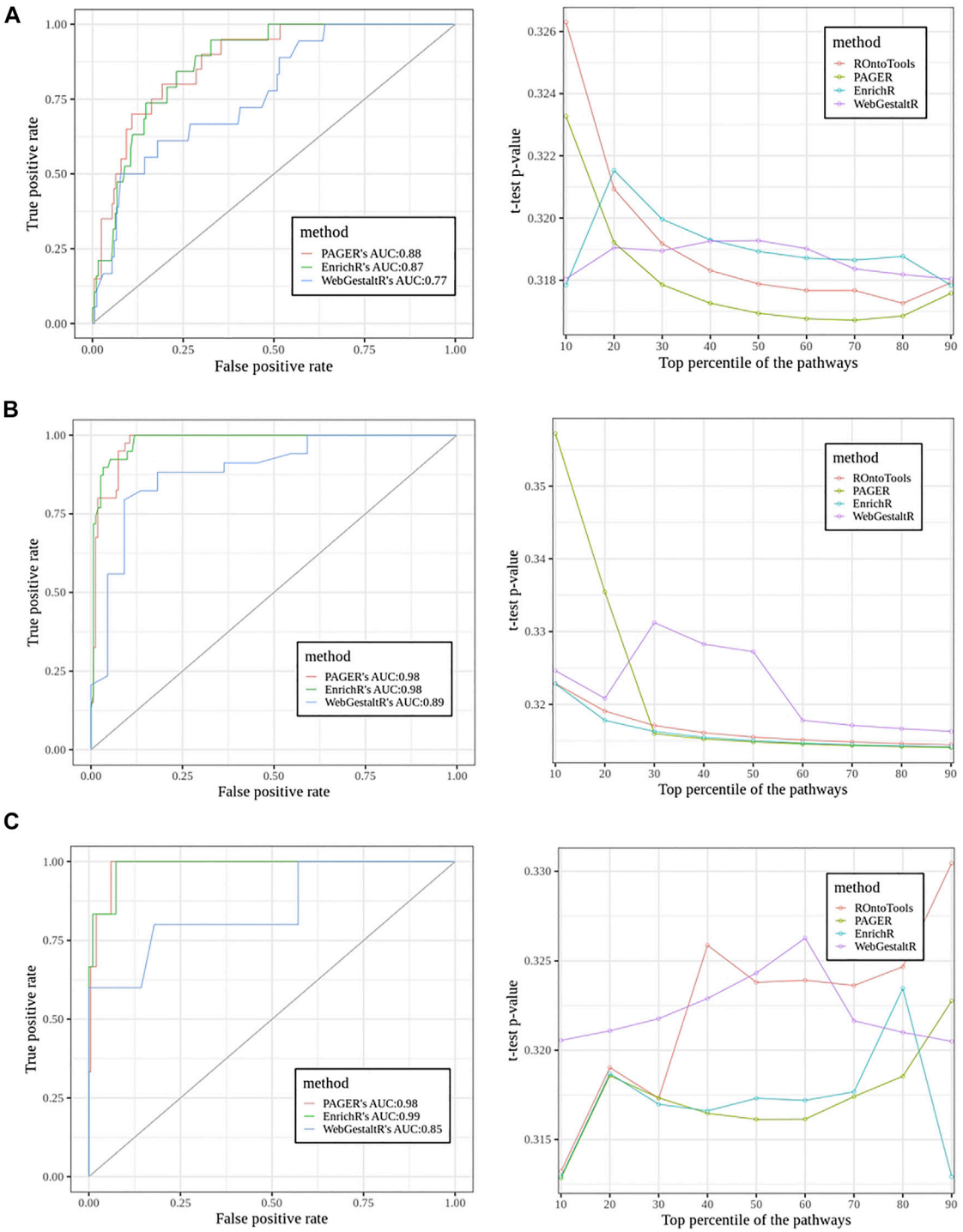

FIGURE 5 | The performance comparisons among PAGER, EnrichrR and WebGestaltR using Receiver Operator Characteristic (ROC) curve and the t-test curve. The pathways' adjusted *p*-values were applied to generate the ROC curves. The PubMed scores were used for the t-test curve. **(A)** The sclerosis study (E-GEOD-21942). **(B)** The inflamed colonic mucosa vs. non-inflamed colonic mucosa in Crohn's disease study. **(C)** The inflamed colonic mucosa vs. non-inflamed colonic mucosa in the ulcerative colitis study.

members, the five genes, IL6 (Interleukin 6), CDKN1A (Cyclin Dependent Kinase Inhibitor 1A), FGF7 (Fibroblast Growth Factor 7), BCL3 (B-Cell Lymphoma 3-Encoded Protein) and PDGFRA (Platelet Derived Growth Factor Receptor Alpha) were under-expressed in the drug-resistant patients, and the three genes, IL6, CDKN1A and FGF7 were connected in the gene

regulatory network from PAGER (de Waal Malefyt et al., 1991) (**Figure 4**). The pathway, "pi3k-akt signaling pathway" in **Table 3**, contained twelve overlapped genes. Similarly, among the under-expressed genes, the six genes, IL6 (Interleukin 6), CDKN1A (Cyclin Dependent Kinase Inhibitor 1A), FGF7 (Fibroblast Growth Factor 7), OSM (Oncostatin M), COL1A1 (Collagen Type I Alpha 1 Chain) and COL1A2 (Collagen Type I Alpha 2 Chain) were connected in the gene regulatory network (**Figure 4**). OSM gene is upstream and stimulates the other five genes. Since OSM is an interleukin-6 (IL-6) type cytokine to inhibit melanoma proliferation, the loss of OSM gene expression in drug-resistant patients may inhibit the activity of collagen biosynthesis and interleukin-6 family signaling. Lacreusette A et al. (Lacreusette et al., 2007) reported that the histone deacetylase inhibitor (HDACi) Trichostatin A (TSA), increased OSM protein activity and histone acetylation of the OSM receptor-beta (OSMRbeta) promoter as well as expression of OSMRbeta mRNA and protein. Therefore, Trichostatin A (TSA) allows the OSM protein to activate the signal transducer and activator of transcription 3 (STAT3) and inhibit proliferation. Thus, OSM/IL-6 resistance of melanoma cells in the late-stage patients may benefit from histone deacetylase inhibitor Trichostatin A.

Another intriguing pathway, the interleukin 10 (IL-10) signaling pathway, reported in PAGER also shows how literature supports its involvement in melanoma. IL-10's role in immune system biology is that it acts as an immunomodulator, which means that it regulates how the immune system behaves (Terai et al., 2012). Terai et al. found that metastatic melanoma cells can produce IL-10 and that this product can prevent the immune cells from attacking it (Terai et al., 2012). The group also found that IL-6 may play a role in the stimulation of IL-10 production in melanoma cells (Terai et al., 2012). Thus, the PAGER analysis can help give hints to researchers as far as finding potential disease mechanisms is concerned.

We applied precision to measure the performance among the three tools using different cutoffs (**Table 5**). To evaluate the co-citation coverage in the literature, we tested the result's precision using different cutoffs. When we set the co-citation (k) cutoff to be 1, PAGER's precision is 0.95 as a little lower than WebGestalt's precision is 0.99. When the odds ratio cutoff is set to be 0.1, PAGER has the best precision, which is 0.75, and when the PubMed score cutoff is set to be 10e-5, PAGER still leads, giving precision to be 0.30.

Validation of the Enriched Pathways Using the Topology-Based Method and Literature Support in Multiple Sclerosis (MS), Colonic Mucosa in Crohn's Disease (CD), and Ulcerative Colitis (UC) Studies

In the sclerosis study, we found 20 pathways in the true set and 203 pathways in the false set using ground truth discovered by the topology-based method, ROntoTools (**Figure 5A**). The PAGER led by giving the AUC 0.88, EnrichR came the next with AUC to be 0.87, and the WebGestaltR's AUC was 0.77. In the t-test curve, We found PAGER had the lowest average p-value (0.318) compared with ROntoTools (0.319), EnrichR (0.319) and WebgestaltR (0.319). In the inflamed colonic mucosa vs. non-inflamed colonic mucosa in Crohn's disease study (**Figure 5B**), we found 40 pathways in the true set and 161 pathways in the false set. Both EnrichR and PAGER had the highest AUC of 0.98, and the WebGestaltR's AUC was 0.87. We found EnrichR had the lowest average p-value (0.316) compared with ROntoTools (0.317), PAGER (0.322) and WebgestaltR (0.322). In the inflamed colonic mucosa vs. non-inflamed colonic mucosa in the ulcerative colitis study (**Figure 5C**), we found 6 pathways in the true set and 199 pathways in the false set. The EnrichR had the highest AUC of 0.99, PAGER came the next with AUC to be 0.98, and the WebGestaltR's AUC was 0.85. We found PAGER and EnrichR tied with the lowest average p-value (0.317) compared with ROntoTools (0.321) and WebgestaltR (0.322). Overall, PAGER was among the best.

DISCUSSION AND CONCLUSION

To summarize, we developed an interactive online functional genomics analysis tool, PAGER Web APP. The tool can provide new and significant insights into functional genomics studies and may support precision medicine in delivering the candidate targets. In the melanoma drug-resistant-sensitive case study, we observed that the P-type PAGs (pathways) reported in PAGER lead to insights into molecular mechanisms validated in literature support. PAGER web server supports the feature of r-type PAG-to-PAG network generation.

There are two potential explanations for the differences in the enrichment results among the three tools. First, we noticed that the database versions might vary. As reported in the EnrichR and PAGER Web APP, the KEGG data was processed in 2021, and the WebgestaltR's KEGG data was processed in 2018. Newer database processing time may suggest more freshly updated content of databases—variability that we couldn't control in this case study. Second, the enrichment algorithms used in these three tools are different. PAGER adapts hypergeometric test to perform the enrichment analysis and applies adjusted p-value using $p_0 * (m + n_{p_i \le p_0} - 1)$, where p_0 is the original p-value, m is the number of p-value's multiple tests from the PAGs under the constraints of PAG source, overlaps, PAG size, and similarity score, and the $n_{p_i \le p_0}$ is the number of p-values in the multiple test that has less than or equal to the original p-value. EnrichR uses fuzzy enrichment analysis, and applies Benjamini–Hochberg for FDR, according to the documentation online. The WebgestaltR uses hypergeometric test to evaluate the significance of enrichment and uses Bonferroni for p-value adjustment. To construct the ground truth in assessing the performance of functional genomics analysis tools, many data-driven approaches can be applied, such as target pathway (Tarca et al., 2012; Tarca et al., 2013) or gene knockout (KO) dataset (Nguyen et al., 2019). In the target pathway approach, the datasets from diseases have a pathway describing the underlying mechanisms, and hence this pathway is implicated in this phenotype. Therefore, a pathway analysis method is assessed based on the ranking and significance of these target pathways. We explored the feasibility of using either

pathway ground truth or gene knockout (KO) data sets for our case study, i.e., the study of late-stage drug-resistant melanoma. However, we could not find "target pathway" (Tarca et al., 2012; Tarca et al., 2013) as ground truth or pathways that are not directly related to the dataset to build all the counts in a confusion table. For genes discovered in the enriched pathway, the "pi3k-akt signaling pathway", we didn't find any OSM gene-KO Melanoma dataset in the GEO database. We also could not use non-melanoma KO experiments for fear of introducing additional noises. Thus, we used BEERE to extract those semantic relationships that co-mention melanoma and the pathways' keywords with a statistical evaluation to assess the statistical significance of the PubMed literature reference count above a random model. As for NCBI e-utils literature retrieval, we also applied the PubMed score to evaluate the statistical significance of the literature counts to conquer the literature volume and breadth.

In the future, we expect to implement features to enhance the usage of PAGER Web APP, which can be plugged in geneset, network, and pathway analysis (GNPA) to improve the use. In the current version, we observed that the user interface, especially, in the enriched results, the enriched PAGs' result is not that interactive enough for users to select a certain number of PAGs or arbitrarily remove some of the records to generate PAG-to-PAG networks. We will implement the interactive panels in the future release. PAGER Web APP calls the PAGER API, which implements an over-representation analysis (ORA) technique by default. In general, advanced functional class scoring (FCS) techniques, e.g., Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), Gene Set Analysis (GSA) (Efron and Tibshirani, 2007) and Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) (Tarca et al., 2012), can better detect the significant effects on pathways led by large changes in individual genes and the weaker coordinated. Other pathway analysis tools may also incorporate network topology information to integrate signaling interactions among genes in a pathway, e.g., Pathway-Express (Khatri et al., 2007), SPIA (Tarca et al., 2009), Pathway-Guide (Advaita Bioinformatics, http://www.advaitabio.com), TopoGSA (Glaab et al., 2010), Bayesian Pathway Analysis (BPA) (Isci et al., 2011), and PathNet (Dutta et al., 2012), etc. We plan to implement additional advanced topology-based pathway GSEA analysis techniques into the PAGER APIs, and adopt comprehensive benchmark data sets (Tarca et al., 2012; Tarca et al., 2013; Nguyen et al., 2019) to guide users in selecting the proper method for the right application scenario in future releases. Thus, PAGER Web APP will offer users more expanded analysis choices than today.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ZY developed the PAGER Web APP, processed the melanoma data, performed all the case studies, and wrote the manuscript. RS processed the melanoma data, performed the literature validation of the results. SB helped in the framework design and processed the multiple sclerosis, Crohn's disease and ulcerative colitis case studies. JC conceptualized the ideas, helped design the analytical experiments, provided feedback throughout the project, and revised the final manuscript. All authors read, edited, and approved the manuscript.

FUNDING

The work was in part supported by the internal University of Alabama at Birmingham research grants to JC and the National Institutes of Health grant awards U54TR001005 in which JC serves as a co-investigator.

ACKNOWLEDGMENTS

All authors thank the following general technical support that made case studies included for this work possible: Saghapour Ehsan for testing the PAGER Web APP and providing suggestions. Nishant Batra for testing scripts in the project. Jelai Wang for managing the data management and data analysis computing framework. Hiren Desai from UAB Information Technology groups for supporting the backend Oracle database 19c management.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.820361/full#supplementary-material

REFERENCES

Abramova, O. B., Kaplan, M. A., Grin, M. A., Yuzhakov, V. V., Suvorov, N. V., Mironov, A. F., et al. (2021). Photodynamic Therapy of Melanoma B16 with Chlorin E6 Conjugated with a PSMA-Ligand. *Bull. Exp. Biol. Med.* 171, 468–471. doi:10.1007/s10517-021-05252-x

Andrieux, G., and Chakraborty, S. (2021). Editorial: Integration of Multi-Omics Techniques in Cancer. *Front. Genet.* 12, 733965. doi:10.3389/fgene.2021.733965

Angeloni, M., Thievessen, I., Engel, F. B., Magni, P., and Ferrazzi, F. (2021). Functional Genomics Meta-Analysis to Identify Gene Set Enrichment Networks in Cardiac Hypertrophy. *Biol. Chem.* 402, 953–972. doi:10.1515/hsz-2020-0378

Ansari, S., Voichita, C., Donato, M., Tagett, R., and Draghici, S. (2016). A Novel Pathway Analysis Approach Based on the Unexplained Disregulation of Genes. *Proc. IEEE* 105, 1–14. doi:10.1109/JPROC.2016.2531000

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008

Bock, J.-O., and Ortea, I. (2020). Re-analysis of SARS-CoV-2-Infected Host Cell Proteomics Time-Course Data by Impact Pathway Analysis and Network Analysis: a Potential Link with Inflammatory Response. *Aging* 12, 11277–11286. doi:10.18632/aging.103524

Bokanizad, B., Tagett, R., Ansari, S., Helmi, B. H., and Draghici, S. (2016). SPATIAL: A System-Level PAThway Impact AnaLysis Approach. *Nucleic Acids Res.* 44, 5034–5044. doi:10.1093/nar/gkw429

Boytsov, L. (2011). Indexing Methods for Approximate Dictionary Searching: Comparative Analysis. *ACM J. Exp. Algorithmics* 16, 1–91. Article 1.1. doi:10.1145/1963190.1963191

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V. S. K., et al. (2017). UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* 19, 649–658. doi:10.1016/j.neo.2017.05.002

Chen, J. Y., Yan, Z., Shen, C., Fitzpatrick, D. P. G., and Wang, M. (2007). A Systems Biology Approach to the Study of Cisplatin Drug Resistance in Ovarian Cancers. *J. Bioinform. Comput. Biol.* 05, 383–405. doi:10.1142/s0219720007002606

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC Bioinformatics* 14, 128. doi:10.1186/1471-2105-14-128

Chen, J. Y., Pandey, R., and Nguyen, T. M. (2017). HAPPI-2: a Comprehensive and High-Quality Map of Human Annotated and Predicted Protein Interactions. *BMC Genomics* 18, 182. doi:10.1186/s12864-017-3512-1

de Waal Malefyt, R., Haanen, J., Spits, H., Roncarolo, M. G., te Velde, A., Figdor, C., et al. (1991). Interleukin 10 (IL-10) and Viral IL-10 Strongly Reduce Antigen-specific Human T Cell Proliferation by Diminishing the Antigen-Presenting Capacity of Monocytes *via* Downregulation of Class II Major Histocompatibility Complex Expression. *J. Exp. Med.* 174, 915–924. doi:10.1084/jem.174.4.915

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, P3. doi:10.1186/gb-2003-4-5-p3

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., et al. (2007). A Systems Biology Approach for Pathway Level Analysis. *Genome Res.* 17, 1537–1545. doi:10.1101/gr.6202607

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

Dutta, B., Wallqvist, A., and Reifman, J. (2012). PathNet: a Tool for Pathway Analysis Using Topological Information. *Source Code Biol. Med.* 7, 10. doi:10.1186/1751-0473-7-10

Efron, B., and Tibshirani, R. (2007). On Testing the Significance of Sets of Genes. *Ann. Appl. Stat.* 1, 107–129. doi:10.1214/07-aoas101

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, pl1. doi:10.1126/scisignal.2004088

Glaab, E., Baudot, A., Krasnogor, N., and Valencia, A. (2010). TopoGSA: Network Topological Gene Set Analysis. *Bioinformatics* 26, 1271–1272. doi:10.1093/bioinformatics/btq131

Hamburg, M. A., and Collins, F. S. (2010). The Path to Personalized Medicine. *N. Engl. J. Med.* 363, 301–304. doi:10.1056/NEJMp1006304

Harini, N. K., Xiaogang, W., and Jake Yue, C. (2008). "Towards an Integrative Human Pathway Database for Systems Biology Applications," in Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil, March, 2008 (ACM), 1297–1301.

Huang, H., Wu, X., Sonachalam, M., Mandape, S. N., Pandey, R., MacDorman, K. F., et al. (2012). PAGED: a Pathway and Gene-Set Enrichment Database to Enable Molecular Phenotype Discoveries. *BMC Bioinformatics* 13 (Suppl. 15), S2. doi:10.1186/1471-2105-13-S15-S2

Isci, S., Ozturk, C., Jones, J., and Otu, H. H. (2011). Pathway Analysis of High-Throughput Biological Data within a Bayesian Network Framework. *Bioinformatics* 27, 1667–1674. doi:10.1093/bioinformatics/btr269

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a Stateful Web Service to Facilitate Gene/protein List Analysis. *Bioinformatics* 28, 1805–1806. doi:10.1093/bioinformatics/bts251

Kemppinen, A. K., Kaprio, J., Palotie, A., and Saarela, J. (2011). Systematic Review of Genome-wide Expression Studies in Multiple Sclerosis. *BMJ Open* 1, e000053. doi:10.1136/bmjopen-2011-000053

Khatri, P., Draghici, S., Tarca, A. L., Hassan, S. S., and Romero, R. (2007). "A System Biology Approach for the Steady-State Analysis of Gene Signaling Networks," in Iberoamerican Congress on Pattern Recognition, Valparaiso, Chile, November 13–16, 2007, 32–41.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *Plos Comput. Biol.* 8, e1002375. doi:10.1371/journal.pcbi.1002375

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

Lacreusette, A., Nguyen, J.-M., Pandolfino, M.-C., Khammari, A., Dreno, B., Jacques, Y., et al. (2007). Loss of Oncostatin M Receptor β in Metastatic Melanoma Cells. *Oncogene* 26, 881–892. doi:10.1038/sj.onc.1209844

Lafferty, A., O'Farrell, A. C., Migliardi, G., Khemka, N., Lindner, A. U., Sassi, F., et al. (2021). Molecular Subtyping Combined with Biological Pathway Analyses to Study Regorafenib Response in Clinically Relevant Mouse Models of Colorectal Cancer. *Clin. Cancer Res.* 27, 5979–5992. doi:10.1158/1078-0432.CCR-21-0818

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205. doi:10.1093/nar/gkz401

Livshits, A., Git, A., Fuks, G., Caldas, C., and Domany, E. (2015). Pathway-Based Personalized Analysis of Breast Cancer Expression Data. *Mol. Oncol.* 9, 1471–1483. doi:10.1016/j.molonc.2015.04.006

Mallavarapu, T., Hao, J., Kim, Y., Oh, J. H., and Kang, M. (2020). Pathway-Based Deep Clustering for Molecular Subtyping of Cancer. *Methods* 173, 24–31. doi:10.1016/j.ymeth.2019.06.017

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., et al. (2013). Methods and Approaches in the Topology-Based Analysis of Biological Pathways. *Front. Physiol.* 4, 278. doi:10.3389/fphys.2013.00278

Nguyen, T., Mitrea, C., and Draghici, S. (2018). Network-Based Approaches for Pathway Level Analysis. *Curr. Protoc. Bioinformatics* 61, 1–8. doi:10.1002/cpbi.42

Nguyen, T.-M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Identifying Significantly Impacted Pathways: a Comprehensive Review and Assessment. *Genome Biol.* 20, 203. doi:10.1186/s13059-019-1790-4

Pian, C., He, M., and Chen, Y. (2021). Pathway-Based Personalized Analysis of Pan-Cancer Transcriptomic Data. *Biomedicines* 9, 1502. doi:10.3390/biomedicines9111502

Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., et al. (2001). A Functional Genomics Strategy that Uses Metabolome Data to Reveal the Phenotype of Silent Mutations. *Nat. Biotechnol.* 19, 45–50. doi:10.1038/83496

Raghavan, S., Mehta, P., Ward, M. R., Bregenzer, M. E., Fleck, E. M. A., Tan, L., et al. (2017). Personalized Medicine-Based Approach to Model Patterns of Chemoresistance and Tumor Recurrence Using Ovarian Cancer Stem Cell Spheroids. *Clin. Cancer Res.* 23, 6934–6945. doi:10.1158/1078-0432.CCR-17-0133

Rahaman, M. M., Chen, D., Gillani, Z., Klukas, C., and Chen, M. (2015). Advanced Phenotyping and Phenotype Data Analysis for the Study of Plant Growth and Development. *Front. Plant Sci.* 6, 619. doi:10.3389/fpls.2015.00619

Sayers, E. (2008). "E-utilities Quick Start," in *Entrez Programming Utilities Help*. [Internet].

Shivashankarappa, A., and Sanjay, K. R. (2019). Photodynamic Therapy on Skin Melanoma and Epidermoid Carcinoma Cells Using Conjugated 5-aminolevulinic Acid with Microbial Synthesised Silver Nanoparticles. *J. Drug Target.* 27, 434–441. doi:10.1080/1061186X.2018.1531418

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* 371, 2189–2199. doi:10.1056/NEJMoa1406498

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based

Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and its Application. *Bioinform Biol. Insights* 14, 117793221989905. doi:10.1177/1177932219899051

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics* 25, 75–82. doi:10.1093/bioinformatics/btn577

Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting Overlapping Genes Improves Gene Set Analysis. *BMC Bioinformatics* 13, 136. doi:10.1186/1471-2105-13-136

Tarca, A. L., Bhatti, G., and Romero, R. (2013). A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS One* 8, e79217. doi:10.1371/journal.pone.0079217

Terai, M., Eto, M., Young, G. D., Berd, D., Mastrangelo, M. J., Tamura, Y., et al. (2012). Interleukin 6 Mediates Production of Interleukin 10 in Metastatic Melanoma. *Cancer Immunol. Immunother.* 61, 145–155. doi:10.1007/s00262-011-1084-5

Turkoglu, E. B., Pointdujour-Lim, R., Mashayekhi, A., and Shields, C. L. (2019). Photodynamic Therapy as Primary Treatment for Small Choroidal Melanoma. *Retina* 39, 1319–1325. doi:10.1097/IAE.0000000000002169

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a Web-Based Text Mining Tool for Assisting Biocuration. *Nucleic Acids Res.* 41, W518–W522. doi:10.1093/nar/gkt441

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: Automated Concept Annotation for Biomedical Full Text Articles. *Nucleic Acids Res.* 47, W587–W593. doi:10.1093/nar/gkz389

Wu, X., Hasan, M. A., and Chen, J. Y. (2014). Pathway and Network Analysis in Proteomics. *J. Theor. Biol.* 362, 44–52. doi:10.1016/j.jtbi.2014.05.031

Yang, T.-L., Shen, H., Liu, A., Dong, S.-S., Zhang, L., Deng, F.-Y., et al. (2020). A Road Map for Understanding Molecular and Genetic Determinants of Osteoporosis. *Nat. Rev. Endocrinol.* 16, 91–103. doi:10.1038/s41574-019-0282-7

Yordi, S., Soto, H., Bowen, R. C., and Singh, A. D. (2021). Photodynamic Therapy for Choroidal Melanoma: What Is the Response Rate? *Surv. Ophthalmol.* 66, 552–559. doi:10.1016/j.survophthal.2020.09.006

Yue, Z., Kshirsagar, M. M., Nguyen, T., Suphavilai, C., Neylon, M. T., Zhu, L., et al. (2015). PAGER: Constructing PAGs and New PAG-PAG Relationships for Network Biology. *Bioinformatics* 31, i250–i257. doi:10.1093/bioinformatics/btv265

Yue, Z., Zheng, Q., Neylon, M. T., Yoo, M., Shin, J., Zhao, Z., et al. (2018). PAGER 2.0: an Update to the Pathway, Annotated-List and Gene-Signature Electronic Repository for Human Network Biology. *Nucleic Acids Res.* 46, D668–D676. doi:10.1093/nar/gkx1040

Yue, Z., Nguyen, T., Zhang, E., Zhang, J., and Chen, J. Y. (2019a). WIPER: Weighted In-Path Edge Ranking for Biomolecular Association Networks. *Quant Biol.* 7, 313–326. doi:10.1007/s40484-019-0180-y

Yue, Z., Willey, C. D., Hjelmeland, A. B., and Chen, J. Y. (2019b). BEERE: a Web Server for Biomedical Entity Expansion, Ranking and Explorations. *Nucleic Acids Res.* 47, W578–W586. doi:10.1093/nar/gkz428

Zhang, F., and Chen, J. Y. (2010). Discovery of Pathway Biomarkers from Coupled Proteomics and Systems Biology Methods. *BMC Genomics* 11 (Suppl. 2), S12. doi:10.1186/1471-2164-11-S2-S12

Zhang, F., and Chen, J. Y. (2013). Breast Cancer Subtyping from Plasma Proteins. *BMC Med. Genomics* 6 (Suppl. 1), S6. doi:10.1186/1755-8794-6-S1-S6

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an Integrated System for Exploring Gene Sets in Various Biological Contexts. *Nucleic Acids Res.* 33, W741–W748. doi:10.1093/nar/gki475

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

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

Copyright © 2022 Yue, Slominski, Bharti and Chen. 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.

frontiers | Frontiers in Genetics

METHODS
published: 14 April 2022
doi: 10.3389/fgene.2022.818683

Venn Diagrams May Indicate Erroneous Statistical Reasoning in Transcriptomics

January Weiner 3rd, Benedikt Obermayer and Dieter Beule*

Core Unit Bioinformatics, Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Berlin, Germany

A common application of differential expression analysis is finding genes that are differentially expressed upon treatment in only one out of several groups of samples. One of the approaches is to test for significant difference in expression between treatment and control separately in the two groups, and then select genes that show statistical significance in one group only. This approach is then often combined with a gene set enrichment analysis to find pathways and gene sets regulated by treatment in only this group. Here we show that this procedure is statistically incorrect and that the interaction between treatment and group should be tested instead. Moreover, we show that gene set enrichment analysis applied to such incorrectly defined genes group-specific genes may result in misleading artifacts. Due to the presence of false negatives, genes significant in one, but not the other group are enriched in gene sets which correspond to the overall effect of the treatment. Thus, the results appear related to the problem at hand, but do not reflect the group-specific effect of a treatment. A literature search revealed that more than a quarter of papers which used a Venn diagram to illustrate the results of separate differential analysis have also applied this incorrect reasoning.

Keywords: gene set analysis, functional genomics, gene set enrichment, transcriptomics, venn diagram

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Peter Chapman,
Edinburgh Napier University, United Kingdom
Jüri Reimand,
University Health Network, Canada

****Correspondence:***
January Weiner
january.weiner@bih-charite.de

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

Received: *19 November 2021*
Accepted: *17 March 2022*
Published: *14 April 2022*

Citation:
Weiner J, Obermayer B and Beule D (2022) Venn Diagrams May Indicate Erroneous Statistical Reasoning in Transcriptomics. Front. Genet. 13:818683. doi: 10.3389/fgene.2022.818683

INTRODUCTION

Experimental designs for transcriptomic analyses frequently include more than one factor. Often, the question asked is whether there is a difference between groups (first factor) with respect to reaction to a particular treatment (second factor). For example, we may ask whether there are differentially expressed genes (DEGs) which are specific to a particular group of patients, e.g., interferon response elicited by a virus in one group, but absent in another group of patients. In other words, we ask whether the difference between the control group (healthy subjects) and the treatment group (infected patients) is different between two groups of individuals. This "difference of differences" is known in statistics as an interaction (Blalock, 1965). To find out whether it is statistically significant, an appropriate statistical test for interaction should be employed.

However, another approach is widely spread (Nieuwenhuis, Forstmann, and Wagenmakers 2011). Instead of testing the interaction, the effect of the treatment is tested separately in both groups. Next, a difference between groups is inferred if the effect of treatment is significant in one comparison, but not significant in the other. This approach is not correct from statistical point of view, as "the difference between significant and not significant is not itself statistically significant" (Gelman and Stern 2006). For example, the p-value in the first comparison may be 0.009, and in the other comparison 0.011. At an alpha level of 0.01 the difference will be statistically significant in the first, but not significant in the other comparison.

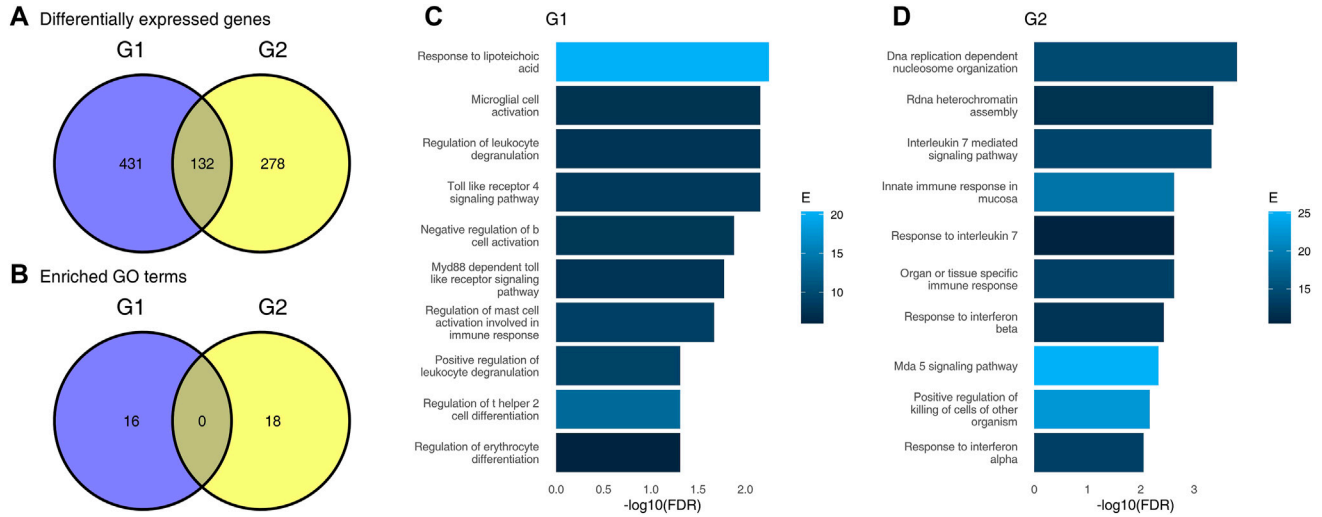

FIGURE 1 | Results of differential gene expression analysis and gene set enrichment analysis using an incorrect approach. **(A)**, Venn diagram showing numbers of differentially expressed genes (DEG) in each of the two groups, G1 and G2; **(B)**, Venn diagram showing numbers of significantly enriched GO terms in each of the two groups; **(C)** results of gene set enrichment analysis for genes "specific" to group G1; **(D)**, results of gene set enrichment analysis for genes "specific" to group G2 (only top 10 terms are shown).

In transcriptomics, statistical tests are performed for thousands of genes. As in the general case, the inference of differences between the groups should correctly be done by testing the significance of interaction between the group and the treatment. In practice, the differences between treatment and control are frequently tested in the two groups separately. This can be visualized using a Venn diagram (VD, **Figure 1**) showing the overall number of DEGs significant in both comparisons (the intersection in the VD) or significant in only one comparison (the remaining two fields on a VD). The genes which are significant in only one comparison are sometimes incorrectly considered as specific for the corresponding group. Following this, gene set enrichment analysis may be used in an attempt to test which pathways are specific to one, but not the other group.

In this paper, we show that under reasonable assumptions this approach may result in apparent enrichments even if there are no real statistically significant differences between the groups. To this end, we randomly split a cohort into two groups, compared the treatment (viral infection) with controls in each of the groups separately and then applied gene set enrichment to the sets of genes significantly different in one, but not the other group. Moreover, we show that the resulting gene set enrichments correspond to the differential expression between treatment and control. Thus, the enriched terms are relevant to the biological question at hand, yet while they do reflect real processes linked to viral infection, they do not correspond to the differences between the study groups. Finally, we use literature search to show that this incorrect approach to study group-specific treatment effects is not uncommon. In fact, while VDs are a useful visualization tool also in transcriptomics, in more than quarter of the papers where VDs were used, group-specific genes were defined as significant in one, but not other groups, and in 19% of the papers a gene set enrichment was performed.

TABLE 1 | Overall design in the case study: transcriptomic changes due to Sars-Cov-2 infection. The table shows number of patients in each combination of study group/disease status.

		Study Group	
		Group 1 (G1)	**Group 2 (G2)**
Disease status	Sars-Cov-2 infection	20	20
	Another infection	20	20

RESULTS

Transcriptomic Changes due to Sars-Cov-2 Infection

Consider two group of patients, G1 and G2 (**Table 1**). Each group contains 40 individuals. In both groups, there is an equal number of healthy individuals (labeled "Ctrl" on figures below) or patients infected with Sars-Cov-2 (labeled "SC2"). Our aim is to understand the differences between G1 and G2 in the response to infection. For example, we ask which genes or pathways are specifically upregulated by SC2 infection in G1 as compared to G2, and vice versa. In the following, we used the data set GSE156063 (Mick et al., 2020) in two approaches (an incorrect and the correct one) to arrive at opposite conclusions.

First, we have performed differential gene expression analysis for each of the groups G1 and G2 separately using standard bioinformatic tools. For each comparison, we defined DEGs as genes for which the false discovery rate (FDR) was lower than 0.05 and absolute $\log_2$ fold change (LFC) was higher than 1. There

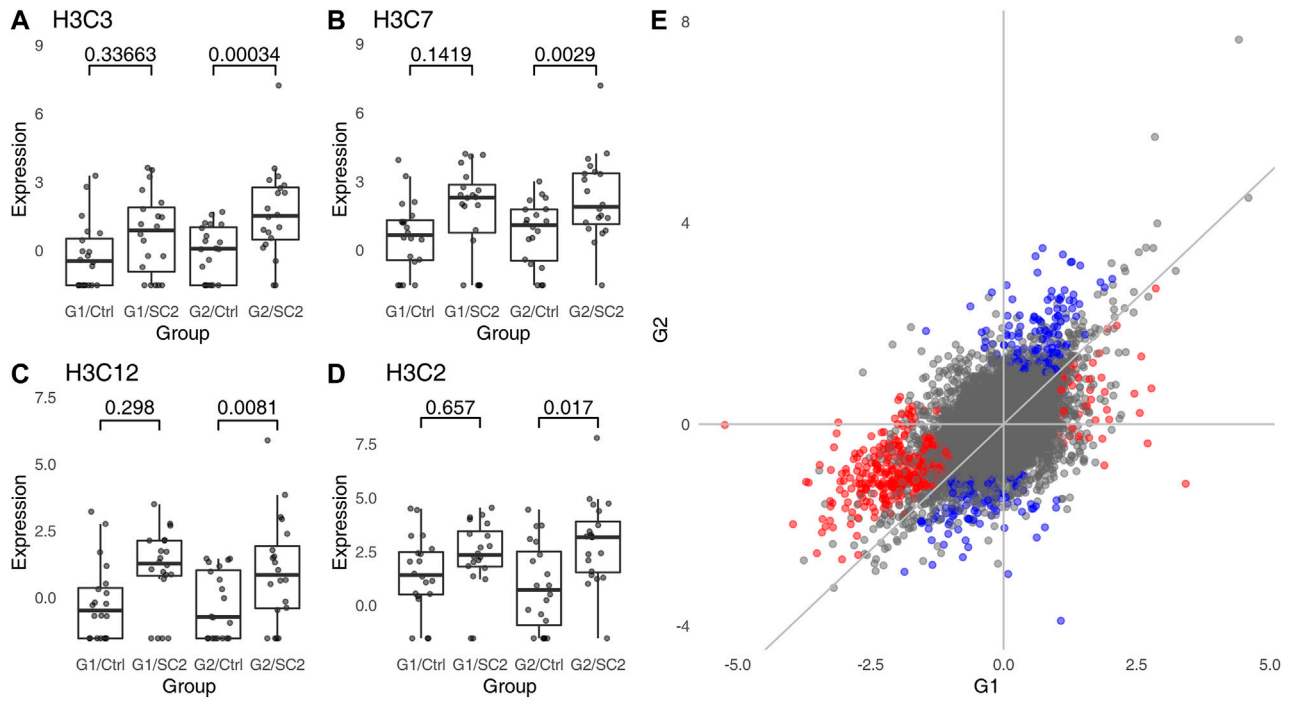

FIGURE 2 | Genes which are significant in one comparison, but not the other do not show a statistically significant interaction. **(A–D)**, examples of genes which are DEG in one group, but are not significantly different in the other group. "Ctrl," healthy individuals; "SC2," Sars-Cov-2 infected patients. Values above the plot indicate FDR (*p*-values corrected for multiple testing). **(E)**, correlation between $\log_2$ fold changes in G1 and G2. Color indicates genes which are significant in one, but not significant in the other comparison; red indicates genes significant in G1, blue indicates genes significant in G2. The overall Pearson correlation coefficient between $\log_2$ fold changes is 0.55.

were 563 DEGs in the G1 group, and 410 in the G2 group. In total, 132 DEGs were common for G1 and G2, 431 DEGs were significant in G1 only ("specific" for G1), and 278 were significant in G2 only (see **Figure 1A**). A naive interpretation of these results implies that there is a substantial difference between these two groups of individuals, as evidenced by a small overlap in commonly regulated genes. The majority of DEGs is significant in one comparison only.

To understand which pathways are upregulated in each of the two groups, we used a standard generation I gene set enrichment analysis—a hypergeometric test—on the DEGs in each group. Gene sets for the gene set enrichment analysis were taken from the Gene Ontology (GO) database. Gene sets with more than 50 or fewer than 10 genes were removed. For each group, we have selected only genes which are DEGs in that group, but not the other, mimicking a naive approach for finding pathways regulated in one patient group only. Here, a similar picture emerged. Overall, 16 gene sets were significantly enriched in G1, and 18 gene sets were significantly enriched in G2. Both the Venn diagram (**Figure 1B**) and the results of enrichments (**Figures 1C,D**) suggest that there is a fundamental difference between the groups, and that the groups have little in common in their response to the virus.

Importantly, the different GO terms enriched in the two groups were related to infection, and may tempt to speculate about the underlying biological differences between these two groups. For example, the significance of Toll like receptor 4 pathway in G1, but not G2; and, vice versa, significance of response to interleukin 7 in G2, but not in G1 may be considered as evidence of altered immune response to the virus in G2 as compared to G1.

However, the groups G1 and G2 were randomly sampled from the same data set. In fact, repeated re-sampling always results in some genes being found to be significantly different in one group, but not the other, despite the fact that one does not expect any major differences between sets of individuals randomly drawn from a single population. Thus, the conclusions drawn from a Venn diagram-driven gene set enrichment analysis are based on artifacts. Closer inspection of genes which are DEGs in one group, but not the other reveals the underlying statistical fallacy (**Figures 2A–D**), that is, that difference between significant and non-significant is, in itself, not statistically significant (Gelman and Stern 2006). This does not necessarily mean that there are no differences at all between these two groups, but that lack of significance in one group and significance in the other group does not correctly identify differences between groups.

To find genes which are differentially regulated in the two groups, the correct statistical approach is to calculate interaction between groups (G1, G2) and disease status (no disease vs. COVID). While it may be argued that a test for interaction has lower power than a test for a simple contrast, no genes show a significant interaction even at FDR <0.1. In fact, this is not

surprising. The $\log_2$ fold changes for comparisons withing G1 and G2 are strongly correlated (**Figure 2E**). For all significant genes, the Pearson correlation coefficient is 0.72, while for genes exclusively significant in G1 or G2 (genes "specific" to G1 or G2), it is 0.7 and 0.73, respectively. Thus, genes which are significant in one, but not in the other comparison tend to have similar $\log_2$ fold changes in both groups (e.g., **Figures 2A,C,D**).

Consequently, it is not possible to calculate gene set enrichment for the interaction using a hypergeometric test, as there are no DEGs for the interaction contrast. Gene set enrichment using a second generation algorithm (CERNO), relying on the ordering of genes according to their raw *p*-values from the interaction contrast rather than selecting a set of DEGs (Zyla et al., 2019), does not show any significant enrichment.

Artifacts Arise Because of False Negatives

It is worth noting that in the gene set enrichment analysis of the genes "specific" for a given comparison—i.e., genes which are significant in that comparison, but not significant in others—we have observed a number of terms associated with immune response. It is a crucial point of this manuscript to note that the spurious enrichments not only show significant *p*-values, but also that the terms or pathways which appear in them are relevant to the research hypothesis being tested. Below, we will show why these terms (rather than random terms which have no obvious relevance to an infectious disease) appear in the results.

To understand how significant results appear in a gene set enrichment analysis in randomly generated groups despite absence of genes with significant interaction, it is first necessary to consider the definition of a differentially expressed gene in this context. More often than not, DEGs are defined by a threshold in *p*-value adjusted for multiple testing, possibly combined with a threshold in $\log_2$ fold change. The commonly used Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) ensures that among genes for which FDR <0.05 there are at most 5% false positives irrespective of the sample size.

This way, we can exert control over the false positive rate (FPR, type I errors), keeping it at a relatively low level. However, we do not control the false negative rate (FNR, type II errors). In a powerful statistical test (such as a t-test), the test power in a typical application will rarely achieve more than 80%. For example, even for large effects (Cohen's d > 0.8) and type I error rate of 0.05, a t-test only achieves 80% power with at least 25 samples per group. For small effects (Cohen's d > 0.2), the required number of samples is at least 393 per group. Even assuming a test power of 80%, the FNR is 20%. Clearly, false negatives (FNs) occur at much higher rates than false positives (FPs). In the case of high throughput data sets, where the FPR is controlled by Bejnamini-Hochberg procedure or a similar technique, the FNR may be even as high as 80% (White, Ende, and Nichols 2019).

These FNs occur at a much higher rate within the sets of DEGs defined by the non-overlapping areas of the VDs, that is DEGs considered to be "specific" for one group or other in a naive approach. To illustrate this phenomenon, we have analyzed the full data set from which G1 and G2 were drawn (**Figure 3**), comparing the 100 healthy controls to 93 COVID-19 patients. Of the 431 genes significant in G1, but not in G2, 199 (46%) are significant in the full data set; of the 278 genes significant in G2, but not in G1, 99 (36%) are significant in the full data set. Given that G1 and G2 were sampled from the total population, and since the FDR was set to 0.05, we do not expect more than 30 FPs in the full data set, which implicates that at least 268 out of the 709 "specific" DEGs are true positives in the full data set. Thus, we can assume that at least a third of the genes that appeared to be "specific" in the initial analysis were, in fact, false negatives in one of the comparisons.

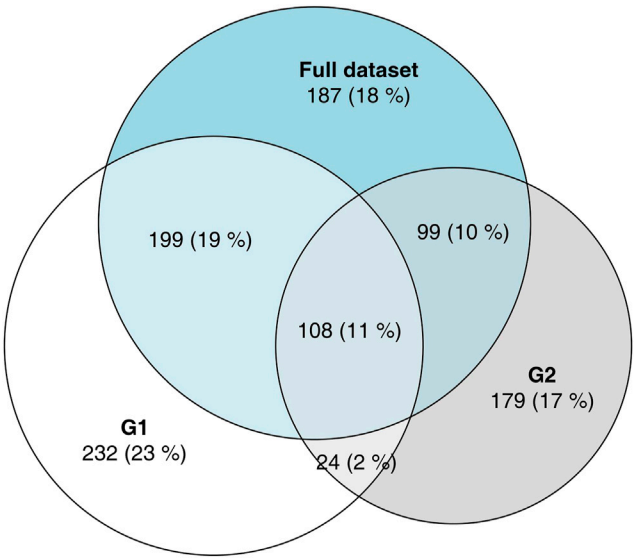

FIGURE 3 | Area-proportional Venn diagram showing overlaps in DEGs between G1, G2 and the full data set. The majority of genes which have been labeled as DEGs in only one of the groups G1, G2 are DEGs when all data were analyzed.

In other words, a substantial fraction of the "specific" genes are genes that are in reality differentially expressed in both groups alike. Therefore, if one is to perform a gene set enrichment analysis on one of these "specific" groups of genes, then the enriched functions will be related to the pathways and processes up- or downregulated in both groups due to the common factor (in this example, the COVID-19 disease), but which are not related to differences between the two groups.

Influence of Sample Size and Cut-Off Thresholds on Number of Artifacts

In the example above, the groups have been randomly sampled from a larger data set only once. Arguably, the observations might differ if the groups were to be resampled. Furthermore, we have chosen a group size of 40 (20 per group/treatment combination). Larger sample sizes are known to increase robustness of gene set enrichment analysis, and group size of 20 has been shown to be relatively robust (Maleki et al., 2019). However, a smaller or larger group size might change the proportion of FNs in the results and

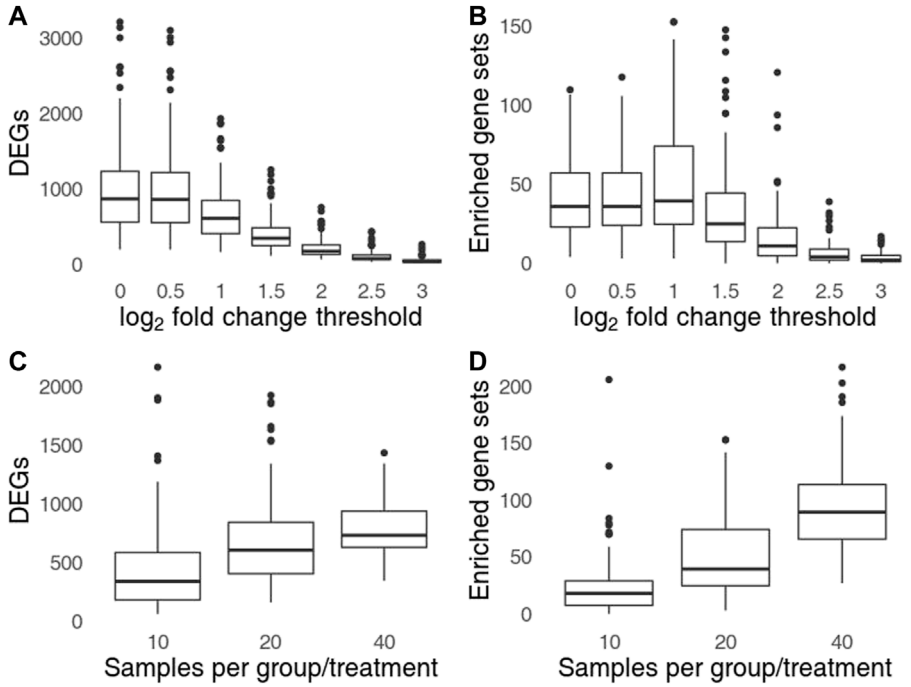

FIGURE 4 | Influence of $\log_2$ fold change treshold **(A,C)** and the sample size **(B,D)** on the number of genes significant in one, but not in the other group **(A,B)** and the number of significant gene sets found in one group, but not the other **(C,D)**. Top row shows the influence of $\log_2$ fold change threshold for sample size of 40 per group (20 per group/treatment combination), bottom row shows the influence of sample size for a $\log_2$ fold change threshold set to 1.

thus influence the results of gene set enrichment. Finally, we have used a $\log_2$ fold change threshold of 1, because raising it would increase the fraction of FNs. However, filtering for biologically relevant genes with a substantially higher effect size may influence the observed enrichments. On the other hand, raising the $\log_2$ fold change threshold or lowering the sample size may lead to a smaller number of defined DEGs, thus making the hypergeometric test less powerful and lead to fewer gene set enrichment.

To investigate whether the sampling had an impact on the artifacts, we have repeated the above procedure for 100 replicates, each containing a different set of samples randomly assigned to the two groups. Furthermore, we tested whether the selection of the $\log_2$ fold change threshold or different sample size might have an impact on the extent of the arising artifacts. To this end, we have tested 7 different threshold values for the $\log_2$ fold changes and three sample sizes: 20, 40 or 80 samples per group (corresponding to 10, 20 or 40 samples per group/treatment combination).

Setting a higher $\log_2$ fold change threshold reduces the number of DEGs as well as of the observed artifactual gene set enrichments (**Figures 4A,B**). However, even for $\log_2$ threshold of 3 (DEGs defined by 8-fold change and FDR <0.05) the number of replicates in which artifacts can be observed is 78 (out of 100), and in at least 31 replicates, 5 or more gene sets were significantly enriched. Thus, setting a more conservative threshold while retaining the incorrect procedure cannot fully protect from the arising artifacts. The number of artifacts rose with sample size (**Figures 4C,D**). For total sample size of 160 (40 samples per group/treatment combination) the number of artifacts was almost 2 times higher than for total sample size of 80.

Incorrect Analysis of Interactions is Common in Transcriptomics

Nieuwenhuis, Forstmann, and Wagenmakers (2011) observed incorrect analyses of interactions in about half of the analyzed papers from top neuroscience journals where the authors considered an experimental design allowing for a test for interaction. We wanted to know if this problem is common in transcriptomics, too. To this end, we have searched three journals—from broad to specialized—for the occurrence of the terms "differential expression" with "venn diagrams" (**Table 2**). Next, we analyzed the selected papers from year 2020 (and years 2015–2020 in one case) to decide whether the VD was described or referred to as showing genes "specific" or "unique" to a particular group or whether a test for interaction was performed. Finally, we checked whether gene set enrichment analysis was applied to genes significant in one, but not the other group in order to find group-specific differences.

We found that of the 282 analyzed articles which used the terms "venn diagram" and "differential expression," at least 88 (31%) were using Venn diagrams to compare statistical significance with lack thereof by referring to "unique,"

TABLE 2 | Results of the informal literature survey. We searched for papers using Google Scholar and the keywords "venn diagram" and "differential expression." Journal, journal title; Years, publication dates; Total, total number of papers found using the search phrase; Analyzed, total number of papers which have been analyzed for correctness; Incorrect, total number (percentage) of papers in which the Venn diagram was combined with comparing significance to non-significance; Enrichment, total number (percentage) of papers which combined the Venn diagram with a gene set enrichment analysis.

Journal	Years	Total	Analyzed	Incorrect	Enrichment
Nature Communications	2020	127	30	9 (30%)	6 (20%)
Science Immunology	2015–2020	14	14	6 (43%)	5 (36%)
Scientific Reports	2020	238	238	73 (31%)	42 (18%)
Total		379	282	88 (31%)	53 (19%)

"specific," "solely regulated" or "exclusive" DEGs. Out of these, at least 53 coupled the VDs with some form of gene set enrichment analysis on the set of supposedly "specific" DEGs. In summary, in at least a quarter of the papers on differential expression in which a Venn diagram was used, it was illustrating an incorrect statistical procedure which may result in artifactual gene set enrichments.

DISCUSSION

Drawing conclusions from comparing significance with lack thereof is a common statistical fallacy (Gelman and Stern 2006). Just as absence of evidence is not evidence of absence, the failure to reject the null hypothesis does not constitute the same level of evidence as rejecting it. However, when such an incorrect analysis is combined with downstream functional analysis, the resulting pathways or gene ontologies are misleadingly relevant. For example, the identified gene sets are associated with immune response for a research hypotheses involving an infectious disease, or cancer pathways if the underlying research hypothesis involved cancer treatment. Such results may appear reasonable in the given context, especially if the correct analysis of interactions does not show any significant differences. This effect is persistent or even exacerbated for larger sample sizes (**Figure 4**).

We found that this type of incorrect analysis occurs in more than a quarter of papers where the procedure was illustrated with a VD. That is not to say that VDs are not a useful tool, even in the context of transcriptomics and gene set enrichments, if used correctly. For example, gene set enrichment analysis of an intersection of DEGs (i.e., by considering genes from the overlap in a VD) is not an incorrect procedure. Genes in the overlapping part of a VD are significant in both (or all) comparisons, hence no comparison between significance and non-significance is made.

While VDs appear to be frequently associated with an incorrect statistical reasoning, the use of VDs is not the cause. In the absence of a VD illustrating the DEGs common and unique to the different study groups, two incorrect approaches may still be found. Firstly, the direct comparison of gene set enrichment results: that is, drawing conclusions from the fact that a gene set enrichment result was significant in one comparison only. Second, while VDs are often used to illustrate the numbers of "specific" DEGs and so present a mean to find examples of this fallacy in scientific literature, researches test for enrichment these "specific" genes without using the phrase "Venn diagram" or even clearly stating how the lists of "specific" genes were derived. In all these cases, the analysis boils down to comparing results significant in one, but not in another comparison.

As an alternative to Venn diagrams and the downstream gene set enrichment analysis, two approaches can be considered. The correct statistical approach, as shown above, involves a test for interaction which can reveal genes for which the impact of treatment significantly differ between the groups. The results can then be plugged into a gene set enrichment analysis the usual way. Unfortunately, this has two major drawbacks. Firstly, effect sizes ($\log_2$ fold changes) of the interaction term are harder to interpret than $\log_2$ fold changes in a direct, group vs. group comparison. The effect size in an interaction is negative if the $\log_2$ fold change in the first comparison is larger than the $\log_2$ fold change in the second comparison; this is, however, irrespective of whether the differences in the individual comparisons are negative or positive, which makes it harder to separate the differences in genes upregulated in one or both groups.

The second problem may arise if the changes are similar in both comparisons, but of larger magnitude in one of them. For example, in a time series context, the changes may be more pronounced at a later time point. In this case, the analysis will show that the processes enriched for the interaction term are the same as those enriched in each of the comparisons individually. While the results of the gene set enrichment analysis in this context are correct, the result may not be what the researchers intended—processes which qualitatively (rather than quantitatively) differ between the comparisons.

An alternative approach, discordance/concordance analysis, has been proposed by Domaszewska et al. (2017), aiming at identifying processes which qualitatively differ between the two comparisons. Here, a heuristic score ("disco score") has been defined which depends on the effect sizes and *p*-values in both comparisons. The sign of the score depends on whether the effects have the same sign (concordant; genes upregulated in both comparisons or downregulated in both comparisons) or opposite signs (discordant; genes upregulated in one, but downregulated in the other comparison, and genes downregulated in one, but upregulated in the other comparison). While the score does not allow the calculation of a *p*-value and does not present an alternative to an analysis of

interaction, it can facilitate both visualization and further analysis using a gene set enrichment algorithm.

Visualization of interaction for individual genes is straightforward (see for example **Figures 2A–D**). However, the point of VDs is to show a grand overview of the whole analysis summarizing thousands of results for the analyzed genes. As an alternative of such an overview, we suggest plotting the log~2 fold changes in one comparison against $\log_2$ fold changes in the second comparison. This allows an intuitive assessment of the differences between the two comparisons, in especially in combination with color coding the genes which either are significant in the interaction or by coloring using the disco score (see **Figure 2E** for an example).

Defining group-specific genes based on significant difference in one, but no significant difference in another comparison is thus more than only a statistical fallacy leading to erroneous results. When combined with gene set enrichment analysis it can lead to potentially sound-looking, and therefore particularly misleading results. This method of obtaining specific differences between groups should therefore be abandoned in favor of statistically correct approaches. Furthermore, gene set enrichment analysis must never be applied to sets of genes defined as significant in one comparison, but not the other.

METHODS

Methods Availability

This manuscript has been written in R markdown (Xie, Allaire, and Grolemund 2018).All statistical calculations required to replicate the findings and figures are contained in the source R markdown file. The R markdown file, along with additional files required to recreate this manuscript as well as the results of literature survey have been uploaded to https://github.com/bihealth/manuscript_venn_diagrams.

Data

The expression data as a count matrix has been downloaded from GEO, accession GSE156063.

Statistical Analyses

Power calculation was done using the R package pwr, version 1.3.0. For differential gene expression, the R package DESeq2, version 1.32.0 has been used. Gene set enrichments were done using either hypergeometric test (where stated) or the CERNO test using the package tmod (Zyla et al., 2019), version 0.50.1. GO terms have been sourced from the R package msigdbr, version 7.4.1.

Simulation Study

We have generated replicates of the example study for different $\log_2$ fold change thresholds (0, 0.5, 1, 1.5, 2, 2.5, 3) and three different sample sizes (40, 80 and 160, corresponding to sample size per group/treatment combination of 10, 20 and 40). For each replicate, the full procedure as described above was repeated, and numbers of DEGs and significantly enriched terms were collected.

Literature Survey

A literature survey was performed using Google Scholar to estimate the frequency of the incorrect use of Venn diagrams. We searched for articles including the phrases "differential expression" and "venn diagram" in three journals: Scientific Reports (2020), Nature Communications (2020) and Science Immunology (2015–2020). For each of the papers identified, we checked whether 1) the authors used the VD to show differentially expressed transcripts significant in one comparison, but not another, 2) the authors discussed "unique," "non-overlapping" or "specific" regions of the Venn diagram and 3) whether this was coupled to gene set enrichment analysis is any form. Articles which 1) focused only on the intersections of the Venn diagrams (genes common to all groups), or 2) which used the Venn diagrams for a purpose other than to compare genes significant in one groups, but not significant in other groups or 3) for which a clear-cut error could not be identified past any reasonable doubt were not considered incorrect. The results of the literature survey (including links to papers classified as incorrectly analysing the interaction) are included in the manuscript sources.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156063.

AUTHOR CONTRIBUTIONS

Study design, coding and execution: JW. Manuscript writing: JW, DB and BO.

REFERENCES

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 (Methodological)* 57 (1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Blalock, H. M., Jr (1965). Theory Building and the Statistical Concept of Interaction. *Am. Sociological Rev.* 30, 374–380. doi:10.2307/2090718

Domaszewska, T., Scheuermann, L., Hahnke, K., Mollenkopf, H., Dorhoi, A., Kaufmann, S. H. E., et al. (2017). Concordant and Discordant Gene Expression Patterns in Mouse Strains Identify Best-Fit Animal Model for Human Tuberculosis. *Sci. Rep.* 7 (1), 12094–12113. doi:10.1038/s41598-017-11812-x

Gelman, A., and Stern, H. (2006). The Difference between "Significant" and "Not Significant" Is Not Itself Statistically Significant. *The Am. Statistician* 60 (4), 328–331. doi:10.1198/000313006x152649

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2019). Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. *Hum. Genomics* 13 (1), 42–12. doi:10.1186/s40246-019-0226-2

Mick, E., Kamm, J., Pisco, A. O., Ratnasiri, K., BabikBabik, J. M., Castañeda, G., et al. (2020). Upper Airway Gene Expression Reveals Suppressed

Immune Responses to SARS-CoV-2 Compared with Other Respiratory Viruses. *Nat. Commun.* 11 (1), 5854–5857. doi:10.1038/s41467-020-19587-y

Nieuwenhuis, S., ForstmannForstmann, B. U., and Wagenmakers, E.-J. (2011). Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance. *Nat. Neurosci.* 14 (9), 1105–1107. doi:10.1038/nn.2886

White, T., van der Ende, J., and Nichols, T. E. (2019). Beyond Bonferroni Revisited: Concerns over Inflated False Positive Research Findings in the Fields of Conservation Genetics, Biology, and Medicine. *Conserv Genet.* 20 (4), 927–937. doi:10.1007/s10592-019-01178-0

Xie, Y., Joseph, J. A., and Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Boka Raton, London and New York: Chapman; Hall/CRC.

Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S. H. E., Polanska, J., and Weiner, J., 3rd (2019). Gene Set Enrichment for Reproducible Science: Comparison of CERNO and Eight Other Algorithms. *Bioinformatics* 35 (24), 5146–5154. doi:10.1093/bioinformatics/btz447

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

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

Copyright © 2022 Weiner, Obermayer and Beule. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 February 2022
doi: 10.3389/fgene.2021.820154

A New Prognostic Risk Score: Based on the Analysis of Autophagy-Related Genes and Renal Cell Carcinoma

*Minxin He[1,2], Mingrui Li[1,2], Yibing Guan[1,2], Ziyan Wan[1,2], Juanhua Tian[1,2], Fangshi Xu[1,2], Haibin Zhou[1,2], Mei Gao[1], Hang Bi[1,2] and Tie Chong[1]**

[1]Department of Urology, The Second Afilliated Hospital, School of Medicine, Xi'an Jiaotong University, Xi'an, China, [2]School of Medicine, Xi'an Jiaotong University, Xi'an, China

Introduction: Clear cell renal cell carcinoma (ccRCC) patients suffer from its high recurrence and metastasis rate, and a new prognostic risk score to predict individuals with high possibility of recurrence or metastasis is in urgent need. Autophagy has been found to have a dual influence on tumorigenesis. In this study we aim to analyze autophagy related genes (ATGs) and ccRCC patients and find a new prognostic risk score. Method: Analyzing differential expression genes (DEGs) in TCGA-KIRC dataset, and took intersection with ATGs. Through lasso, univariate, and multivariate cox regression, DEGs were chosen, and the coefficients and expression levels of them were components constructing the formula of risk score. We analyzed mRNA expression of DEGs in tumor and normal tissue in ONCOMINE database and TCGA-KIRC dataset. The Human Protein Atlas (HPA) was used to analyze protein levels of DEGs. The protein-protein interaction (PPI) network was examined in STRING and visualized in cytoscape. Functional enrichment analysis was performed in RStudio. To prove the ability and practicibility of risk score, we analyzed univariate and multivariate cox regression, Kaplan-Meier curve (K-M curve), risk factor association diagram, receiver operating characteristic curve (ROC curve) of survival and nomogram, and the performance of nomogram was evaluated by calibration curve. Then we further explored functional enrichment related to risk groups through Gene Set Enrichment Analysis (GSEA), weighted gene co-expression network analysis (WGCNA), and Metascape database. At last, we investigated immune cell infiltration of DEGs and two risk groups through TIMER database and "Cibersort" algorithm.

Result: We identified 7 DEGs (BIRC5, CAPS, CLDN7, CLVS1, GMIP, IFI16, and TCIRG1) as components of construction of risk score. All 7 DEGs were differently expressed in ccRCC and normal tissue according to ONCOMINE database and TCGA-KIRC dataset. Functional enrichment analysis indicated DEGs, and their most associated genes were shown to be abundant in autophagy-related pathways and played roles in tumorigenesis and progression processes. A serious analysis proved that this risk score is independent from the risk signature of ccRCC patients.

Conclusion: The risk score constructed by 7 DEGs had the ability of predicting prognosis of ccRCC patients and was conducive to the identification of novel prognostic molecular markers. However, further experiment is still needed to verify its ability and practicability.

Keywords: risk score, prognosis, bioinformactics analysis, renal cell carcinoma, autophagy

OPEN ACCESS

Edited by:
Farhad Maleki,
McGill University, Canada

Reviewed by:
Mahmoud Ahmed,
Gyeongsang National University, South Korea
Dipayan Roy,
All India Institute of Medical Sciences Jodhpur, India
Trang Huyen Lai,
Gyeongsang National University, South Korea

****Correspondence:***
Tie Chong
chongtie@126.com

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

Received: *22 November 2021*
Accepted: *30 December 2021*
Published: *14 February 2022*

Citation:
He M, Li M, Guan Y, Wan Z, Tian J, Xu F, Zhou H, Gao M, Bi H and Chong T (2022) A New Prognostic Risk Score: Based on the Analysis of Autophagy-Related Genes and Renal Cell Carcinoma. Front. Genet. 12:820154. doi: 10.3389/fgene.2021.820154

INTRODUCTION

Renal cell carcinoma (RCC) is a prevalent tumor of the urinary system as it was reported by GLOBOCAN in 2020, with an incidence of 2.2% and mortality of 1.8% annually (Sung et al., 2021). Clinically, the main treatment of RCC patients is radical or unitary partial nephrectomy; however, about 30% of postoperative patients have the potential to be found with recurrence or metastasis (Capitanio et al., 2019). RCC is insensitive to chemotherapy or radiation, although in recent years anti-angiogenesis molecular targeted therapy has become the standard of care for advanced RCC, and most patients have developed drug resistance after 5–11 months (Khattak and Larkin, 2014). Up to now, the diagnosis of recurrence or metastasis of RCC still relies on imaging, but it is always too late, and patients who were found recurrence or metastasis by imaging have a poor prognosis. Thus, it's of great significance for early diagnosis and treatment of RCC patients to find new biomarkers.

Autophagy refers to the process by which lysosomes decompose cellular materials to provide cells with biosynthetic components and energy sources (Glick et al., 2010). This process has been found relevant to many human diseases (Mizushima and Levine, 2020) such as cardiovascular disease (Gatica et al., 2021), Parkinson's disease (Lizama and Chu, 2021), Alzheimer's disease (Zhang et al., 2021), and so on. In the process of tumorigenesis and development, researches found autophagy had dual roles; in the earlier stage autophagy inhibits tumors from happening, while in the later stage it facilitates the progression of tumor (Kimmelman, 2011; Rangel et al., 2021). Clear cell RCC (ccRCC) accounts for the majority of RCC and had poorer prognosis; as a result, our study aims to combine ccRCC and autophagy and investigate how autophagy affects ccRCC, then build a risk score and provide insight for prognosis and treatment of ccRCC.

MATERIALS AND METHODS

Data Source

We obtained the clinical information, raw counts of RNA-sequencing data, overall survival (OS), and disease free survival (DFS) of 537 ccRCC patients and 74 paracancerous samples in the cancer genome atlas-kidney renal clear cell carcinoma (TCGA-KIRC) dataset from the TCGA database (Cancer Genome Atlas Research et al., 2013) (http://portal.gdc.cancer.gov) through R package "TCGAbiolinks" (Colaprico et al., 2016). Gene IDs conversion were finished with the assistance of a GTF file which were downloaded from GENCODE (http://www.gencodegenes.org/), and 18,569 protein-coding genes were annotated by gene IDs and were selected for subsequent analysis. To meet the requirement of data integrality, patients with the following criteria were excluded from subsequent analysis: (1) patients with OS less than 1 month, (2) patients with inadequate clinical information. Finally, a total of 515 ccRCC patients were selected for further analysis. A total of 531 autophagy related genes (ATGs) was gathered from the human autophagy database (HADb, http://www.autophagy.lu/index.html) and GO_AUTOPHAGY dataset from The Molecular Signatures Database (MsigDB) (http://www.gsea-msigdb.org/gsea/msigdb/index.jsp) (Wang Y. et al., 2020).

Selecting DEGs

Variation analysis of gene expression in TCGA-KIRC dataset was accomplished by R package "Deseq2" (Love et al., 2014), genes with $|\log_2$ Fold Change$|$ ($|\log_2$ FC$|$) ≥ 1, and adjusted p value < 0.05 were regarded as differentially expressed genes (DEGs). Take the intersection of DEGs and ATGs. R package "ezcox" (http://github.com/ShixiangWang/ezcox/issue/23) was used for univariate cox regression of the intersection, then genes with $p < 0.05$ in univariate cox regression underwent lasso regression and multivariate cox regression. Finally, genes with $p < 0.05$ were selected as DEGs that were selected to construct a new risk score formula.

Analysis of mRNA and Protein Expression Levels of DEGs

mRNA expression levels of DEGs were analyzed based on the data from TCGA-KIRC dataset and visualized by RStudio. Meanwhile, we explored mRNA expression levels of DEGs in different datasets through the ONCOMINE database (Rhodes et al., 2004) (http://www.oncomine.org). The protein levels of DEGs were tested through The Human Protein Atlas (Uhlen et al., 2017) (HPA, http://www.proteinatlas.org).

Protein-Protein Interaction Network and Enrichment Analysis

Protein-protein interaction (PPI) network analysis was finished in STRING (Szklarczyk et al., 2019) (http://string-db.org), selecting the top 50 closest genes with DEGs and visualized by cytoscape. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of DEGs and their closest genes was finished by R package "Clusterprofiler" (Wu et al., 2021).

Construction of Risk Score

Performing multivariate cox regression of DEGs and collecting the expression levels of DEGs and their coefficients to construct the formula of risk score:

$$\text{Risk score} = \sum_{i=1}^{n} coef_i * Exp_i$$

Correlation analysis of risk score and other clinical signatures was performed by the method of "pearson." We divided patients from the TCGA-KIRC dataset into two cohorts, train cohort and validation cohort with R package "caret." Depict the receiver operating characteristic curve (ROC curve), Kaplan-Meier curve (K-M curve), and risk factor association diagram of risk score and calculate its area under the curve (AUC) in train cohort. Furthermore, univariate and multivariate cox regression was performed to prove risk

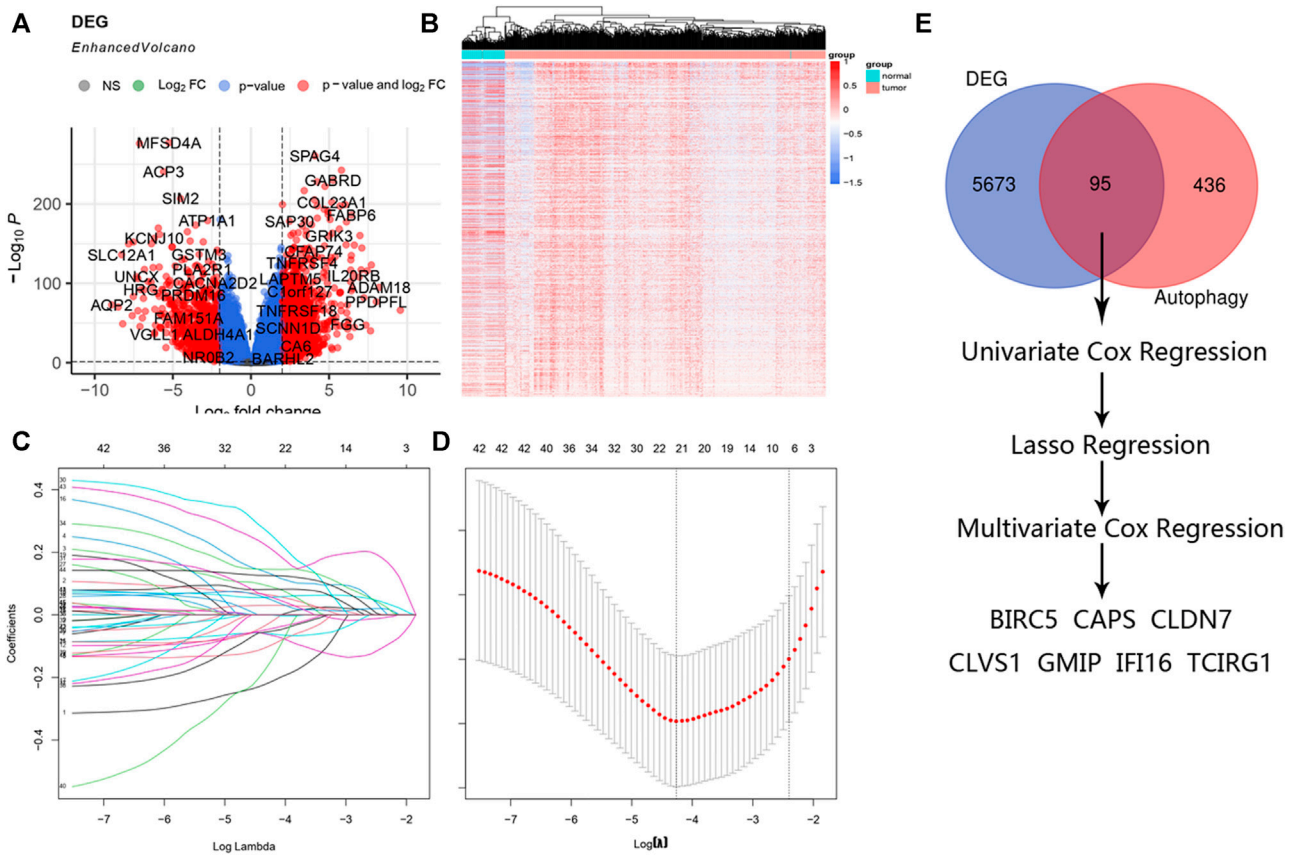

FIGURE 1 | Selection of DEGs. **(A)** Enhanced volcano plot of DEGs when comparing ccRCC with normal tissue. Red nodes represented genes with $|\log_2 FC| \geq 1$ and adjusted $p < 0.05$, blue nodes represented genes with $p < 0.05$ only, and grey nodes represented genes that were neither eligible in conditions of p value nor $|\log_2 FC|$. **(B)** Heatmap of DEGs in ccRCC. **(C)** Lasso coefficients profiles of 95 genes significant in univariate cox regression. **(D)** Lasso regression obtained 21 prognostic genes using minimum lambda value. **(E)** Selecting procedure of DEGs, venn gram showed 95 genes in the intersection of 5,768 DEGs and 531 ATGs. These genes then underwent univariate cox regression, lasso cox regression, multivariate cox regression, and finally 7 DEGs were selected to construct the risk score formula.

TABLE 1 | Multivariate cox regression of DEGs.

Gene symbol	Coefficient	*p* Value	HR (95%CI)
BIRC5	0.000355	0.010	1.00036 (1.00009, 1.00062)
CAPS	0.000382	0.014	1.00038 (1.00008, 1.00069)
CLDN7	−0.000131	0.010	0.99987 (0.99977, 0.99997)
CLVS1	0.001747	<0.001	1.00175 (1.00105, 1.00245)
GMIP	−0.00033	0.048	0.99967 (0.99934, 1.00000)
IFI16	0.000082	0.007	1.00008 (1.00002, 1.00014)
TCIRG1	0.000118	0.021	1.00012 (1.00002, 1.00022)

HR, hazard ratio.

score as an independent risk factor of ccRCC patients. Likewise, we tested results above through data from validation cohort and total cohort. Nomogram was to predict the probability of 1-, 3-, and 5-years survival for ccRCC patients according to the results from multivariate cox regression, and calibration curves were drawn to evaluate the nomogram. We've published a glycolysis-related risk score signature before, since both of our studies were metabolism-related, and we then compared their ability of predicting prognosis of ccRCC patients by depicting ROC curve and circulating AUC. All analytical methods above were finished in Rstudio with R packages such as "timeROC," "survival," "survminer," "rms," and "ggrisk," and $p < 0.05$ was considered as statistically significant.

To further explore pathways related with risk score we then performed in Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) in high and low risk groups; | Normalized Enrichment Score| ($|NES|$) ≥ 1.5, $p < 0.05$ and false discovery rate (FDR) < 0.25 were set as threshold. Additionally, to find out the genes that were connected with risk score and their function, weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) was performed, and genes with top 5,000 median absolute deviation were analyzed and soft threshold was selected when scale free $R^2 = 0.9$. Genes in the most significant co-expression module were then analyzed in Metascape (Zhou et al., 2019) (http://metascape.org). The

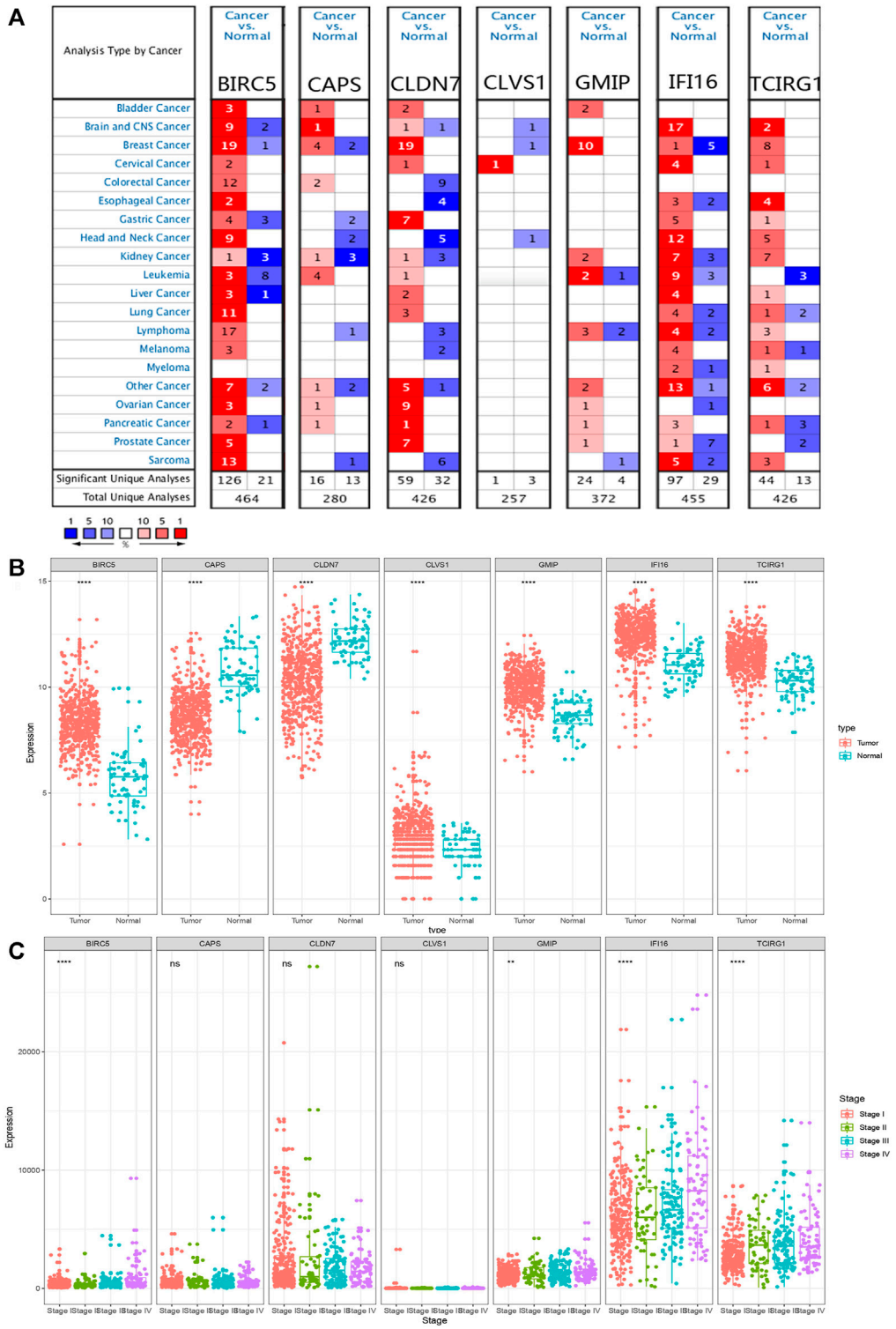

FIGURE 2 | mRNA expression levels of DEGs. **(A)** mRNA expression levels of DEGs from ONCOMINE database. The threshold ($p < 0.05$, $|\log_2 FC| \geq 1.5$, gene rank: Top 10% datatype: mRNA) was indicated in the colored cells. The red cells indicated the target gene was overexpressed in ccRCC while blue cells represented downregulated in ccRCC. Gene rank was depicted in the color depth in the cells. **(B)** mRNA expression levels of DEGs in ccRCC and normal tissue from TCGA database. *p* value was replaced by "*," $^{ns}p > 0.05$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ and $^{****}p < 0.0001$. **(C)** Different mRNA expression levels of DEGs in different stages.

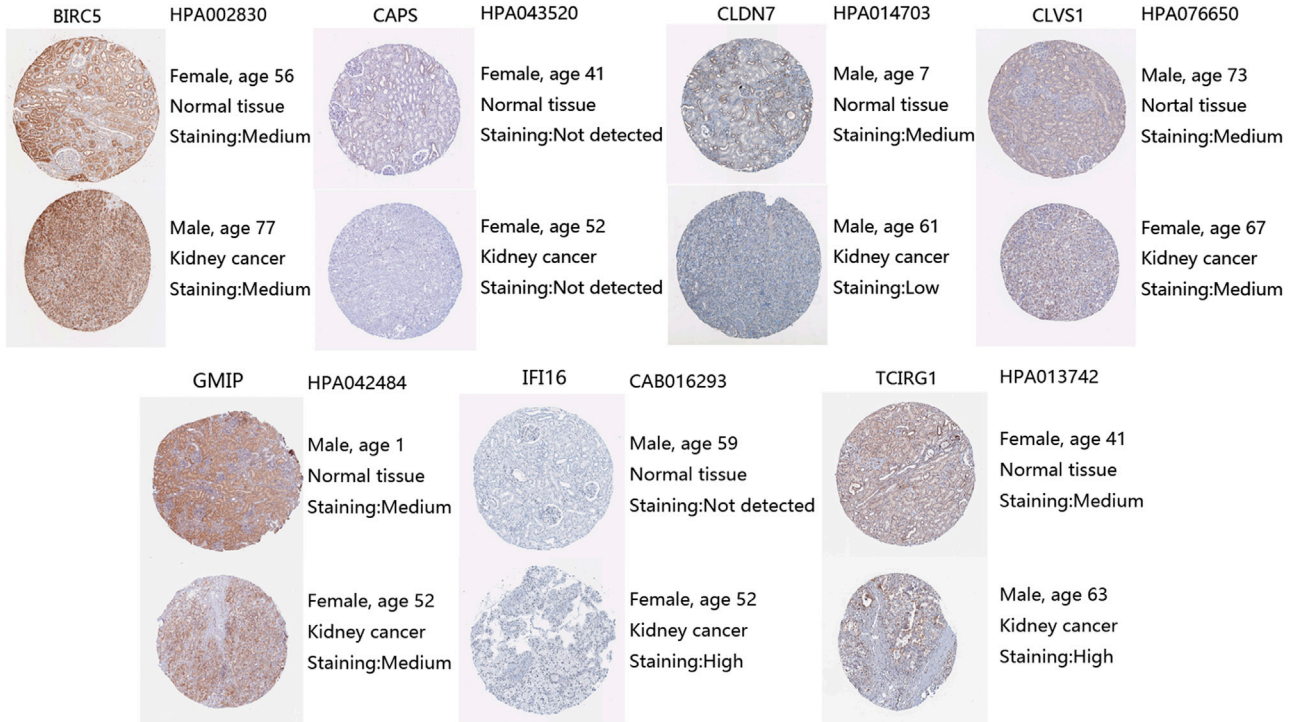

FIGURE 3 | Protein expression levels of DEGs from HPA, antibody, and patient information were listed as well.

threshold of min overlap = 3, p value cutoff = 0.01, and min enrichment = 1.5 was set to select enriched pathways in the module. MCODE was selected with physical score >0.132, min network size = 3, max network size = 500, and databases as physical core.

Immune Cell Infiltration

We explored immune infiltration of DEGs from TIMER 2.0 (Li et al., 2020) (https:// timer.cistrome.org/). As for the analysis of immune infiltration in ccRCC patients, "Cibersort" algorithm (Newman et al., 2015) was performed in RStudio. Also, the difference of immune infiltration in different groups of risk score was analyzed.

RESULTS

Acquisition of DEGs

The raw counts data was downloaded through R package "TCGAbiolinks" from TCGA-KIRC with setting data category as "Transcriptome Profiling" and data type as "Gene expression Quantigication." Up to 56,612 genes were downloaded, and after gene ID conversion we obtained 18,569 mRNA. They were then estimated with differential analysis between ccRCC patients and paracancerous patients by R package "Deseq2." With the threshold of $|\log FC| \geq 1$, adjusted p value < 0.05, we obtained 5,768 DEGs among which 471 were up-regulated, 388 were down-regulated, and volcano plot and heatmap were drawn for better understanding (**Figures 1A,B**).

Selection of DEGs

A total of 95 genes were in the intersection between DEGs and ATGs, and they were analyzed with univariate cox regression. Among 95 genes there were 46 genes with $p < 0.05$. We then performed lasso regression analysis and got 21 genes (**Figures 1C,D**). These 21 genes were analyzed by multivariate cox regression, among which 7 genes were found significant ($p < 0.05$). Thus, we obtained 7 genes (BIRC5, CAPS, CLDN7, CLVS1, GMIP, IFI16, and TCIRG1) as DEGs to construct the formula of new risk score (**Figure 1E**; **Table1**).

Analysis of mRNA and Protein Expression Levels of DEGs

Comparing the mRNA expression levels of BIRC5, CAPS, CLDN7, CLVS1, GMIP, IFI16, and TCIRG1 in ccRCC patients and normal people, we found that BIRC5, CLVS1, GMIP, IFI16, and TCIRG1 were significantly overexpressed in ccRCC patients from TCGA-KIRC dataset and CAPS and CLDN7 had lower expression level in ccRCC patients than

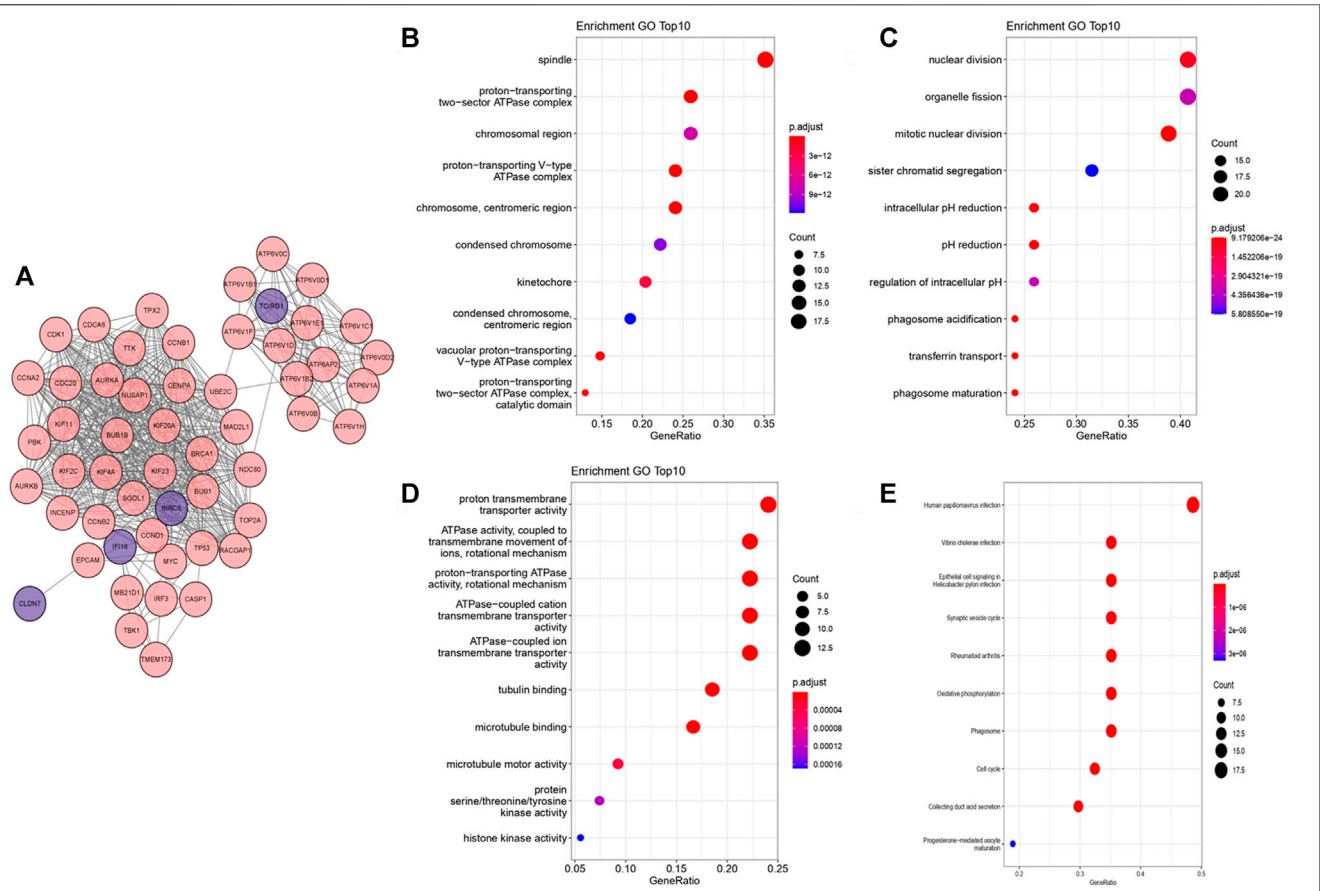

FIGURE 4 | PPI network and functional enrichment analysis from "Clusterprofiler." The deeper the color is, the more significant the enrichment is. The bigger the bubble is, the more genes are participated in the term. **(A)** PPI network from STRING visualized by cytoscape. **(B)** Bubble map of top 10 enriched GO terms in CC. **(C)** Bubble map of top 10 enriched GO terms in BP. **(D)** Bubble map of top 10 enriched GO terms in MF. **(E)** Bubble map of top 10 KEGG enriched pathways.

in normal people. Apart from BIRC5, GMIP, IFI16, and TCIRG1, other genes didn't show significant differences in individual cancer stages (**Figures 2A–C**). Results from ONCOMINE database partly coordinated with what we found before, that BIRC5 was found overexpressed in Gumz Renal ($FC = 2.753$, p value < 0.001), IFI16 was highly expressed in ccRCC patients from Gumz Renal ($FC = 4.454$, $p < 0.001$), Yusenko Renal ($FC = 5.099$, $p < 0.001$), Lenburg Renal ($FC = 2.160$, $p < 0.001$) and Jones Renal ($FC = 3.863$, $p < 0.001$). GMIP was overexpressed in Yusenko Renal ($FC = 4.020$, $p < 0.001$). TCIRG1 was overexpressed in Yusenko Renal ($FC = 2.516$, $p = 0.001$), Jones Renal ($FC = 2.153$. $p < 0.001$), and Lenburg Renal ($FC = 1.860$, $p = 0.001$) (**Supplementary Material S1**).

Moreover, we explored protein expression levels of DEGs in the HPA website, compared with normal kidney tissue, IFI16 and TCIRG1 were highly expressed in ccRCC kidney tissue. CLDN7 were found lower expressed in ccRCC tissue than in normal tissue. BIRC5, CLVS1, and GMIP were found the same level in ccRCC tissue as in normal tissue. Unfortunately, CAPS was not detected in ccRCC kidney tissue nor normal kidney tissue (**Figure 3**).

PPI Network and Functional Enrichment Analysis

We performed PPI network in STRING and selected the top 50 closest genes with DEGs (**Supplementary Material S2**, **Figure 4A**). After GO function enrichment analysis and KEGG pathway analysis, we found the top 10 significant items of 57 genes in cellular component (CC) were spindle, protein-transporting two-secor ATPase complex, chromosomal region, proton-transporting V-type ATPase complex, chromosome, centromeric region, condensed chromosome, kinetochore, condensed chromosome, centromeric region, vacuolar protein-transporting V-type ATPase complex, proton-transporting two-sector ATPase complex, and catalytic domain (**Figure 4B**). The top 10 significant items of biological process (BP) included nuclear division, organelle fission, mitotic nuclear division,

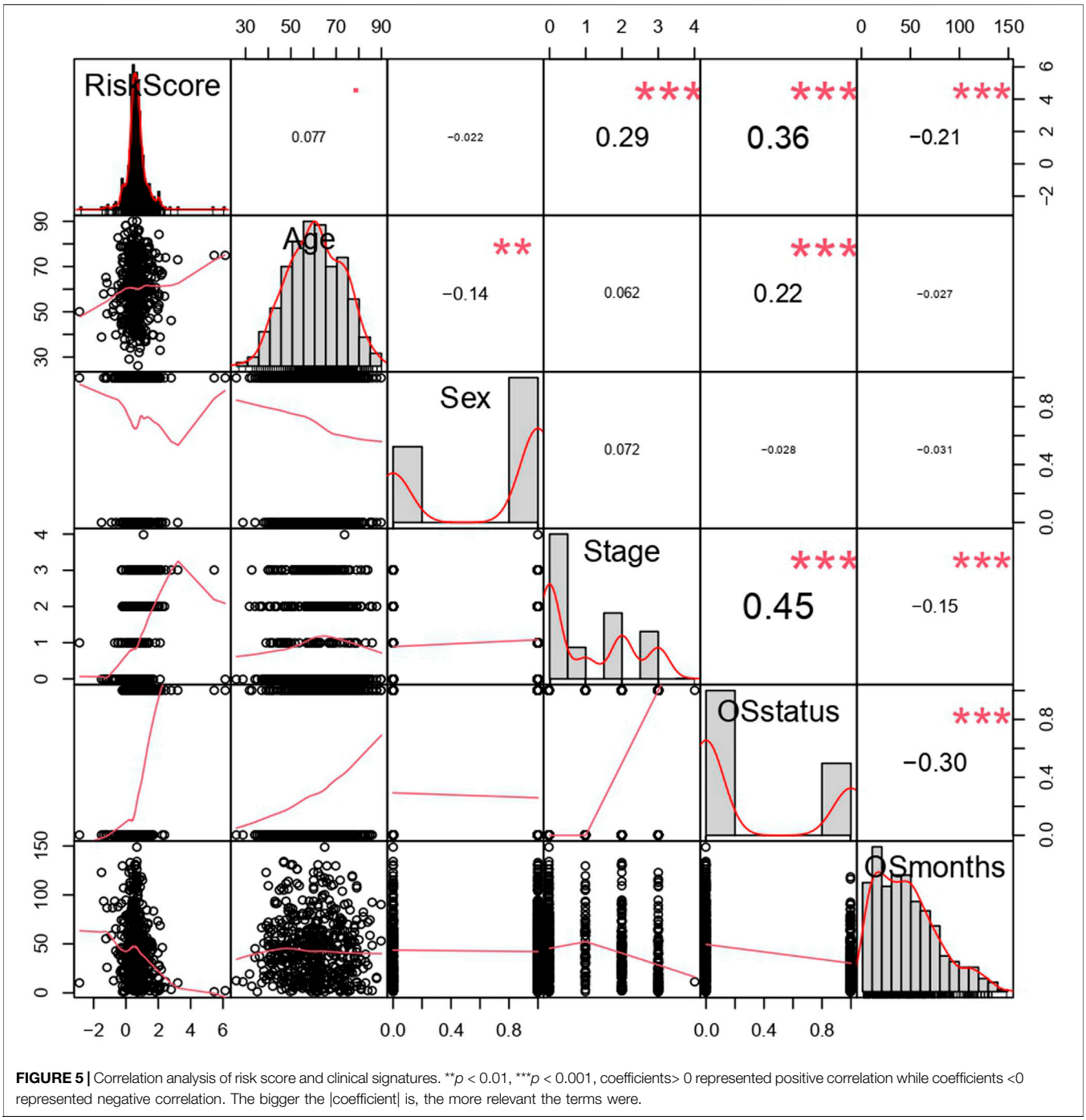

FIGURE 5 | Correlation analysis of risk score and clinical signatures. $^{**}p < 0.01$, $^{***}p < 0.001$, coefficients> 0 represented positive correlation while coefficients <0 represented negative correlation. The bigger the |coefficient| is, the more relevant the terms were.

sister chromatid segregation, intracellular pH reduction, pH reduction, regulation of intracellular pH, phagosome acidification, transferrin transport, and phagosome maturation (**Figure 4C**). Top molecular functions (MFs) were mainly associated with energy metabolism including proton transmembrane transporter activity, ATPase activity, coupled to transmembrane movement of ions, rotational mechanism, ATPase-coupled cation transmembrane transporter activity, ATPase-coupled ion transmembrane transporter activity, tubulin binding, microtubule binding, microtubule motor binding, protein serine/threonine/tyrosine kinase activity, and histone kinase activity (**Figure 4D**). In KEGG pathway analysis, several pathways were found related with autophagy, such as rheumatoid arthritis, phagosome, cell cycle, and oxidative phosphorylation (**Figure 4E**).

Construction of Risk Score

According to expression levels of 7 DEGs and their coefficients, we constructed the formula of risk score (**Table 1**):

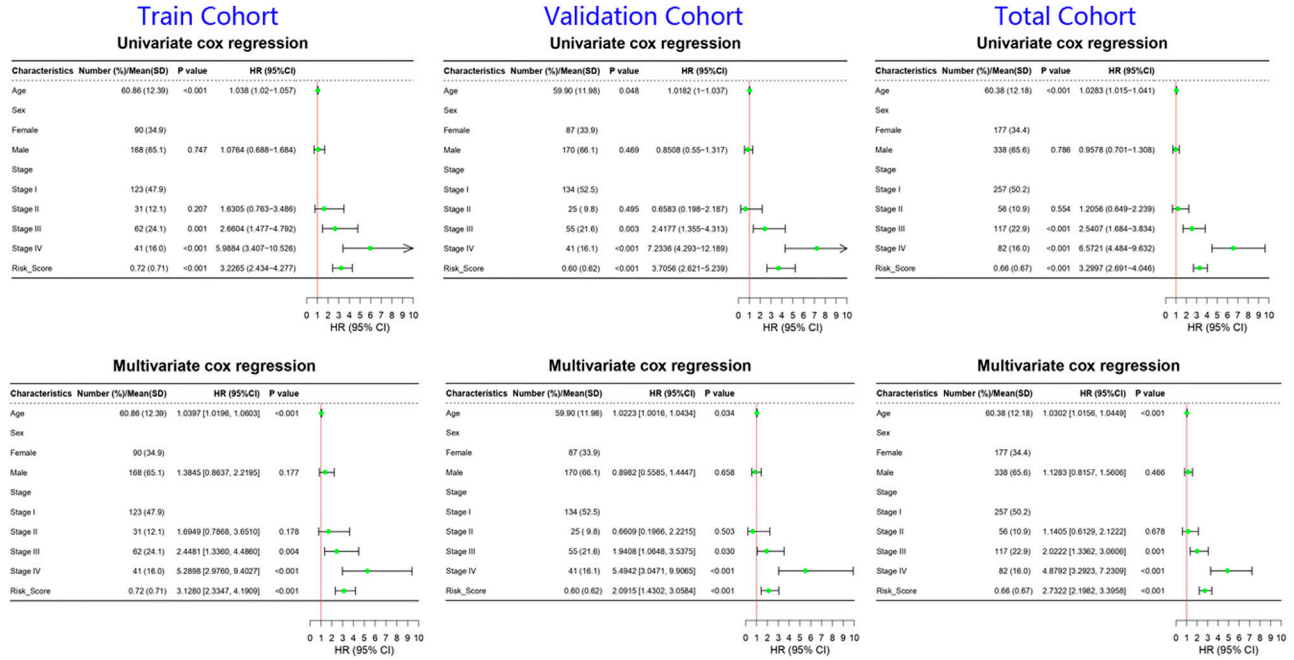

FIGURE 6 | Forrest plots of univariate and multivariate cox regression of ccRCC patients in train cohort, validation cohort, and total cohort. The green nodes represented HR and the line extending from the nodes indicated 95% CI.

$$\text{Risk score} = \sum_{i=1}^{7} coef_i * Exp_i$$

By analyzing correlation between risk score and other clinical signatures, we found that risk score was positively associated with stage and OS status while negatively related with OS months (**Figure 5**). Analyzing clinical characteristics and risk score of patients from TCGA-KIRC in train cohort, validation cohort, and total cohort, we found age of diagnosis, stage, and risk score were independent risk factors of prognosis of ccRCC patients (**Figure 6**). Dividing all patients from train cohort into two groups (high risk and low risk) according to median of risk score, K-M curve, and Log-rank test of OS demonstrated significant difference, and the high risk group had shorter OS than patients in low risk group (**Figure 7A**). Time-dependent ROC curve showed the AUC of the first year, the third year, and the fifth year was 0.75, 0.68, and 0.70, respectively, which indicated risk score had a good predictive ability (**Figure 7B**). Depicting risk factor association diagram, it was clear to see that as risk score rose, mortality grew, and survival time was reduced (**Figure 7C**). Combining with all the clinical signatures that mattered, we constructed a nomogram to predict the survival rate of ccRCC patients. Calibration curve verified the accuracy of its ability to predict prognosis (**Figures 7D,E**). Similar analyses were performed in the validation cohort which provided stronger evidence of our risk score having significant value in predicting prognosis of ccRCC patients with AUC of the first year, the third year, and the fifth year of 0.73, 0.71, and 0.78, respectively. Nomogram and calibration in validation cohort also validated the practicability of the model (**Figures 8A–E**). In the total cohort, time-dependent ROC curve showed the ability of risk score predicting prognosis with AUC of the first year, the third year, and the fifth year at 0.74, 0.70, and 0.74, respectively (**Figure 9A**). Comparing with other clinical signatures we found the AUC of stage was 0.75, AUC of risk score was 0.72, AUC of age was 0.63, and AUC of sex was 0.50. We've published a glycolysis-related risk score before, and since they were both metabolic-related signatures we compared their ability of predicting prognosis of ccRCC patients by AUC. It turned out that as the AUC of glycolysis-related risk score was 0.66, the autophagy-related risk score had better ability of prediction (**Figure 9B**). Risk factor association diagram showed as risk score rose, mortality grew, and survival time reduced (**Figure 9C**). Calibration of nomogram of total cohort perfectly consisted with results in train cohort and validation cohort (**Figures 9D,E**). Moreover, K-M curves indicated patients in the low risk group had longer OS and DFS than in the high risk group. In addition, to eliminate influences from clinical characteristics, we grouped all the patients by age, gender, and stage and proved significant difference in survival time for patients with different levels of risk score (**Figure 10**). Thus, although not as efficient as stage, risk score could still be a reliable index to predict prognosis of ccRCC patients without concern about clinical characteristics.

In order to figure out different signatures underlying two risk groups, we further performed GSEA and WGCNA analysis, respectively. Dividing patients into high and low risk groups, a total of 18,569 genes were analyzed through GSEA. We regarded KEGG pathways with $|NES| \geq 1.5$, FDR < 0.25 as significant, and the results indicated the high risk group

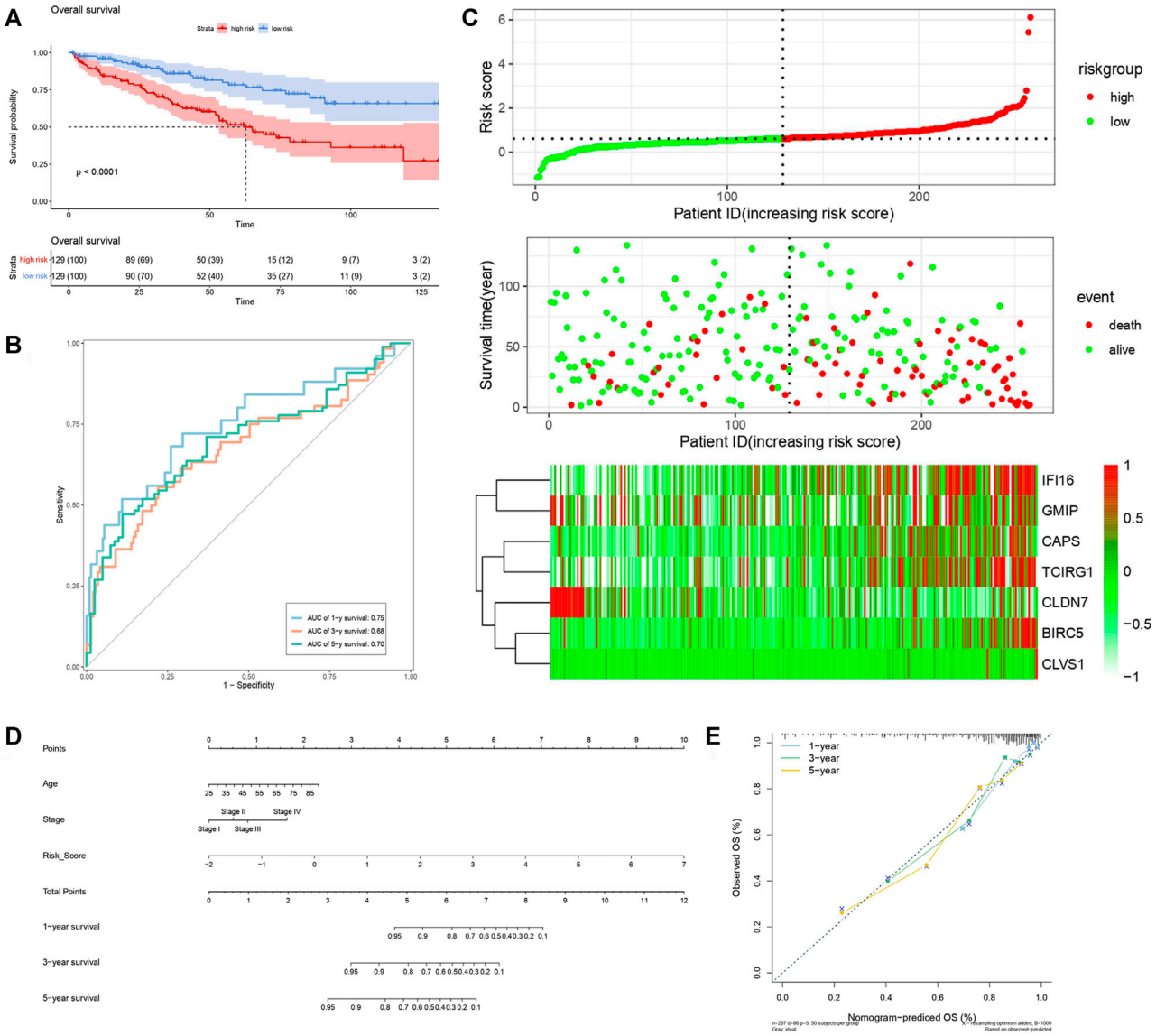

FIGURE 7 | Verification of risk score as predicting factor of ccRCC patients in train cohort. **(A)** Survival curve of different groups. Red line represented patients in high risk group, blue line represented patients in low risk group. **(B)** Time-dependent ROC curve of 1-, 3-, 5-years. **(C)** Risk factor association diagram. Red nodes in the upper and middle graph represented patients with high risk score, green nodes represented patients with low risk score. Cells in the gram below represented each patients and color of cells indicated up- or down-regulation of genes. **(D)** Nomogram included age, stage, and 7 gene-based risk score. **(E)** Calibration curve was depicted for verification accurancy of nomogram predicting 1-, 3-, 5-years OS rate.

was connected with cytokine-cytokine receptor interaction, cytosolic DNA sensing pathway, glycosaminoglycan biosynthesis chondroitin sulfate, JAK-STAT signaling pathway, NOD-like receptor signaling pathway, RIG-I like receptor signaling pathway, RNA degradation, spliceosome, and viral myocarditis (**Figure 11A**). Meanwhile, butanoate metabolism, citrate cycle tca cycle, fatty acid metabolism, glycine serine and threonine metabolism, glycolysis gluconeogenesis, peroxisome, propanoate metabolism, proximal tubule bicarbonate reclamation, pyruvate metabolism and valine leucine, and isoleucine degradation were found enriched in the low risk group (**Figure 11B**). Genes with top 5,000 median absolute deviation were analyzed in WGCNA, and soft threshold was set as 9 (**Figure 12A**). Correlation between risk score and modules was calculated, as the figure shows that a black module was related with high risk score closely (**Figures 12B,C**). As a result, we analyzed the relation among genes in black module and high risk score and found they were positively related (cor = 0.52, $p < 0.001$) (**Figure 12D**). Finally, we performed functional enrichment analysis in Metascape, with the threshold of min overlap = 3, p value cutoff = 0.01, and min enrichment = 1.5,

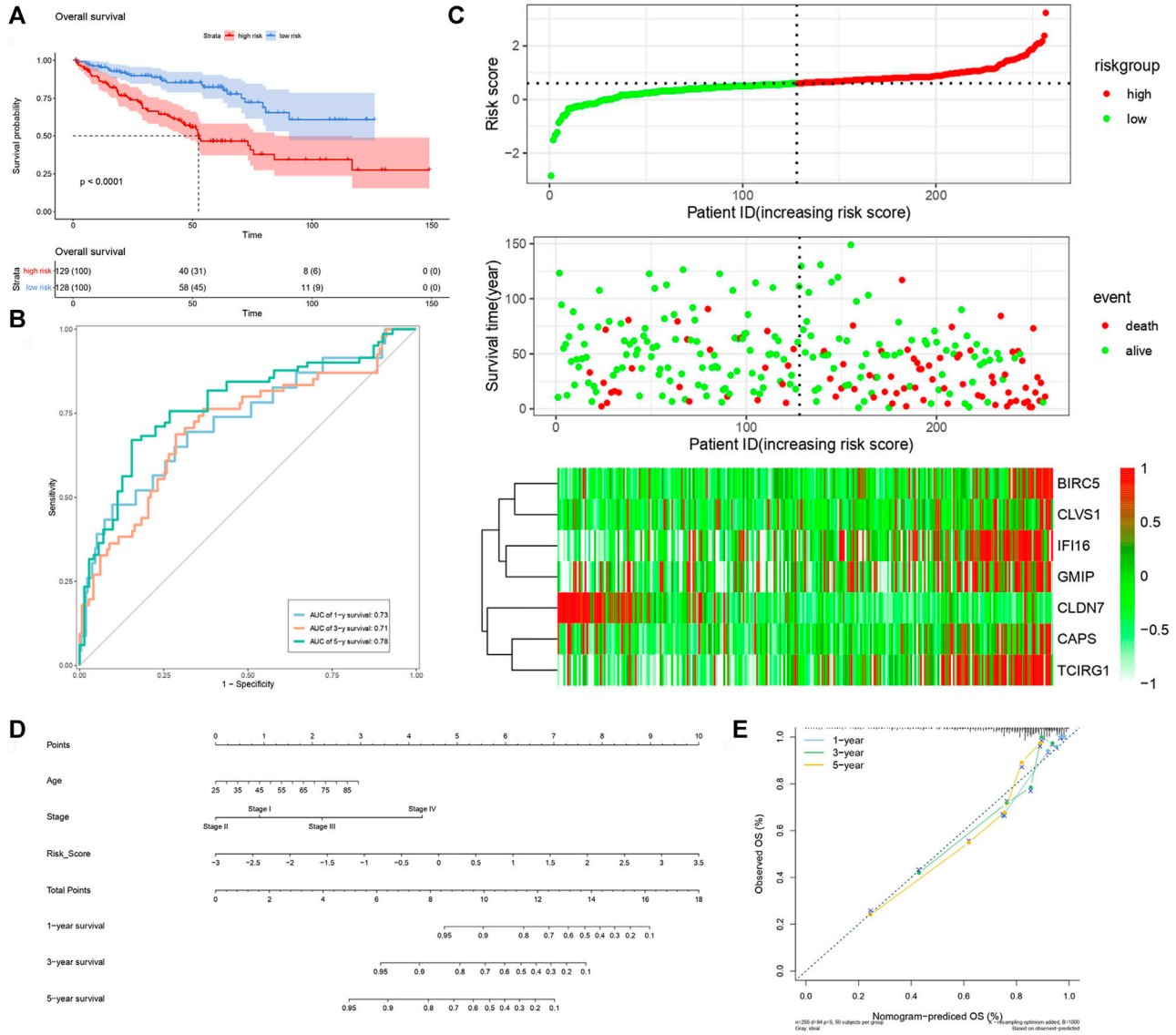

FIGURE 8 | Verification of risk score as predicting factor of ccRCC patients in validation cohort. **(A)** Survival curve of different groups. Red line represented patients in high risk group, blue line represented patients in low risk group. **(B)** Time-dependent ROC curve of 1-, 3-, 5-years. **(C)** Risk factor association diagram. Red nodes in the upper and middle graph represented patients with high risk score, green nodes represented patients with low risk score. Cells in the gram below represented each patient, and color of cells indicated up- or down-regulation of genes. **(D)** Nomogram included age, stage, and 7 gene-based risk score. **(E)** Calibration curve was depicted for verification accurancy of nomogram predicting 1-, 3-, 5-years OS rate.

and 20 pathways were found enriched in black module (**Figure 12E**). MCODE was selected with physical score > 0.132, min network size = 3, max network size = 500, and databases as physical core (**Figure 12F**).

Immune Cell Infiltration

Concerning the significance of immunity on tumorigenesis and progression, we analyzed immune cell infiltration of these seven DEGs on TIMER (**Figure 13**). We found expression of BIRC5 was positively associated with B Cell, $CD8^+$ T Cell, Marcophage, Neutrophill and Dendritic Cell. CAPS was negatively associated with B Cell, $CD8^+$ T Cell, Marcophage and Dendritic Cell and positively associated with $CD4^+$ T Cell. CLDN7 was found positively related with B Cell, and CLVS1 was found negatively related with B Cell, Macrophage and Dendritic Cell. High expression levels of GMIP and IFI16 were highly related with B Cell, $CD8^+$ T Cell, $CD4^+$ T Cell, Marcophage, Neutrophill and Dendritic Cell. Similarly, TCIRG1 was found positively related with B Cell, $CD8^+$ T Cell, $CD4^+$ T Cell, Neutrophill and Dendritic Cell.

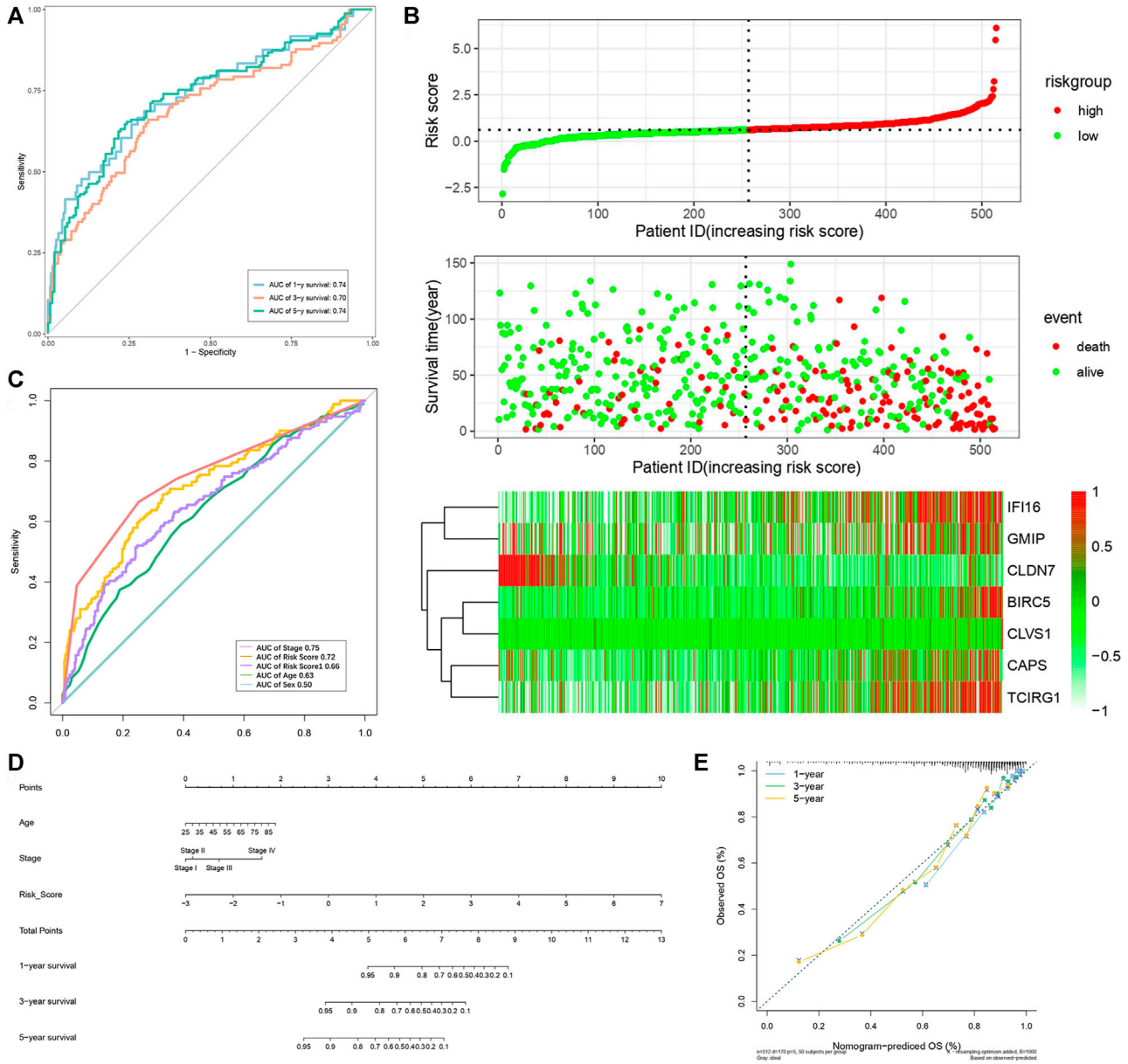

FIGURE 9 | Verification of risk score as predicting factor of ccRCC patients in total cohort. **(A)** Time-dependent ROC curve of 1-, 3-, 5-year. **(B)** ROC curve of clinical signatures of ccRCC patints. **(C)** Risk factor association diagram. Red nodes in the upper and middle graph represented patients with high risk score, green nodes represented patients with low risk score. Cells in the gram below represented each patients and color of cells indicated up- or down-regulation of genes. **(D)** Nomogram included age, stage, and 7 gene-based risk score. **(E)** Calibration curve was depicted for verification accurancy of nomogram predicting 1-, 3-,5-years OS rate.

We performed "Cibersort" on R studio to assess immune cell infiltration level in ccRCC patients and normal patients, macrophages M2, T cells CD8, macrophages M1, T cells gamma delta, T Cells regulatory, macrophages M0, NK cells resting, T cells CD4 activated and T cells follicular are significantly higher infiltrated in ccRCC patients than in normal patients (**Figure 14A**). Meanwhile, comparing immune cell infiltration level between high risk group and low risk group, we found macrophages M2, monocytes, mast cells resting, and neutrophils were significantly lower infiltrated in high risk group, NK cell resting, T cells CD4 memory activated, and T cells regulatory and T cells follicular helper were infiltrated significantly high in high risk group (**Figure 14B**). It seemed that immune cells highly enriched in high risk group and ccRCC patients were not typical immune cells that promoted tumorigenesis and progression. As a

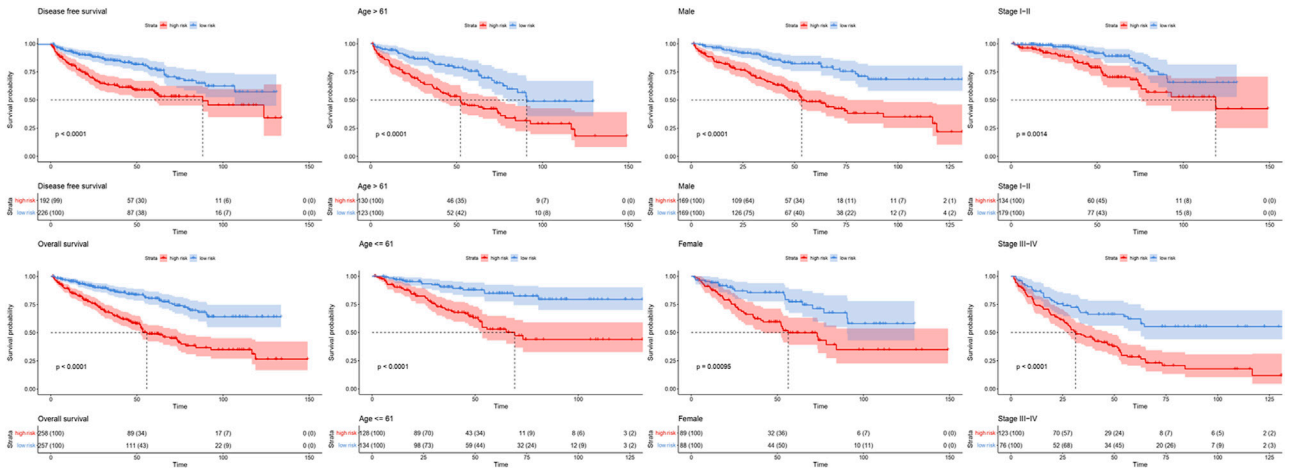

FIGURE 10 | Survival curve of high and low risk groups in different cohorts sorted by clinical signatures. Red line represented patients in high risk group, blue line represented patients in low risk group.

result, to find out immune cells that had significant effects on ccRCC, we performed univariate and multivariate cox regression on immune cell infiltration and the results revealed that in TCGA-KIRC, mast cells resting suppressed tumor progression while macrophages M0, T cells CD4 memory activated, and T cells regulation were risk factors of tumor progression (**Supplementary Material S3**).

DISCUSSION

Autophagy is an indispensable biological process which enables cells to self-degrade and recycle intracellular components. It is well-recognized that in the early stage of tumorigenesis, autophagy represses tumorigenesis by its function of stability and inhibiting genome destruction from metabolic stress and immunoreaction (Karantza-Wadsworth et al., 2007; White, 2012). On the other hand, in the late stage autophagy protects tumor cells from stress to improve tumor progression. Studies indicated basic function of autophagy provides cellular metabolites for tumor cells and regulates mitochondrial metabolism to meet the high metabolic requirements of rapid proliferation of tumor cells (White, 2012; Katheder et al., 2017). In addition, autophagy not only modulates transfer-related biological phenotypes such as resistance to anoikis (Coates et al., 2010), but also stimulates TGF-β and EMT process (Li et al., 2013; Papageorgis, 2015; Yeo et al., 2016). Thus it can be seen that autophagy affects on tumorigenesis and tumor progression through multiple approaches. In our study we aim to explore how autophagy affects progression of ccRCC and seek ATGs that can predict the progression of ccRCC.

Through lasso regression analysis and cox regression analysis, we finally identified 7ATGs (BIRC5, CAPS, CLDN7, CLVS1, GMIP, IFI16, and TCIRG1) associated closely with ccRCC prognosis. BIRC5 is a EMT related gene which prevents cell apoptotic through different approaches and participates in cell cycle regulation, and also in cancer cells it regulates autophagy directly (Lin et al., 2020). High expression of BIRC5 was found to indicate poor prognosis in hepatocellular carcinoma (Xu et al., 2021). Also BIRC5 was found related with prognosis of ccRCC and gastric cancer (Yao et al., 2020; Li et al., 2021). CAPS encodes a calcium-binding protein, which may play a role in the regulation of ion transport; research showed that CAPS might indicate tamoxifen resistance in ER positive breast cancer (Johansson et al., 2015). CLDN7 encodes a member of claudin family and were found expressed in several malignancies such as prostate cancer, lung cancer, urinary tumors, and so on. Overexpression of CLDN7 is closely related to lymph node metastasis (Wu et al., 2018). In addition, CLDN7 was found upregulated in mouse pancreas exposed to caerylein for 12 h and its function concerned tight junction formation, while destruction of tight might be closely related with autophagy's detrimental effects (Nakada et al., 2010; Wang S. et al., 2020). So far CLVS1 wasn't found significant in tumorigenesis and progression, but research found it is involved in lysosome maturation and associated with psychiatric and steroid-sensitive nephrotic syndrome (Corponi et al., 2019; Lane et al., 2021). GMIP is a protein coding gene that encodes ARHGAP family of Rho/Rac/Cdc42-like GTPase activating proteins. In lung cancer, overexpression of GMIP was associated with longer survival; in the null mice model with a xenografted tumor of A549 cells, GMIP treatment has once been proved to induce autophagy and reduce tumor growth (Hsin et al., 2011; Amaar and Reeves, 2020). IFI16 modulates p53 function and inhibits cell growth in the Ras/Raf signaling pathway. It can be induced by AMPK/p53 pathway and the induced levels of IFI16 were associated with the induction of autophagy (Duan et al., 2011). TCIRG1 is involved in

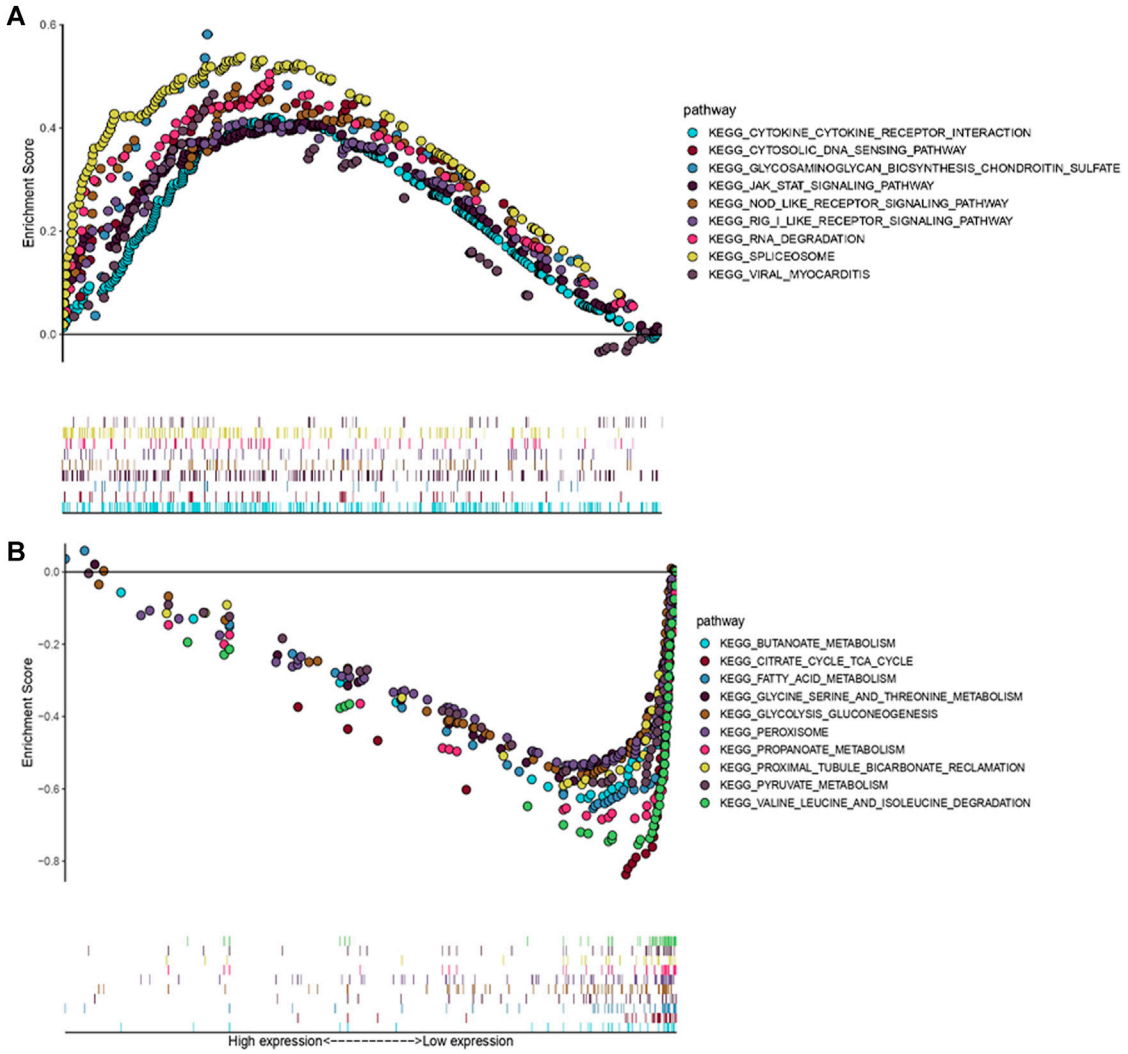

FIGURE 11 | GSEA associated with risk score [Gene matrix: c2.cp.kehh.v7.symbols.gmt (Curated), Number of permutations: 1,000, Permutation type: phenotype]. **(A)** KEGG pathways enriched in high risk group. **(B)** KEGG pathways enriched in low risk group.

autophagosome assembly, and it is usually found relevant with osteopetrosis (Belaid et al., 2013; Chavez-Guitron et al., 2018).

We first analyzed expression levels in ccRCC of these seven genes, and in mRNA level, results from ONCOMINE showed except GMIP, IFI16, and TCIRG1 were overexpressed in ccRCC patients, while analysis based on TCGA indicated expression levels of CAPS and CLDN7 were significantly low in ccRCC patients, and BIRC5, CLVS1, GMIP, IFI16, and TCIRG1 were highly expressed in ccRCC patients. In protein level, we found, IFI16, and TCIRG1 were highly expressed in ccRCC kidney tissue, and others were lower in ccRCC kidney tissue or not detected. Through STRING we found the top 50 related genes of these 7 genes and performed functional enrichment and pathway analysis, and results revealed they were closely related with autophagy process, tumorigenesis, and involved in biological processes of tumor progression. Results from GSEA analyzing functional enrichment of high and low risk group indicated that JAK-STAT signaling pathway, NOD-like receptor signaling pathway, and RIG-I-like receptor signaling pathway might be the cause of poorer prognosis of ccRCC patients (Smith et al., 2018; Mey et al., 2019; Zhou et al., 2020). Actually, these pathways were found directly or indirectly related with autophagy (Chan and Gack, 2015; Velloso et al., 2019; Billah et al., 2020). Further, we explored a module co-expressed with high risk in WGCNA, and genes from the most significant module were found quite closely connected with high risk, too. Functional enrichment analysis was performed in Metascape then.

According to the formula we construct a new ATGs-related risk score in train cohort, and we found the high

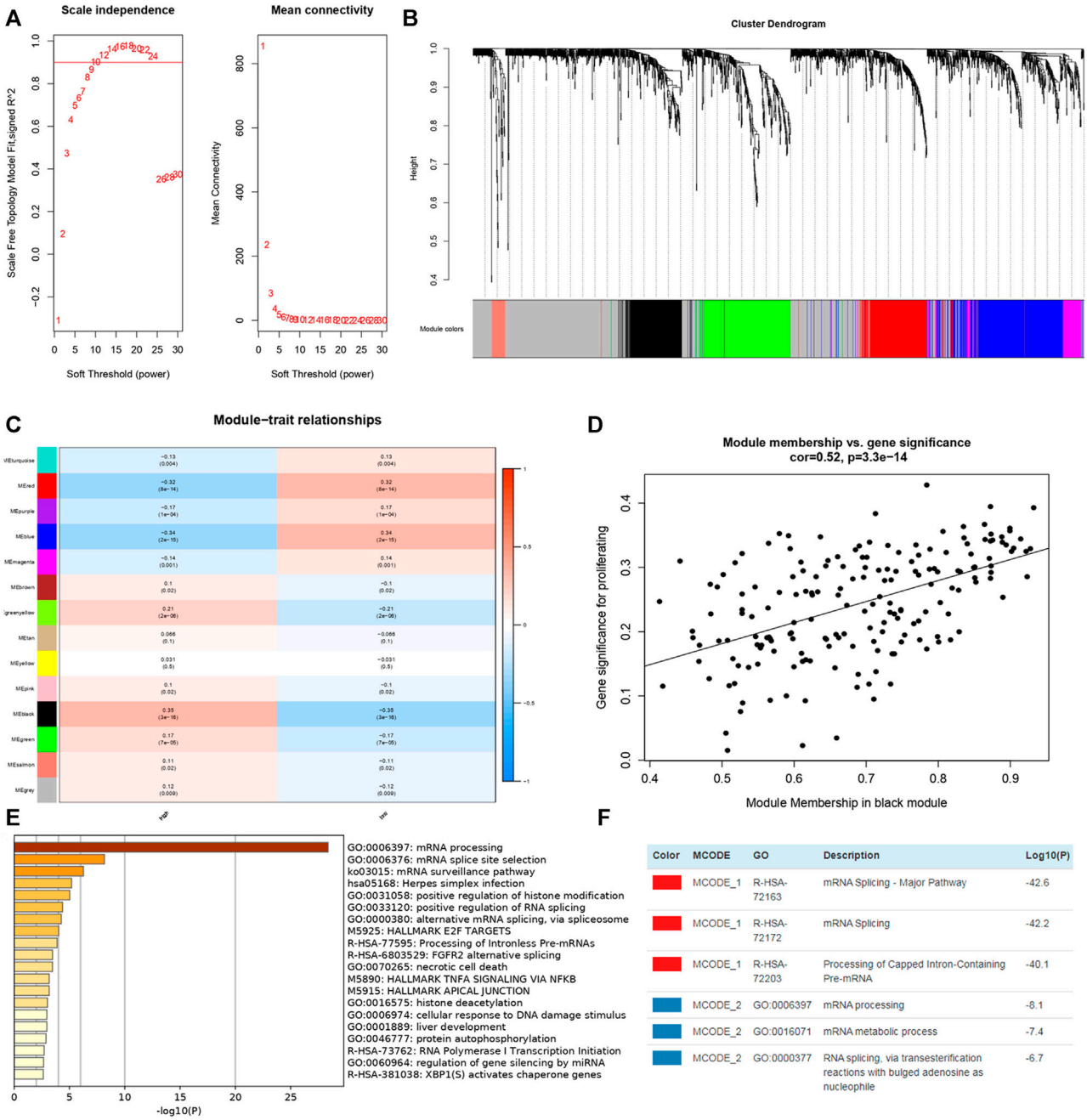

FIGURE 12 | WGCNA in TCGA-KIRC dataset and enrichment analysis of genes in black module from Metascape. **(A)** Soft thresholding filtering. **(B)** Module screening, 14 modules have been identified. **(C)** The relationship between 14 module and different risk groups, the red cells indicated the module was positively related with high/low risk group while blue cells represented negative relationship. **(D)** Correlation plot between genes in black module and high risk score. **(E)** Top 20 pathways enriched in black module. **(F)** MCODEs of black module.

risk score was related to poor prognosis of ccRCC patients. Cox regression analysis indicated together with age of diagnosis and stage, risk score was an independent risk factor of prognosis of ccRCC. All the results above were verified by similar analysis in validation cohort and total cohort. Concerned about the importance of immune response in tumorigenesis and progression, we further explored immune infiltration of 7 genes, and all of them were related with immune cells in varying degrees. The high risk group was highly infiltrated with NK cell resting, T cells CD4 memory activated, T cells regulatory, and T cells follicular helper.

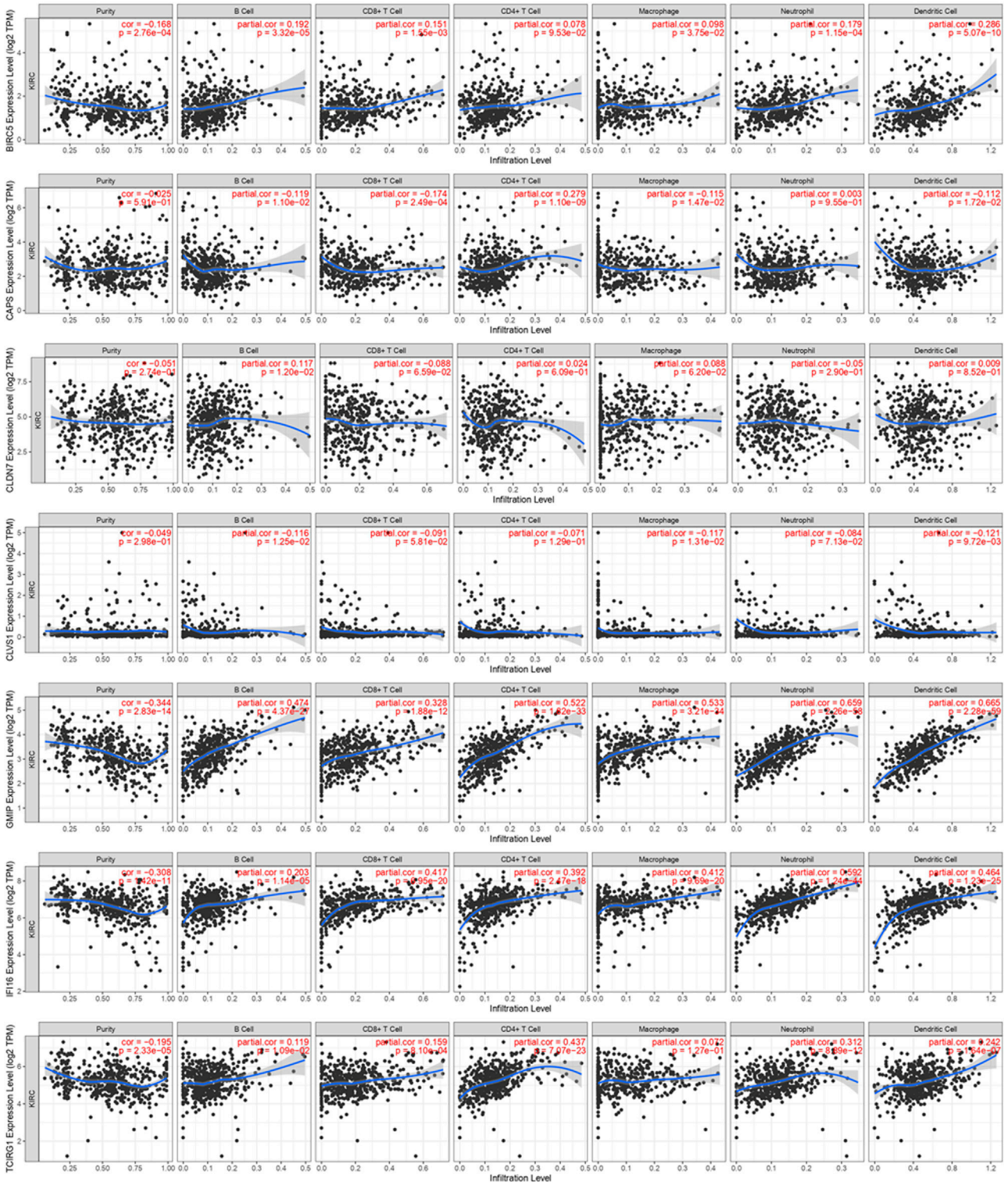

FIGURE 13 | Immune cell infiltration of 7 DEGs in TCGA-KIRC from TIMER.

Although all results above demonstrated that the risk score signatures constructed by 7 DEGs contribute to the progression of ccRCC patients and functional enrichment related with risk score demonstrated risk score had an association with autophagy, whether autophagy itself in our study took the responsibility of tumor progression remained

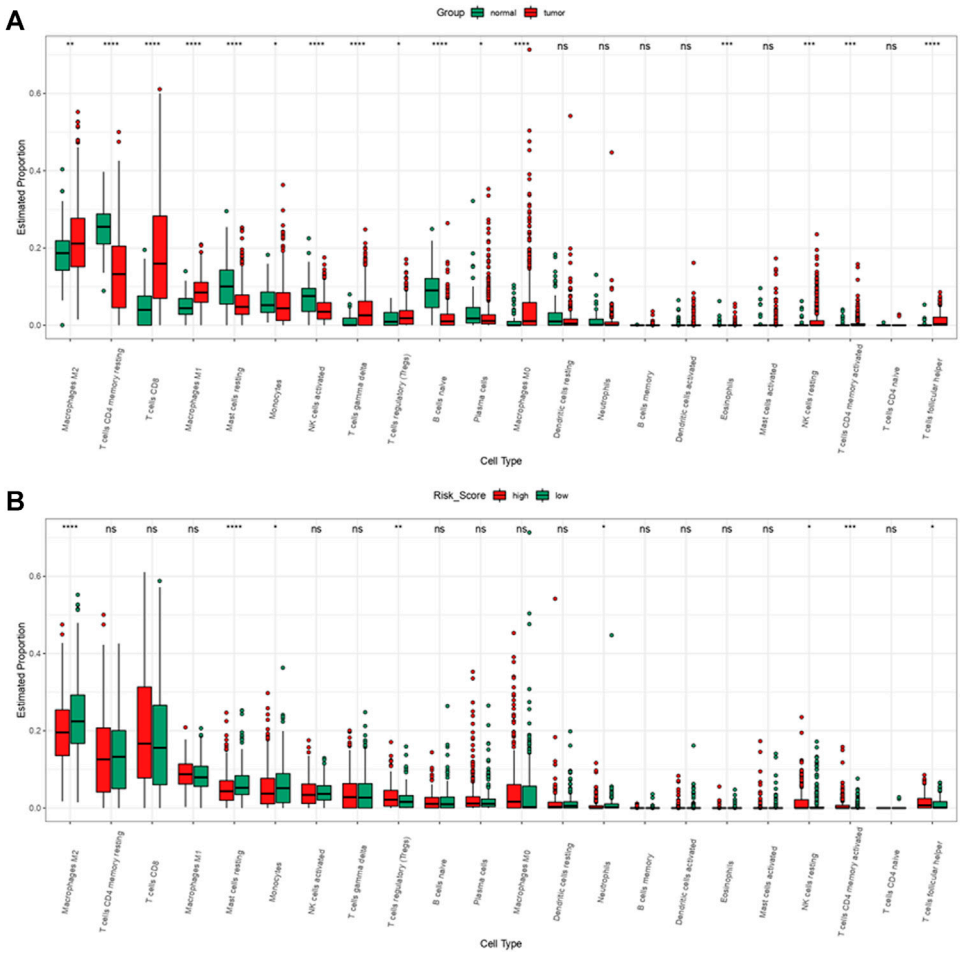

FIGURE 14 | Immune cell infiltration of ccRCC patients from TCGA-KIRC dataset. **(A)** Immune cell infiltration levels in ccRCC and normal tissue. **(B)** Immune cell infiltration levels of high and low risk group.

unknown. Further experiments *in vivo* and *in vitro* are still needed to prove practicality and feasibility of the new risk score.

CONCLUSION

A serious of analysis based on autophagy and DEGs was performed, and it turned out that a new risk score constructed by 7 ATGs (BIRC5, CAPS, CLDN7, CLVS1, GMIP, IFI16, and TCIRG1) could be a potential predictive signature of ccRCC patients. The relevant findings in this study still need mechanism and molecular verification in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/ and remaining datasets are available in **Supplementary Materials**.

AUTHOR CONTRIBUTIONS

TC designed the study and revised the article. MXH, MRL, and YBG performed the statistical analysis, interpreted the data, and was main contributors in writing the manuscript. ZYW, JHT, HBZ, HB, and MG performed the statisitical analyses. All authors have read and approved the manuscript.

ACKNOWLEDGMENTS

The authors appreciate study investigators and staff who participated in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.820154/full#supplementary-material

REFERENCES

Amaar, Y. G., and Reeves, M. E. (2020). The Impact of the RASSF1C and PIWIL1 on DNA Methylation: the Identification of GMIP as a Tumor Suppressor. *Oncotarget* 11 (45), 4082–4092. doi:10.18632/oncotarget.27795

Belaid, A., Cerezo, M., Chargui, A., Corcelle–Termeau, E., Pedeutour, F., Giuliano, S., et al. (2013). Autophagy Plays a Critical Role in the Degradation of Active RHOA, the Control of Cell Cytokinesis, and Genomic Stability. *Cancer Res.* 73 (14), 4311–4322. doi:10.1158/0008-5472.CAN-12-4142

Billah, M., Ridiandries, A., Allahwala, U. K., Mudaliar, H., Dona, A., Hunyor, S., et al. (2020). Remote Ischemic Preconditioning Induces Cardioprotective Autophagy and Signals through the IL-6-Dependent JAK-STAT Pathway. *Int. J. Mol. Sci.* 21 (5), 1692. doi:10.3390/ijms21051692

Cancer Genome Atlas ResearchWeinstein, J. N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., et al. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Capitanio, U., Bensalah, K., Bex, A., Boorjian, S. A., Bray, F., Coleman, J., et al. (2019). Epidemiology of Renal Cell Carcinoma. *Eur. Urol.* 75 (1), 74–84. doi:10.1016/j.eururo.2018.08.036

Chan, Y. K., and Gack, M. U. (2015). RIG-I-like Receptor Regulation in Virus Infection and Immunity. *Curr. Opin. Virol.* 12, 7–14. doi:10.1016/j.coviro.2015.01.004

Chávez-Güitrón, L. E., Cerón-Torres, T., Sobacchi, C., Ochoa-Ruiz, E., and Villegas-Huesca, S. (2019). Autosomal Recessive Osteopetrosis Type I: Description of Pathogenic Variant of TCIRG1 Gene. *Bmhim* 75 (4), 255–259. doi:10.24875/BMHIM.M18000028

Coates, J. M., Galante, J. M., and Bold, R. J. (2010). Cancer Therapy beyond Apoptosis: Autophagy and Anoikis as Mechanisms of Cell Death. *J. Surg. Res.* 164 (2), 301–308. doi:10.1016/j.jss.2009.07.011

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

Corponi, F., Bonassi, S., Vieta, E., Albani, D., Frustaci, A., Ducci, G., et al. (2019). Genetic Basis of Psychopathological Dimensions Shared between Schizophrenia and Bipolar Disorder. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* 89, 23–29. doi:10.1016/j.pnpbp.2018.08.023

Duan, X., Ponomareva, L., Veeranki, S., and Choubey, D. (2011). IFI16 Induction by Glucose Restriction in Human Fibroblasts Contributes to Autophagy through Activation of the ATM/AMPK/p53 Pathway. *PLoS One* 6 (5), e19532. doi:10.1371/journal.pone.0019532

Gatica, D., Chiong, M., Lavandero, S., and Klionsky, D. J. (2021). The Role of Autophagy in Cardiovascular Pathology. *Cardiovasc. Res.* [Epub ahead of print], cvab158. doi:10.1093/cvr/cvab158

Glick, D., Barth, S., and Macleod, K. F. (2010). Autophagy: Cellular and Molecular Mechanisms. *J. Pathol.* 221 (1), 3–12. doi:10.1002/path.2697

Hsin, I.-L., Ou, C.-C., Wu, T.-C., Jan, M.-S., Wu, M.-F., Chiu, L.-Y., et al. (2011). GMI, an Immunomodulatory Protein fromGanoderma Microsporum, Induces Autophagy in Non-small Cell Lung Cancer Cells. *Autophagy* 7 (8), 873–882. doi:10.4161/auto.7.8.15698

Johansson, H. J., Sanchez, B. C., Forshed, J., Stål, O., Fohlin, H., Lewensohn, R., et al. (2015). Proteomics Profiling Identify CAPS as a Potential Predictive Marker of Tamoxifen Resistance in Estrogen Receptor Positive Breast Cancer. *Clin. Proteom* 12 (1), 8. doi:10.1186/s12014-015-9080-y

Karantza-Wadsworth, V., Patel, S., Kravchuk, O., Chen, G., Mathew, R., Jin, S., et al. (2007). Autophagy Mitigates Metabolic Stress and Genome Damage in Mammary Tumorigenesis. *Genes Dev.* 21 (13), 1621–1635. doi:10.1101/gad.1565707

Katheder, N. S., Khezri, R., O'Farrell, F., Schultz, S. W., Jain, A., Rahman, M. M., et al. (2017). Microenvironmental Autophagy Promotes Tumour Growth. *Nature* 541 (7637), 417–420. doi:10.1038/nature20815

Khattak, M., and Larkin, J. (2014). Sequential Therapy with Targeted Agents in Metastatic Renal Cell Carcinoma: beyond Second-Line and Overcoming Drug Resistance. *World J. Urol.* 32 (1), 19–29. doi:10.1007/s00345-012-1013-z

Kimmelman, A. C. (2011). The Dynamic Nature of Autophagy in Cancer. *Genes Dev.* 25 (19), 1999–2010. doi:10.1101/gad.17558811

Lane, B. M., Chryst-Stangl, M., Wu, G., Shalaby, M., El Desoky, S., Middleton, C. C., et al. (2021). Steroid-sensitive Nephrotic Syndrome Candidate Gene CLVS1 Regulates Podocyte Oxidative Stress and Endocytosis. *JCI Insight* [Epub ahead of print], e152102. doi:10.1172/jci.insight.152102

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

Li, F., Aljahdali, I. A. M., Zhang, R., Nastiuk, K. L., Krolewski, J. J., and Ling, X. (2021). Kidney Cancer Biomarkers and Targets for Therapeutics: Survivin (BIRC5), XIAP, MCL-1, HIF1α, HIF2α, NRF2, MDM2, MDM4, P53, KRAS and AKT in Renal Cell Carcinoma. *J. Exp. Clin. Cancer Res.* 40 (1), 254. doi:10.1186/s13046-021-02026-1

Li, J., Yang, B., Zhou, Q., Wu, Y., Shang, D., Guo, Y., et al. (2013). Autophagy Promotes Hepatocellular Carcinoma Cell Invasion through Activation of Epithelial-Mesenchymal Transition. *Carcinogenesis* 34 (6), 1343–1351. doi:10.1093/carcin/bgt063

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48 (W1), W509–W514. doi:10.1093/nar/gkaa407

Lin, T.-Y., Chan, H.-H., Chen, S.-H., Sarvagalla, S., Chen, P.-S., Coumar, M. S., et al. (2020). BIRC5/Survivin Is a Novel ATG12-ATG5 Conjugate Interactor and an Autophagy-Induced DNA Damage Suppressor in Human Cancer and Mouse Embryonic Fibroblast Cells. *Autophagy* 16 (7), 1296–1313. doi:10.1080/15548627.2019.1671643

Lizama, B. N., and Chu, C. T. (2021). Neuronal Autophagy and Mitophagy in Parkinson's Disease. *Mol. Aspects Med.* 82, 100972. doi:10.1016/j.mam.2021.100972

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

Mey, L., Jung, M., Roos, F., Blaheta, R., Hegele, A., Kinscherf, R., et al. (2019). NOD1 and NOD2 of the Innate Immune System Is Differently Expressed in Human clear Cell Renal Cell Carcinoma, Corresponding Healthy Renal Tissue, its Vasculature and Primary Isolated Renal Tubular Epithelial Cells. *J. Cancer Res. Clin. Oncol.* 145 (6), 1405–1416. doi:10.1007/s00432-019-02901-7

Mizushima, N., and Levine, B. (2020). Autophagy in Human Diseases. *N. Engl. J. Med.* 383 (16), 1564–1576. doi:10.1056/NEJMra2022774

Nakada, S., Tsuneyama, K., Kato, I., Tabuchi, Y., Takasaki, I., Furusawa, Y., et al. (2010). Identification of Candidate Genes Involved in Endogenous protection Mechanisms against Acute Pancreatitis in Mice. *Biochem. Biophysical Res. Commun.* 391 (3), 1342–1347. doi:10.1016/j.bbrc.2009.12.047

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Papageorgis, P. (2015). TGFβSignaling in Tumor Initiation, Epithelial-To-Mesenchymal Transition, and Metastasis. *J. Oncol.* 2015, 1–15. doi:10.1155/2015/587193

Rangel, M., Kong, J., Bhatt, V., Khayati, K., and Guo, J. Y. (2021). Autophagy and Tumorigenesis. *FEBS J.* [Epub ahead of print] doi:10.1111/febs.16125

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a Cancer Microarray Database and Integrated Data-Mining Platform. *Neoplasia* 6 (1), 1–6. doi:10.1016/s1476-5586(04)80047-2

Smith, C. C., Beckermann, K. E., Bortone, D. S., De Cubas, A. A., Bixby, L. M., Lee, S. J., et al. (2018). Endogenous Retroviral Signatures Predict Immunotherapy Response in clear Cell Renal Cell Carcinoma. *J. Clin. Invest.* 128 (11), 4804–4820. doi:10.1172/JCI121476

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A Pathology Atlas of the Human Cancer Transcriptome. *Science* 357 (6352), eaan2507. doi:10.1126/science.aan2507

Velloso, F. J., Trombetta-Lima, M., Anschau, V., Sogayar, M. C., and Correa, R. G. (2019). NOD-like Receptors: Major Players (And Targets) in the Interface between Innate Immunity and Cancer. *Biosci. Rep.* 39 (4), BSR20181709. doi:10.1042/BSR20181709

Wang, S., Xu, J., Xi, J., Grothusen, J. R., and Liu, R. (2020). Autophagy Inhibition Preserves Tight Junction of Human Cerebral Microvascular Endothelium under Oxygen Glucose Deprivation. *Curr. Neurovasc. Res.* 17 (5), 644–651. doi:10.2174/1567202617999201103200705

Wang, Y., Zhao, W., Xiao, Z., Guan, G., Liu, X., and Zhuang, M. (2020). A Risk Signature with Four Autophagy-related Genes for Predicting Survival of Glioblastoma Multiforme. *J. Cel Mol Med* 24 (7), 3807–3821. doi:10.1111/jcmm.14938

White, E. (2012). Deconvoluting the Context-dependent Role for Autophagy in Cancer. *Nat. Rev. Cancer* 12 (6), 401–410. doi:10.1038/nrc3262

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *The Innovation* 2 (3), 100141. doi:10.1016/j.xinn.2021.100141

Wu, Z., Shi, J., Song, Y., Zhao, J., Sun, J., Chen, X., et al. (2018). Claudin-7 (CLDN7) Is Overexpressed in Gastric Cancer and Promotes Gastric Cancer Cell Proliferation, Invasion and Maintains Mesenchymal State. *Neoplasma.* 65 (3), 349–359. doi:10.4149/neo_2018_170320N200

Xu, R., Lin, L., Zhang, B., Wang, J., Zhao, F., Liu, X., et al. (2021). Identification of Prognostic Markers for Hepatocellular Carcinoma Based on the Epithelial-Mesenchymal Transition-Related Gene BIRC5. *BMC Cancer* 21 (1), 687. doi:10.1186/s12885-021-08390-7

Yao, Y., Liu, Z., Cao, Y., Guo, H., Jiang, B., Deng, J., et al. (2020). Downregulation of TRIM27 Suppresses Gastric Cancer Cell Proliferation via Inhibition of the Hippo-BIRC5 Pathway. *Pathol. - Res. Pract.* 216 (9), 153048. doi:10.1016/j.prp.2020.153048

Yeo, S. K., Wen, J., Chen, S., and Guan, J.-L. (2016). Autophagy Differentially Regulates Distinct Breast Cancer Stem-like Cells in Murine Models via EGFR/Stat3 and Tgfβ/Smad Signaling. *Cancer Res.* 76 (11), 3397–3410. doi:10.1158/0008-5472.CAN-15-2946

Zhang, Z., Yang, X., Song, Y.-Q., and Tu, J. (2021). Autophagy in Alzheimer's Disease Pathogenesis: Therapeutic Potential and Future Perspectives. *Ageing Res. Rev.* 72, 101464. doi:10.1016/j.arr.2021.101464

Zhou, L., Li, Y., Li, Z., and Huang, Q. (2020). Mining Therapeutic and Prognostic Significance of STATs in Renal Cell Carcinoma with Bioinformatics Analysis. *Genomics* 112 (6), 4100–4114. doi:10.1016/j.ygeno.2020.06.032

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10 (1), 1523. doi:10.1038/s41467-019-09234-6

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

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

Copyright © 2022 He, Li, Guan, Wan, Tian, Xu, Zhou, Gao, Bi and Chong. 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.

frontiers | Frontiers in Genetics

CORRECTION
published: 04 July 2022
doi: 10.3389/fgene.2022.904512

Corrigendum: A New Prognostic Risk Score: Based on the Analysis of Autophagy-Related Genes and Renal Cell Carcinoma

Minxin He[1,2], *Mingrui Li*[1,2], *Yibing Guan*[1,2], *Ziyan Wan*[1,2], *Juanhua Tian*[1,2], *Fangshi Xu*[1,2], *Haibin Zhou*[1,2], *Mei Gao*[1], *Hang Bi*[1,2] *and Tie Chong*[1]*

[1]*Department of Urology, The Second Affiliated Hospital, School of Medicine, Xi'an Jiaotong University, Xi'an, China,* [2]*School of Medicine, Xi'an Jiaotong University, Xi'an, China*

OPEN ACCESS

Edited and reviewed by:
Farhad Maleki,
McGill University, Canada

***Correspondence:**
Tie Chong
chongtie@126.com

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

Received: *25 March 2022*
Accepted: *23 May 2022*
Published: *04 July 2022*

Citation:
He M, Li M, Guan Y, Wan Z, Tian J, Xu F, Zhou H, Gao M, Bi H and Chong T (2022) Corrigendum: A New Prognostic Risk Score: Based on the Analysis of Autophagy-Related Genes and Renal Cell Carcinoma. Front. Genet. 13:904512. doi: 10.3389/fgene.2022.904512

Keywords: risk score, prognosis, bioinformatics analysis, renal cell carcinoma, autophagy

A Corrigendum on

A New Prognostic Risk Score: Based on the Analysis of Autophagy-Related Genes and Renal Cell Carcinoma

by He, M, Li, M, Guan, Y, Wan, Z, Tian, J, Xu, F, Zhou, H, Gao, M, Bi, H and Chong, T (2022). Front. Genet. 12:820154. doi: 10.3389/fgene.2021.820154

In the original article, there was a mistake in the legend for **Figure 1A**. The threshold of DEGs was wrongly depicted in the legend. The correct legend is presented as follows:

"(A) Enhanced volcano plot of DEGs when comparing ccRCC with normal tissue. Red nodes represented genes with |log 2 FC| ≥ 1 and adjusted $p < 0.05$, blue nodes represented genes with adjusted $p < 0.05$ only, and gray nodes represented genes that were neither eligible in conditions of adjusted p-value nor |log 2 FC|."

In the original article, there was a mistake in **Figure 1** as published. The threshold of DEGs was wrongly set when drawing enhanced volcano plots the corrected **Figure 1** is included here.

In the original article, the method of correlation analysis was wrongly typed as "pearson" in "Correlation analysis of risk score and other clinical signatures was performed by the method of "pearson"." A correction has been made to **Materials and Methods, Construction of Risk Score**, Paragraph 2:

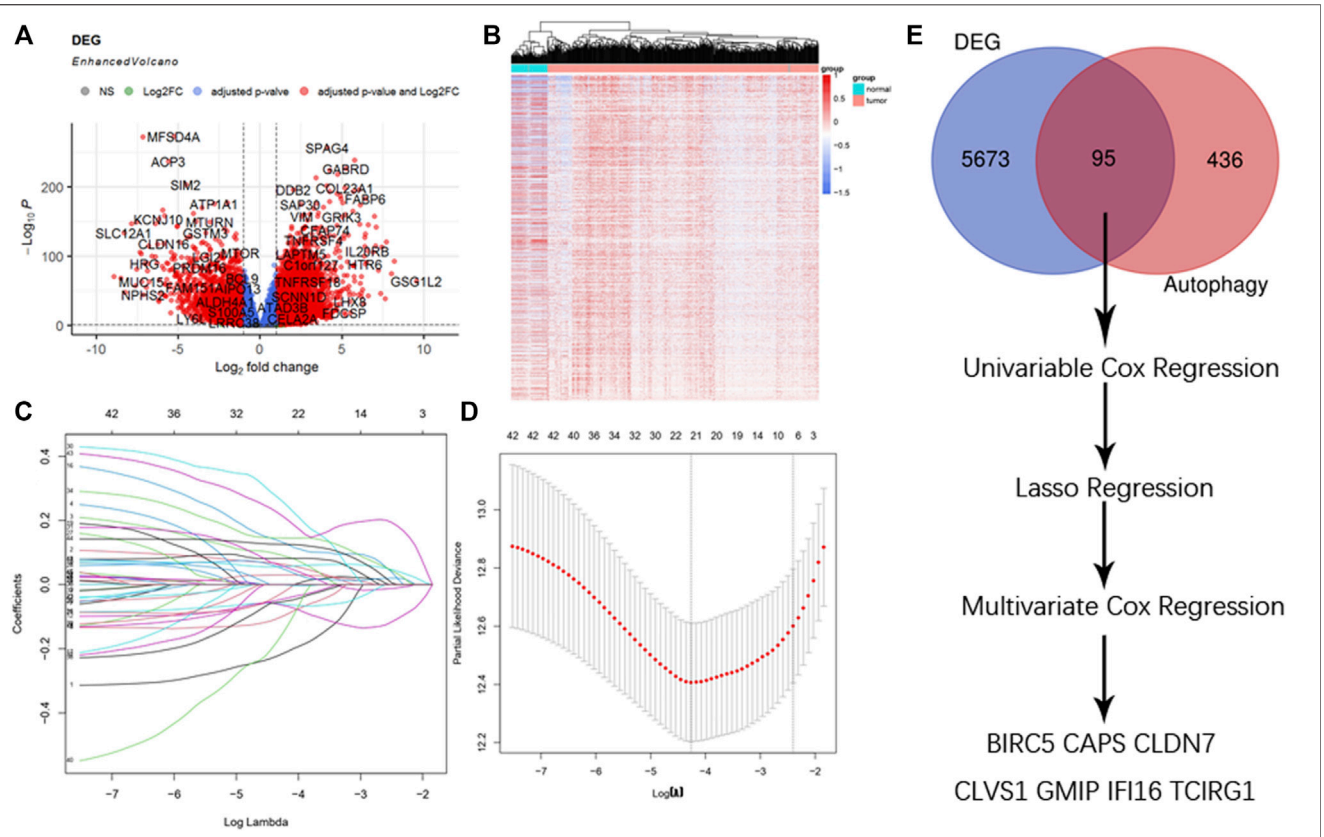

FIGURE 1 | Selection of DEGs. **(A)** Enhanced volcano plot of DEGs when comparing ccRCC with normal tissue. Red nodes represented genes with ||log 2 FC| ≥ 1 and adjusted $p < 0.05$, blue nodes represented genes with adjusted $p < 0.05$ only, and grey nodes represented genes that were neither eligible in conditions of adjusted p value nor ||log 2 FC|. **(B)** Heatmap of DEGs in ccRCC. **(C)** Lasso coefficients profiles of 95 genes significant in univariate cox regression. **(D)** Lasso regression obtained 21 prognostic genes using minimum lambda value. **(E)** Selecting procedure of DEGs, venn gram showed 95 genes in the intersection of 5,768 DEGs and 531 ATGs. These genes then underwent univariate cox regression, lasso cox regression, and multivariate cox regression, and finally 7 DEGs were selected to construct the risk score formula.

The sentence "Correlation analysis of risk score and other clinical signatures was performed by the method of "pearson"." should be corrected as "Correlation analysis of risk score and other clinical signatures was performed by the method of "Spearman"."

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

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

Copyright © 2022 He, Li, Guan, Wan, Tian, Xu, Zhou, Gao, Bi and Chong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership