APPLICATION OF SYSTEMS BIOLOGY IN MOLECULAR CHARACTERIZATION AND DIAGNOSIS OF CANCER

EDITED BY: Cheng Zhang, Adil Mardinoglu, Yongjun Wei and Peng Zhang
PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

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

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

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

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

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

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

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

ISSN 1664-8714
ISBN 978-2-88971-089-8
DOI 10.3389/978-2-88971-089-8

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

APPLICATION OF SYSTEMS BIOLOGY IN MOLECULAR CHARACTERIZATION AND DIAGNOSIS OF CANCER

Topic Editors:
Cheng Zhang, Royal Institute of Technology, Sweden
Adil Mardinoglu, King's College London, United Kingdom
Yongjun Wei, Zhengzhou University, China
Peng Zhang, University of Maryland, United States

Citation: Zhang, C., Mardinoglu, A., Wei, Y., Zhang, P., eds. (2021). Application of Systems Biology in Molecular Characterization and Diagnosis of Cancer. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-089-8

Table of Contents

05 ***Editorial: Application of Systems Biology in Molecular Characterization and Diagnosis of Cancer***
Cheng Zhang, Yongjun Wei, Adil Mardinoglu and Peng Zhang

07 ***Validation of Gene Profiles for Analysis of Regional Lymphatic Metastases in Head and Neck Squamous Cell Carcinoma***
Zhenrong Hu, Ranran Yang, Li Li, Lu Mao, Shuli Liu, Shichong Qiao, Guoxin Ren and Jingzhou Hu

19 ***Systematical Analysis of the Cancer Genome Atlas Database Reveals* EMCN/MUC15 *Combination as a Prognostic Signature for Gastric Cancer***
Wentao Dai, Jixiang Liu, Bingya Liu, Quanxue Li, Qingqing Sang and Yuan-Yuan Li

28 ***A Qualitative Transcriptional Signature for Predicting Extreme Resistance of ER-Negative Breast Cancer to Paclitaxel, Doxorubicin, and Cyclophosphamide Neoadjuvant Chemotherapy***
Yanhua Chen, Hao Cai, Wannan Chen, Qingzhou Guan, Jun He, Zheng Guo and Jing Li

37 ***Systems Biology Integration and Screening of Reliable Prognostic Markers to Create Synergies in the Control of Lung Cancer Patients***
Aman Chandra Kaushik, Aamir Mehmood, Dong-Qing Wei and Xiaofeng Dai

47 ***Intrinsic Genetic and Transcriptomic Patterns Reflect Tumor Immune Subtypes Facilitating Exploring Possible Combinatory Therapy***
Yong Xu, Daixi Li, Zhenhao Liu, David L. Gibbs, Lu Xie and Guangrong Qin

59 ***XRCC1 Is a Promising Predictive Biomarker and Facilitates Chemo-Resistance in Gallbladder Cancer***
Zhengchun Wu, Xiongying Miao, Yuanfang Zhang, Daiqiang Li, Qiong Zou, Yuan Yuan, Rushi Liu and Zhulin Yang

71 ***Systems Biology of Gastric Cancer: Perspectives on the Omics-Based Diagnosis and Treatment***
Xiao-Jing Shi, Yongjun Wei and Boyang Ji

79 ***Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy***
Pei Zhang, Qin Xia, Liqun Liu, Shouwei Li and Lei Dong

92 ***A Theoretical Approach for Correlating Proteins to Malignant Diseases***
Rasha Elnemr, Mohammed M. Nasef, Passant Elkafrawy, Mahmoud Rafea and Amani Tariq Jamal

102 ***Identification of Tamoxifen-Resistant Breast Cancer Cell Lines and Drug Response Signature***
Qingzhou Guan, Xuekun Song, Zhenzhen Zhang, Yizhi Zhang, Yating Chen and Jing Li

112 ***Biased Influences of Low Tumor Purity on Mutation Detection in Cancer***
Jun Cheng, Jun He, Shanshan Wang, Zhangxiang Zhao, Haidan Yan, Qingzhou Guan, Jing Li, Zheng Guo and Lu Ao

122 *A Novel Prognostic Model of Endometrial Carcinoma Based on Clinical Variables and Oncogenomic Gene Signature*
Fang Deng, Jing Mu, Chiwen Qu, Fang Yang, Xing Liu, Xiaomin Zeng and Xiaoning Peng

frontiers in Molecular Biosciences

EDITORIAL
published: 28 May 2021
doi: 10.3389/fmolb.2021.668146

Editorial: Application of Systems Biology in Molecular Characterization and Diagnosis of Cancer

Cheng Zhang[1,2]*, *Yongjun Wei*[2], *Adil Mardinoglu*[2,3] and *Peng Zhang*[4]

[1] Science for Life Laboratory, KTH - Royal Institute of Technology, Stockholm, Sweden, [2] School of Pharmaceutical Sciences & Key Laboratory of Advanced Drug Preparation Technologies, Ministry of Education, Zhengzhou University, Zhengzhou, China, [3] Faculty of Dentistry, Oral & Craniofacial Sciences, Centre for Host-Microbiome Interactions, King's College London, London, United Kingdom, [4] Department of Surgery, School of Medicine, University of Maryland, College Park, MD, United States

Keywords: systems biology, cancer, multi omics, cancer diagnoses, molecular subtype classification

Editorial on the Research Topic

Application of Systems Biology in Molecular Characterization and Diagnosis of Cancer

OPEN ACCESS

Edited by:
William C. Cho, QEH, China

Reviewed by:
Alexandros G. Georgakilas, National Technical University of Athens, Greece

***Correspondence:**
Cheng Zhang
cheng.zhang@scilifelab.se

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *15 February 2021*
Accepted: *19 April 2021*
Published: *28 May 2021*

Citation:
Zhang C, Wei Y, Mardinoglu A and Zhang P (2021) Editorial: Application of Systems Biology in Molecular Characterization and Diagnosis of Cancer. Front. Mol. Biosci. 8:668146. doi: 10.3389/fmolb.2021.668146

Cancer is one of the top killers of human beings, causing ∼10 million deaths in 2020 alone, which makes it either the first or second leading cause of death for people under 70-years-old (Sung et al., 2021). Therefore, there is an increasing need for an effective diagnosis and treatment for cancer, and researchers have spent a huge amount of resources and efforts to define the molecular mechanisms driving its development and progression. However, studying cancer is quite difficult as there is huge heterogeneity among human cancers, which means the variation of different individuals diagnosed with the same cancer type can sometimes be greater than that of patients from different types of cancers (Uhlén et al., 2017). As a result, most current cancer drugs are only effective in a certain subgroup of patients, and there is still a huge gap in our understanding of other treatment approaches and cancer pathogenesis (Brennan et al., 2010).

To address this huge heterogeneity among cancer patients, there is an urgent need to develop personalized diagnostic strategies to characterize cancer patients with different molecular profiles which could consequently facilitate the development of personalized and precision medicine for better treatment strategies of individual cancer. Systems biology has been a powerful tool in the integration of omics data and the characterization of different cancers. The cancer research community is increasingly using systems biology approaches to understand the complex molecular profile of cancers and decipher the mechanisms of tumor progression for the development of more effective cancer therapies (Du and Elemento, 2015). Creighton et al. characterized four different subtypes of clear cell renal cell carcinoma based on multi-omic molecular profile of the tumor (Creighton et al., 2013). Bidkhori et al. employed a metabolic network to stratify hepatocellular carcinoma and revealed three molecular subtypes relying on alternative enzymes to catalyze the same metabolic reactions (Bidkhori et al., 2018). In addition, Toy et al. performed a meta-analysis on long non-coding RNA HOX transcript antisense RNA using publically available data and identified potential prognostic biomarkers for the prediction of the survival of different cancers (Toy et al., 2019). In this Research Topic, Shi et al. reviewed the recent progression of multi-omic data integration for the study of gastric cancer. The authors specifically focused on systems biology approaches for integration of multi-omics data, and also discussed the association between gastrointestinal microbiota and gastric cancer. In addition, Zhang et al. summarized recent

understanding of four existing molecular subtypes of glioblastoma, and the indication of these classifications in guiding diagnosis, prognosis, and treatment of cancer.

Thanks to the rapid development of Next-Generation Sequencing technology, the cost and time needed for the generation of RNA-sequencing data have been significantly reduced in the past few years. As a result, a huge amount of transcriptomic data has been generated from different cancer patients and made publically available, which greatly facilitated the molecular characterization of subtypes based on cancer transcriptomic profile. In this context, many transcriptomic based models have been developed for the diagnosis of cancer molecular subtypes. Hu et al. evaluated the transcriptomic profile of tumor and adjacent normal tissue samples as well as lymph nodes from Head and Neck Squamous Cell Carcinoma patients, and identified a list of gene markers for metastasis of the tumor. Based on the transcriptomic profiles of gastric cancer patients, Dai et al. revealed the association between mucins and clinical outcomes. In addition, they proposed a prognostic marker by combining the transcriptomic expression of two mucin related genes. Kaushik et al. who are more interested in non-small cell lung cancer, focused on the commonly differentially expressed genes among several patient cohorts, and proposed a combined prognostic model which can stratify patients into different molecular subgroups with different survival outcomes.

The canonical transcriptomic and survival analyses are sensitive to the batch effect, and they may also mask the heterogeneity of individual cancer patients. In order to address these issues, Chen et al. and Guan et al. both applied a method (Wang et al., 2015) that uses gene ranking within each individual sample for patient classification to study the molecular subtypes of breast cancers. Both of these studies obtained biomarkers that could classify the individual patient into different subtypes which are either resistant or sensitive to a certain treatment.

Apart from transcriptomics, other omics profiles can also be integrated and applied in characterizing cancer molecular subtypes. Xu et al. integrated both genetic and transcriptomic information from breast cancer patients and identified immune subtypes among the patients. Wu et al. analyzed both proteomics and transcriptomics data from patients and identified XRCC1 as a promising predictive biomarker and therapeutic target for gallbladder cancer. Deng et al. integrated clinical and comprehensive molecular information from patients diagnosed with endometrial carcinoma and built a prognosis model to predict the prognosis of the patients from different identified subgroups. Elnemr et al. developed a machine learning method to identify biological causes of malignant diseases based on protein correlations. Moreover, Cheng et al. reported that cancer purity correlates with the number of mutations in tumors and will affect the genomic mutation profile in pathological analyses.

The work presented in this Research Topic highlights the importance of the characterization of the molecular subtypes of different cancers and presents many recent studies that identify different cancer subtypes based on transcriptomics, proteomics, and/or genomics using systems biology approaches. These studies provide valuable insights and extend understanding of the complexity of cancer pathogenesis and progression, and will accelerate the development of personalized and precision medicine for cancer treatment.

AUTHOR CONTRIBUTIONS

CZ, YW, AM, and PZ wrote the editorial together and approved its final version. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Knut and Alice Wallenberg Foundation.

REFERENCES

Bidkhori, G., Benfeitas, R., Klevstig, M., Zhang, C., Nielsen, J., Uhlén, M., et al. (2018). Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11874–E11883. doi: 10.1073/pnas.1807305115

Brennan, D. J., O'Connor, D. P., Rexhepaj, E., Pontén, F., and Gallagher, W. M. (2010). Antibody-based proteomics: fast-tracking molecular diagnostics in oncology. *Nat. Rev. Cancer* 10, 605–617. doi: 10.1038/nrc2902

Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. A., Gibbs, R. A., Muzny, D., et al. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *The Cancer Genome Atlas Research Network. Analysis Working Group: Baylor College of Medicine. Nature* 499, 43–49. doi: 10.1038/nature12222

Du, W., and Elemento, O. (2015). Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. *Oncogene* 34, 3215–3225. doi: 10.1038/onc.2014.291

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

Toy, H. I., Okmen, D., Kontou, P. I., Georgakilas, A. G., and Pavlopoulou, A. (2019). HOTAIR as a prognostic predictor for diverse human cancers: a meta- and bioinformatics analysis. *Cancers* 11:778. doi: 10.3390/cancers11060778

Uhlén, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357:eaan2507. doi: 10.1126/science.aan2507

Wang, H., Sun, Q., Zhao, W., Qi, L., Gu, Y., Li, P., et al. (2015). Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* 31, 62–68. doi: 10.1093/bioinformatics/btu522

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

Copyright © 2021 Zhang, Wei, Mardinoglu and Zhang. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fmolb.2020.00003

Validation of Gene Profiles for Analysis of Regional Lymphatic Metastases in Head and Neck Squamous Cell Carcinoma

Zhenrong Hu[1,2†], *Ranran Yang*[1,2†], *Li Li*[2†], *Lu Mao*[1,2], *Shuli Liu*[2,3,4], *Shichong Qiao*[5*], *Guoxin Ren*[1,2,3,4*] *and Jingzhou Hu*[1,2,3,4*]

[1] School of Stomatology, Weifang Medical University, Weifang, China, [2] Department of Oral Maxillofacial-Head and Neck Oncology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3] Shanghai Key Laboratory of Stomatology & Shanghai Research Institute of Stomatology, Shanghai, China, [4] National Clinical Research Center of Stomatology, Shanghai, China, [5] Shanghai Key Laboratory of Stomatology, Department of Oral and Maxillo-facial Implantology, School of Medicine, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Cheng Zhang,
Royal Institute of Technology, Sweden

Reviewed by:
Changli Qian,
Dana–Farber Cancer Institute, United States
Sunjae Lee,
Science for Life Laboratory, Sweden

***Correspondence:**
Shichong Qiao
shichong_qiao@hotmail.com
Guoxin Ren
renguoxincn@sina.com
Jingzhou Hu
huyayi@shsmu.edu.cn

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

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: 27 October 2019
Accepted: 10 January 2020
Published: 04 February 2020

Citation:
Hu Z, Yang R, Li L, Mao L, Liu S, Qiao S, Ren G and Hu J (2020) Validation of Gene Profiles for Analysis of Regional Lymphatic Metastases in Head and Neck Squamous Cell Carcinoma. Front. Mol. Biosci. 7:3. doi: 10.3389/fmolb.2020.00003

The progress of Head and Neck Squamous Cell Carcinoma (HNSCC) is dependent on both cancer stem cells (CSCs) and immune suppression. This study was designed to evaluate the distribution of CSCs and the characteristic immune suppression status in HNSCC primary tumors and lymph nodes. A total of 303 lymph nodes from 25 patients, as well as tumor and adjacent normal tissue samples, were evaluated by a quantitative PCR assay of the markers of CSCs and the characteristic immune suppression. Expressions of selected genes in The Cancer Genome Atlas (TCGA) datasets were also analyzed. In the primary tumors, we found that expressions of CSCs markers *(ALDH1L1, PECAM1, PROM1)* were down-regulated, while immune suppression *markers FOXP3, CD47, EGFR, SOX2,* and *TGFB1* were up-regulated significantly when compared to that in adjacent normal tissues. In the lymph nodes, expressions of both CSCs, and immune suppression *markers* were upregulated significantly compared with that in primary tumors. The mRNA expression of selected CSCs and immune suppression markers exhibited the highest expression in the level II of metastasis, then declined in the level III and remained constant at a reduced value in levels IV and V of metastases. These results reveal a comprehensive understanding of the unique genetic characteristics associated with metastatic loci and potential routes of lymphatic dissemination of HNSCC, which helps to explain why the level II has a high incidence of lymph node metastasis, and why skip metastasis straight to the level IV or level V is rarely found in the clinic.

Keywords: head and neck squamous cell carcinoma (HNSCC), lymph nodes, metastasis, mRNA expression, CSCs and immune suppression markers

INTRODUCTION

Head and Neck squamous cell carcinoma (HNSCC) is ranked as the sixth most common cancer in the world, with almost 600,000 new cases occurring every year (Bray et al., 2018). More than 50% of HNSCC patients present with metastasis to local lymph nodes at the time of diagnosis. Regional lymph nodes metastasis not only indicates poor survival, but is also a major prognostic factor for the determination of the appropriate treatment (Ozdek et al., 2000; Michikawa et al., 2012).

Patients with regional lymph nodes metastasis present a 30–60% 5-year survival rate compared with ∼85% for patients without synchronous nodal dissemination (de Juan et al., 2013). Furthermore, regional or metastatic recurrence are more likely to arise in the patients with synchronous nodal dissemination after completing synthetic serial therapy (Wan et al., 2012). However, such regional or metastatic recurrence in HNSCC is generally considered to be incurable and resistant to conventional treatment, with almost 22 months median survival rate in patients receiving salvage surgery or irradiation, and ∼12 months for those undergoing palliative chemotherapy alone (Leon et al., 2012; Ho et al., 2014).

Increasing evidence demonstrates a complex, nonlinear, branched evolution model of subclonal populations in cancers (Ginos et al., 2004). The model characterized as a dynamic process that minor subclones likely expand under the selective pressure of therapy (Ginos et al., 2004). It has been reported that hematological malignancies reveal a distinct pattern of clonal evolution in the development of therapeutic resistance and relapse (Tomasson et al., 2012). A microarray-based study has identified a number of HNSCC metastasis and recurrence associated genes in the tissue of primary or recurrent tumors, unmatched normal mucosa and lymph node metastases, but the clinical implication of these observations remains unknown (Lacko et al., 2012). Another study revealed that the mRNA expression of HNSCC in primary tumors are similar to their respective matched metastatic lymph nodes (Reis et al., 2011). These studies have provided a great insight into the genetic alterations underlying the process of metastasis in HNSCC, which will help us to identify novel therapeutic targets.

The extent of lymph node metastasis is an important prognostic factor for locally advanced HNSCC, and previous studies have reported that skip metastasis to inferior cervical lymph nodes at levels III or IV in the absence of demonstrable involvement of levels I and II is rarely found in HNSCC (Amin et al., 2017). However, the genetic alterations and underlying mechanism in the process of nodal dissemination are poorly understood and little is known about the impact of the level of lymph nodes metastasis (LNM) for patients with HNSCC. It presents a great challenge to develop more effective therapeutic strategies to prevent metastases and recurrence. For the purpose of uncovering the CSCs and immune suppression-related genetic alterations underlying metastasis in HNSCC, we chose 15 of the CSCs and immune suppression-related genes that have been reported in the primary expression of HNSCC before (Kosan and Kunz, 2002; Grosse-Gehling et al., 2013; Nor et al., 2014; Prakasam et al., 2014; Yang et al., 2014; Hartomo et al., 2015; Wu et al., 2015; Ji, 2016; Ren et al., 2016) performed RT-PCR to examine the expression levels of in cancer tissues, lymph nodes, and the matched normal tissues from the same patients with synchronous nodal metastases.

MATERIALS AND METHODS

Cohorts of Enrolled HNSCC Patients

Twenty-five patients diagnosed with HNSCC and subjected to primary operation in Oral and Maxillofacial–Head and Neck Oncology Department of Shanghai Ninth People's Hospital between 2015 and 2016 were screened for the experiments. All patients recruited had not had chemotherapy or radiotherapy prior to the surgical treatment and the patients who underwent neo-adjuvant chemo or radiotherapy were excluded from our study. The mean age of 25 participants (17 men and 8 women) was 57, ranging from 52 to 74. The samples were collected during the surgery and immediately frozen, including cancer tissues, adjacent normal tissues and lymph nodes from the same patient. Adjacent normal tissues were required to be more than 2 cm from the tumor margins in the same patient. The categorization of neck dissection samples was in accordance with the topographic classification of cervical lymph node levels suggested by Gregoire et al. (2014) as IA, IB, IIA, IIB, III, IV, V. Metastatic lymph nodes were confirmed by hematoxylin-eosin (HE) staining and observed by two pathologists independently. The written informed consent of all patients was obtained, following the protocols approved by Shanghai Ninth People's Hospital Ethical Committee and the study was performed in accordance with the Declaration of Helsinki.

Exclusion Criteria

The patients in one of the following situations were excluded in this study (Zhi et al., 2015).

Patients with local recurrences or second primary tumors.

Patients who were HPV positive.

Patients who experienced chemo- or radiotherapy prior to this study.

The Cancer Genome Atlas (TCGA) Datasets Analysis

The selected gene expression was downloaded from mRNA expression detection platform RNA-seq version 2 (level 3) in TCGA data portal (http://cancergenome.nih.gov/). The final number of HNSCC patients included was 509. The statistical programing software R (version 2.14.1) was used to analyze the datasets with the statistical significance at $P < 0.05$. The normalized counts (cancer and adjacent normal) were used to compare the gene expression (RNA-Seq version 2).

mRNA Expression Profiling

Trizol reagent (Life Technologies, USA) was used to extract the total RNAs of all the acquired samples, and iScriptTMcDNA synthesis kit (Bio-Rad, CA) was used to reverse transcription (RT). The FastSYBR Green master mix with Rox (Life Technologies, USA) was used to perform the quantitative PCR. Fifteen genes were evaluated by quantitative RT-PCR. The primers for *IRF1, IFNAR2, FOXP3, TMEM173, CD47, PECAM1, BMI-1, TWISTNB, ALDH1L1, PROM1, EGFR, SOX2, TGFB1, SMAD3,* and *STAT3* were purchased from SABiosciences. The primer sequences of the genes are listed in **Table S1**. The gene β-actin was chosen to be the control gene for normalization.

Statistical Analysis

Data from more than three independent experiments are represented as the mean ± standard deviation (SD). Pared t test was used to do statistical analyses comparing the genes expression between primary and adjacent normal tissues of HNSCC patients, Mann Whitney test was performed to analyze

TABLE 1 | Clinicopathological parameters of enrolled HNSCC patients.

No.	Tobacco	Alcohol	Site of tumor origin	Pathological stage	AJCC stage
1	YES	YES	Tongue (oro-pharyngeal)	pT4aN0M0	IVa
2	YES	YES	Mouth floor	pT2N0M0	II
3	NO	NO	Buccal	pT4aN0M0	IVa
4	NO	NO	Tongue	pT2N1M0	III
5	YES	YES	Mouth floor	pT2N0M0	II
6	YES	YES	Gingiva	pT2N2aMO	IVa
7	NO	NO	Gingiva	pT2N2bM0	IVa
8	YES	YES	Mouth floor	pT3N1MO	III
9	YES	NO	Tongue	pT3N0M0	III
10	NO	NO	Palate	pT4N0M0	IVa
11	YES	YES	Mouth floor	pT3N1MO	III
12	YES	YES	Mouth floor	pT4aN2cM0	IVa
13	NO	NO	Tongue	pT3N1M0	III
14	YES	YES	Tongue (oro-pharyngeal)	pT4N0M0	IVa
15	NO	NO	Tongue	pT3N1M0	III
16	NO	NO	Tongue (oro-pharyngeal)	pT4N1M0	IVa
17	YES	YES	Mouth floor	pT4N2bM0	IVa
18	NO	YES	Gingiva	pT2N2bM0	IVa
19	NO	NO	Tongue	pT2N0M0	II
20	YES	YES	Tongue	pT1N0M0	I
21	NO	NO	Tongue (oro-pharyngeal)	pT3N1M0	III
22	YES	NO	Tongue (oro-pharyngeal)	pT4N0M0	IVa
23	NO	NO	Tongue (oro-pharyngeal)	pT4N1M0	IVa
24	NO	NO	Buccal	pT2N0M0	II
25	NO	NO	Buccal	pT2N0M0	II

TABLE 2 | Results of histopathologic examination of various cervical lymph node levels.

	Lymph node levels							
Numbers	**Level IA**	**Level IB**	**Level IIA**	**Level IIB**	**Level III**	**Level IV**	**Level V**	**Total**
Metastasis	4	10	3	2	5	2	1	27
Non-metastasis	39	40	25	38	62	52	20	276
Total	43	50	28	40	67	54	21	303

the genes expression in metastatic nodes and non-metastatic nodes and data from TCGA, Kruskal–Wallis test was used to access the expression of metastatic nodes in different levels. Mean-normalized mRNA expression value in non-metastasis lymph nodes were chosen to be control for the entire cohort of lymph nodes. The relative value of each gene was calculated as follow: The $\Delta\Delta$Ct = tumor ΔCt-control ΔCt, fold change of mRNA was obtained as $2^{(-\Delta\Delta Ct)}$. We considered the data was significant at $P < 0.05$.

RESULTS

Clinicopathologic Characteristics of HNSCC Patient Samples

Clinicopathologic characteristics for the study groups ($n = 25$) including age, gender, tobacco and alcohol history, tumor site, and AJCC stage [which was according to AJCC 8th Edition (Tao et al., 2006)] are summarized in **Table 1**. Patients' ages ranged from 52 to 74 with a mean age of 57 $\pm$ 7.8 years. The gender distribution was 68% male (17/25) and 32% female (8/25). Forty eight percent (12/25) had a history of substantial tobacco exposure (generally >20 pack years), and 44% (11/25) had documented alcohol use. The lesion sites involved were oropharynx (24%) and oral cavity (76%). 4, 36, 24, and 36% of the patients had T1, T2, T3, and T4 tumors, respectively, while 4, 20, 28, and 48% were staged to I, II, III, and IV, accordingly. The majority of the patients had subsequent therapy with radiation (92%) after the surgery. Of all the 25 patients, 13 patients were diagnosed with synchronous nodal metastasis. Metastatic carcinomatous cells were observed in 27 (2.5%) of the 303 lymph nodes. However, there were only 2 cases of skip metastases to level IV and 1 case of skip metastasis for level V. The details of these results are presented in **Table 2**.

TABLE 3 | Clinicopathological parameters of HNSCC patients in TCGA dataset.

Characteristic	Number	Characteristic	Number
Gender		Age	
Male	381	< 60	288
Female	141	≥ 60	233
		Not available	1
Smoking status		Alcohol	
Smoker	388	Yes	351
Nonsmoker	121	No	162
Not available	13	Not available	9
Tumor stage		Lymph node stage	
I-II	186	N0-1	327
III-IV	320	N2-3	173
Tx	12	Nx	18
Not available	4	Not available	4
Metastasis stage		AJCC stage	
M0	492	I-II	120
M1	4	III-IV	388
Mx	21	Not available	14
Not available	5		
Primary region		Race	
Gingiva	18	American Indian	2
Tongue	131	Asian	11
Hard palate	7	Black American	45
Floor of mouth	61	White American	447
Bucca	22	Not Available	17
Tonsil	45	HPV	
oropharynx	37	Positive	44
hypopharynx	10	Negative	81
Larynx	115	Not Available	397
Lip	3		
Oral Cavity	73		

Gene Expression Profiles in HNSCC TCGA Dataset

Since differentially expressed genes of the transcriptome data of HNSCC in TCGA dataset have been reported (Cancer Genome Atlas, 2015; Yan et al., 2016), we further examined the expression profiles of the selected genes (i.e., *IRF1*, *IFNAR2*, *FOXP3*, *TMEM173*, *CD47*, *PECAM1*, *BMI-1*, *TWISTNB*, *ALDH1L1*, *PROM1*, *EGFR*, *SOX2*, *TGFB1*, *SMAD3,* and *STAT3*) in the primary tumors of HNSCC. The Clinicopathologic characteristics of TCGA dataset was shown in **Table 3**. It enrolled 522 HNSCC patients and lesion sites involved 312 in oral cavity (59.8%), 115 in Larynx (22.0%), 82 in oropharynx (15.7%), 10 in hypopharynx (1.9%), and 3 in Lip (0.6%). Forty four of them were diagnosed with HPV infection. This is shown in **Figure 1**, in which *ALDH1L1*, *PECAM1*, *SMAD3*, *TMEM173*, *PROM1,* and *STAT3* were expressed lower, and the rest of the genes were expressed higher in primary tumors than in adjacent normal tissues. Among them, the expression levels of *IRF1*, *IFNAR2*, *FOXP3*, *CD47*, *ALDH1L1*, *PROM1*, *EGFR*, *SOX2*, *TGFB1,* and *STAT3* were significant ($P < 0.05$) between tumors and normal tissues (**Figure 1**).

Examination of mRNA Expression in Primary Tumors by RT-PCR Detection

To further investigate the above-mentioned gene expression in our HNSCC samples, we validated the mRNA expression using real time RT-PCR detection. The differential levels of mRNA expression ($-\Delta\Delta$Ct) of the 15 genes in each of the 25 individuals HNSCC sample are shown in **Figure 2**. Except for *PECAM1,* which showed a significant decrease ($P < 0.05$), while the TCGA dataset revealed no significant difference, the expression levels (up or down) of the rest 14 selected genes were consistent with TCGA profiles as validated by real time RT-PCR assay for all recruited patients and a clustering analysis was performed in the primary of HNSCC (**Figure 2**). Furthermore, there were no significant differences between men and women in normal tissue, but it showed gender difference with male dominancy in the primary (**Figure 3** and **Figure S1**). We also found there were no significant differences between sex and risk factors (such as smoking or drinking) ($p = 0.896$ and 0.694, respectively). It suggested that sex factor might play a role in the formation of CSCs and immune suppression and might be more susceptible to metastasis.

Gene Expression in Different Lymph Nodes

We further compared the mRNA expression of the selected genes between metastatic nodes and non-metastatic nodes in cervical lymph nodes using RT-PCR assay. Similar to the expression levels in primary tumors, the expressions of *CD47*, *EGFR*, *FOXP3*, *IFNAR2*, *IRF1*, *SOX2*, *BMI-1*, *PROM1,* and *TGFB1* were upregulated significantly in metastatic nodes when compared to non-metastatic nodes (**Figure 4**). *STAT3* and *PECAM1* were decreased in primary but upregulated in lymph nodes ($P > 0.05$) (**Figures 2**, **3**). One exception is *ALDH1L1* that exhibited a low mRNA expression level in metastatic nodes (**Figure 3**). Finally, we analyzed the mRNA expression of the selected genes among different levels of metastatic lymph nodes. We found that all the selected genes except *ALDH1L1* followed a similar change from level I–level V metastatic nodes, in which all the genes reached the highest expression in level II, declined in level III and remained at a constant low value in level IV and level V of metastatic nodes (**Figure 5**). Also, it revealed gender difference with male dominancy in the lymphatic loci of HNSCC (**Figure S2**).

DISCUSSION

The nodal dissemination of carcinomatous cell is a vital determinant of prognosis in patients with HNSCC, as the overall survival rate of patients with metastatic lymph nodes decreased by a half compared with that without nodal implicated (Yan et al., 2014). Keeping track of tumor cells through lymphatic metastasis and microenvironment in lymph nodes is particularly vital for treatment (Owens et al., 2014). As far as we know, there were no studies that investigated the mRNA expression in an individual cervical lymph node level. Hence, the TCGA

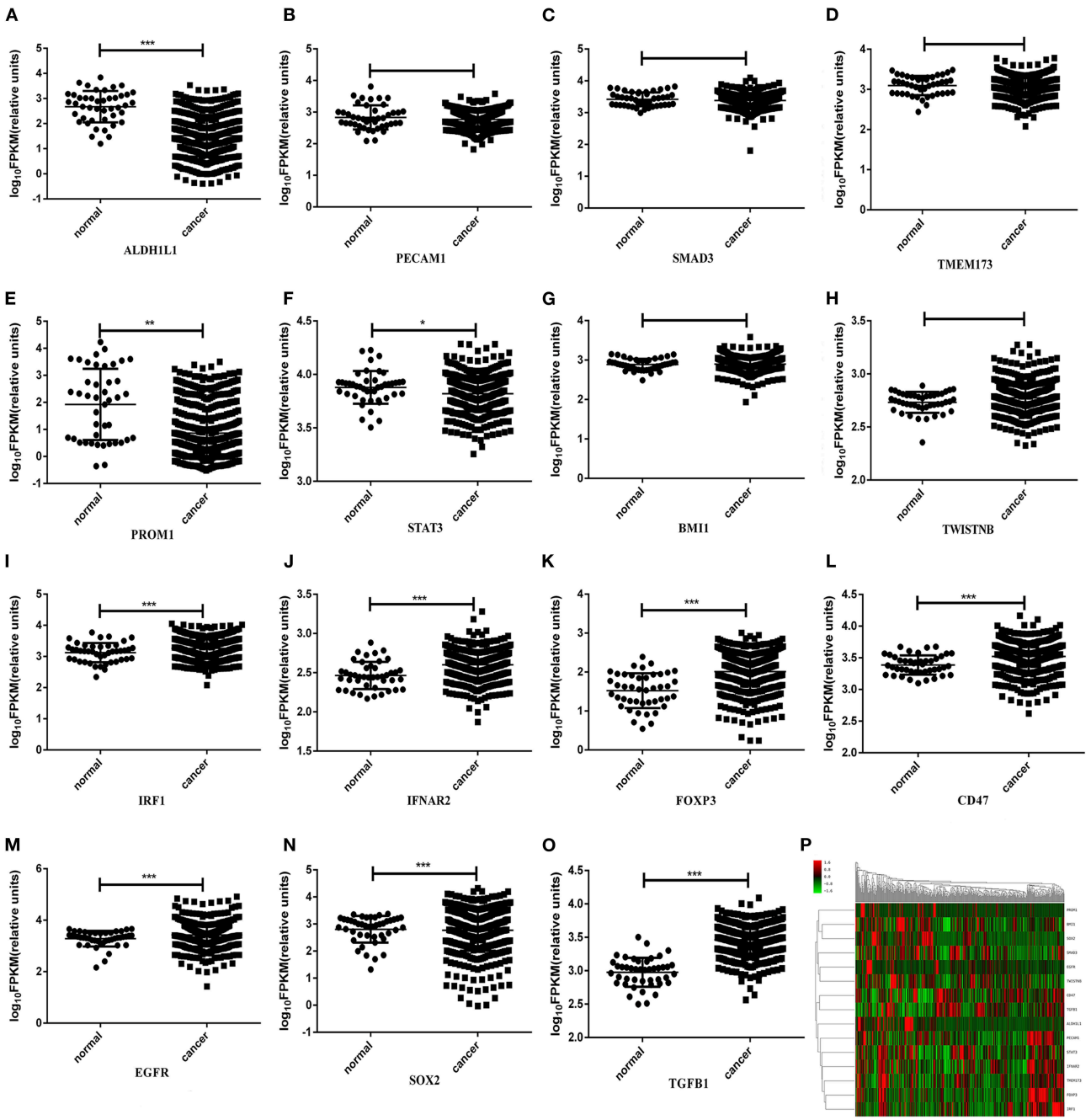


FIGURE 1 | FPKM values of *ALDH1L1* **(A)**, *PECAM1* **(B)**, *SMAD3* **(C)**, *TMEM173* **(D)**, *PROM1* **(E)**, *STAT3* **(F)**, *BMI-1* **(G)**, *TWISTNB* **(H)**, *IRF1* **(I)**, *IFNAR2* **(J)**, *FOXP3* **(K)**, *CD47* **(L)**, *EGFR* **(M)**, *SOX2* **(N)**, and *TGFβ1* **(O)** expression of HNSCC patients in TCGA datasets; **(P)** Heatmap of gene expression profiles of HNSCC patients in TCGA datasets. Gene expression values are calculated based on the log transform of FPKM values determined from RNA sequencing analyses. High expression is depicted by red; low expression is depicted by green, normalized by row z scores (legend). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

database was used in our study to perform the unbiased, large-scale analysis of 15 metastasis and tumor microenvironment associated genes. The results showed a significant change of related gene expression through both the HNSCC TCGA-based and our tumor-based analyses. Furthermore, the head and neck squamous cancer (HNSCC) with HPV-negative (HPV^{-}) always occurred in an older patient population and their clinical outcomes were unquestionably worse than HPV^{+} HNSCC, and few HNSCCs are associated with HPV infection. Therefore, our results provide a model to determine gene expression patterns in both the primary tumors and metastatic lymph nodes of HPV^{-} HNSCC.

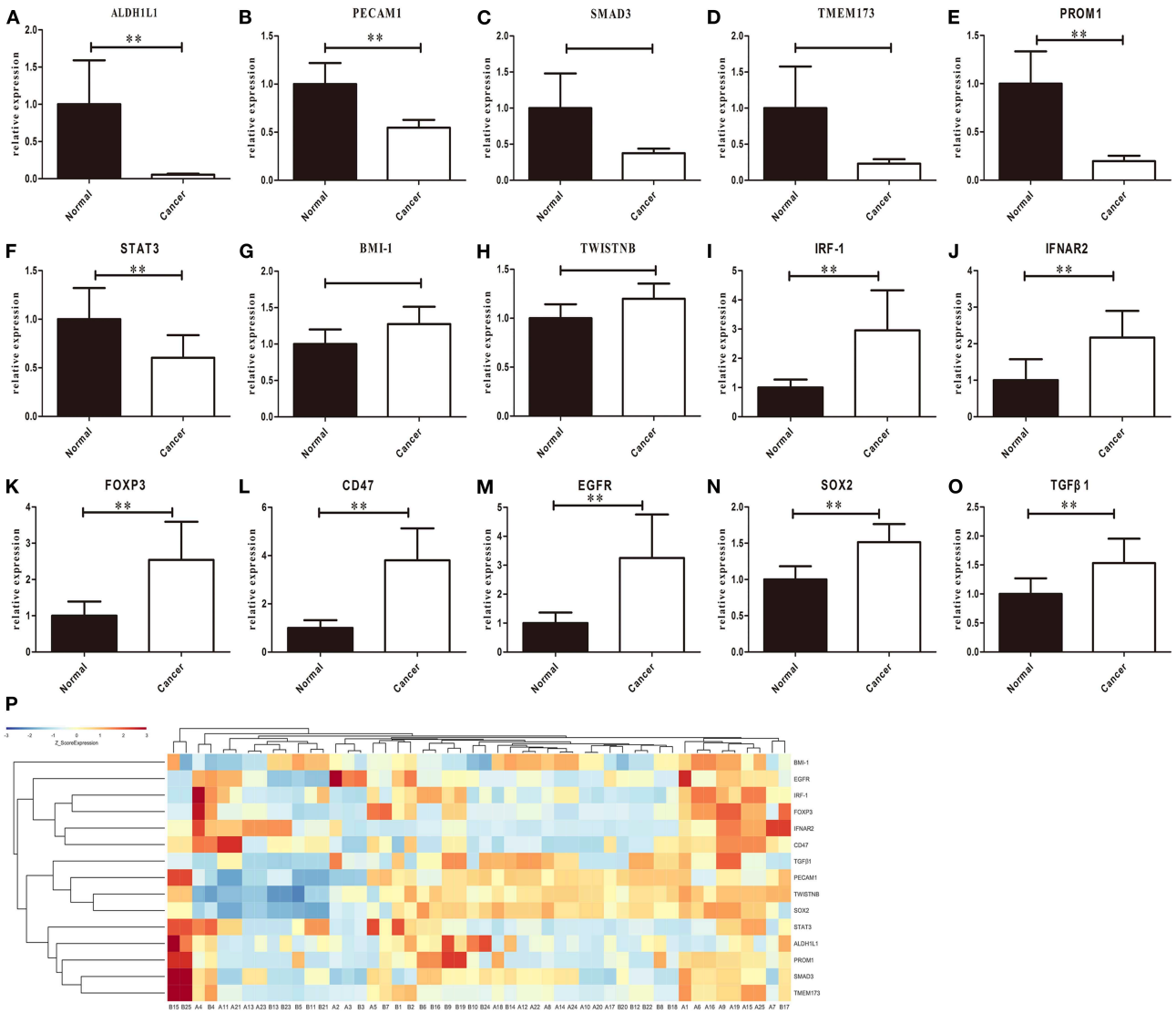


FIGURE 2 | RT-PCR results of *ALDH1L1* **(A)**, *PECAM1* **(B)**, *SMAD3* **(C)**, *TMEM173* **(D)**, *PROM1* **(E)**, *STAT3* **(F)**, *BMI-1* **(G)**, *TWISTNB* **(H)**, *IRF1* **(I)**, *IFNAR2* **(J)**, *FOXP3* **(K)**, *CD47* **(L)**, *EGFR* **(M)**, *SOX2* **(N)**, and *TGFβ1* **(O)** expression in primary tumors. Each bar represents the average log2 gene mRNA expression level of the paired tumor and adjacent normal tissues ($n = 25$ per group). Eleven genes showed significant changes between paired tumor and adjacent normal tissues **(P)** Heatmap of gene expression profiles of HNSCC patients in the primary of HNSCC, Gene expression values are calculated based on the log transform of FPKM values determined from RNA sequencing analyses. High expression is depicted by red; low expression is depicted by green, normalized by row z scores (legend) (**$p < 0.05$, *t*-test).

IFNAR2 encodes a protein that forms one of the two chains of a receptor for interferons alpha and beta (Zhang et al., 2015). The protein encoded by IRF1 functions as a transcription activator of interferons α-,β-, and γ-induced transcription (Choe et al., 2015) *TMEM173* encodes a transmembrane protein that functions as a pattern recognition receptor that activates type I interferon responses (Schreiber and Piehler, 2015; West et al., 2015). All three genes promote the IFN-signaling in the lymph node of HNSCC (Li and Flavell, 2008). TGF-β1 encodes a secreted ligand of the transforming growth factor-beta superfamily proteins (Kudinov et al., 2016). The band of these ligands and various TGF-β receptors result in the recruitment and stimulation of *SMAD3* that promotes the process of carcinogenesis (Chaturvedi et al., 2014; Wang et al., 2016). IFN-α, IFN-β and TGF-β play important roles in regulating the activity of lymph node stromal cells embracing lymphatic endothelial cells (LECs), follicular dendritic cells (DCs), and fibroblastic reticular cells (FRCs) (Yang et al., 2014; Ji, 2016). The functional stromal cells may reconstruct and remodel the lymph node, which would produce a unique microenvironment benefit for cancer metastasis (Hartomo et al., 2015). Our RT-PCR results show that are both upregulated in the lymph nodes of HNSCC patients and suggest that these hyperactive lymph node stromal cells may provide a suitable microenvironment for the metastases.

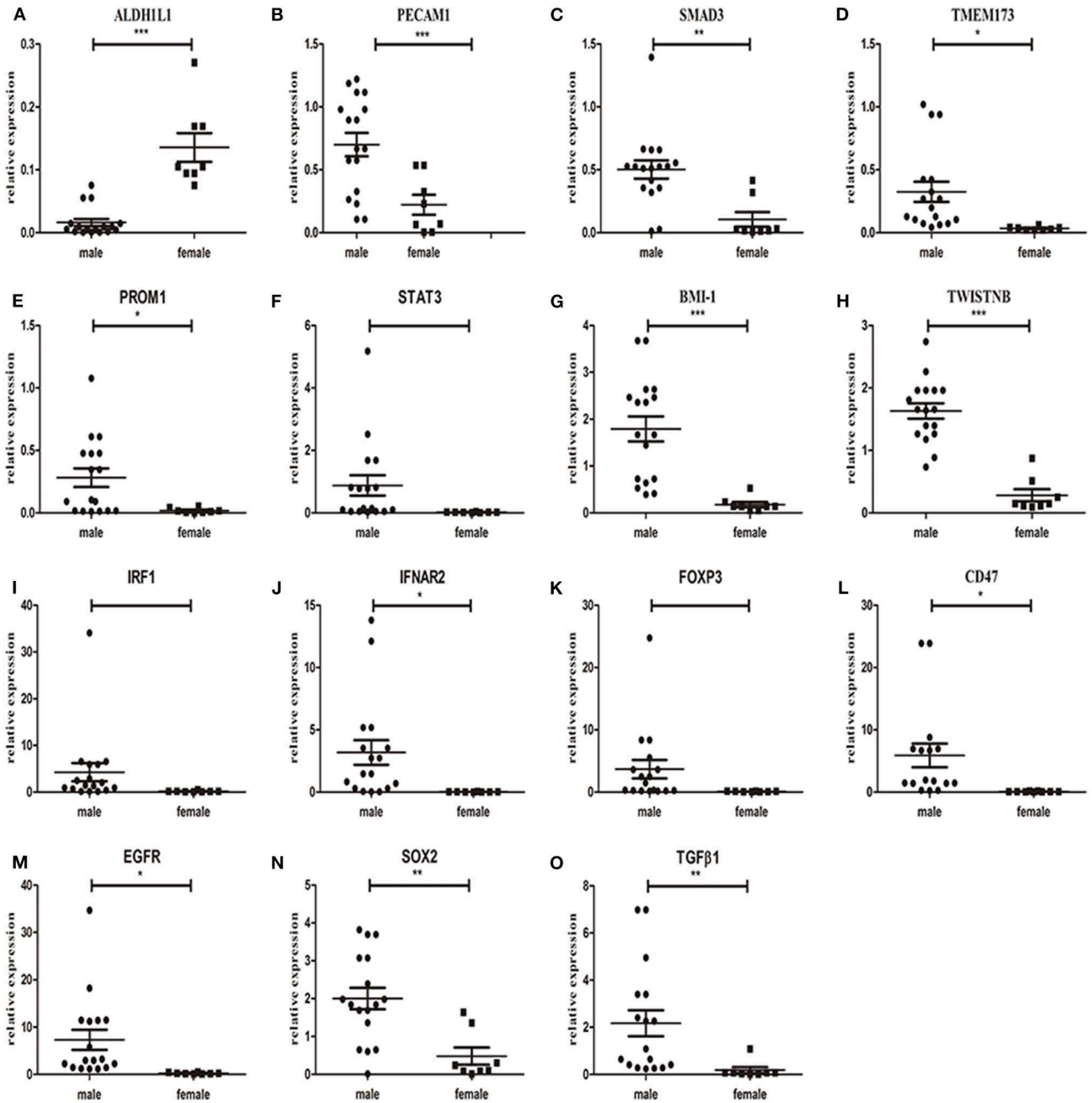


FIGURE 3 | mRNA expression of *ALDH1L1* **(A)**, *PECAM1* **(B)**, *SMAD3* **(C)**, *TMEM173* **(D)**, *PROM1* **(E)**, *STAT3* **(F)**, *BMI-1* **(G)**, *TWISTNB* **(H)**, *IRF1* **(I)**, *IFNAR2* **(J)**, *FOXP3* **(K)**, *CD47* **(L)**, *EGFR* **(M)**, *SOX2* **(N)**, and *TGFβ1* **(O)** in the primary of tumor samples between male and female patients. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

It has been reported that ALDH1 has higher activity in the stem cell subclone of leukemia and some solid tumors (Li et al., 2016). However, *ALDH1L1* showed totally different presentation in distinct types of cancers; mRNA high expressions of *ALDH1L1* were reported to be correlated to higher overall survival rate for breast cancer patients but were revealed as a poor prognostic factor in gastric and prostatic cancers (Prakasam et al., 2014; Wu et al., 2015; Ren et al., 2016). Our result revealed *ALDH1L1* had a lower mRNA expression level in primary and metastatic nodes and exhibited an entirely different behavior compared with other selected genes, indicated that *ALDH1L1* mRNA low expressions is related with metastasis of HNSCC. Our observations showed *SOX2*, *BMI-1*, *PROM1* (CD133), and *TWISTNB (TWIST NEIGHBOR)* upregulated in lymph nodes. *SOX2*, *BMI-1,* and *PROM1* are cancer stem cells-related genes (Kosan and Kunz, 2002; Grosse-Gehling et al., 2013; Nor et al., 2014), and *TWISTNB* is implicated in EMT (Li and Li, 2015). Previous studies indicated that cancer stem cells (CSCs) *in situ*

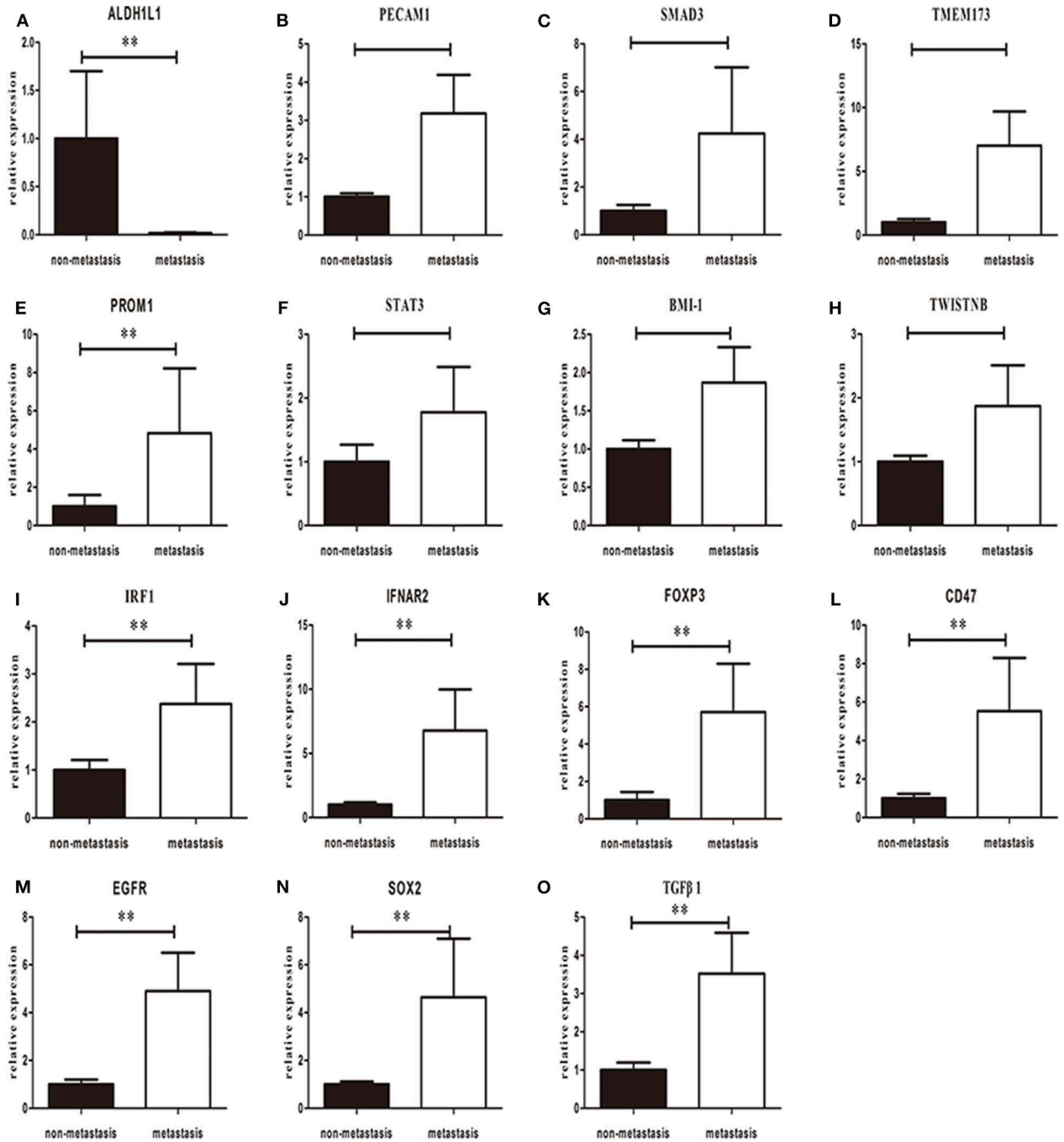


FIGURE 4 | Overall mean-normalized mRNA expression of *ALDH1L1* **(A)**, *PECAM1* **(B)**, *SMAD3* **(C)**, *TMEM173* **(D)**, *PROM1* **(E)**, *STAT3* **(F)**, *BMI-1* **(G)**, *TWISTNB* **(H)**, *IRF1* **(I)**, *IFNAR2* **(J)**, *FOXP3* **(K)**, *CD47* **(L)**, *EGFR* **(M)**, *SOX2* **(N)**, and *TGFβ1* **(O)** in lymph nodes. Each bar represents the average log2 gene mRNA expression level of the non-metastatic and metastatic lymph nodes. *P*-values are calculated by Wilcoxon rank-sum test. (**$p < 0.05$).

can transform into migrating cancer stem cells (MCSCs) by EMT (Schlereth et al., 2014). Subsequently, the MCSCs disseminate and form metastatic colonies (Li and Li, 2014). Furthermore, direct trans-differentiation of CSCs into LECs occurs during tumor lymphatic metastasis (Chao et al., 2012; Semenza, 2013). As previously mentioned, LECs remodel the lymph nodes that provide a comfortable microenvironment for cancer metastasis. Our observations of gene expression profiles in lymph nodes suggest that CSC may directly convert into LECs and contribute to tumor neovascularization in the metastasis process of HNSCC.

CD47 is likely involved in the process of evading immunological eradication (Chattopadhyay et al., 2005; Matlung

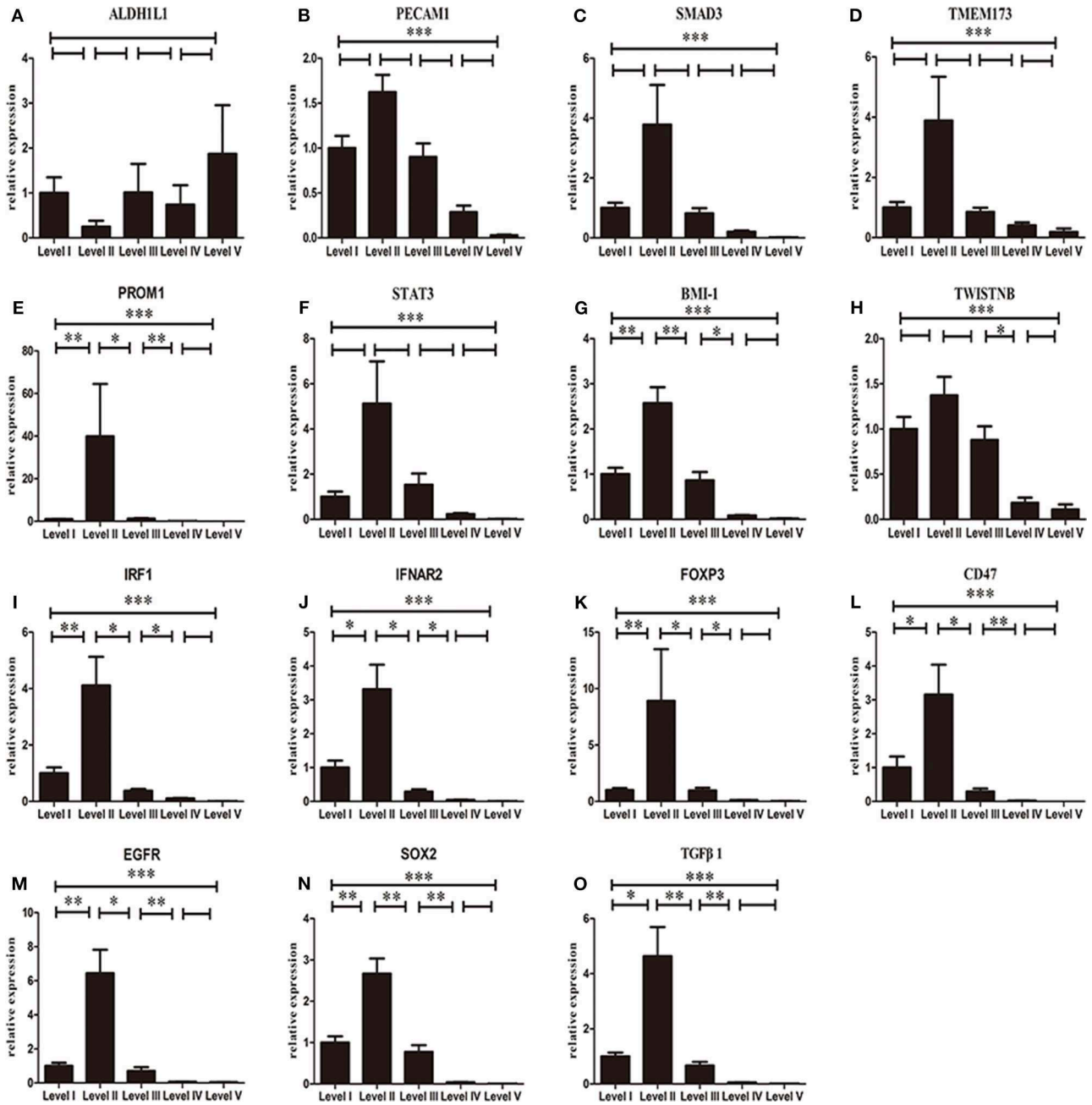


FIGURE 5 | Normalized 2 $^{(-\Delta\Delta Ct)}$ mRNA expression of *ALDH1L1* **(A)**, *PECAM1* **(B)**, *SMAD3* **(C)**, *TMEM173* **(D)**, *PROM1* **(E)**, *STAT3* **(F)**, *BMI-1* **(G)**, *TWISTNB* **(H)**, *IRF1* **(I)**, *IFNAR2* **(J)**, *FOXP3* **(K)**, *CD47* **(L)**, *EGFR* **(M)**, *SOX2* **(N)**, and *TGFβ1* **(O)** in different levels of lymph nodes. Each bar represents the average log2 gene mRNA expression level of the metastatic lymph nodes in each individual level. *P*-values are calculated by Kruskal–Wallis test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

et al., 2017), and FOXP3 is mainly considered as a biomarker of Treg cells that impede the antitumor immune responses in cancer patients (Quante et al., 2013; Triulzi et al., 2013). *STAT3* overexpression in metastatic sites may restrain the immune responses to render an immunosuppressive environment (Punt et al., 2015). PECAM1 (CD31) makes up a large portion of cell intercellular junctions and loss of its function may disrupt cell adhesion (ElShamy et al., 2016). Interaction of hypoxia-surviving cells with the immunosuppressive environment influenced by newly recruited tumor-associated macrophages (TAMs), mesenchymal stromal cells (MSCs), and other types of immune cells most likely form and maintain a necrotic/hypoxic core called "aggressiveness niche" that will be the foster ground for cancer metastasis precursors (Johnson, 2001). In accordance with above reports, our study demonstrates that *CD47* and *FOXP3* were significantly upregulated both in primary site and

lymphatic loci, but *STAT3,* and *PECAM1* were only upregulated in lymph nodes. It indicates that there may form a more immunosuppressive environment in the lymph nodes than primary site and may exist an aggressiveness niche to foster metastasis in HNSCC patients.

A previous study revealed that male patients may have higher tendency of cancer-associated gene expression in the primary than matched adjacent normal tissue (Shores et al., 2004). Similarly, the expression of 15 selected genes showed gender differences with male dominancy not only in the primary but also in metastatic loci, with no significant differences between men and women in matched adjacent normal tissue. It indicated that males may be at a higher risk of metastasis than females in HNSCC. The mechanism underlying these expression patterns will be further studied in the future.

It have been reported that a relatively constant and sequential route might exist in the lymphatic drainage of the head and neck region (Ji, 2016). Level I and level II are the most common region of metastases in cancers of the oral cavity, while level III is most likely area of metastases in cancers of hypopharynx (Vishak and Rohan, 2014). Considering the potential mechanisms of the route of lymphatic metastasis remains unknown, our observations of the differential mRNA expressions in respective cervical levels showed that CSC and immunosuppression-related genes achieve a peak value in the level II, but maintain a constant low value in the levels IV and V, which suggests that the level II may mostly provide a necrosis-induced inflammation and hypoxia-induced immunosuppressive environment that seems to be the most fertile ground to generate the tumor cells with metastatic potentials. This supports that carcinomatous cells in the level II tend to have the highest metastatic potency, which may also offer a genetic explanation why rare skip metastases of the level IV or level V were found in HNSCC (Lydiatt et al., 2017).

The relatively small quantity of lymphatic metastasis samples, the variations of tumor sample purity, and the intratumor heterogeneity limit our current study, especially given that the expression of *SMAD3, PECAM1, TMEM173, STAT3, and TWISTNB* reveals no statistically significant difference among different groups. Meanwhile, our observation revealed *PECAM1* was significantly decreased but it showed no significant difference in TCGA datasets. It may result from that the squamous cell cancer in oropharynx, hypopharynx, and larynx were included in TCGA dataset, and they had an entirely different biological behavior and prognosis from HNSCC. While these findings will require further validation in larger cohorts of patient samples in the future, we believe this first gene expression analysis of cervical lymph nodes in the individual level of cancer patients will provide us an important opportunity to guide the future investigations of HNSCC.

Through gene expression analyses, we found that the mRNA expression of selected CSCs and immune suppression markers exhibit the highest expression in the level II metastatic lymph nodes, then declined in the level III and remained constant at a reduced value in levels IV and V metastatic lymph nodes. These results help to explain the reason why the level II has a high incidence of lymph node metastasis, and skip metastasis to the level IV or level V is rarely found in the clinic. It will help to increase the understanding of the genetic characteristics associated with metastatic loci and potential routes of lymphatic dissemination of HNSCC, and may aid the clinical diagnosis and treatment of HNSCC.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Shanghai Ninth People's Hospital Ethical Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SQ wrote the paper. ZH, RY and LL carried out experimental studies. LM and SL were involved in statistical analysis. GR and JH modified the paper and designed this study concepts. All authors read and approved the final manuscript.

FUNDING

Support for this work was obtained from the National Natural Science Foundation of China (Grant Nos. 31140007, 81472516, 81600902).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.00003/full#supplementary-material

REFERENCES

Amin, M. B., Greene, F. L., Edge, S. B., Compton, C. C., Gershenwald, J. E., Brookland, R. K., et al. (2017). The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J. Clin.* 67, 93–99. doi: 10.3322/caac.21388

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

Cancer Genome Atlas, N. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576–582. doi: 10.1038/nature14129

Chao, M. P., Weissman, I. L., and Majeti, R. (2012). The CD47-SIRPalpha pathway in cancer immune evasion and potential therapeutic implications. *Curr. Opin. Immunol.* 24, 225–232. doi: 10.1016/j.coi.2012.01.010

Chattopadhyay, S., Chakraborty, N. G., and Mukherji, B. (2005). Regulatory T cells and tumor immunity. *Cancer Immunol. Immunother.* 54, 1153–1161. doi: 10.1007/s00262-005-0699-9

Chaturvedi, P., Gilkes, D. M., Takano, N., and Semenza, G. L. (2014). Hypoxia-inducible factor-dependent signaling between triple-negative breast cancer cells and mesenchymal stem cells promotes macrophage recruitment. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2120–E2129. doi: 10.1073/pnas.1406655111

Choe, C. H., Park, I. S., Park, J., Yu, K. Y., Jang, H., Kim, J., et al. (2015). Transmembrane protein 173 inhibits RANKL-induced osteoclast differentiation. *FEBS Lett.* 589, 836–841. doi: 10.1016/j.febslet.2015.02.018

de Juan, J., Garcia, J., Lopez, M., Orus, C., Esteller, E., Quer, M., et al. (2013). Inclusion of extracapsular spread in the pTNM classification system: a proposal for patients with head and neck carcinoma. *JAMA Otolaryngol. Head Neck Surg.* 139, 483–488. doi: 10.1001/jamaoto.2013.2666

ElShamy, W. M., Sinha, A., and Said, N. (2016). Aggressiveness niche, can it be the foster ground for cancer metastasis precursors? *Stem Cells Int.* 2016:4829106. doi: 10.1155/2016/4829106

Ginos, M. A., Page, G. P., Michalowicz, B. S., Patel, K. J., Volker, S. E., Pambuccian, S. E., et al. (2004). Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res.* 64, 55–63. doi: 10.1158/0008-5472.CAN-03-2144

Gregoire, V., Ang, K., Budach, W., Grau, C., Hamoir, M., Langendijk, J. A., et al. (2014). Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiother. Oncol.* 110, 172–181. doi: 10.1016/j.radonc.2013.10.010

Grosse-Gehling, P., Fargeas, C. A., Dittfeld, C., Garbe, Y., Alison, M. R., Corbeil, D., et al. (2013). CD133 as a biomarker for putative cancer stem cells in solid tumours: limitations, problems and challenges. *J. Pathol.* 229, 355–378. doi: 10.1002/path.4086

Hartomo, T. B., Van Huyen Pham, T., Yamamoto, N., Hirase, S., Hasegawa, D., Kosaka, Y., et al. (2015). Involvement of aldehyde dehydrogenase 1A2 in the regulation of cancer stem cell properties in neuroblastoma. *Int. J. Oncol.* 46, 1089–1098. doi: 10.3892/ijo.2014.2801

Ho, A. S., Kraus, D. H., Ganly, I., Lee, N. Y., Shah, J. P., and Morris, L. G. (2014). Decision making in the management of recurrent head and neck cancer. *Head Neck* 36, 144–151. doi: 10.1002/hed.23227

Ji, R. C. (2016). Lymph nodes and cancer metastasis: new perspectives on the role of intranodal lymphatic sinuses. *Int. J. Mol. Sci.* 18:E51. doi: 10.3390/ijms18010051

Johnson, N. (2001). Tobacco use and oral cancer: a global perspective. *J. Dent. Educ.* 65, 328–339.

Kosan, C., and Kunz, J. (2002). Identification and characterisation of the gene TWIST NEIGHBOR (TWISTNB) located in the microdeletion syndrome 7p21 region. *Cytogenet. Genome Res.* 97, 167–170. doi: 10.1159/000066618

Kudinov, A. E., Deneka, A., Nikonova, A. S., Beck, T. N., Ahn, Y. H., Liu, X., et al. (2016). Musashi-2 (MSI2) supports TGF-beta signaling and inhibits claudins to promote non-small cell lung cancer (NSCLC) metastasis. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6955–6960. doi: 10.1073/pnas.1513616113

Lacko, M., De Herdt, M. J., Jansen, J. C., Brakenhoff, R. H., Slootweg, P. J., Takes, R. P., et al. (2012). Validation of a gene expression signature for assessment of lymph node metastasis in oral squamous cell carcinoma. *J. Clin. Oncol.* 30, 4104–4110. doi: 10.1200/JCO.2011.40.4509

Leon, X., Martinez, V., Lopez, M., Garcia, J., Venegas Mdel, P., Esteller, E., et al. (2012). Second, third, and fourth head and neck tumors. A progressive decrease in survival. *Head Neck* 34, 1716–1719. doi: 10.1002/hed.21977

Li, K., Guo, X., Wang, Z., Li, X., Bu, Y., Bai, X., et al. (2016). The prognostic roles of ALDH1 isoenzymes in gastric cancer. *Onco. Targets Ther.* 9, 3405–3414. doi: 10.2147/OTT.S102314

Li, M. O., and Flavell, R. A. (2008). Contextual regulation of inflammation: a duet by transforming growth factor-beta and interleukin-10. *Immunity* 28, 468–476. doi: 10.1016/j.immuni.2008.03.003

Li, S., and Li, Q. (2014). Cancer stem cells and tumor metastasis (Review). *Int. J. Oncol.* 44, 1806–1812. doi: 10.3892/ijo.2014.2362

Li, S., and Li, Q. (2015). Cancer stem cells, lymphangiogenesis, and lymphatic metastasis. *Cancer Lett.* 357, 438–447. doi: 10.1016/j.canlet.2014.12.013

Lydiatt, W. M., Patel, S. G., O'Sullivan, B., Brandwein, M. S., Ridge, J. A., Migliacci, J. C., et al. (2017). Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J. Clin.* 67, 122–137. doi: 10.3322/caac.21389

Matlung, H. L., Szilagyi, K., Barclay, N. A., and van den Berg, T. K. (2017). The CD47-SIRPalpha signaling axis as an innate immune checkpoint in cancer. *Immunol. Rev.* 276, 145–164. doi: 10.1111/imr.12527

Michikawa, C., Uzawa, N., Kayamori, K., Sonoda, I., Ohyama, Y., Okada, N., et al. (2012). Clinical significance of lymphatic and blood vessel invasion in oral tongue squamous cell carcinomas. *Oral. Oncol.* 48, 320–324. doi: 10.1016/j.oraloncology.2011.11.014

Nor, C., Zhang, Z., Warner, K. A., Bernardi, L., Visioli, F., Helman, J. I., et al. (2014). Cisplatin induces Bmi-1 and enhances the stem cell fraction in head and neck cancer. *Neoplasia* 16, 137–146. doi: 10.1593/neo.131744

Owens, T., Khorooshi, R., Wlodarczyk, A., and Asgari, N. (2014). Interferons in the central nervous system: a few instruments play many tunes. *Glia* 62, 339–355. doi: 10.1002/glia.22608

Ozdek, A., Sarac, S., Akyol, M. U., Unal, O. F., and Sungur, A. (2000). Histopathological predictors of occult lymph node metastases in supraglottic squamous cell carcinomas. *Eur. Arch. Otorhinolaryngol.* 257, 389–392. doi: 10.1007/s004050000231

Prakasam, A., Ghose, S., Oleinik, N. V., Bethard, J. R., Peterson, Y. K., Krupenko, N. I., et al. (2014). JNK1/2 regulate Bid by direct phosphorylation at Thr59 in response to ALDH1L1. *Cell Death Dis.* 5:e1358. doi: 10.1038/cddis.2014.316

Punt, S., Houwing-Duistermaat, J. J., Schulkens, I. A., Thijssen, V. L., Osse, E. M., de Kroon, C. D., et al. (2015). Correlations between immune response and vascularization qRT-PCR gene expression clusters in squamous cervical cancer. *Mol. Cancer* 14:71. doi: 10.1186/s12943-015-0350-0

Quante, M., Varga, J., Wang, T. C., and Greten, F. R. (2013). The gastrointestinal tumor microenvironment. *Gastroenterology* 145, 63–78. doi: 10.1053/j.gastro.2013.03.052

Reis, P. P., Waldron, L., Perez-Ordonez, B., Pintilie, M., Galloni, N. N., Xuan, Y., et al. (2011). A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC Cancer* 11:437. doi: 10.1186/1471-2407-11-437

Ren, Z. H., Zhang, C. P., and Ji, T. (2016). Expression of SOX2 in oral squamous cell carcinoma and the association with lymph node metastasis. *Oncol. Lett.* 11, 1973–1979. doi: 10.3892/ol.2016.4207

Schlereth, S. L., Refaian, N., Iden, S., Cursiefen, C., and Heindl, L. M. (2014). Impact of the prolymphangiogenic crosstalk in the tumor microenvironment on lymphatic cancer metastasis. *Biomed. Res. Int.* 2014:639058. doi: 10.1155/2014/639058

Schreiber, G., and Piehler, J. (2015). The molecular basis for functional plasticity in type I interferon signaling. *Trends Immunol.* 36, 139–149. doi: 10.1016/j.it.2015.01.002

Semenza, G. L. (2013). Cancer-stromal cell interactions mediated by hypoxia-inducible factors promote angiogenesis, lymphangiogenesis, and metastasis. *Oncogene* 32, 4057–4063. doi: 10.1038/onc.2012.578

Shores, C. G., Yin, X., Funkhouser, W., and Yarbrough, W. (2004). Clinical evaluation of a new molecular method for detection of micrometastases in head and neck squamous cell carcinoma. *Arch. Otolaryngol. Head Neck Surg.* 130, 937–942. doi: 10.1001/archotol.130.8.937

Tao, L., Lefevre, M., Ricci, S., Saintigny, P., Callard, P., Perie, S., et al. (2006). Detection of occult carcinomatous diffusion in lymph nodes from head and neck squamous cell carcinoma using real-time RT-PCR detection of cytokeratin 19 mRNA. *Br. J. Cancer* 94, 1164–1169. doi: 10.1038/sj.bjc.6603073

Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481, 506–510. doi: 10.1038/nature10738

Triulzi, T., Tagliabue, E., Balsari, A., and Casalini, P. (2013). FOXP3 expression in tumor cells and implications for cancer progression. *J. Cell Physiol.* 228, 30–35. doi: 10.1002/jcp.24125

Vishak, S., and Rohan, V (2014). Cervical node metastasis in T1 squamous cell carcinoma of oral tongue- pattern and the predictive factors. *Indian J. Surg. Oncol.* 5, 104–108. doi: 10.1007/s13193-014-0301-z

Wan, X. C., Egloff, A. M., and Johnson, J. (2012). Histological assessment of cervical lymph node identifies patients with head and neck squamous cell carcinoma (HNSCC): who would benefit from chemoradiation after surgery? *Laryngoscope* 122, 2712–2722. doi: 10.1002/lary.23572

Wang, H., Qiu, T., Shi, J., Liang, J., Wang, Y., Quan, L., et al. (2016). Gene expression profiling analysis contributes to understanding the association between non-syndromic cleft lip and palate, and cancer. *Mol. Med. Rep.* 13, 2110–2116. doi: 10.3892/mmr.2016.4802

West, A. P., Khoury-Hanold, W., Staron, M., Tal, M. C., Pineda, C. M., Lang, S. M., et al. (2015). Mitochondrial DNA stress primes the antiviral innate immune response. *Nature* 520, 553–557. doi: 10.1038/nature14156

Wu, S., Xue, W., Huang, X., Yu, X., Luo, M., Huang, Y., et al. (2015). Distinct prognostic values of ALDH1 isoenzymes in breast cancer. *Tumour Biol.* 36, 2421–2426. doi: 10.1007/s13277-014-2852-6

Yan, J., Xue, F., Chen, H., Wu, X., Zhang, H., Chen, G., et al. (2014). A multi-center study of using carbon nanoparticles to track lymph node metastasis in T1-2 colorectal cancer. *Surg. Endosc.* 28, 3315–3321. doi: 10.1007/s00464-014-3608-5

Yan, L., Zhan, C., Wu, J., and Wang, S. (2016). Expression profile analysis of head and neck squamous cell carcinomas using data from the cancer genome Atlas. *Mol. Med. Rep.* 13, 4259–4265. doi: 10.3892/mmr.2016.5054

Yang, C. Y., Vogt, T. K., Favre, S., Scarpellino, L., Huang, H. Y., Tacchini-Cottier, F., et al. (2014). Trapping of naive lymphocytes triggers rapid growth and remodeling of the fibroblast network in reactive murine lymph nodes. *Proc. Natl. Acad. Sci. U.S.A.* 111, E109–E118. doi: 10.1073/pnas.1312585111

Zhang, X. J., Jiang, D. S., and Li, H. (2015). The interferon regulatory factors as novel potential targets in the treatment of cardiovascular diseases. *Br. J. Pharmacol.* 172, 5457–5476. doi: 10.1111/bph.12881

Zhi, X., Lamperska, K., Golusinski, P., Schork, N. J., Luczewski, L., Kolenda, T., et al. (2015). Gene expression analysis of head and neck squamous cell carcinoma survival and recurrence. *Oncotarget* 6, 547–555. doi: 10.18632/oncotarget.2772

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

Copyright © 2020 Hu, Yang, Li, Mao, Liu, Qiao, Ren and Hu. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 25 February 2020
doi: 10.3389/fmolb.2020.00019

Systematical Analysis of the Cancer Genome Atlas Database Reveals *EMCN/MUC15* Combination as a Prognostic Signature for Gastric Cancer

Wentao Dai[1,2,3†], Jixiang Liu[1,3†], Bingya Liu[2], Quanxue Li[1,3,4], Qingqing Sang[2] and Yuan-Yuan Li[1,2,3]*

[1] Shanghai Center for Bioinformation Technology, Shanghai, China, [2] Shanghai Key Laboratory of Gastric Neoplasms, Department of Surgery, Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3] Shanghai Engineering Research Center of Pharmaceutical Translation, Shanghai Industrial Technology Institute, Shanghai, China, [4] School of Biotechnology, East China University of Science and Technology, Shanghai, China

OPEN ACCESS

Edited by:
Peng Zhang,
University of Maryland, United States

Reviewed by:
Zhi-Qiang Ye,
Peking University, China
Guangrong Qin,
Institute for Systems Biology (ISB), United States

***Correspondence:**
Yuan-Yuan Li
yyli@scbit.org

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *15 December 2019*
Accepted: *04 February 2020*
Published: *25 February 2020*

Citation:
Dai W, Liu J, Liu B, Li Q, Sang Q and Li Y-Y (2020) Systematical Analysis of the Cancer Genome Atlas Database Reveals EMCN/MUC15 Combination as a Prognostic Signature for Gastric Cancer. Front. Mol. Biosci. 7:19. doi: 10.3389/fmolb.2020.00019

Digestive cancers-including gastric cancer (GC), colorectal cancer, hepatocellular carcinoma, esophageal cancer, and pancreatic cancer-accounted for 26% of cancer cases and 35% of cancer deaths worldwide in 2018. It is crucial and urgent to develop biomarkers for the diagnosis, prognosis, and therapeutic benefits of digestive cancers, especially for GC, since the incidence of GC is lower only than lung cancer in China, is hard to detect at an early stage, and is associated with poor prognosis. Mucins, glycoproteins encoded by MUC family genes, act as a part of a physical barrier in the digestive tract and participate in various signaling pathways. Some mucins have been used or proposed as biomarkers for carcinomas, such as MUC16 (CA125) and MUC4. However, there are no systematic investigations on the association of MUC family members with diagnoses and clinical outcomes even though relevant data have been largely accumulated in the past decade. By analyzing transcriptomic and clinical data of digestive cancer samples from TCGA involving colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), and pancreatic adenocarcinoma (PAAD), it was found that expressions levels of *MUC15*, *MUC13,* and *MUC21* were individually associated with survival for digestive cancers, and high expressions of *EMCN* (MUC14) and *MUC15* were correlated with poor survival for STAD. Cox regression analysis indicated the predictive power of an *EMCN/MUC15* combination for overall survival (OS) of GC patients, which was validated on an independent dataset from GEO. EMCN/MUC15 correlated genes were identified to be enriched in cancer-related processes, such as vasculature development, mitosis, and immunity. Therefore, we propose that an *EMCN/MUC15* combination could be a potential prognostic signature for gastric cancer.

Keywords: MUC family, EMCN, MUC15, prognostic, gastric cancer

INTRODUCTION

Digestive cancers are a group of cancers that occur in the digestive tract, and include gastric cancer (GC), colorectal cancer, hepatocellular carcinoma, esophageal cancer, and pancreatic cancer. Digestive cancers accounted for around 26% of cancer cases and 35% of cancer deaths in the world in 2018 (Bray et al., 2018). Among them, the morbidity and mortality of GC in Eastern Asia is much higher than the worldwide average level. In China, the incidence of GC is only lower than lung cancer, and the mortality is third to lung cancer and liver cancer (Chen et al., 2014). Most patients suffering from early stage GC are asymptomatic and always develop distant metastasis at the time of diagnosis (Van Cutsem et al., 2016; Bray et al., 2018). Surgery is the main treatment for GC. Adjuvant or neoadjuvant therapy combined with surgery is commonly used to treat advanced GC, while targeted drugs for advanced GC, such as the HER2 (also known as ERBB2) antibody trastuzumab, and the VEGFR-2 antibody ramucirumab, are still in clinical trials (Van Cutsem et al., 2016). Therefore, developing biomarkers for the diagnosis, prognosis, and therapeutic response of digestive cancers, especially of GC, is necessary and urgent for reducing the mortality rate.

Mucins represent a group of glycoproteins encoded by MUC family genes. These high-molecular weight and filamentous glycoproteins could be classified into secreted mucins and membrane-bound mucins. In the digestive tract, secreted mucins form a mucus layer and act as part of a physical defensive barrier against external aggressive forces (Dekker et al., 2002; Dhanisha et al., 2018); membrane-bound mucins possess membrane specific domains which enable their diverse roles in signaling pathways (Dekker et al., 2002; Dhanisha et al., 2018). Not surprisingly, dysfunction of mucins in their fundamental roles is implicated in disease development at mucosal surfaces (Corfield, 2015; Dhanisha et al., 2018), and some mucins have been reported to display diagnostic or prognostic significance in different types of cancer. For example, MUC16, also known as CA125, is a widely used biomarker for the diagnosis of ovarian cancer (Yonezawa et al., 2011; Jonckheere and Van Seuningen, 2018) and was also found to be over-expressed in several other human malignancies, including pancreas, breast, and lung (Aithal et al., 2018). MUC4 promotes carcinogenetic progression and has been proposed as a promising biomarker for pancreatic, ovarian, esophagus, and lung cancers (Kaur et al., 2013; Jonckheere and Van Seuningen, 2018). MUC15 overexpression is significantly correlated with several types of cancers, including colon cancer, hepatocellular carcinoma, and thyroid cancer (Huang et al., 2009; Nam et al., 2011; Wang et al., 2013; Choi et al., 2018). Moreover, *MUC4*/*MUC16*/*MUC20* high-expression signature was very recently reported to be correlated with poor overall survival (OS) in several types of digestive cancers including pancreatic, colon, and GCs (Jonckheere and Van Seuningen, 2018). However, there are no systematic investigations, so far, on the association of MUC family members with diagnosis, prognosis, and/or therapeutic benefits, even though the Cancer Genome Atlas (TCGA) project is producing massive genomic, transcriptomic, proteomic, and clinical data involving more than 11,000 patients of 33 different types of tumors (Weinstein et al., 2013), and meanwhile, a number of web tools, such as GEPIA (Tang et al., 2017) and cBioPortal for Cancer Genomics (Cerami et al., 2012; Gao et al., 2013), have been developed that enable users to easily and effectively mine TCGA data.

In the present study, by analyzing digestive cancer samples from TCGA involving colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), and pancreatic adenocarcinoma (PAAD), we found that expression levels of *MUC15*, *MUC13,* and *MUC21* were individually associated with survival for all these digestive cancers, and high expressions of *EMCN* (MUC14) and *MUC15* were correlated with poor survival for STAD. Cox regression analysis showed that *EMCN*/*MUC15* combination still exhibited a significant correlation with the OS of GC patients. The prognostic prediction power of signature *EMCN*/*MUC15* was further validated on an independent GC dataset, GSE84437. *EMCN*/*MUC15* top 50 correlated genes were identified to be enriched in cancer-related processes, including vasculature development, mitosis, immunity, and so on. Taken together, we propose *EMCN*/*MUC15* combination as a potential prognostic signature for GC.

MATERIALS AND METHODS

Datasets

Datasets were collected from TCGA[1] and GEO[2] (Barrett et al., 2012). Specifically, gene expression data (TPM, Transcripts Per Kilobase Million) and clinical data for digestive cancers including COAD, ESCA, LIHC, STAD, and PAAD, were analyzed with the online webserver GEPIA 1.0 (Tang et al., 2017). Among them, MUC family mRNA expression data (mRNA expression z-scores, which is based on RNASeqV2 processed and normalized using RSEM) and clinical profiles involving 407 STAD samples were extracted by using an online web tool cBioPortal for Cancer Genomics (Cerami et al., 2012; Gao et al., 2013). Additionally, GSE84437 were extracted from the GEO database, which involves mRNA microarray data and clinical profiles of 433 GC samples.

Survival Analysis

Kaplan–Meier (KM) survival analysis for digestive cancer samples as a whole was carried out by using the webserver GEPIA 1.0 (Tang et al., 2017), and for GC samples (TCGA-STAD from cBioPortal and GSE84437 R package *survival*[3] was used. KM analysis was based on individual gene expression value and survival data. By using the median expression value of a query gene in a certain sample group as a cutoff, the samples were split into high and low expression groups with the expression level of the query gene not less than and less than the cutoff. The Cox proportional hazard model was built by using R package

[1]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[2]https://www.ncbi.nlm.nih.gov/geo/

[3]https://cran.r-project.org/web/packages/survival/

survival, fitted with two genes' expression values for OS or disease free survival (DFS). Similar to the individual gene analysis, the median value of weighted expression value (shortened as WEV) of a gene combination in a certain cohort were used as a group cutoff, where WEV was calculated as the sum of cox-regression coefficient weighted expression value of each gene involved in the combination. Log rank *p*-values, cox proportional hazard ratios (HRs), and HR *p*-values were calculated to compare the survival between two groups split by the median value of gene expression or WEV. A *p*-value of less than 0.05 and HR greater than 1.05 or less than 0.95 suggest statistical significance of the survival difference between high and low groups, which indicates the corresponding gene or gene combination has a prognostic potential.

Gene Co-expression Analysis and Enrichment Analysis

Gene co-expression analysis was carried out using webserver cBioPortal, and the top 25 positively correlated and top 25 negatively correlated genes were selected according to Spearman correlation coefficients, which were taken together and simplified as "top 50 correlated genes" in our results. Here, correlated genes met two criteria: the absolute value of Spearman correlation coefficient is greater than 0.25, and the *p*-value is less than 0.01. Gene set enrichment analysis (GSEA) was performed by using R package *clusterProfiler* (Yu et al., 2012). The pathways enriched for GO (Gene Ontology) (Ashburner et al., 2000; The Gene Ontology Consortium, 2019) were plotted based on the negative logarithm of *p*-value.

RESULTS

MUC15, 13, and 21 Display Prognostic Potential for Digestive Cancer on TCGA

Aiming to assess the prognostic potentials of every MUC gene, KM survival analysis was applied to TCGA digestive cancer samples as a whole involving COAD, ESCA, LIHC, STAD, and PAAD by using the webserver GEPIA 1.0 (Tang et al., 2017). Among the 14 MUC family members with expression data available, the expression levels of MUC1, MUC5AC, MUC6, OVGP1 (MUC9), MUC13, EMCN (MUC14), MUC15, MUC16, MUC17, and MUC21 individually exhibited significant correlations with OS, with HR *p*-values less than 0.05 and HR greater than 1.05 or less than 0.95; similarly, MUC2, MUC3A, MUC12, MUC13, MUC15, MUC17, MUC20, and MUC21 were significantly correlated with DFS (**Table 1** and **Supplementary Figure S1**). *MUC13*, *MUC15*, *MUC17*, and *MUC21* were significant for both OS and DFS, among which MUC15 performed best for OS correlation and the second best for DFS correlation. In comparison, MUC13 displayed the best performance in DFS analysis, while ranked relatively lower (9th) in OS analysis; MUC21 ranked 3rd for OS, and 8th for DFS (**Table 1** and **Supplementary Figure S1**). These indicate that *MUC15* represents a promising candidate for developing strategies for prognosis prediction for digestive cancers.

MUC14 (EMCN) and 15 Display Prognostic Potential for Gastric Cancer on TCGA-STAD

To investigate the prognostic potentials of MUC family genes for STAD, we performed KM survival analysis exclusively on STAD samples from TCGA with R package *survival*. It was found that the expression levels of *EMCN* (MUC14) and *MUC15* individually showed significant correlations with both OS and DFS, and MCAM (MUC18) was significant only with OS (**Table 2**). KM survival plots, together with log rank *p*-values,

TABLE 1 | Survival analysis of TCGA digestive cancer samples for prognostic potentials of MUC family genes.

Gene	HR *p*-value for OS	OS *p*-value rank	HR *p*-value for DFS	DFS *p*-value rank
MUC1	**1.3E−05**	6	0.23	11
MUC2	0.69	14	**5.2E−08**	3
MUC3A	0.49	12	**1.7E−06**	4
MUC5AC	**7.9E−06**	5	0.74	14
MUC6	**2.7E−07**	3	0.41	12
OVGP1 (MUC9)	**0.0021**	7	0.094	10
MUC12	0.58	13	**0.00012**	5
MUC13	**0.032**	9	**2.1E−08**	1
EMCN (MUC14)	**0.044**	10	0.71	13
MUC15	**1.7E−09**	1	**3.6E−08**	2
MUC16	**6.4E−09**	2	0.059	9
MUC17	**0.0053**	8	**0.00051**	6
MUC20	0.36	11	**0.01**	7
MUC21	**9.8E−07**	4	**0.01**	8

OS stands for overall survival and DFS stands for disease free survival (DFS). The p-values less than 0.05 are displayed in bold.

TABLE 2 | Survival analysis of TCGA STAD samples for prognostic potentials of MUC family genes.

Gene	HR *p*-value for OS	HR *p*-value for DFS
MUC1	0.654	0.591
MUC2	0.129	0.364
MUC4	0.9	0.203
MUC5B	0.441	0.753
MUC6	0.67	0.0854
OVGP1 (MUC9)	0.662	0.925
MUC12	0.957	0.637
MUC13	0.0511	0.234
EMCN (MUC14)	**0.00154**	**0.00737**
MUC15	**0.0185**	**0.0141**
MUC16	0.825	0.0975
MUC17	0.145	0.406
MCAM (MUC18)	**0.0167**	0.323
MUC20	0.891	0.62
MUC21	0.224	0.745

OS stands for overall survival and DFS stands for disease free survival. The p-values less than 0.05 are displayed in bold.

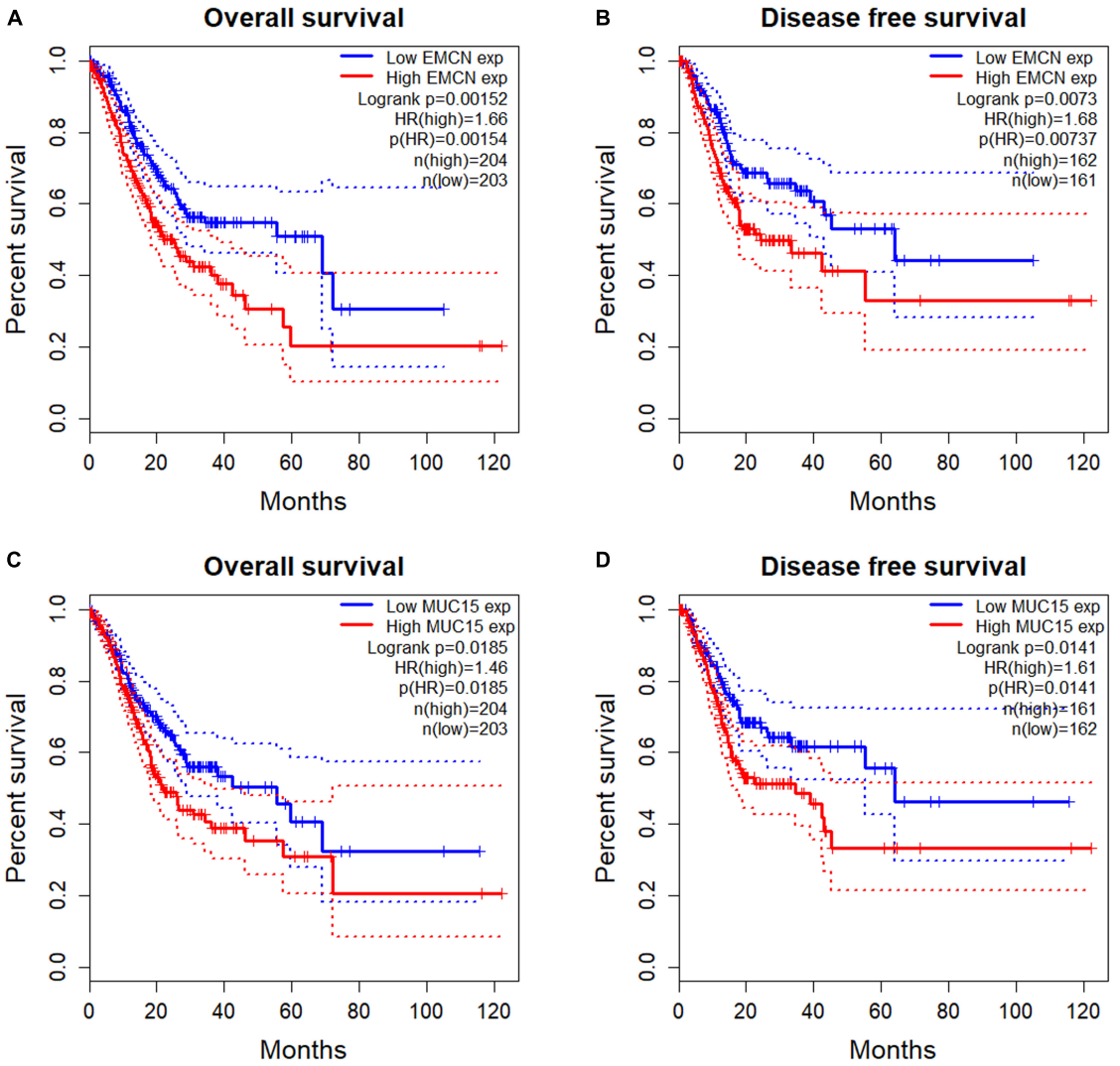


FIGURE 1 | Survival analysis of TCGA STAD samples for prognostic potentials of EMCN (MUC14) and MUC15. **(A)** Overall Survival (OS) of EMCN. **(B)** Disease Free Survival (DFS) of EMCN. **(C)** Overall Survival of MUC15. **(D)** Disease Free Survival of MUC15. Log rank *p*-values, hazard ratios (HRs) and hazard ratio *p*-values were calculated. The 95% confidence intervals for survival time were shown in as dotted lines in the Kaplan–Meier (KM) survival plot.

cox proportional HRs, and HR *p*-values summarized in **Figure 1** indicated that *EMCN* performed better than *MUC15* in both OS and DFS analyses. Overall, *EMCN* and *MUC15* could be potential biomarkers for STAD prognosis.

EMCN/MUC15 Combination Could Serve as Prognostic Signature for Gastric Cancer

So far we have observed that high expressions of both EMCN and MUC15 were associated with poor prognosis in GC, and that EMCN and MUC15 displayed the strongest correlation to survival for GC and digestive cancers, respectively (**Table 2** and **Figure 1**). Thus, we set out to investigate whether *EMCN/MUC15* combination could be a prognostic signature for GC. Cox proportional hazards regression analysis was performed based on the two genes' expression values and OS data derived from TCGA STAD dataset. As expected, the expression of *EMCN/MUC15* combination exhibited significant correlation with OS, with log rank *p*-value of 0.00299 and HR *p*-value of 0.00301 (**Figure 2A**).

We then separately tested the prognostic prediction power of *EMCN*, *MUC15* and their combination on an independent dataset, GSE84437, which involved 433 GC samples. Again, significant results of *EMCN/MUC15* combination (HR = 1.33) were obtained with log rank *p*-value being 0.0419 and HR *p*-value being 0.0413 (**Figure 2B**); while one single gene, *EMCN* (HR *p*-value of 0.0807, HR = 1.27) or *MUC15* (HR *p*-value of 0.156, HR = 0.82), had no significant prognostic prediction power, as shown in **Supplementary Figure S2**. We therefore proposed that *EMCN/MUC15* combination could be a potential prognostic signature for GC.

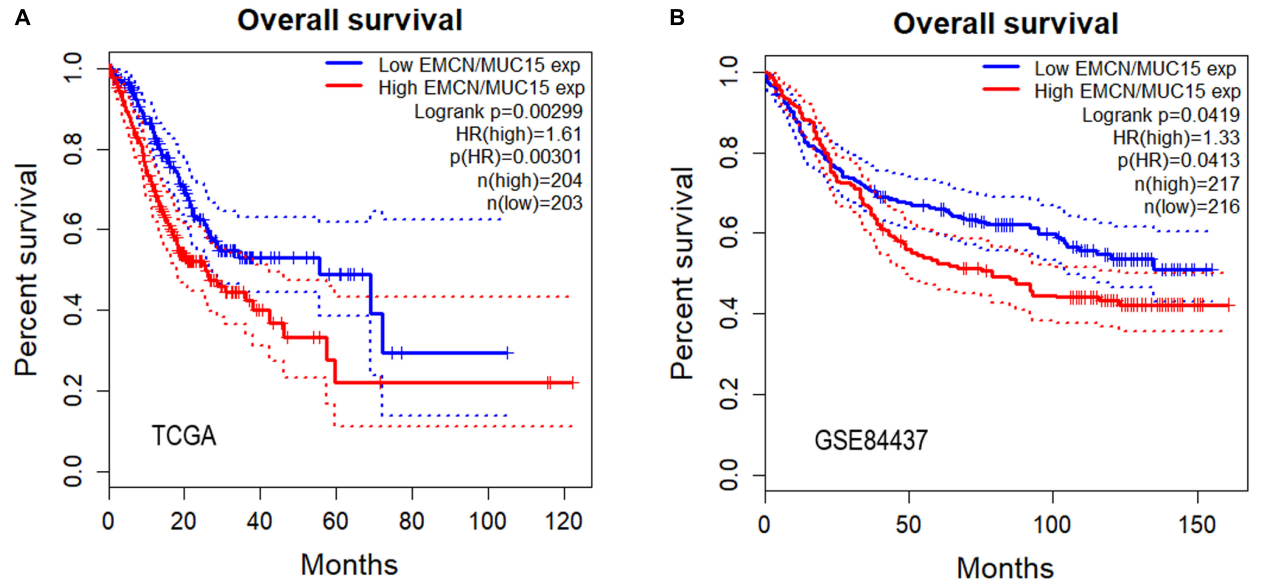


FIGURE 2 | Overall survival analyses of gastric cancer (GC) samples from TCGA STAD **(A)** and GSE84437 **(B)** for predictive power of EMCN/MUC15 signature. Log rank *p*-values, hazard ratios (HRs) and hazard ratio *p*-values were calculated. The 95% confidence intervals for survival time were shown as dotted lines in the Kaplan–Meier survival plot.

EMCN/MUC15 Correlated Genes Are Functionally Enriched in Cancer Related Processes

By using webserver cBioPortal, the top 50 EMCN- (**Table 3**) or MUC15- (**Table 4**) correlated genes were identified based on mRNA expression data of TCGA STAD samples, including the top 25 positively correlated genes and top 25 negatively correlated genes. It is noticeable that there is no intersection between the two top 50 gene lists at all and no co-expression between *EMCN* and *MUC15* (Spearman's Correlation of 0.0264 with *p*-value of 0.592) either, implying the functional complementarity between *EMCN* and *MUC15* and thus the rationality of the combination of the two genes in predicting prognosis for GC.

We then performed functional enrichment analysis with the two top 50 correlated genes as a whole. GSEA identified a total of 22 GO terms (**Figure 3** and **Supplementary Table S1**). Among them, the most significant pathways were associated with vasculature development, such as glomerulus vasculature development and renal system vasculature development. Some enriched pathways are associated with mitosis, such as mitotic sister chromatid segregation and mitotic metaphase plate congression. Some pathways were associated with immunity, such as inflammatory cell apoptotic process and response to interferon-gamma. The other enriched pathways were involved in DNA binding, cell cycle phase transition, cell polarity, phosphatase activity, and side of plasma membrane (**Figure 3** and **Supplementary Table S1**). These indicate that genes correlated with EMCN and MUC15 in GC tend to be enriched in cancer related processes, such as vasculature development, mitosis, and immunity.

DISCUSSION

In the present study, by systematically analyzing mRNA expression and clinical data of TCGA digestive cancer samples and GEO GC samples, we propose MUC15 as a promising candidate for prognosis prediction of digestive cancers, and *EMCN/MUC15* combination as a potential prognostic signature for GC.

Gene signature identification is essentially a process of dimension reduction of high dimensional data. On one hand, a signature involving less features or genes obviously has more practicality; on the other hand, a signature is also expected to have sufficient interpretability, although it is far from achieved. In this sense, a good signature is supposed to consist of orthogonal or mutually exclusive features which are able to hold a testable hypothesis from a systematic viewpoint while also sustaining the robustness and reliability of the signature. However, most current efforts in this field focus on reducing dimension over enhancing explanatory power of the signature. In our work, although *EMCN* and *MUC15* coding genes belong to the same gene family, it is noted that there is no expression correlation between the two genes and no intersection between their top 50 correlated genes, implying the orthogonality and functional complementarity between EMCN and MUC15. As we expected, the combination of *EMCN/MUC15* shows more robust prognostic power than the individual genes in GC according to the testing result implemented on an independent dataset GSE84437. These observations not only support the rationality of the combination of the two genes in predicting prognosis, but also indicate the explanatory power of *EMCN/MUC15* signature, which is supposed to play an important role in the robustness improvement.

TABLE 3 | Top 50 genes correlated with EMCN based on TCGA STAD dataset.

Correlated gene	Cytoband	Spearman correlation	*p*-value
CYYR1	21q21.3	0.931414	2.19E−183
MYCT1	6q25.2	0.929044	1.90E−180
ERG	21q22.2	0.894179	3.19E−146
DIPK2B	Xp11.3	0.887525	4.57E−141
ADGRL4	1p31.1	0.886757	1.71E−140
CD34	1q32.2	0.880383	6.99E−136
TEK	9p21.2	0.873397	4.03E−131
PECAM1	17q23.3	0.871639	5.73E−130
S1PR1	1p21.2	0.870224	4.72E−129
LDB2	4p15.32	0.860092	8.59E−123
RHOJ	14q23.2	0.859913	1.10E−122
CLEC14A	14q21.1	0.854201	2.25E−119
GNG11	7q21.3	0.853027	1.03E−118
EBF1	5q33.3	0.846286	5.16E−115
MMRN2	10q23.2	0.846005	7.29E−115
CLEC1A	12p13.2	0.843416	1.71E−113
CALCRL	2q32.1	0.841594	1.53E−112
LRRC70	5q12.1	0.84015	8.47E−112
MEF2C	5q14.3	0.839354	2.16E−111
ARHGEF15	17p13.1	0.836065	9.86E−110
CDH5	16q21	0.828483	4.80E−106
PALMD	1p21.2	0.828283	5.97E−106
SHE	1q21.3	0.826792	3.01E−105
SPARCL1	4q22.1	0.823121	1.52E−103
JAM2	21q21.3	0.821442	8.85E−103
RAD54L	1p34.1	−0.53926	1.10E−32
CDCA5	11q13.1	−0.53612	2.96E−32
PKP3	11p15.5	−0.53108	1.41E−31
CDCA8	1p34.3	−0.5303	1.79E−31
ZWINT	10q21.1	−0.52817	3.44E−31
KIF2C	1p34.1	−0.52339	1.46E−30
HJURP	2q37.1	−0.51982	4.21E−30
MCM2	3q21.3	−0.51829	6.63E−30
CDT1	16q24.3	−0.51369	2.54E−29
MYO19	17q12	−0.51058	6.24E−29
TONSL	8q24.3	−0.50684	1.82E−28
CCNA2	4q27	−0.5056	2.58E−28
NCAPH	2q11.2	−0.5018	7.48E−28
POC1A	3p21.2	−0.50165	7.81E−28
NELFA	4p16.3	−0.50116	8.95E−28
UBE2T	1q32.1	−0.50026	1.15E−27
POLD2	7p13	−0.49997	1.25E−27
DTL	1q32.3	−0.49967	1.35E−27
PTBP1	19p13.3	−0.49959	1.38E−27
CNOT11	2q11.2	−0.49871	1.76E−27
STIP1	11q13.1	−0.49718	2.69E−27
MAP7	6q23.3	−0.49631	3.41E−27
ESPL1	12q13.13	−0.49591	3.81E−27
TBRG4	7p13	−0.49548	4.29E−27
CDC25A	3p21.31	−0.49474	5.24E−27

Genes mentioned in Discussion section are highlighted in bold and italic.

TABLE 4 | Top 50 genes correlated with MUC15 based on TCGA STAD dataset.

Correlated gene	Cytoband	Spearman correlation	*p*-value
ANO3	11p14.3-p14.2	0.558879	1.82E−35
FSTL4	5q31.1	0.4959	3.82E−27
TMPRSS13	11q23.3	0.469609	3.76E−24
ZNF750	17q25.3	0.464898	1.21E−23
LGALS7	19q13.2	0.454428	1.54E−22
NCCRP1	19q13.2	0.452369	2.52E−22
PCLO	7q21.11	0.449054	5.50E−22
GABRA3	Xq28	0.446711	9.51E−22
DLX3	17q21.33	0.443637	1.94E−21
LIN28B	6q16.3-q21	0.440243	4.21E−21
ADGRV1	5q14.3	0.439028	5.55E−21
USH1G	17q25.1	0.436641	9.52E−21
C12ORF56	12q14.2	0.429849	4.32E−20
RSPO4	20p13	0.428819	5.41E−20
SPAG17	1p12	0.425992	1.00E−19
MARK1	1q41	0.424353	1.43E−19
HTR2C	Xq23	0.423044	1.90E−19
CT45A5	Xq26.3	0.420712	3.13E−19
PRPF40B	12q13.12	0.419994	3.64E−19
C3ORF67	3p14.2	0.419376	4.16E−19
RIPPLY3	21q22.13	0.417437	6.27E−19
CNGB3	8q21.3	0.417398	6.32E−19
ATP6V0A4	7q34	0.413452	1.45E−18
LINC00964	8q24.13	0.412548	1.74E−18
VGLL1	Xq26.3	0.409463	3.30E−18
MCUB	4q25	−0.35985	3.92E−14
FAS	10q23.31	−0.32779	7.51E−12
IRF1	5q31.1	−0.32732	8.08E−12
ZIC2	13q32.3	−0.31402	5.99E−11
CDC42SE2	5q31.1	−0.31243	7.55E−11
HK3	5q35.2	−0.30198	3.37E−10
NUB1	7q36.1	−0.30007	4.41E−10
GBP4	1p22.2	−0.29733	6.45E−10
BBC3	19q13.32	−0.29722	6.55E−10
AIM2	1q23.1-q23.2	−0.29707	6.68E−10
NLRC5	16q13	−0.29669	7.04E−10
MAX	14q23.3	−0.29642	7.30E−10
MTHFD1	14q23.3	−0.29437	9.67E−10
AGAP2	12q14.1	−0.29096	1.54E−09
IFNG	12q15	−0.29068	1.59E−09
RASSF1	3p21.31	−0.28787	2.32E−09
GZMA	5q11.2	−0.28696	2.62E−09
CCL4	17q12	−0.28515	3.32E−09
MAT2B	5q34	−0.28231	4.82E−09
FCGR3A	1q23.3	−0.28226	4.85E−09
THG1L	5q33.3	−0.28207	4.97E−09
TK2	16q21	−0.28202	5.01E−09
PRKX	Xp22.33	−0.27772	8.71E−09
JAK2	9p24.1	−0.27752	8.94E−09
EEF2	19p13.3	−0.2774	9.07E−09

Genes mentioned in Discussion section are highlighted in bold and italic.

EMCN, i.e. MUC14, encodes a membrane-bound protein, endothelial sialomucin or mucin-like sialo glycoprotein, which was reported to inhibit cell and extracellular matrix interaction, interfere with leukocyte-endothelial cell adhesion, and even promote the peritoneal metastasis process of GC cells (Liu et al., 2001; Zahr et al., 2016; Dhanisha et al., 2018;

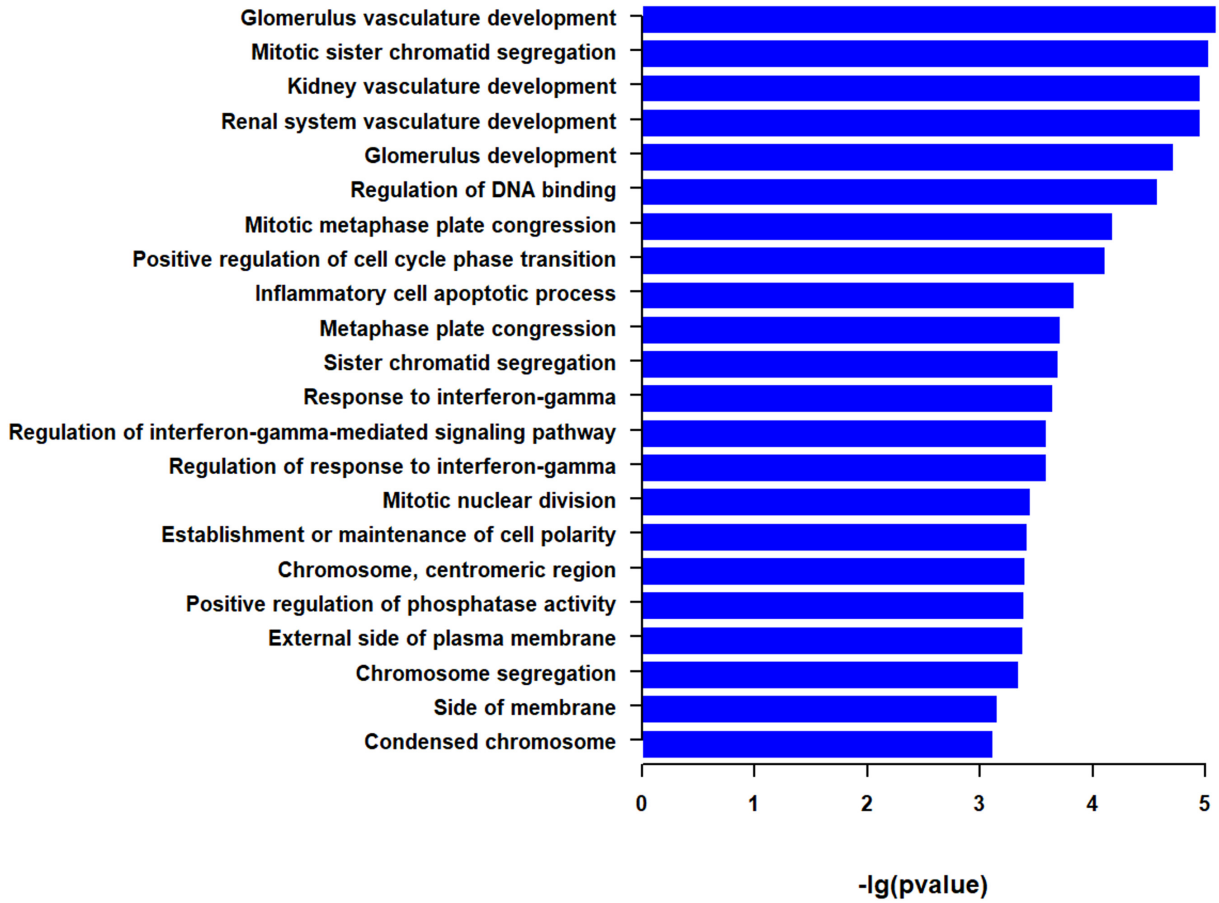


FIGURE 3 | Go terms derived from gene set enrichment analysis (GSEA) for top 50 EMCN- and MUC15-correlated genes. The pathways are ranked by -log *p*-value. The 95% confidence intervals for survival time were shown as dotted lines in the Kaplan–Meier survival plot.

Bao et al., 2019). Among the 22 enriched functions for top 50 EMCN-correlated genes and top 50 MUC15-correlated genes, the most significant one is glomerulus vasculature development that is associated with four *EMCN*/*MUC15* correlated genes including *CD34*, *TEK*, *PECAM1*, and *IFNG* (**Tables 3**, **4** and **Supplementary Table S1**). After carefully checking functional annotations of the four genes, we focused on two cancer relevant genes, *CD34* and *PECAM1*. Both genes are significantly coexpressed with *EMCN* with correlation coefficients of 0.880 and 0.871, respectively (**Table 3**). *CD34*, a marker of vascular endothelial cells, is capable of supporting cell adhesion by increasing surface expression (Nielsen and McNagny, 2008). *PECAM1*, also known as *CD31*, encodes platelet endothelial cell adhesion molecule 1 that is necessary for leukocyte transendothelial migration (TEM) (Dasgupta et al., 2009). It is noteworthy that *EMCN*/*COL4A5*/*CCL11* combination was very recently reported as a prognostic signature for diffuse type GC (Bao et al., 2019). In our study, among MUC family members, *EMCN* exhibits the strongest correlation with survival for GC. Taken together, *EMCN* may play crucial roles in tumorigenesis and progression of GC via cell adhesion and TEM of lymphocytes.

MUC15 also encodes a membrane-bound protein, which could promote cell proliferation, cell-extracellular matrix adhesion, colony forming ability, and invasion in colon cancer cells (Huang et al., 2009). Its overexpression is significantly correlated with diverse cancers (Pallesen et al., 2002; Shyu et al., 2007; Huang et al., 2009; Nam et al., 2011; Wang et al., 2013; Choi et al., 2018). However, it was also found that the expression of *MUC15* decreased in hepatocellular carcinoma cells and negatively regulated metastasis of hepatocellular carcinoma (Wang et al., 2013). This suggests that *MUC15* may perform diverse functions in tumorigenesis and progression. In our study, *MUC15* displays the strongest correlation among the MUC family with survival for digestive cancers and *MUC15* overexpression seems to be a promising candidate for a prognosis biomarker of digestive cancers. Combined with *EMCN*, the two genes provide a potential prognostic signature for GC and show more robustness in the prognostic prediction power than individual genes. As far as we know, the association of *MUC15* with GC is rarely reported.

In summary, we propose *EMCN*/*MUC15* combination as a prognostic signature with mechanistic interpretability. It not only possesses prognostic capability for GC, but also offers clues for further exploring systematic mechanisms of carcinogenesis of GC and other digestive cancers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA-STAD, TCGA-COAD, TCGA-ESCA, TCGA-LIHC, TCGA-PAAD, and GSE84437.

AUTHOR CONTRIBUTIONS

Y-YL and WD designed the study. WD and JL implemented the data analysis. BL, QL, and QS provided the valuable suggestions. JL and WD drafted the manuscript. Y-YL revised the manuscript and coordinated the study. All authors read and approved the final manuscript.

FUNDING

This work was supported by the grants from the National Key R&D Program of China (2018YFC0910500), the National Natural Science Foundation of China (81672736 and 31600750), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01 and 18DZ2294200), and the NIH CPTAC (Cancer Proteomic Tumor Analysis Consortium) program.

ACKNOWLEDGMENTS

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.00019/full#supplementary-material

REFERENCES

Aithal, A., Rauth, S., Kshirsagar, P., Shah, A., Lakshmanan, I., Junker, W. M., et al. (2018). MUC16 as a novel target for cancer therapy. *Expert Opin. Ther. Targets* 22, 675–686. doi: 10.1080/14728222.2018.1498845

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

Bao, B., Zheng, C., Yang, B., Jin, Y., Hou, K., Li, Z., et al. (2019). Identification of subtype-specific three-gene signature for prognostic prediction in diffuse type gastric cancer. *Front. Oncol.* 9:1243. doi: 10.3389/fonc.2019.01243

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

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

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data: figure 1. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.cd-12-0095

Chen, W., Zheng, R., Zhang, S., Zhao, P., Zeng, H., and Zou, X. (2014). Report of cancer incidence and mortality in China, 2010. *Ann. Transl. Med.* 2:61. doi: 10.3978/j.issn.2305-5839.2014.04.05

Choi, C., Thi Thao Tran, N., Van Ngu, T., Park, S. W., Song, M. S., Kim, S. H., et al. (2018). Promotion of tumor progression and cancer stemness by MUC15 in thyroid cancer via the GPCR/ERK and integrin-FAK signaling pathways. *Oncogenesis* 7:85. doi: 10.1038/s41389-018-0094-y

Corfield, A. P. (2015). Mucins: a biologically relevant glycan barrier in mucosal protection. *Biochim. Biophys. Acta* 1850, 236–252. doi: 10.1016/j.bbagen.2014.05.003

Dasgupta, B., Dufour, E., Mamdouh, Z., and Muller, W. A. (2009). A novel and critical role for tyrosine 663 in platelet endothelial cell adhesion molecule-1 trafficking and transendothelial migration. *J. Immunol.* 182, 5041–5051. doi: 10.4049/jimmunol.0803192

Dekker, J., Rossen, J. W. A., Büller, H. A., and Einerhand, A. W. C. (2002). The MUC family: an obituary. *Trends Biochem. Sci.* 27, 126–131. doi: 10.1016/s0968-0004(01)02052-2057

Dhanisha, S. S., Guruvayoorappan, C., Drishya, S., and Abeesh, P. (2018). Mucins: structural diversity, biosynthesis, its role in pathogenesis and as possible therapeutic targets. *Crit. Rev. Oncol.* 122, 98–122. doi: 10.1016/j.critrevonc.2017.12.006

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:l1. doi: 10.1126/scisignal.2004088

Huang, J., Che, M. I., Huang, Y. T., Shyu, M. K., Huang, Y. M., Wu, Y. M., et al. (2009). Overexpression of MUC15 activates extracellular signal-regulated kinase 1/2 and promotes the oncogenic potential of human colon cancer cells. *Carcinogenesis* 30, 1452–1458. doi: 10.1093/carcin/bgp137

Jonckheere, N., and Van Seuningen, I. (2018). Integrative analysis of the cancer genome atlas and cancer cell lines encyclopedia large-scale genomic databases: MUC4/MUC16/MUC20 signature is associated with poor survival in human carcinomas. *J. Transl. Med.* 16:259. doi: 10.1186/s12967-018-1632-1632

Kaur, S., Kumar, S., Momi, N., Sasson, A. R., and Batra, S. K. (2013). Mucins in pancreatic cancer and its microenvironment. *Nat. Rev. Gastroenterol. Hepatol.* 10, 607–620. doi: 10.1038/nrgastro.2013.120

Liu, C., Shao, Z.-M., Zhang, L., Beatty, P., Sartippour, M., Lane, T., et al. (2001). Human endomucin is an endothelial marker. *Biochem. Biophys. Res. Commun.* 288, 129–136. doi: 10.1006/bbrc.2001.5737

Nam, K.-H., Noh, T.-W., Chung, S.-H., Lee, S. H., Lee, M. K., Won Hong, S., et al. (2011). Expression of the membrane mucins MUC4 and MUC15, potential markers of malignancy and prognosis, in papillary thyroid carcinoma. *Thyroid* 21, 745–750. doi: 10.1089/thy.2010.0339

Nielsen, J. S., and McNagny, K. M. (2008). Novel functions of the CD34 family. *J. Cell Sci.* 121, 4145–4145. doi: 10.1242/jcs.03504

Pallesen, L. T., Berglund, L., Rasmussen, L. K., Petersen, T. E., and Rasmussen, J. T. (2002). Isolation and characterization of MUC15, a novel cell membrane-associated mucin. *Eur. J. Biochem.* 269, 2755–2763. doi: 10.1046/j.1432-1033.2002.02949.x

Shyu, M. K., Lin, M. C., Shih, J. C., Lee, C. N., Huang, J., Liao, C. H., et al. (2007). Mucin 15 is expressed in human placenta and suppresses invasion of trophoblast-like cells in vitro. *Hum. Reprod.* 22, 2723–2732. doi: 10.1093/humrep/dem249

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, W98–W102. doi: 10.1093/nar/gkx247

The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K., and Prenen, H. (2016). Gastric cancer. *Lancet* 388, 2654–2664. doi: 10.1016/s0140-6736(16)30354-30353

Wang, R. Y., Chen, L., Chen, H. Y., Hu, L., Li, L., Sun, H. Y., et al. (2013). MUC15 inhibits dimerization of EGFR and PI3K–AKT signaling and is associated with aggressive hepatocellular carcinomas in patients. *Gastroenterology* 145, 1436–1448.e1-12. doi: 10.1053/j.gastro.2013.08.009

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Yonezawa, S., Higashi, M., Yamada, N., Yokoyama, S., Kitamoto, S., Kitajima, S., et al. (2011). Mucins in human neoplasms: clinical pathology, gene expression and diagnostic application. *Pathol. Intern.* 61, 697–716. doi: 10.1111/j.1440-1827.2011.02734.x

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Zahr, A., Alcaide, P., Yang, J., Jones, A., Gregory, M., dela Paz, N. G., et al. (2016). Endomucin prevents leukocyte–endothelial cell adhesion and has a critical role under resting and inflammatory conditions. *Nat. Commun.* 7:10363. doi: 10.1038/ncomms10363

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

Copyright © 2020 Dai, Liu, Liu, Li, Sang 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 25 March 2020
doi: 10.3389/fmolb.2020.00034

A Qualitative Transcriptional Signature for Predicting Extreme Resistance of ER-Negative Breast Cancer to Paclitaxel, Doxorubicin, and Cyclophosphamide Neoadjuvant Chemotherapy

Yanhua Chen[1,2†], Hao Cai[3†], Wannan Chen[2,4], Qingzhou Guan[5,6,7], Jun He[1,2], Zheng Guo[1,2] and Jing Li[1,2*]*

[1] Fujian Key Laboratory of Medical Bioinformatics, Department of Bioinformatics, The School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China, [2] Key Laboratory of Gastrointestinal Cancer (Fujian Medical University), Ministry of Education, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China, [3] Medical Big Data and Bioinformatics Research Center, First Affiliated Hospital of Gannan Medical University, Ganzhou, China, [4] Fujian Key Laboratory of Tumor Microbiology, Department of Medical Microbiology, Fujian Medical University, Fuzhou, China, [5] Henan Key Laboratory of Chinese Medicine for Respiratory Disease, Henan University of Chinese Medicine, Zhengzhou, China, [6] Co-construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan & Education Ministry of P.R. China, Henan University of Chinese Medicine, Zhengzhou, China, [7] Academy of Sciences of Chinese Medicine, Henan University of Chinese Medicine, Zhengzhou, China

For estrogen receptor (ER)-negative breast cancer patients, paclitaxel (P), doxorubicin (A) and cyclophosphamide (C) neoadjuvant chemotherapy (NAC) is the standard therapeutic regimen. Pathologic complete response (pCR) and residual disease (RD) are common surrogate measures of chemosensitivity. After NAC, most patients still have RD; of these, some partially respond to NAC, whereas others show extreme resistance and cannot benefit from NAC but only suffer complications resulting from drug toxicity. Here we developed a qualitative transcriptional signature, based on the within-sample relative expression ordering (REO) of gene pairs, to identify extremely resistant samples to PAC NAC. Using gene expression data for ER-negative breast cancer patients including 113 pCR samples and 137 RD samples from four datasets, we selected 61 gene pairs with reversal REO patterns between the two groups as the resistance signature, denoted as NR61. Samples with more than 37 signature gene pairs that had the same REO patterns within the extremely resistant group were defined as having extreme resistance; otherwise, they were considered responders. In the GSE25055 and GSE25065 dataset, the NR61 signature could correctly identify 44 (97.8%) of the 45 pCR samples and 22 (95.7%) of the 23 pCR samples as responder samples, respectively; it also identified 13 (16.9%) of 77 RD samples and 8 (21.1%) of 38 RD samples as extremely resistant samples, respectively. Survival analysis showed that the distant relapse-free survival (DRFS) time of the 14 extremely resistant cases was significantly shorter than that of the 108 responders ($P < 0.01$; HR = 3.84; 95% CI = 1.91–7.70) in GSE25055. Similar results were obtained in GSE25065. Moreover, in

OPEN ACCESS

Edited by:
Adil Mardinoglu, King's College London, United Kingdom

Reviewed by:
Austin Shull, Presbyterian College, United States
Ankita Thakkar, Burke Medical Research Institute, United States

***Correspondence:**
Jing Li
haerbinlisa@hotmail.com
Zheng Guo
guoz@ems.hrbmu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *03 November 2019*
Accepted: *13 February 2020*
Published: *25 March 2020*

Citation:
Chen Y, Cai H, Chen W, Guan Q, He J, Guo Z and Li J (2020) A Qualitative Transcriptional Signature for Predicting Extreme Resistance of ER-Negative Breast Cancer to Paclitaxel, Doxorubicin, and Cyclophosphamide Neoadjuvant Chemotherapy. Front. Mol. Biosci. 7:34. doi: 10.3389/fmolb.2020.00034

the integrated data of the two datasets with 94 responders and 21 extremely resistant samples identified from RD patients, the former had significantly longer DRFS than the latter ($P < 0.01$; HR = 2.22; 95% CI = 1.26–3.90). In summary, our signature could effectively identify patients who completely respond to PAC NAC, as well as cases of extreme resistance, which can assist decision-making on the clinical therapy for these patients.

Keywords: breast cancer, neoadjuvant chemotherapy, pathological complete response, extreme resistance, relative expression ordering

INTRODUCTION

Breast cancer is a common malignancy with the highest incidence and mortality among females (Ferlay et al., 2015; Jia et al., 2015). A standard regimen for estrogen receptor (ER)-negative breast cancer patients, accounting for 30% of breast cancer patients, is paclitaxel (P), doxorubicin (A), and cyclophosphamide (C) neoadjuvant chemotherapy (NAC) (Jemal et al., 2011). However, the heterogeneity of breast cancer can result in different responses to standard therapy (Rouzier et al., 2005; Carey et al., 2007).

In clinical practice, a pathologic complete response (pCR) is defined as a non-viable invasive cancer in the breast and lymph nodes after the completion of NAC, indicating a complete response to NAC and a favorable outcome (Kaufmann, 2003; Guarneri, 2006; Mieog et al., 2007; Liedtke et al., 2008; Rastogi et al., 2008). However, the proportion of pCR is quite low among patients accepting NAC, and most patients have residual disease (RD) (Popovici et al., 2010). Among patients with RD, accounting for a great proportion of patients treated with NAC, most are partial responders, whereas the others are extremely resistant to NAC. These extremely resistant patients cannot benefit from NAC, but only suffer complications resulting from the toxic effects of NAC. More seriously, these patients may lose the best treatment time because clinicians would evaluate the feasibility of curative or conservative surgery after finishing chemotherapy and a series of examinations (Helene et al., 2012). Therefore, the development of a predictor to identify extremely resistant patients who cannot benefit from NAC is of great significance.

Up to now, many signatures have been developed for pCR prediction (Hess et al., 2006; Thuerigen, 2006; Liedtke et al., 2009), but few studies have focused on the identification of extremely resistant patients. The pCR predictive signatures are based on risk scores summarized from quantitative transcriptional data, which have poor reproducibility (Borst and Wessels, 2010; Tabchy et al., 2010; Zhang et al., 2013; Qi et al., 2016) due to widespread batch effects and the uncertain quality of clinical samples. Although several reported quantitative transcriptional disease signatures – including AlloMap® (Pham et al., 2010) – have been approved by the Food and Drug Administration, the tissue samples must be sent to specific laboratories for measurement with strict quality control, which limits their wider applications in clinical practice.

In contrast, qualitative transcriptional signatures based on within-sample relative expression orderings (REOs) are found to be robust against experimental batch effects and can be directly applied to samples at the individualized level (Eddy et al., 2010; Wang et al., 2013; Chen et al., 2017). REO is a binary variable based on comparing the mRNA levels within a single pair of genes (Geman et al., 2004). For a gene pair (i, j), the REO pattern represents whether the expression level of i is higher or lower than that of j in the sample. Additionally, REO-based signatures are also highly robust against common factors that lead to the failure of quantitative transcriptional signatures in clinical applications, such as varied proportions of tumor epithelial cells (Cheng et al., 2017), amplification bias for minimum specimens (Liu et al., 2017), and partial RNA degradation (Freidin et al., 2012; Chen et al., 2017). Thus, the REO-based method is more practicable for tissue biopsy samples acquired by fine needle aspiration (FNA) or core biopsy (CBX) prior to NAC.

Based on the within-sample REOs of gene pairs, Zhang et al. (2013) have developed a pCR predictor and a prognosis predictor for RD to identify patients who might benefit from NAC. However, this study did not consider the impact of ER subtype. ER-positive patients with good prognosis have a lower pCR rate than that of ER-negative patients with poor prognosis (Guarneri, 2006). Meanwhile, for the same set of breast cancer patients approximately 20% of ER states determined by immunohistochemical (IHC) methods gave different results for different pathologists (Dubowitz, 1991; Arihiro et al., 2007), especially for weak ER-positive samples (Hammond et al., 2010; Sheffield et al., 2016), which may reduce the accuracy of pCR prediction. Thus, we re-determined the ER status of breast cancer patients using the 112-gene-pair signature for ER status developed by Cai et al. (2018) to reduce misjudgments of ER status by IHC.

In this study, we used the gene expression data of ER-negative samples reclassified by the 112-gene-pair signature to identify a qualitative transcriptional signature consisting of 61 gene pairs to predict patients with extreme resistance to PAC chemotherapy. Our signature was well-verified in two independent datasets with survival information.

MATERIALS AND METHODS

Data and Preprocessing

We collected four expression datasets (GSE20194, GSE20271, GSE41998, and MDA133) including 250 IHC-determined ER-negative breast cancer patients in total, who accepted PAC NAC, from the Gene Expression Omnibus (GEO[1]) and the MD

[1]http://www.ncbi.nlm.nih.gov/geo/

Anderson Cancer Center[2] databases. In the datasets of GSE20194 and GSE20271, we only used the expression data of patients who received paclitaxel followed by fluorouracil (F), doxorubicin [or epirubicin (E)], and cyclophosphamide. In the GSE41998 and MDA133 datasets, the treatment regimens for these patients were PFAC and PAC, respectively.

Two other independent expression datasets (GSE25055 and GSE25065) were used to evaluate whether there was a difference in survival between the responsive and the resistant groups. The treatment regimen for patients in GSE25055 was PAC or PA and the treatment regimen for patients in GSE25065 was PA.

Although PAC NAC is a very common chemotherapeutic regimen, doctors design individual drug delivery schemes for each patient, depending on their condition. Some patients received 6 months of NAC including PFAC (e.g., GSE20194), whereas others received sequential NAC starting with 4 cycles of AC administered every 3 weeks, followed by paclitaxel weekly for 12 weeks (e.g., GSE41998). In this study, we only considered the drug type and not the dose of each drug or the duration of chemotherapy. The clinical characteristics for each dataset are summarized in **Table 1**.

For the Affymetrix array data, the raw intensity files (.cel), downloaded from the GEO database were processed using the Robust Multichip Average algorithm (RMA) for background adjustment without quantile normalization. The probe identity documents (ID) were mapped to the Entrez gene ID according to the corresponding platform annotation files. If a probe did not map to a gene or was mapped to multiple genes, the data for this probe were deleted. If multiple probes mapped to the same gene, the arithmetic mean of the expression values for the multiple probes was taken as the final expression value for this gene.

ER Status Re-determination

We used the 112-gene-pair signature developed by Cai et al. (2018) to reclassify the ER-negative samples. An IHC-determined ER-negative patient was reclassified as ER-negative if more than 68 gene pairs match the REOs of the ER-negative signature.

Identification of the REO-Based Resistant Signature

For each RD (or pCR) sample, the gene expression profile was first converted into a rank profile according to measured expression levels in ascending order (the lowest expression value corresponds to the smallest rank). Then, pair-wise combinations of all genes were examined to determine the REO pattern of each gene pair within the sample. The within-sample REO of a gene pair (i, j) has only two possibilities, $G_i > G_j$ or $G_i < G_j$, where G_i and G_j denote the expression values. If the number of RD samples with a certain REO pattern ($G_i > G_j$ or $G_i < G_j$) is significantly more than expected by chance, we define this gene pair as a stable gene pair of RD samples; stable gene pairs of pCR samples are defined in a similar manner. The significance of a REO in RD (or pCR) samples was determined using a binomial test (Bahn, 1969) as follows:

$$P = 1 - \sum_{i=0}^{k-1} \binom{n}{i} {p_0}^i (1 - p_0)^{(n-i)} \quad (1)$$

where n is the total number of samples with the RD (or pCR) status, k denotes the number of samples that have a certain REO pattern ($G_i > G_j$ or $G_i < G_j$), and p_0 denotes the probability of observing a gene pair with a certain REO pattern by chance (here, $p_0 = 0.5$). Then the P-values were adjusted using the Benjamini and Hochberg (1995) procedure to control the false discovery rate (FDR).

We then defined stable-reversal gene pairs as pairs that had a significantly stable REO pattern in the pCR samples and RD samples, respectively, but had a reversal REO pattern between the two groups.

Significant Majority Vote Rule

Based on the stable-reversal gene pairs between the pCR and RD, we developed an extremely resistant signature. A sample was identified as an extremely resistant sample, if the number of REOs of the signature gene pairs matching that of the extremely resistant group was significantly more than expected by chance. The threshold for identifying an extremely resistant sample was determined according to a binomial test as follows:

$$P = 1 - \sum_{i=0}^{k-1} \binom{n}{i} {p_0}^i (1 - p_0)^{(n-i)} \quad (2)$$

where n is the number of signature gene pairs and k is the number of gene pairs in the sample that match the REOs

[2]https://bioinformatics.mdanderson.org/pubdata.html

TABLE 1 | Description of all datasets collected in this study.

Usage	Dataset	Regimen	ER-negative sample size	pCR	RD	With DRFS information
Training	GSE20194 Popovici et al. (2010)	T-FA(E)C[a]	114	46	68	no
	GSE20271 Tabchy et al. (2010)	T-FA(E)C	79	19	60	no
	MDA133 Hess et al. (2006)	T-FAC	51	27	24	no
	GSE41998 Horak et al. (2013)	T-AC[b]	48	29	19[d]	no
Validation	GSE25055 Hatzis et al. (2011)	T-AC;TA[c]	129	45	84	yes
	GSE25065 Hatzis et al. (2011)	TA	68	23	45	yes

[a] T-FA(E)C paclitaxel (T) followed by fluorouracil (F), doxorubicin (A) [or epirubicin (E)] and cyclophosphamide (C). [b] T-AC doxorubicin (A) and cyclophosphamide (C) followed by paclitaxel (T). [c] TA taxane (T) and anthracycline (A) based regimens. [d] PD and SD samples representing tumor residuals screened from RD samples.

for the extremely resistant group. p_0 (here, p_0 = 0.5) is the probability of a gene pair having a certain REO pattern in a sample by chance.

Survival Analysis

The distant relapse-free survival (DRFS), defined as the time from surgery to distant recurrence or the final documented date (censored), was used as a surrogate assessment of tumor response status (Liedtke et al., 2008). A log-rank test was used to assess the difference between the Kaplan–Meier estimates of DRFS in two different groups. The univariate Cox proportional-hazards regression model was used to calculate the hazard ratios (HRs) and their 95% confidence intervals (CIs).

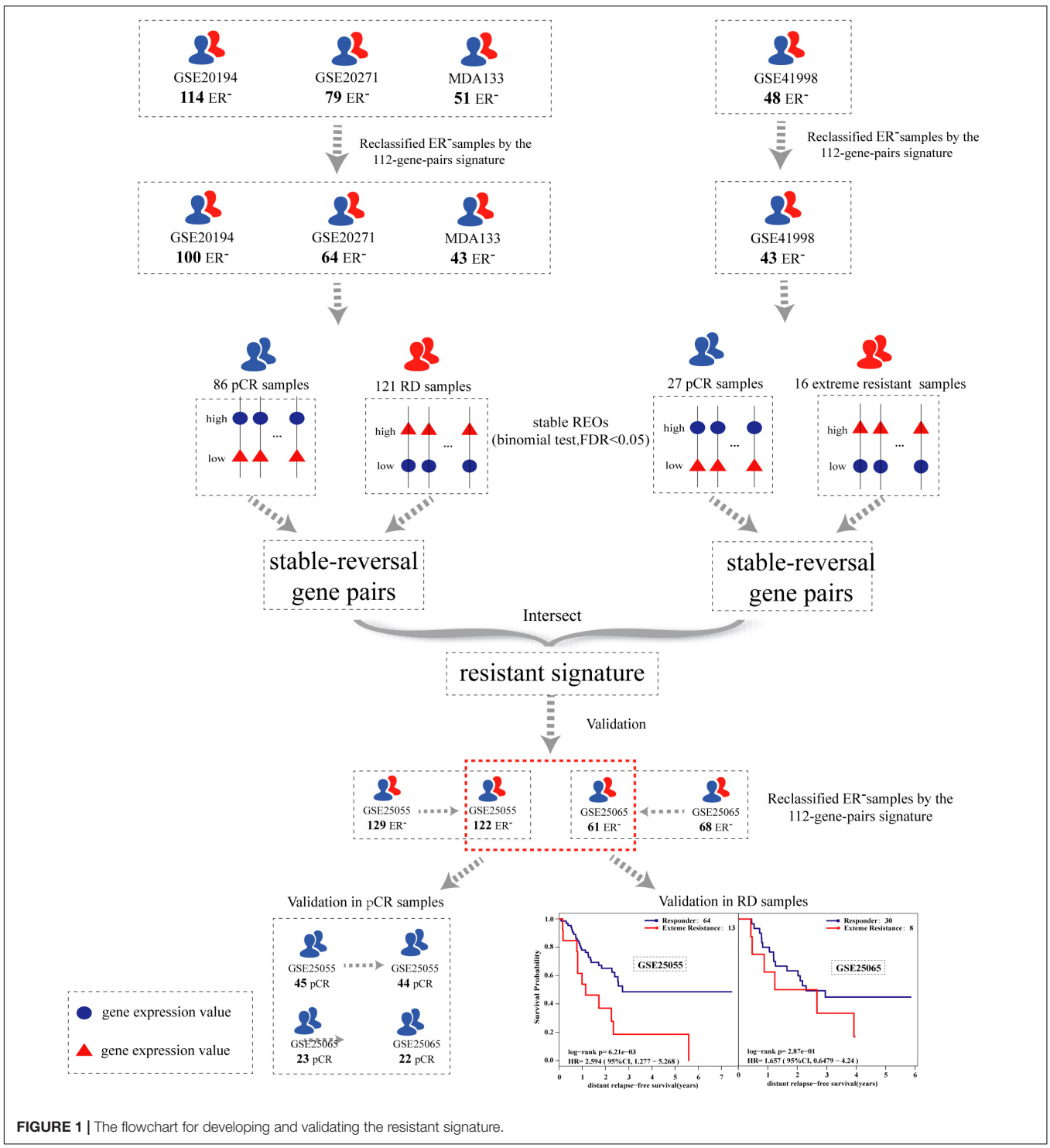


FIGURE 1 | The flowchart for developing and validating the resistant signature.

RESULTS

Development of the Resistant Signature

The flowchart of the process used for developing and validating the resistance signature is shown in **Figure 1**. Using the 112-gene-pair signature for the ER status, 100, 64, 43, and 43 samples were re-determined as ER-negative samples from the GSE20194, GSE20271, MDA133, and GSE41998 dataset, respectively (**Table 2**).

To identify an extremely resistant signature, we first extracted 169,222 gene pairs with stable (binomial test, FDR < 0.05) but reversed REOs between the pCR and RD group from the integrated data of the GSE20194, GSE20271, and MDA133 datasets, with 86 pCR samples and 121 RD samples in total. We then used the GSE41998 dataset to optimize the signature. Beside the pCR and RD category, the GSE41998 dataset also provided an evaluation of drug response criteria in solid tumors (RECIST), which divided the patients into four groups: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD) (Watanabe et al., 2009). Among them, SD and PD indicated that the tumor area of the patients did not improve significantly but was increased after receiving NAC; therefore, we screened out the PD and SD samples from the RD samples as extremely resistant samples. We then extracted 30,588 stable-reversal gene pairs between the 16 extremely resistant samples and 27 pCR samples. Finally, 61 gene pairs that had consistent REO patterns between the above two lists of stable-reversal gene pairs were selected as the resistance signature, denoted as NR61. The details of NR61 are shown in **Table 3**. Each gene pair has a certain REO pattern in extremely resistant patients and a reversal REO pattern in responsive patients. Based on the significant majority vote rule (see section "Materials and Methods"), if more than 37 gene pairs ($P < 0.05$) of NR61 showed the same REO patterns as observed in extreme resistance, the sample was identified as extremely resistant; otherwise, it was considered a responder.

Researchers (Tong et al., 2015) have proven that if two different regimens share one or several drugs, then the overlaps of the clinically relevant drug resistance genes (CRGs) for the two different regimens should be considered as the CRGs for the shared drug(s). We speculated that this is similar for clinically relevant drug resistance gene pairs (CRGPs). In this study, overlapping gene pairs between 169,222 CRGPs of PFAC and 30,588 CRGPs of PAC should thus be the CRGPs for PAC. Thus, the resistant signature that we developed is specific for predicting PAC resistance. However, the NR61 signature should be applicable for patients who received any combination of P, A, and C, as the extremely resistant patients identified by this signature showed multidrug resistant to P, A, and C.

TABLE 2 | The ER-negative samples reclassified by the 112-gene-pairs signature from the IHC-determined ER-negative samples.

Usage	Dataset	Reclassified ER-negative sample size	pCR	RD
Training	GSE20194	100	43	57
	GSE20271	64	19	45
	MDA133	43	24	19
	GSE41998	43	27	16[a]
Validation	GSE25055	122	45	77
	GSE25065	61	23	38

[a] *PD and SD samples representing tumor residuals screened from RD samples.*

TABLE 3 | Each pair of genes in NR61.

Gene 1	Gene 2	Gene 1	Gene 2	Gene 1	Gene 2
UBTD1	ACOX1	LAMA5	SMARCC1	LMAN2L	COBL
NOVA2	ADCY2	GPX5	SST	RBP3	PART1
TAS2R1	APLP1	GRIA1	SST	DNAH2	PART1
GCLM	ARL1	TMEM165	VAMP7	GCLM	CHIC2
RASL11B	CKB	STC1	VEGFB	ACKR4	TOX3
PTPRA	RCAN1	TGFB3	AKAP1	ATHL1	SLC43A3
AGPAT2	GNAQ	TPST2	SPOP	SLC30A1	ERGIC2
PLD2	GTF2F1	LETM1	SORBS2	FAM69A	CRNKL1
B4GALT5	HNRNPF	COPZ1	IQGAP1	VRK2	DPM3
NOS2	HSPA1L	GCLM	PRPF4B	SFXN3	CPVL
GCLM	IPO5	MICALL2	PRPF4B	TRAFD1	BSPRY
SYDE1	MAZ	IGSF3	ZNHIT3	GCLM	C5orf22
GTF2H3	NFIB	FAM69A	RNF14	SLC12A4	LMO3
MCAM	NFIB	TJP1	GCC2	SULT2B1	LMO3
TBC1D4	NFIB	LETM1	TOX4	SLC28A1	LMO3
NUAK1	NFIB	C10orf2	DCAF7	SEMA3F	FKBPL
FZD6	NUCB2	CPA3	SPAG5	GCLM	AIDA
CIAPIN1	PBX1	DUOX1	OR7E14P	P3H1	C17orf70
TRIT1	PBX3	MPPE1	KAT7	KREMEN2	IL17RC
GCLM	RBM3	GLTSCR1	XPO7	SMURF1	KLHL22
MMP16	RYR3				

The expression patterns of these gene pairs is Gene 1 > Gene 2 in the extremely resistant samples and Gene 1 < Gene 2 in the response samples.

Performance of the NR61 Signature

In the GSE25055 and GSE25065 datasets, 122 and 61 ER-negative breast cancer samples were separately re-determined using the 112-gene-pair signature (**Table 2**) and were used to validate NR61. Among these re-determined ER-negative samples, the NR61 signature could correctly classify 44 (97.8%) out of 45 pCR samples and 22 (95.7%) out of 23 pCR samples as responder samples, which showed that NR61 can effectively identify patients that completely responded to PAC NAC.

The survival analysis was then used to validate the NR61 signature, assuming that the responsive patients have a better prognosis than the extremely resistant patients. First, the survival analysis was performed in all re-determined ER-negative breast cancer patients. In the GSE25055 dataset with 122 ER-negative breast cancer patients, 108 and 14 patients were classified as responders and extremely resistant, respectively. The extremely resistant patients had a significantly shorter DRFS time than the responders (log-rank $P < 0.01$; HR = 3.84; 95% CI = 1.91–7.70; **Figure 2A**). Similar results were obtained in the GSE25065 dataset with 61 ER-negative breast cancer patients (log-rank $P < 0.01$; HR = 3.07; 95% CI = 1.28–7.36; **Figure 2B**).

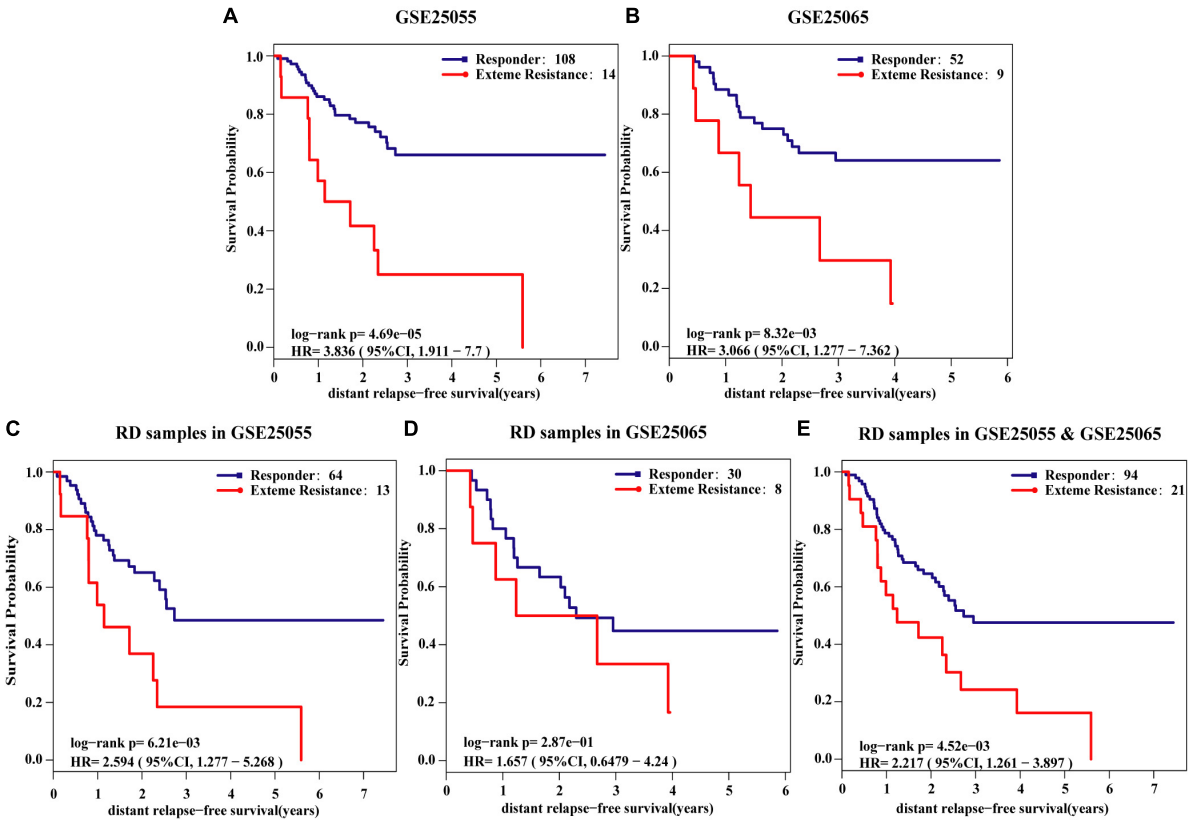


FIGURE 2 | Kaplan–Meier estimates of distant relapse-free survival (DRFS). DRFS curves for responder and extreme resistance in **(A)** GSE25055; **(B)** GSE25065; **(C)** RD samples of GSE25055; **(D)** RD samples of GSE25065; **(E)** integrated RD samples of GSE25055 and GSE25065.

To avoid the impact of pCR patients and further demonstrate the poor prognosis of extremely resistant patients, the survival analysis was limited to RD patients. For the 77 RD patients with ER-negative breast cancer in the GSE25055 dataset, 64 and 13 patients were classified into the responder and extremely resistant group, respectively. Survival analysis showed that the DRFS time of the extremely resistant group was significantly shorter than that of the responders (log-rank $P < 0.01$; HR = 2.59; 95% CI = 1.28–5.27; **Figure 2C**). In the RD samples from the GSE25065 dataset with 38 ER-negative breast cancer patients, the NR61 signature stratified 30 and 8 RD patients into the responder and extremely resistant groups, respectively. Survival analysis showed that, in this dataset with a small sample size (low statistical power) there was a trend of difference in the DRFS time between the responder and extremely resistant groups (log-rank $P = 0.29$; HR = 1.66; 95% CI = 0.65–4.24; **Figure 2D**). In the integrated data of the two datasets with 94 responders and 21 extremely resistant patients in total, identified from the RD patients, the former had significantly longer DRFS than the latter (log-rank $P < 0.01$; HR = 2.22; 95% CI = 1.26–3.90; **Figure 2E**). This result indicates that NR61 well divided the RD samples into two categories, one of which is the PR to NAC with a good prognosis, whereas the other has a very poor prognosis, which is extreme resistance.

In the validation dataset where a number of patients received PA rather than PAC, the extremely resistant patients who were multidrug resistant to P, A, and C should have a poor prognosis, while those patients who were resistant to P and A but sensitive to C would be classified into the response group. The patients under a treatment of PA should have a poor prognosis. However, we still observed the extremely resistant group had a significantly longer survival than the responder group, even though the latter included some patients with poor prognosis.

All the above results indicate that the extremely resistant patients identified by NR61 cannot benefit from the PAC NAC treatment. The NR61 signature is thus expected to assist physicians in choosing treatment plans for ER-negative breast cancer patients in clinical practice. If a patient is judged as extremely resistant by NR61, accepting PAC NAC may only cause complications and a loss of the best time for surgery. For these patients, other chemotherapeutic regimens or direct surgery might be more sensible options.

Correlation of NR61 With HER2 Status and PAM50 Subtype

As HER2 status is an important prognostic and predictive signature, we evaluated the performance of NR61 in HER2−

and HER2+ patients, respectively. We found that all 61 ER-negative breast cancer samples of GSE25065 were HER2− and that 115 of 122 ER-negative breast cancer samples in GSE25055 were HER2−. In the 115 HER2+ patients, the survival of the responder group and the extremely resistant group as identified by NR61 was significantly different (**Supplementary Figure 1A**). A similar result was found in 74 RD samples (**Supplementary Figure 1B**). For another seven patients in GSE25055, the HER2 status of three patients was positive, and four patients were uncertain. All of these seven patients were classified into the responder group by NR61.

In addition, we counted the number of samples for each PAM50 subtype in the responder group and in the extremely resistant group as reclassified by NR61. In the responder group of the GSE25055 dataset, the sample sizes of Normal, Luminal A, Luminal B, HER2, and basal-like were 1, 0, 0, 1, and 12, respectively. In the extremely resistant group, the sample size corresponding to these subtypes was 8, 0, 0, 8, and 92, respectively. A Chi-square test showed no statistically significant difference in the sample distribution of each PAM50 subtype between the responder group and the extremely resistant group ($P = 0.9986$, **Supplementary Figure 2A**). Similar results were also observed in the GSE25065 dataset ($P = 0.1213$, **Supplementary Figure 2B**). This result indicates that there is no relationship between NR61 and PAM50 subtypes.

DISCUSSION

In this study, we developed a qualitative drug resistant signature (NR61), which could well predict the ER-negative breast cancer patients who were extremely resistant to PAC NAC. Based on this signature a total of 183 ER-negative patients in the two validation datasets could be divided into responder and extremely resistant patients. Our research showed that the DRFS time of the extremely resistant group was significantly shorter than that of the responders. Patients identified with extreme resistance should be recommended other treatment schemes to avoid unnecessary suffering and expenses. Additionally, this signature can correctly identify almost all patients who can completely respond to PAC NAC.

Our qualitative transcriptional signature based on the within-sample REOs is robust against batch effects (Chen et al., 2017; Cheng et al., 2017; Guan et al., 2018) and could be performed for the individual analysis of ER-negative breast cancer, which is of great value for clinical application. The REO-based signatures may lose some so-called "subtle" quantitative information of gene expression measurements. However, the "subtle" quantitative information is often unreliable because it is affected by the high variations in measurement and batch effects, the proportions of tumor epithelial cells in clinical tissue samples, partial RNA degradation during specimen preparation and storage, and the amplification bias of low-input RNA (Freidin et al., 2012; Chen et al., 2017). Even the ratios of the expression values of gene pairs are affected by batch effects (Loven et al., 2012; Qi et al., 2016). Thus, this apparent disadvantage of REO analysis is actually a unique advantage in terms of robustness (Chen et al., 2017).

In this study, PD and SD samples screened from RD samples were defined as extremely resistant to PAC NAC. The pCR-RD system is based on microscopic observation and a large number of patients are diagnosed with RD. However, in the image-based RECIST system (Watanabe et al., 2009), PD is defined as at least a 20% increase in the sum of the diameters of target lesions after receiving NAC, and SD is defined as neither sufficient shrinkage to qualify for PR (at least a 30% decrease in the sum of diameters of target lesions) nor sufficient increase to qualify for PD, both of which are less sensitive to NAC. Therefore, it is reasonable to screen PD and SD patients from RD patients as extremely resistant, as used in this study. However, there is only one dataset with information of both RECIST and pCR-RD. Thus, we used the DRFS to evaluate whether the identified patients can benefit from PAC NAC.

Due to the lack of RNA-seq data with suitable drug response information, we only tested the NR61 signature in the microarray data measured on the Affymetrix platform. In future, we will collect the breast cancer expression data from the RNA-seq platform to optimize our signature, in order to improve its cross-platform ability.

CONCLUSION

In summary, the NR61 signature could be used to robustly identify patients who are extremely resistant to PAC NAC among ER-negative breast cancer patients. These patients are highly unlikely to benefit from the PAC NAC regimen and should thus be recommended other therapeutic regimens. The clinical value of the NR61 signature for extreme resistance to the PAC NAC regimen thus deserves further validation.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this study could be found in the Gene Expression Omnibus (GSE20194, GSE20271, GSE41998, GSE25055, and GSE25065) and MD Anderson Cancer Center (https://bioinformatics.mdanderson.org/public-datasets/).

AUTHOR CONTRIBUTIONS

ZG and YC conceived the project. YC and HC performed the computational experiments. YC and JL designed the data analyses. QG and JH interpreted the data. YC, HC, and ZG wrote the manuscript. All authors contributed to the preparation of the manuscript, and read and approved the final manuscript.

FUNDING

This study was funded by the National Natural Science Foundation of China (Grant Nos. 81872396, 61602119, 81602738, and 81903186) and the Joint Technology Innovation Fund of Fujian Province (Grant Nos. 2016Y9044 and 2017Y9109).

ACKNOWLEDGMENTS

We would like to acknowledge the resources at the GEO and MD Anderson Cancer Center that facilitated this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.00034/full#supplementary-material

REFERENCES

Arihiro, K., Umemura, S., Kurosumi, M., Moriya, T., Oyama, T., Yamashita, H., et al. (2007). Comparison of evaluations for hormone receptors in breast carcinoma using two manual and three automated immunohistochemical assays. *Am. J. Clin. Pathol.* 127, 356–365. doi: 10.1309/d7w4-ml22-w228-1484

Bahn, A. K. (1969). Application of binomial distribution to medicine: comparison of one sample proportion to an expected proportion (for small samples). Evaluation of a new treatment. Evaluation of a risk factor. *J Am Med Womens Assoc* 24, 957–966. doi: 10.1309/d7w4-ml22-w228-1484

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300.

Borst, P., and Wessels, L. (2010). Do predictive signatures really predict response to cancer chemotherapy? *Cell Cycle* 9, 4836–4840. doi: 10.4161/cc.9.24.14326

Cai, H., Guo, W., Zhang, S., Li, N., Wang, X., Liu, H., et al. (2018). A qualitative transcriptional signature to reclassify estrogen receptor status of breast cancer patients. *Breast Cancer Res. Treat.* 170, 271–277. doi: 10.1007/s10549-018-4758-2

Carey, L. A., Dees, E. C., Sawyer, L., Gatti, L., Moore, D. T., Collichio, F., et al. (2007). The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clin. Cancer Res.* 13, 2329–2334. doi: 10.1158/1078-0432.ccr-06-1109

Chen, R., Guan, Q., Cheng, J., He, J., Liu, H., Cai, H., et al. (2017). Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. *Oncotarget* 8:6652. doi: 10.18632/oncotarget.14257

Cheng, J., Guo, Y., Gao, Q., Li, H., Yan, H., Li, M., et al. (2017). Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget* 8, 30265–30275. doi: 10.18632/oncotarget.15754

Dubowitz, V. (1991). A new muscle journal for the nineties. *Neuromuscul. Disord.* 1, 1–2. doi: 10.1016/0960-8966(91)90036-r

Eddy, J. A., Sung, J., Geman, D., and Price, N. D. (2010). Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol. Cancer Res. Treat.* 9, 149–159. doi: 10.1177/153303461000900204

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi: 10.1002/ijc.29210

Freidin, M. B., Bhudia, N., Lim, E., Nicholson, A. G., Cookson, W. O., and Moffatt, M. F. (2012). Impact of collection and storage of lung tumor tissue on whole genome expression profiling. *J. Mol. Diagn.* 14, 140–148. doi: 10.1016/j.jmoldx.2011.11.002

Geman, D., D'avignon, C., Naiman, D. Q., and Winslow, R. L. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.* 3:Article19. doi: 10.2202/1544-6115.1071

Guan, Q., Yan, H., Chen, Y., Zheng, B., Cai, H., He, J., et al. (2018). Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer. *BMC Genomics* 19:99. doi: 10.1186/s12864-018-4446-y

Guarneri, V. (2006). Prognostic value of pathologic complete response after primary chemotherapy in relation to hormone receptor status and other factors. *J. Clin. Oncol.* 24, 1037–1044. doi: 10.1200/JCO.2005.02.6914

Hammond, M. E., Hayes, D. F., Wolff, A. C., Mangu, P. B., and Temin, S. (2010). American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J. Oncol. Pract.* 6, 195–197. doi: 10.1016/j.breastdis.2010.10.048

Hatzis, C., Pusztai, L., Valero, V., Booser, D. J., Esserman, L., Lluch, A., et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 305, 1873–1881. doi: 10.1001/jama.2011.593

Helene, K.-G., Laurence, V., and Marie-Christine, B. (2012). Predictive value of neoadjuvant chemotherapy failure in breast cancer using FDG-PET after the first course. *Breast Cancer Res. Treat.* 131, 517–525. doi: 10.1007/s10549-011-1832-4

Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.* 18, 203–203. doi: 10.1016/S1043-321X(07)80272-4

Horak, C. E., Pusztai, L., Xing, G., Trifan, O. C., Saura, C., Tseng, L. M., et al. (2013). Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. *Clin. Cancer Res* 19, 1587–1595. doi: 10.1158/1078-0432.CCR-12-1359

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Jia, M., Zheng, R., Zhang, S., Zeng, H., Zou, X., and Chen, W. (2015). Female breast cancer incidence and mortality in 2011. *China. J. Thorac. Dis.* 7, 1221–1226. doi: 10.3978/j.issn.2072-1439.2015.05.15

Kaufmann, M. (2003). International expert panel on the use of primary (Preoperative) systemic treatment of operable breast cancer: review and recommendations. *J. Clini Oncol. Off. J. Am. Soc. Clin. Oncol.* 21, 2600–2608. doi: 10.1200/JCO.2003.01.136

Liedtke, C., Hatzis, C., Symmans, W. F., Desmedt, C., Haibe-Kains, B., Valero, V., et al. (2009). Genomic grade index is associated with response to chemotherapy in patients with breast cancer. *J. Clin. Oncol.* 27, 3185–3191. doi: 10.1200/JCO.2008.18.5934

Liedtke, C., Mazouni, C., Hess, K. R., Andre, F., Tordai, A., Mejia, J. A., et al. (2008). Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J. Clin. Oncol.* 26, 1275–1281. doi: 10.1200/jco.2007.14.4147

Liu, H., Li, Y., He, J., Guan, Q., Chen, R., Yan, H., et al. (2017). Robust transcriptional signatures for low-input RNA samples based on relative expression orderings. *BMC Genomics* 18:913. doi: 10.1186/s12864-017-4280-7

Loven, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., et al. (2012). Revisiting global gene expression analysis. *Cell* 151, 476–482. doi: 10.1016/j.cell.2012.10.012

Mieog, J. S. D., Hage, J. A. V. D., and Velde, C. J. H. V. D. (2007). Preoperative chemotherapy for women with operable breast cancer. *Cochrane Database. Syst. Rev.* 2:CD005002. doi: 10.1002/14651858.CD005002.pub2

Pham, M. X., Teuteberg, J. J., Kfoury, A. G., Starling, R. C., Deng, M. C., Cappola, T. P., et al. (2010). Gene-expression profiling for rejection surveillance after cardiac transplantation. *N. Engl. J. Med.* 362, 1890–1900. doi: 10.1056/NEJMoa0912965

Popovici, V., Chen, W., Gallas, B. G., and Hatzis, C. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* 12:R5. doi: 10.1186/bcr2468

Qi, L., Chen, L., Yang, L., Yuan, Q., Pan, R., Zhao, W., et al. (2016). Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. *Brief. Bioinform* 2, 233–242. doi: 10.1093/bib/bbv064

Rastogi, P., Anderson, S. J., Bear, H. D., Geyer, C. E., Kahlenberg, M. S., Robidoux, A., et al. (2008). Preoperative chemotherapy: updates of National surgical adjuvant breast and bowel project Protocols B-18 and B-27. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 26, 778–785. doi: 10.1200/JCO.2007.15.0235

Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., and Pusztai, L. (2005). Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin. Cancer Res.* 11, 5678–5685. doi: 10.1158/1078-0432.CCR-04-2421

Sheffield, B. S., Kos, Z., Asleh-Aburaya, K., Wang, X. Q., Leung, S., Gao, D., et al. (2016). Molecular subtype profiling of invasive breast cancers weakly positive for estrogen receptor. *Breast Cancer Res. Treat.* 155, 483–490. doi: 10.1007/s10549-016-3689-z

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

Thuerigen, O. (2006). Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary Breast Cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 24, 1839–1845. doi: 10.1200/JCO.2005.04.7019

Tong, M., Zheng, W., Lu, X., Ao, L., Li, X., Guan, Q., et al. (2015). Identifying clinically relevant drug resistance genes in drug-induced resistant cancer cell lines and post-chemotherapy tissues. *Oncotarget* 6, 41216–41227. doi: 10.18632/oncotarget.5649

Wang, H., Zhang, H., Dai, Z., Chen, M. S., and Yuan, Z. (2013). TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. *BMC Med Genomics* 6(Suppl. 1):S3. doi: 10.1186/1755-8794-6-S1-S3

Watanabe, H., Okada, M., Kaji, Y., Satouchi, M., Sato, Y., Yamabe, Y., et al. (2009). [New response evaluation criteria in solid tumours-revised RECIST guideline (version 1.1)]. *Gan To Kagaku Ryoho* 36, 2495–2501. doi: 10.1016/j.ejca.2008.10.026

Zhang, L., Hao, C., Shen, X., Hong, G., and Guo, Z. (2013). Rank-based predictors for response and prognosis of neoadjuvant taxane-anthracycline-based chemotherapy in breast cancer. *Breast Cancer Res. Treat.* 139, 361–369. doi: 10.1007/s10549-013-2566-2

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

Copyright © 2020 Chen, Cai, Chen, Guan, He, Guo 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 07 April 2020
doi: 10.3389/fmolb.2020.00047

Systems Biology Integration and Screening of Reliable Prognostic Markers to Create Synergies in the Control of Lung Cancer Patients

Aman Chandra Kaushik[1,2], Aamir Mehmood[2], Dong-Qing Wei[2] and Xiaofeng Dai[1]**

[1] Wuxi School of Medicine, Jiangnan University, Wuxi, China, [2] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

This study aims to achieve a clearer and stronger understanding of all the mechanisms involved in the occurrence as well as in the progression of lung cancer along with discovering trustworthy prognostic markers. We combined four gene expression profiles (GSE19188, GSE19804, GSE101929, and GSE18842) from the GEO database and screened the commonly differentially expressed genes (CDEGs). We performed differentially expressed group analysis on CDEGs, alteration and mutational analysis, and expression level verification of core differential genes. Systems biology discoveries in our examination are predictable with past reports. Curiously, our examination revealed that screened biomarker adjustments, for the most part, coexist in lung cancer. After screening 952 CDEGs, we found that the up-regulation of neuromedin U (NMU) and GTSE1 in the case of lung cancer is related to poor prognosis. On the other hand, FOS CDKN1C expression is associated with poor prognosis and is responsible for the down-regulation of CDKN1C and FOS. Changes in these qualities are on free pathways to lung cancer and are not usually of combined quality variety. Even though biomarkers were related to both survival occasions in our examination, it gives us another point of view while playing out the investigation of hereditary changes and clinical highlights employing information mining. Based on our results, we found potential and prospective clinical applications in GTSE1, NMU, FOS, and CDKN1C to act as prognostic markers in case of lung cancer.

Keywords: lung cancer, TCGA, survival, systems biology, prognostic biomarkers

OPEN ACCESS

Edited by:
Cheng Zhang,
KTH Royal Institute of Technology, Sweden

Reviewed by:
Yichao Zheng,
Zhangzhou Municipal Hospital of Fujian Medical University, China
Fu Hui,
Tianjin University of Traditional Chinese Medicine, China

****Correspondence:***
Dong-Qing Wei
dqwei@sjtu.edu.cn
Xiaofeng Dai
1281423490@qq.com

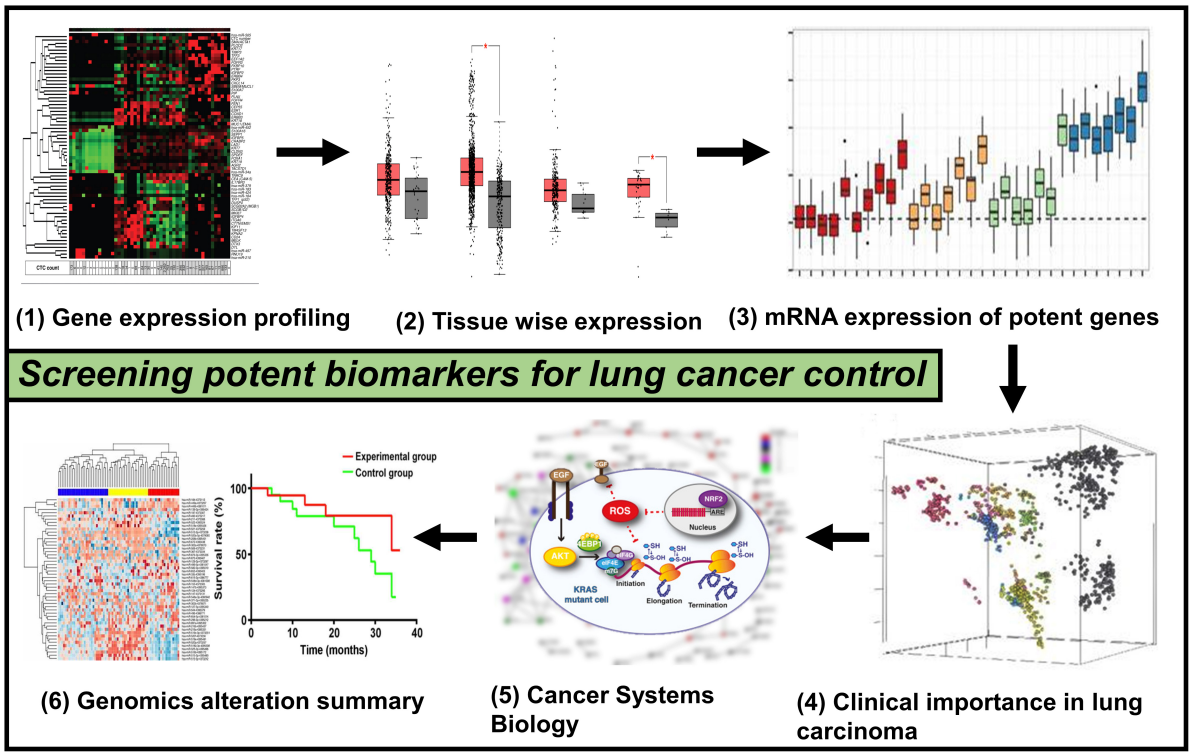
Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *21 December 2019*
Accepted: *05 March 2020*
Published: *07 April 2020*

Citation:
Kaushik AC, Mehmood A, Wei D-Q and Dai X (2020) Systems Biology Integration and Screening of Reliable Prognostic Markers to Create Synergies in the Control of Lung Cancer Patients. Front. Mol. Biosci. 7:47. doi: 10.3389/fmolb.2020.00047

INTRODUCTION

Lung cancer cases are among the most reported tumors that have peak sickness and impermanence rates worldwide (Siegel et al., 2019). Various factors could result in the development of such condition; however, smoking, which may be regular or passive, radon gas, asbestos fibers, familial predisposition, lung diseases, and air pollution is the main cause. The symptoms may vary from person to person and case to case, but some regular signs involve ongoing cough, blood-streaked saliva, puffed hoarseness, or some slight infection that keeps on coming. As per the various diagnostic types, lung cancer is majorly grouped into two types, namely, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), of which nearly 90% of the cases account for



GRAPHICAL ABSTRACT | This picture depicts the whole methodology's pipeline used for the current work that involves genomics and systems biology to scan for potential lung cancer biomarkers.

NSCLC (Siegel et al., 2018). Although a lot of advancement has taken place in the field of science and technology and medical methods, the 5-year existence percentage of patients who have SCLC continues to be merely 30% due to factors such as tumor recurrence and metastasis. Finding tumor markers with accurate prognosis can further aid in understanding the direction and mechanism of tumor progression and provide patients with personalized treatment plans to advance the overall endurance of sufferers.

In the present scenario, the integration of high-throughput omics technology and bioinformatics analysis continues to be a significant and effective research method in clinical research to discover target molecules associated with diseases. Moreover, it is considered to be a reliable technique for bioinformatics analysis of the integration of a huge quantity of omics data to discover targets that have potential application importance, for instance, researches on colorectal cancer (Luca et al., 2019), oral cancer (Di et al., 2019; Pan et al., 2019), ovarian cancer (Hu et al., 2019), osteosarcoma (Ma et al., 2019), and lung cancer (Feng et al., 2019). In the present study, the first step we did was to collect expression profiles of NSCLC mRNA from the GEO database and inspected them for genes that are commonly differentially expressed. The systems biology workbench was used to execute analysis of gene network and visual analysis of network on genes that are commonly differentially expressed, and after this, we chose the major differentially expressed genes from commonly differentially expressed genes Common differentially expressed genes (cDEGs). The prognostic value of major differential genes in NSCLC patients was then analyzed based on metainspection (**Graphical Abstract**).

MATERIALS AND METHODS

Data Retrieval and Acquisition

The gene's countenance contours of [GSE19188 (Hou et al., 2010), GSE19804 (Lu et al., 2010), GSE101929 (Mitchell et al., 2017), and GSE18842 (Sanchez-Palencia et al., 2011)] were acquired from GEO database. All the microarray data belonging to GSE19188, GSE19804, GSE101929, and GSE18842 exist on GPL570 Platforms (Affymetrix Human Genome U133 Plus 2.0 Array), which was inclusive of 54 NSCLC tissues and 49 normal lung matching tissues, 60 tissues of NSCLC and 60 normal lung matching tissues, 30 tissues of NSCLC and 34 normal lung matching tissues, and 46 tissues of NSCLC and 45 normal lung matching tissues, respectively. We used (Teng et al., 2016) R package to plot the gene expression (transcript per million) on the basis of gene length for normalization, where total reads were mapped to gene $\times 10^3$/gene length in base pairs (bp) shown in **Figure 2**.

mRNA Expression Profiling

The microarrays expression in lung tissues was used to identify those genes that are differentially expressed (DEGs). The lung

tissues used in this case were both of the tumor and coordinated head-to-head non-cancerous. In order to scan for genes associated with cancer, detailed literature review was considered, integrating bioinformatics approaches. The GTSE1, neuromedin U (NMU), FOS, and CDKN1C were obtained for confirming the aspirant's gene transcription and expression degree. In the end, Fisher's test was conducted to analyze the connection among the pathological characteristics and aspirant genes.

Functional Identification of GTSE1, NMU, FOS, and CDKN1C Using Systems Biology Approach

To plan and perform the GTSE1, NMU, FOS, and CDKN1C and their associated genes in biological mechanism, the computational systems biology workbench was employed. The required data for both the direct and indirect linkages were collected by conducting a detailed survey of the available information. A complete biological pathway is formed showing all the interacting species. In this constructed pathway, the entities are signified by nodes, while the edges represent the linkage in between a pair of nodes that reveals their close association. A particular concentration was assigned for the time course simulation of the biochemical pathway that was noted from previous reports. Differentially expressed genes of each and every series were taken for analyses, where criterion was set to adjusted $P < 0.01$ and $|\log FC| > 1$. The analysis began by first screening the DEGs present in each of the dataset with standard $P < 0.01$.

Kaplan–Meier Survival Breakdown

The analysis of prognostic value of CDEGs in the patients who were suffering from lung cancer was done by Kaplan–Meier, with 54,000 genes in 21 tumors. In this study, we used information on lung cancer from the database to analyze the prognostic value, inclusive of 675 squamous cell carcinomas and 866 adenocarcinomas.

Validation of CDEGs Expression Levels and Correlation Analysis

For the screening of promising CDEGs, we verified their countenance points in 969 lung cancer models and 735 paracancerous samples where cutoff values were set as $|\log FC| > 1$ and $P < 0.01$. In addition, the correlation between countenance levels and clinical stage of tumors was assessed

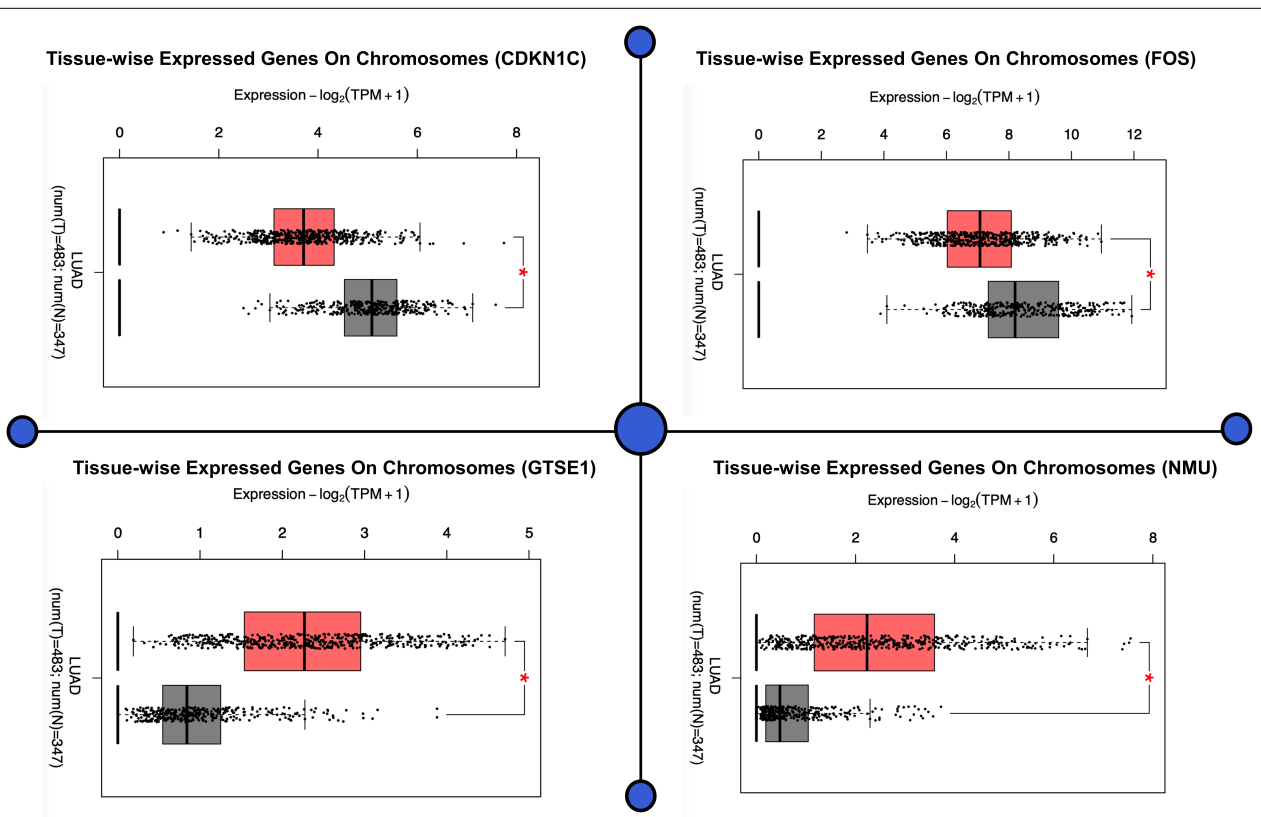


FIGURE 1 | The demonstration of the potential medical value of GTSE1, NMU, FOS, and CDKN1C in the prognosis of patients suffering from lung cancer. This figure shows the difference in the expression pattern on chromosomes of all the four genes in a tissue (tissue-wise expressed genes on chromosomes). The figure demonstrates that four CDEGs were wholly coherent with the results obtained in all the four sets of data, which further demonstrates that the prognostic value of these four targets for lung cancer is good. As per the above results, FOS, GTSE1, CDKN1C, and NMU have good prognostic values for patients with lung cancer, as shown in the figure.

as well, investigating whether valuable CDEGs are independent influencing factors influencing the prognosis of lung cancer.

RESULTS

Genomic Landscape of GTSE1, NMU, FOS, and CDKN1C in Prognosis of Patients Suffering From Lung Cancer

The demonstration of the potential medical value of GTSE1, NMU, FOS, and CDKN1C in the prognosis of patients suffering from lung cancer was done by us in order to verify the levels of expression of FOS, CDKN1C, NMU, and GTSE1. The results illustrated that expression levels of GTSE1 ($P < 0.05$) and NMU ($P < 0.05$) were notably up-regulated, and expression levels of CDKN1C ($P < 0.05$) and FOS ($P < 0.05$) were notably down-regulated in lung adenocarcinoma (LUAD), as well as in lung squamous cell carcinoma, and significantly statistically significant. The results of verification demonstrated that four CDEGs were wholly coherent with the results obtained in all the four sets of data, which further demonstrate that the prognostic value of these four targets for lung cancer is good. As per the above results, FOS, GTSE1, CDKN1C, and NMU have good prognostic values for patients with lung cancer, as shown in **Figure 1**.

Expression of CDEGs

We executed differential expression screening on four datasets of lung cancer (GSE101929, GSE18842, GSE19188, and GSE19804), which we collected from the GEO database. These datasets consist of 3,179, 3,162, 2,601, and 1,404 genes, which are differentially expressed, of which 952 genes were CDEGs, inclusive of 256 up-regulated genes and 696 down-regulated genes shown in **Figure 2**.

mRNA Expression Profiling

Analyzing the mRNA expression microarrays showed FOS, GTSE1, CDKN1C, and NMU are greatly harbored by the patients of lung cancer and are observed to be differentially expressed (fold change $\geq$2.0) of these genes. Of these differentially expressed qualities, GTSE1 and NMU were overexpressed, whereas the down-regulated qualities were CDKN1C and FOS. The computational techniques used here revealed that CDKN1C and FOS, which are down-regulated genes in lung cancer-positive

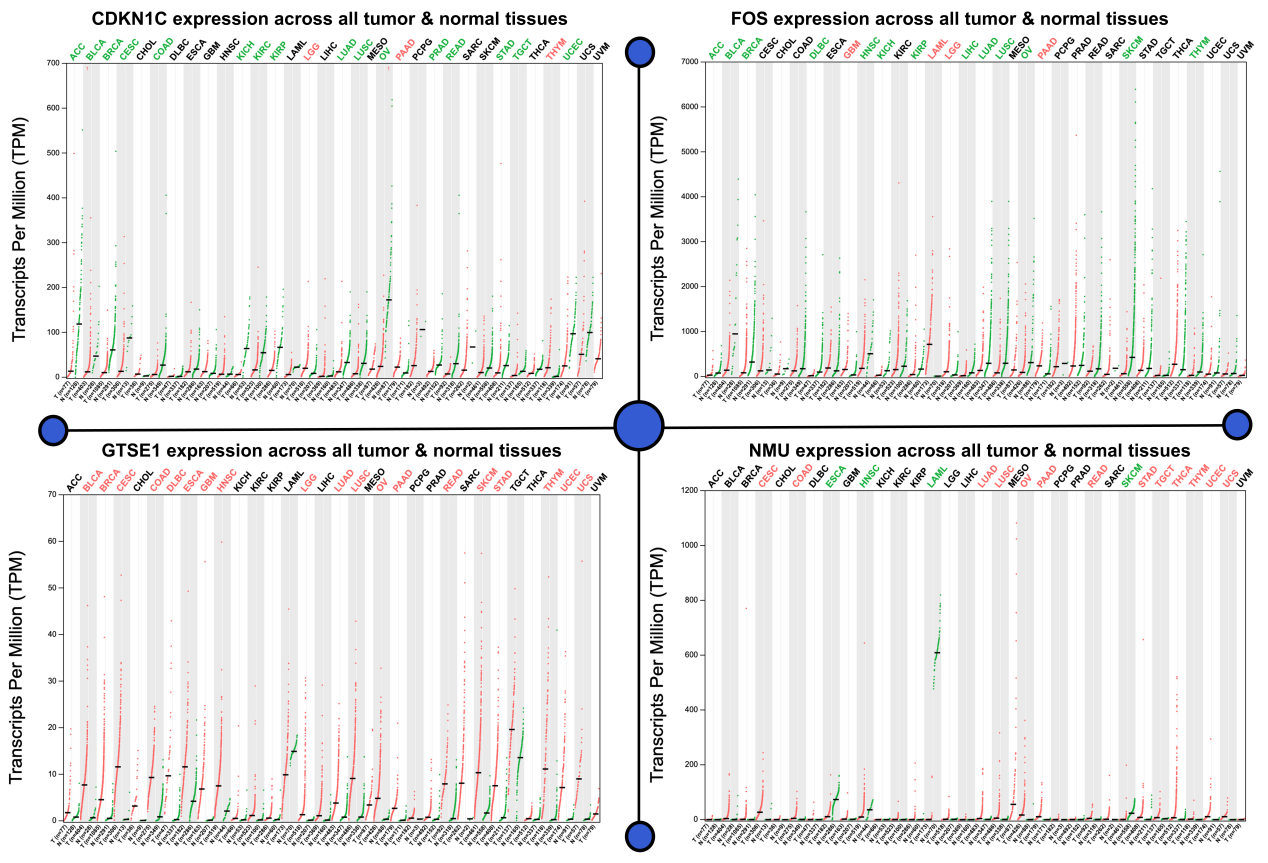


FIGURE 2 | Expression analysis of potent biomarkers. Differential expression screening on four datasets of lung cancer GSE101929, GSE18842, GSE19188, and GSE19804. Gene expression (transcript per million) on the basis of gene length for normalization, where total reads were mapped to gene $\times 10^3$/gene length in bp. The level of expression accords that all the considered tumor samples are given for each up-regulated and down-regulated genes where the CDKN1C is considered to be the least expressed, whereas FOS is observed to be highly overexpressed.

patients, are closely associated with the compulsive situation and are responsible for the regulation number of biological comebacks. Information assessment regarding CDKN1C and FOS exposed its down-regulation as a separate countenance outline in lung cancer happening in other geographical locations shown in **Figure 3**. The association breakdown of experimental features disclosed down-regulation of CDKN1C and FOS and its correlation with the development of the diseased condition. In the end, Fisher's test was conducted to analyze the connection between the pathological characteristics and aspirant genes shown in **Figure 4**.

Functional Identification of DEGs Using Systems Biology Approach

Systems biology is an interdisciplinary field that involves computational and mathematical investigations for modeling a complex biological mechanism. It mainly addresses the linkages formed within the living systems and tracks changes upon the incorporation of any non-native event using an all-inclusive technique. Narrowing down to cancer systems biology, it mainly involves the use of systems biology's technology in cancer research, for the sake of examining an ailment as a challenging adaptive system having evolved characteristics at various biological parameters, as shown in **Figure 5**.

To obtain the synopsis of the role and contribution of 952 CDEGs in the enhancement of lung cancer, we observed that biological processes significantly related to the advancement of lung cancer, angiogenesis, cells' outside medium association, collagen catabolic progression, and positive regulation of angiogenesis. Moreover, cell components such as cells' outside medium, protein-rich medium surrounding the cell, and cells' outer area; collagen trimer and extracellular region; molecular function; integrin binding; and protein-binding and heparin-binding activities of metalloendopeptidase were also found to be tightly linked with the progression of the lung cancer. The

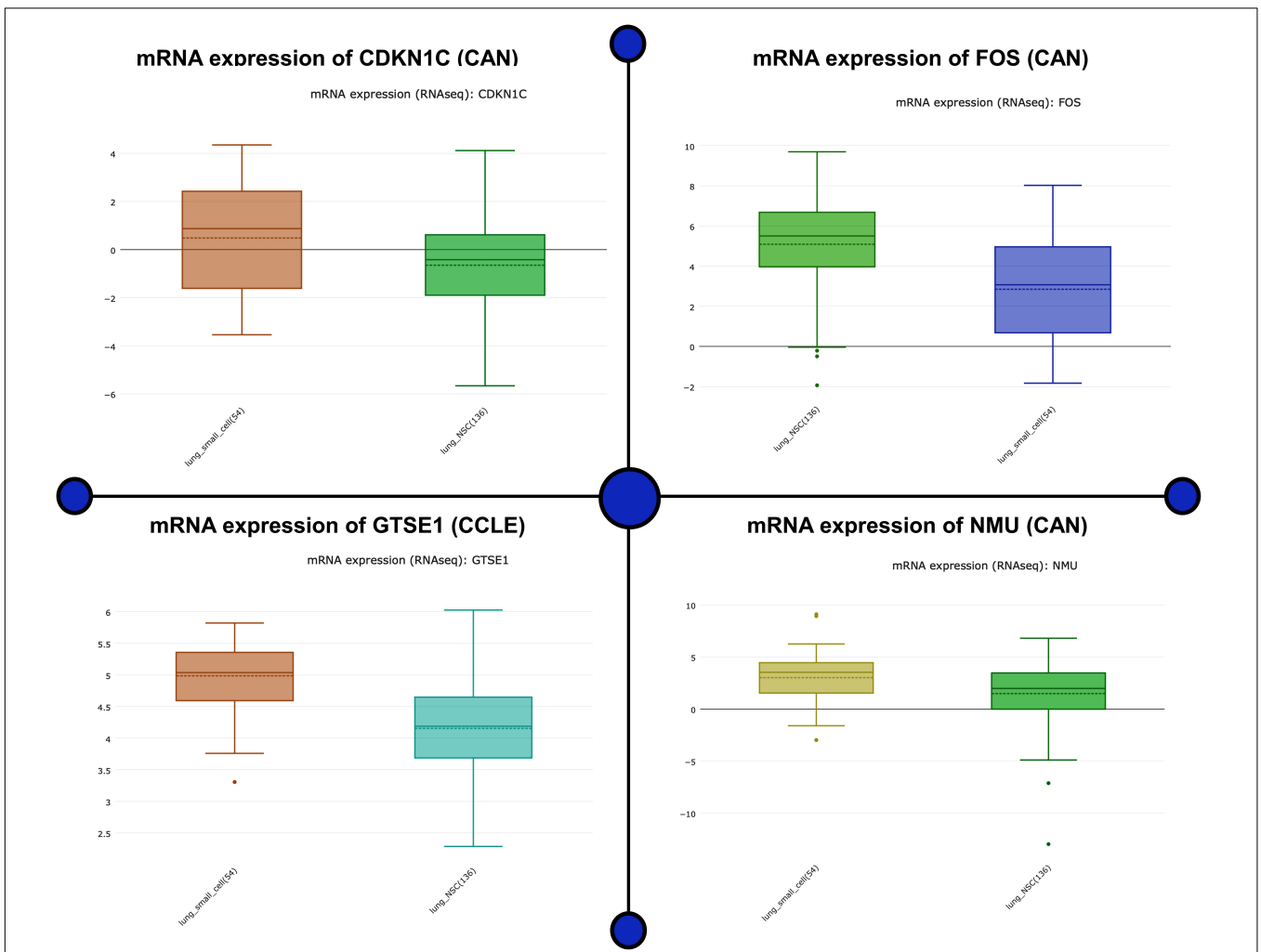


FIGURE 3 | mRNA expression microarrays showed FOS, GTSE1, CDKN1C, and NMU are greatly harbored in lung cancer patients and are observed to be differentially expressed (fold change ≥2.0) in these genes. The up- and down-regulated genes as a potential biomarker are analyzed for its mRNA expression. This figure depicts the mRNA biological impact of this gene by virtue of its intrinsic regulatory nature in patients.

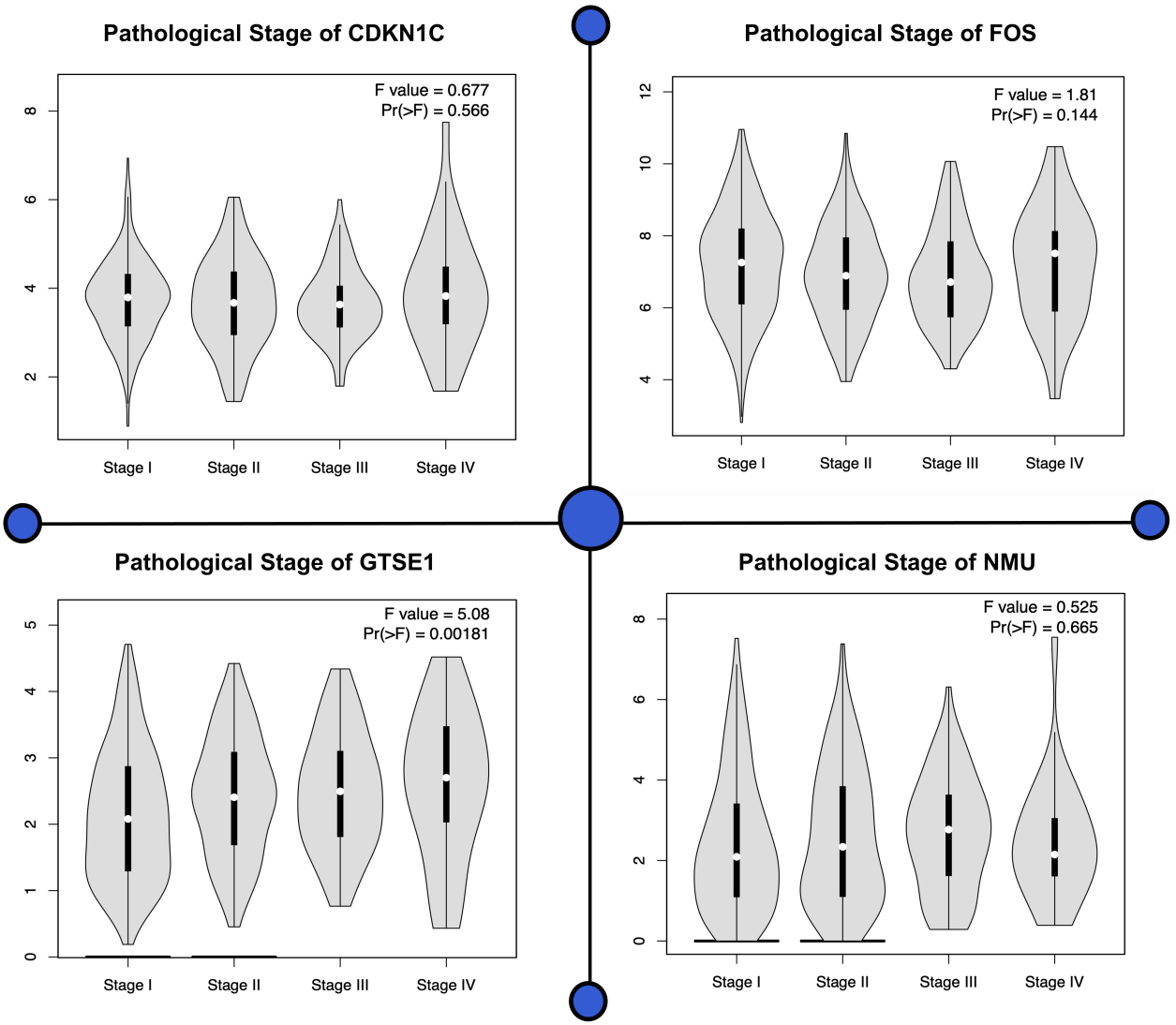


FIGURE 4 | Various pathological stages of all four genes are shown in this plot. For each individual biomarker, four stages can be observed showing significant difference present among all the genes.

results obtained from the signaling pathways and cell adhesion molecules (CAMs) were observed to be most closely related to the occurrence of lung cancer. The ratio of DEGs in lung cancer patients was high and was observed to be responsible for the loss of function in many biological mechanisms. Therefore, it is initially proposed that FOS, GTSE1, CDKN1C, and NMU could be the potential genes for further evaluation in this regard and may play an important role as biomarkers of lung cancer diagnosis, as shown in **Figure 5**.

The CDKN1C, which is also known as $p57^{kip2}$, is claimed to be a tumor suppressor considered to be a potential tumor suppressor caught up in several types of cancer in humans. However, recent reports (Qiu et al., 2018) demonstrate CDKN1C in breast cancer is observed to be intensely down-regulated equated with normal tissue. Furthermore, the CDKN1C expression is detected to be associated with age and tumor size in The Cancer Genome Atlas (TCGA) cohort containing 708 cases of breast cancer. The low expression of CDKN1C is expressively connected with overall poor survival rate, as shown in **Figure 5**.

On the other hand, reports on the FOS maintained that its down-regulation might be associated with the pathogenesis of lung cancer (Mahner et al., 2008). This gene and TP53 play as transcription factors and also as target genes in this system, having the ability to self-regulate. The need for transcription factor for TP53 is fulfilled by the FOS (Levin et al., 1995), as shown in **Figure 5**.

An ample amount of expression is detected regarding NMU in the vast majority of lung cancers. Various analyses have unveiled a substantial connotation of NMU expression with a minor prognosis of patients suffering from NSCLC. The expression of this gene can be suppressed when short interfering RNAs are cast off to treat NSCLC that retards the cell's development. In contrast,

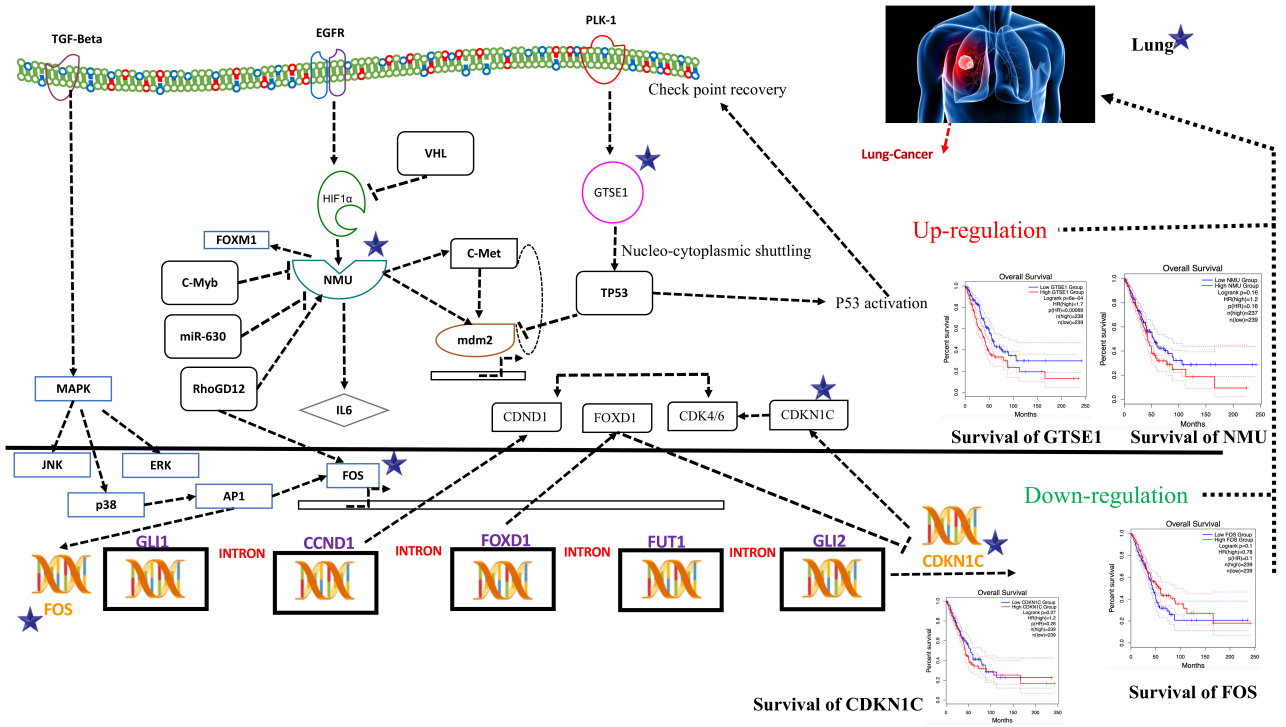


FIGURE 5 | Systems biology is an interdisciplinary field that involves computational and mathematical investigations for modeling a complex biological mechanism. This figure is a depiction of the genes in a lung cancer biochemical pathway and how they interact in a system. FOS, GTSE1, CDKN1C, and NMU could be the potential genes for further evaluation in this regard and may play an important role as biomarkers of lung cancer diagnosis. The CDKN1C, which is also known as $p57^{kip2}$, is claimed to be a tumor suppressor considered to be a potential tumor suppressor caught up in lung cancer in humans. However, recent reports demonstrate CDKN1C in lung cancer is observed to be intensely down-regulated. The low expression of CDKN1C is expressively connected with overall poor survival rate shown in **Figure 5**. In contrast, the growth-encouraging movement and augmented mobilization of cells are conferred by inducing an exogenous expression of NMU shown in **Figure 5**. The GTSE1 gene is claimed to be involved in lung cancerous pathways because of its overexpression and negative regulation of p53 expression. It is also observed that its expression is linked with venous invasion, global short survival, and size of the tumor. FOS can be utilized as markers of poor prognosis of lung cancer shown in **Figures 5**, **6** as pathway for each species.

the growth-encouraging movement and augmented mobilization of cells are conferred by inducing an exogenous expression of NMU shown in **Figure 5**.

The GTSE1 gene is claimed to be involved in many cancerous pathways due to its overexpression and negative regulation p53 expression. Recent reports also claimed that this gene both at its mRNA and protein levels is extremely up-regulated in hepatocellular carcinoma specimens (silencing GTSE-1 expression inhibits proliferation and invasion of hepatocellular carcinoma cells). It is also observed that its expression is linked with venous invasion, global short survival, and size of the tumor shown in **Figure 5**. The time course simulation's illustrate elevated NMU and GTSE1 countenance, reduced expression of CDKN1C and FOS, can be utilized as markers of poor prognosis of lung cancer, as shown in **Figures 5**, **6** showing pathway for each species.

Survival Analysis

We selected four major genes, of which GTSE1 (logFC = 1.32, adjusted $P < 0.001$) and NMU (logFC = 2.81, adjusted $P < 0.001$) were found to be up-regulated in the tissues of those patients who had lung cancer, and expression levels of FOS (logFC = -2.22, adjusted $P < 0.001$) and CDKN1C (logFC = -1.56, adjusted $P < 0.001$) were found to be down-regulated. For the assessment of prognostic value belonging to GTSE1, NMU, FOS, and CDKN1C in patients with NSCLC, we analyzed 1,926 NSCLC cases from TCGA, GEO, and EGA databases. It was illustrated through the results that high levels of expression of GTSE1 ($P < 0.01$) and NMU ($P < 0.01$) were very closely associated with shorter complete survival of NSCLC patients, with statistical importance. On the contrary, high levels of expression of CDKN1C [$P < 0.01$ and FOS ($P < 0.01$)] were significantly related to longer survival in patients with NSCLC; these findings illustrate elevated NMU and GTSE1 countenance and reduced expression of CDKN1C and FOS can be utilized as markers of poor prognosis of lung cancer as shown in **Figure 7**.

DISCUSSION

The mortality rate of lung cancer is high because it easily metastasizes and lapses in between treatments.

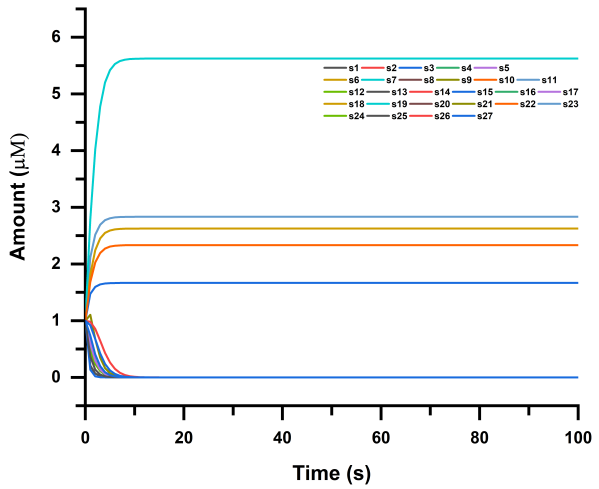


FIGURE 6 | Time course simulation of NMU-, GTSE1-, CDKN1C-, and FOS-associated pathways, where simulation was done in four phases for each species (genes and proteins in the pathway). Narrowing down to cancer systems biology, it mainly involves the use of systems biology's technology in cancer research, for the sake of examining an ailment as a challenging adaptive system having evolved characteristics at various biological parameters shown in the figure.

Hence, it is of utmost importance to overcome the medical obstruction by accurate prediction of potential prognostic markers of the status of tumor progression. Transcription omics that have high-throughput benefits can be of great help for those who are in medical research to facilitate the screening of the target molecules. So, we combined four mRNA expression profiles belonging to lung cancer.

By carrying out a comparison between NSCLC tissues and paired paracancerous tissues, 952 CDEGs were screened from the four expression profiles, among which had up-regulated expression and 696 CDEGs had down-regulated expression, respectively. It was seen by performing Gene Ontology analysis that CDEGs were majorly supplemented in biological processes such as cell adhering, as well as positive modulation of angiogenesis, and KEGG pathways, such as ECM–receptor interaction and CAMs. In the same way, this result was also reported by Piao et al. (2018).

GTSE1 has been found to be highly expressed in the tumors such as melanoma and lung cancer and is related to the weak prognosis of the patients (Wu et al., 2017; Xu et al., 2018). Additionally, GTSE1 might be participating in tumorigenesis and progression by modulating p53 phosphorylation (Liu et al., 2010, 2019).

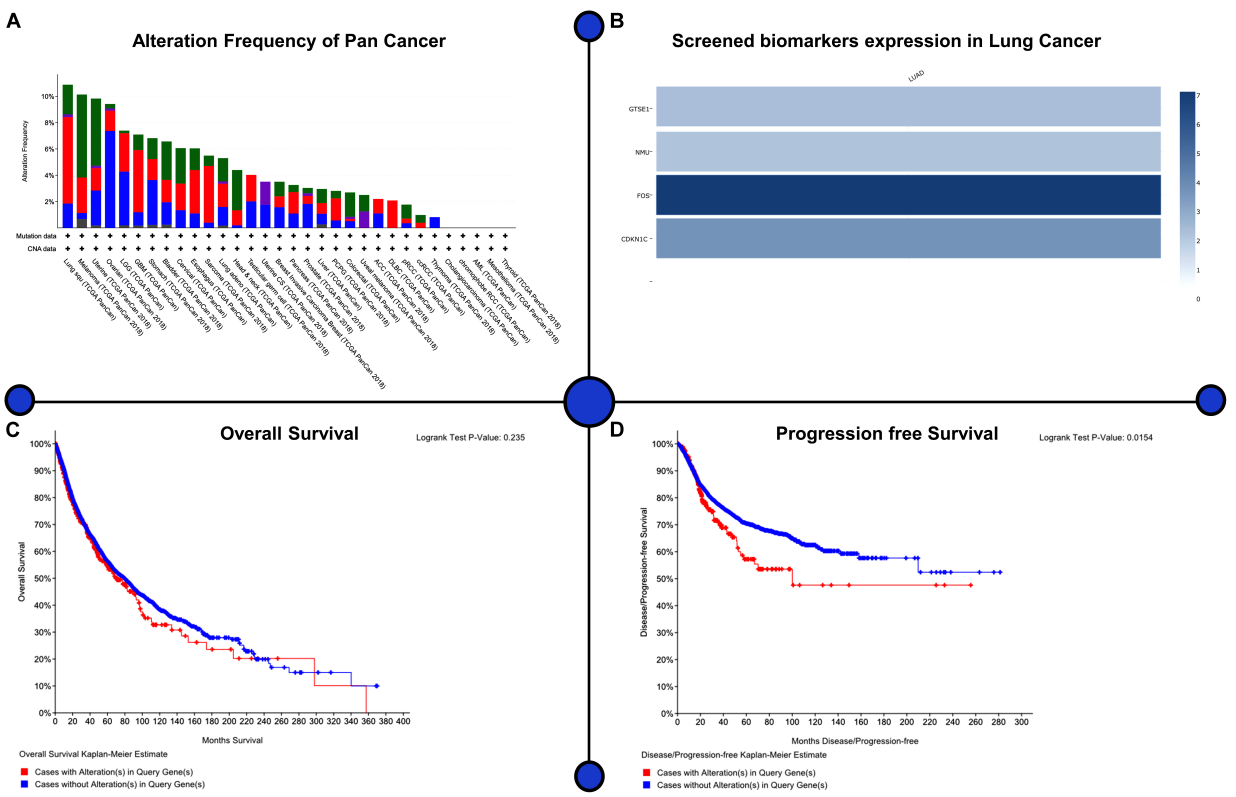


FIGURE 7 | Assessment of prognostic value belonging to GTSE1, NMU, FOS, and CDKN1C in patients with NSCLC. The Kaplan–Meier survival curves of LUAD and their major significance have been drawn to explore their effects on prognosis. Patients with LUAD showed significantly worse prognosis than patients without biomarkers. Lung adenocarcinoma showed worse prognosis than patients with four biomarkers. **(A)** Alteration frequency of pan-cancer, **(B)** screened biomarker expression in lung cancer, **(C)** overall survival, **(D)** progression-free survival.

Neuromedin U is very well known for its uterine smooth muscle contraction inducer. In the meantime, it also contributes to the process of formation and enlargement of various kinds of tumors. For instance, it was reported by Takahashi et al. (2006) that the positive rate of NMU in NSCLC and SCLC was as high as 68 and 82%, and the overexpression of NMU was validated at the transcriptional level and protein level (Takahashi et al., 2006). Moreover, studies have demonstrated that overexpression of NMU is also produced in HER2-overexpressing breast cancer, and overexpression of NMU in breast cancer is linked with reduced prognosis in sufferers (Shetzline et al., 2004; Wu et al., 2007). Similar reports have been reported in the study of clear cell renal cell carcinoma and endometrial carcinoma (Ketterer et al., 2009; Przygodzka et al., 2016; Zhang et al., 2019). CDKN1C is a cancerous lump restrainer gene, which is down-regulated in studies related to gastric cancer (Shin et al., 2000), bladder cancer (Oya and Schulz, 2000), pancreatic cancer (Sato et al., 2005), lung cancer (Sun et al., 2017), and breast cancer (Qiu et al., 2018), and low expression points are connected with reduced prediction in sufferers. Importantly, all of the above research results strongly support our analysis results. Additionally, CDKN1C, GTSE1, NMU, and FOS are not correlated with one another, which indicates that every target can individually be cast off as a predictive marker for lung cancer. Ultimately, the aforementioned studies show the capability of FOS, CDKN1C, GTSE1, and NMU as prognostic markers of lung cancer.

CONCLUSION

The molecular specification of lung cancer has considerably changed the categorization and treatment of tumors, becoming a crucial component of diagnosis and oncologic therapy. We executed differential analysis of samples of lung cancer and matching tissues of paracancer and noticed four core CDEGs could be utilized as prognostic markers of lung cancer via correlation analysis and expression level verification. We strongly believe that GTSE1, NMU, FOS, and CDKN1C have potential and clinical application values to act as prognostic markers of lung cancer.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

AK and D-QW designed the experiments. D-QW, AK, and XD, performed the entire computational experiments and assisted in writing the manuscript. XD, AK analyzed the data and wrote the manuscript. AK, D-QW, XD, and AM read the manuscript and advised on method development. All authors have given approval to the final version of the manuscript.

FUNDING

This work is supported by the grants from the Key Research Area Grant 2016YFA0501703, 2018ZX10302205-004-002, and 2018ZX10302-205-004 of the Ministry of Science and Technology of China, the National Natural Science Foundation of China (Contract Nos. 61832019, 61503244, BK20161130, and 81972789), the State Key Lab of Microbial Metabolism and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2017ZD14), The Six Talent Peaks Project in Jiangsu Province (Grant No. SWYY-128), Major Project of Science and Technology in Henan Province (Grant No. 161100311400), The Technology Development Funding of Wuxi (Grant No. WX18IVJN017), Research Funds for the Medical School of Jiangnan University ESI special cultivation project (Grant No. 1286010241170320), National Science and Technology Major Project (Grant No. 2018ZX10302205-004-002), and the Fundamental Research Funds for the Central Universities (Grant No. JUSRP22011). These funding sources have no role in the writing of the manuscript or the decision to submit it for publication.

ACKNOWLEDGMENTS

The simulations in this work were supported by the Center for High-Performance Computing, Shanghai Jiao Tong University.

REFERENCES

Di, J., Shenglan, L., Dali, L., Haipeng, X., Dan, Y., and Ying, L. (2019). Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. *Aging* 10, 592–605. doi: 10.18632/aging.101415

Feng, H., Gu, Z. Y., Li, Q., Liu, Q. H., Yang, X. Y., and Zhang, J. J. (2019). Identification of significant genes with poor prognosis in ovarian cancer via bioinformatical analysis. *J. Ovarian Res.* 12:35. doi: 10.1186/s13048-019-0508-2

Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., et al. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One* 5:e10312. doi: 10.1371/journal.pone.0010312

Hu, G., Cheng, Z., Wu, Z., and Wang, H. (2019). Identification of potential key genes associated with osteosarcoma based on integrated bioinformatics analyses. *J. Cell Biochem.* 120, 13554–13561. doi: 10.1002/jcb.28630

Ketterer, K., Kong, B., Frank, D., Giese, N. A., Bauer, A., Hoheisel, J., et al. (2009). Neuromedin U is overexpressed in pancreatic cancer and increases invasiveness via the hepatocyte growth factor c-Met pathway. *Cancer Lett.* 277, 72–81. doi: 10.1016/j.canlet.2008.11.028

Levin, W. J., Press, M. F., Gaynor, R. B., Sukhatme, V. P., Boone, T. C., Reissmann, P. T., et al. (1995). Expression patterns of immediate early transcription factors in human non-small cell lung cancer. The lung cancer study group. *Oncogene* 11, 1261–1269.

Liu, A., Zeng, S., Lu, X., Xiong, Q., Xue, Y., Tong, L., et al. (2019). Overexpression of G2 and S phase-expressed-1 contributes to cell proliferation, migration, and invasion via regulating p53/FoxM1/CCNB1 pathway and predicts poor

prognosis in bladder cancer. *Int. J. Biol. Macromol.* 123, 322–334. doi: 10.1016/j.ijbiomac.2018.11.032

Liu, X. S., Li, H., Song, B., and Liu, X. (2010). Polo-like kinase 1 phosphorylation of G2 and S-phase-expressed 1 protein is essential for p53 inactivation during G2 checkpoint recovery. *EMBO Rep.* 11, 626–632. doi: 10.1038/embor.2010.90

Lu, T. P., Tsai, M. H., Lee, J. M., Hsu, C. P., Chen, P. C., Lin, C. W., et al. (2010). Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol. Biomarkers Prev.* 19, 2590–2597. doi: 10.1158/1055-9965.EPI-10-0332

Luca, F., Gabriella, L., Giusy, R. M. L. R., Salvatore, C., Carmelina, D. A., Rossella, S., et al. (2019). Identification of novel MicroRNAs and their diagnostic and prognostic significance in oral cancer. *Cancers* 11:610. doi: 10.3390/cancers11050610

Ma, W., Wang, B., Zhang, Y., Wang, Z., Niu, D., Chen, S., et al. (2019). Prognostic significance of TOP2A in non-small cell lung cancer revealed by bioinformatic analysis. *Cancer Cell Int.* 19:239. doi: 10.1186/s12935-019-0956-1

Mahner, S., Baasch, C., Schwarz, J., Hein, S., Wölber, L., Jänicke, F., et al. (2008). C-Fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma. *Br. J. Cancer* 99:1269. doi: 10.1038/sj.bjc.6604650

Mitchell, K. A., Zingone, A., Toulabi, L., Boeckelman, J., and Ryan, B. M. (2017). Comparative transcriptome profiling reveals coding and noncoding RNA differences in NSCLC from African Americans and European Americans. *Clin. Cancer Res.* 23, 7412–7425. doi: 10.1158/1078-0432.CCR-17-0527

Oya, M., and Schulz, W. A. (2000). Decreased expression of p57(KIP2)mRNA in human bladder cancer. *Br. J. Cancer* 83, 626–631. doi: 10.1054/bjoc.2000.1298

Pan, J. H., Zhou, H., Cooper, L., Huang, J. L., Zhu, S. B., Zhao, X. X., et al. (2019). LAYN is a prognostic biomarker and correlated with immune infiltrates in gastric and colon cancers. *Front. Immunol.* 10:6. doi: 10.3389/fimmu.2019.00006

Piao, J., Sun, J., Yang, Y., Jin, T., Chen, L., and Lin, Z. (2018). Target gene screening and evaluation of prognostic values in non-small cell lung cancers by bioinformatics analysis. *Gene* 647, 306–311. doi: 10.1016/j.gene.2018.01.003

Przygodzka, P., Papiewska-Pajak, I., Bogusz, H., Kryczka, J., Sobierajska, K., Kowalska, M. A., et al. (2016). Neuromedin U is upregulated by Snail at early stages of EMT in HT29 colon cancer cells. *Biochim. Biophys. Acta* 1860(11 Pt A), 2445–2453. doi: 10.1016/j.bbagen.2016.07.012

Qiu, Z., Li, Y., Zeng, B., Guan, X., and Li, H. (2018). Downregulated CDKN1C/p57(kip2) drives tumorigenesis and associates with poor overall survival in breast cancer. *Biochem. Biophys. Res. Commun.* 497, 187–193. doi: 10.1016/j.bbrc.2018.02.052

Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J. A., Pedraza, V., Boyero, L., Rosell, R., et al. (2011). Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer* 129, 355–364. doi: 10.1002/ijc.25704

Sato, N., Matsubayashi, H., Abe, T., Fukushima, N., and Goggins, M. (2005). Epigenetic down-regulation of CDKN1C/p57KIP2 in pancreatic ductal neoplasms identified by gene expression profiling. *Clin. Cancer Res.* 11, 4681–4688. doi: 10.1158/1078-0432.Ccr-04-2471

Shetzline, S. E., Rallapalli, R., Dowd, K. J., Zou, S., Nakata, Y., Swider, C. R., et al. (2004). Neuromedin U: a Myb-regulated autocrine growth factor for human myeloid leukemias. *Blood* 104, 1833–1840. doi: 10.1182/blood-2003-10-3577

Shin, J. Y., Kim, H. S., Lee, K. S., Kim, J., Park, J. B., Won, M. H., et al. (2000). Mutation and expression of the p27KIP1 and p57KIP2 genes in human gastric cancer. *Exp. Mol. Med.* 32, 79–83. doi: 10.1038/emm.2000.14

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA Cancer J Clin* 68, 7–30. doi: 10.3322/caac.21442

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Sun, Y., Jin, S. D., Zhu, Q., Han, L., Feng, J., Lu, X. Y., et al. (2017). Long non-coding RNA LUCAT1 is associated with poor prognosis in human non-small lung cancer and regulates cell proliferation via epigenetically repressing p21 and p57 expression. *Oncotarget* 8, 28297–28311. doi: 10.18632/oncotarget.16044

Takahashi, K., Furukawa, C., Takano, A., Ishikawa, N., Kato, T., Hayama, S., et al. (2006). The neuromedin U-growth hormone secretagogue receptor 1b/neurotensin receptor 1 oncogenic signaling pathway as a therapeutic target for lung cancer. *Cancer Res.* 66, 9408–9419. doi: 10.1158/0008-5472.CAN-06-1349

Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., et al. (2016). A benchmark for RNA-seq quantification pipelines. *Genome biol.* 17:74. doi: 10.1186/s13059-016-0940-1

Wu, X., Wang, H., Lian, Y., Chen, L., Gu, L., Wang, J., et al. (2017). GTSE1 promotes cell migration and invasion by regulating EMT in hepatocellular carcinoma and is associated with poor prognosis. *Sci. Rep.* 7:5129. doi: 10.1038/s41598-017-05311-2

Wu, Y., McRoberts, K., Berr, S. S., Frierson, H. F. Jr., Conaway, M., and Theodorescu, D. (2007). Neuromedin U is regulated by the metastasis suppressor RhoGDI2 and is a novel promoter of tumor formation, lung metastasis and cancer cachexia. *Oncogene* 26, 765–773. doi: 10.1038/sj.onc.1209835

Xu, T., Ma, M., Chi, Z., Si, L., Sheng, X., Cui, C., et al. (2018). High G2 and S-phase expressed 1 expression promotes acral melanoma progression and correlates with poor clinical prognosis. *Cancer Sci.* 109, 1787–1798. doi: 10.1111/cas.13607

Zhang, S., Wang, Q., Han, Q., Han, H., and Lu, P. (2019). Identification and analysis of genes associated with papillary thyroid carcinoma by bioinformatics methods. *Biosci. Rep.* 39:BSR20190083. doi: 10.1042/bsr20190083

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

Copyright © 2020 Kaushik, Mehmood, Wei and Dai. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 23 April 2020
doi: 10.3389/fmolb.2020.00053

Intrinsic Genetic and Transcriptomic Patterns Reflect Tumor Immune Subtypes Facilitating Exploring Possible Combinatory Therapy

Yong Xu[1,2], Daixi Li[1], Zhenhao Liu[2], David L. Gibbs[3], Lu Xie[2]* and Guangrong Qin[2,3]**

[1] Laboratory for Computational Biology, University of Shanghai for Science and Technology, Shanghai, China, [2] Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai, China, [3] Institute for Systems Biology, Seattle, WA, United States

OPEN ACCESS

Edited by:
Peng Zhang,
University of Maryland, United States

Reviewed by:
Ya Cui,
University of California, Irvine, United States
Jiao Yuan,
University of Pennsylvania, United States

***Correspondence:**
Daixi Li
dxli75@126.com
Lu Xie
luxiex2017@outlook.com
Guangrong Qin
guangrong.qin@isbscience.org

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *29 January 2020*
Accepted: *17 March 2020*
Published: *23 April 2020*

Citation:
Xu Y, Li D, Liu Z, Gibbs DL, Xie L and Qin G (2020) Intrinsic Genetic and Transcriptomic Patterns Reflect Tumor Immune Subtypes Facilitating Exploring Possible Combinatory Therapy. Front. Mol. Biosci. 7:53. doi: 10.3389/fmolb.2020.00053

The classification of immune subtypes was based on immune signatures highlighting the tumor immuno-microenvironment. It was found that immune subtypes associated with mutation and expression patterns in the tumor. How the intrinsic genetic and transcriptomic alterations contribute to the immune subtypes and how to select drug combinations from both targeted drugs and immune therapeutic drugs according to different immune subtypes are still not clear. Through statistical analysis of genetic alterations and transcriptional profiles of breast invasive carcinoma (BRCA) samples, we found significant differences in the number of somatic missense mutations and frameshift deletions among the different immune subtypes. The high mutation load for somatic missense mutations and frameshift deletions may be explained by the high frequency of mutations and high expression of DNA double-strand break repair pathway genes. Extensive analysis of signaling pathways in both the genetic and transcriptomic levels reveals significantly altered pathways such as tumor protein Tumor Protein P53 (TP53) and receptor tyrosine kinase (RTK)/RAS signaling pathways among different subtypes. Drug targets in the signaling pathways such as mitogen-activated protein kinase kinase kinase 1 (MAP3K1) and Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA) show genetic alteration in specific subtypes, which may be potential targets for patients of a specific subtype. More drug targets which show transcriptional difference among immune subtypes were discovered, such as cyclin-dependent kinase (CDK)4, CDK6, Erb-B2 receptor tyrosine kinase 2 (ERBB2), etc. Moreover, differences in functional activity between tumor growth and immune-related pathways also elucidate the extrinsic factors of differences in prognosis and suggest potential drug combinations for different immune subtypes. These results help to explain how intrinsic alterations are associated with the immune subtypes and provide clues for possible combination therapy for different immune subtypes.

Keywords: immune subtypes, breast invasive carcinoma, target, signaling pathways, combination therapy

INTRODUCTION

In recent years, growing evidence has shown that immunosuppression in the tumor microenvironment (TME) is a major obstacle for effective antitumor therapy in patients (Munn and Bronte, 2016). The relationship between the infiltration level of immune cells in solid tumors and prognosis has been reported (Fridman et al., 2017). Solid tumors from diverse tissues of origin in The Cancer Genome Atlas (TCGA) have been classified into six immune subtypes, namely, wound healing (C1), interferon IFN-γ dominant (C2), inflammatory (C3), lymphocyte depleted (C4), immunologically quiet (C5), and transforming growth factor TGF-β dominant (C6) (Thorsson et al., 2018). The immune subtypes are associated with different prognoses and provide clues for immunotherapy response.

Some clinical successes are due to patient stratification, typically according to either the genetic features or immune environment, where it is hoped that more precise treatments can be delivered. Cancer immunotherapy demonstrates tremendous success in improving prognosis of some cancer types, including breast invasive carcinoma and melanomas (Li et al., 2016; Naik et al., 2019). Programmed cell death protein 1 (PD-1) and programmed death ligand 1 (PD-L1) antibodies have been shown to be effective in treating multiple malignancies (Gang et al., 2018); however, the drug efficacy depends on the mutational load (Hugo et al., 2016). Breast invasive carcinoma is a widely investigated tumor type with targeted drugs for different genetic subtypes. For example, the Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA) inhibitor alpelisib was approved in 2019 by the US Food and Drug Administration (FDA) for the treatment of PIK3CA-mutated hormone receptor-positive advanced breast invasive carcinoma as it significantly increases the progression-free survival for patients (Turner et al., 2015); human epidermal growth factor (EGF) receptor 2 human epidermal growth factor receptor-2 (HER2) antibodies such as trastuzumab and lapatinib can be used to treat HER2-positive patients; talazoparib, a poly (ADP-ribose) polymerase (PARP) inhibitor was approved in 2018 for the treatment of patients with breast cancer gene (BRCA) mutations and HER2-negative advanced or metastatic breast invasive carcinoma.

Although tumor patients can be classified into different immune subtypes, the biological mechanism that drives the differences in the immune microenvironment is not fully understood. How alterations in tumor cells induce specific tumor immuno-microenvironments or are associated with each other is still not clear. With multiple drug options for both immune therapy and targeted therapy, the understanding of the associated genetic factors of immune subtypes may provide new clues for precise drug or drug combinations.

For over 10 years, the TCGA has profoundly illuminated the multiple omics landscape of human malignancy. Cell composition methods, such as CIBERSORT (Newman et al., 2015) and TIMER (Li et al., 2016), have been developed to characterize complex tissue cell compositions. Using the genomic and transcriptomic data derived from bulk tumor samples, both TME and tumor genetic features from cancer cells can be explored, which can help to understand the association of tumor genetic features with TMEs, as well as exploring new drug combinations for different tumor subtypes.

In this study, based on the tumor immune subtypes identified in literature (Thorsson et al., 2018), we explored the genetic and transcriptional features for different immune subtypes. Through the integrative analysis of gene mutations, DNA damage response, and oncogenic signaling, we find an association of these pathways with immune subtypes and identified targeted drugs which are associated with different immune subtypes in breast invasive carcinoma. We also analyzed the interactions between key immune-related altered pathways and tumor growth pathways to explain the significant differences in prognosis among different immune subtypes.

MATERIALS AND METHODS

Data Acquisition

Multiple omics data including gene expression data normalized by RSEM from Illumina HiSeq RNASeq, DNA somatic mutation data, and clinical data were downloaded from UCSC Xena (2018)[1]. In this study, two solid tumor types were selected, with breast invasive carcinoma (BRCA) as the research subject and lung adenocarcinoma (LUAD) as the comparative analysis and verification.

Mutation Signature Analysis

Different mutational processes generate unique combinations of mutation types, termed "Mutational Signatures," which have been classified based on the analysis of somatic mutation spectrum (Alexandrov et al., 2013). Based on tumor somatic mutation data in the TCGA database, the weights of mutation signatures (Alexandrov et al., 2013) for each tumor sample were calculated using the R packages "deconstructSigs" and "maftools" (Mayakonda et al., 2018). Kruskal–Wallis test was performed to estimate the difference of mutation signature weights among immune subtypes, and significant mutation signatures (P-value < 0.05) were selected among the immune subtypes. For pairwise analysis of mutation signature weights between immune subtypes, Wilcoxon rank-sum test was used. Significant results (P-value < 0.05) were shown in the boxplot using the R package "ggpubr."

Prognosis Analysis of Immune Subtypes

Survival analysis of tumors and relapse-free survival were performed using the R packages of "survminer" and "survival." According to the overall survival time and relapse-free survival time in TCGA clinical data, the survival rates of different immune subtypes were compared and the survival curves were drawn. Log-rank test was performed to compare the difference of survival distribution between immune subtypes, P-values smaller than 0.05 were considered as significant difference in the survival rate of immune subtypes.

[1] https://tcga.xenahubs.net/

Gene Set Variation Analysis

To calculate single-sample gene set enrichment, we used the gene set variation analysis (GSVA) program (Hanzelmann et al., 2013) to derive the absolute enrichment scores of previous literature reported DNA damage repair (DDR) (Knijnenburg et al., 2018) gene signatures as follows: (1) Base Excision Repair (BER), (2) Nucleotide Excision Repair (NER; including TC-NER and GC-NER), (3) Mismatch Repair (MMR), (4) Fanconi Anemia (FA), (5) Homologous Recombination (HR), (6) Non-Homologous End Joining (NHEJ), (7) Direct Repair (DR), (8) Translesion Synthesis (TLS), (9) Damage Sensor, etc., and oncogenic signaling pathway (Sanchezvega et al., 2018) gene signatures as follows: (1) cell cycle, (2) Hippo signaling, (3) Myc signaling, (4) Notch signaling, (5) oxidative stress response/Nrf2, (6) phosphatidylinositol 3-kinase (PI3K) signaling, (7) receptor tyrosine kinase (RTK)/RAS/mitogen-activated protein (MAP) kinase signaling, (8) TGF-β signaling, (9) tumor protein (TP)53 signaling, (10) b-catenin/Wnt signaling, and (11) Erb-B2 receptor tyrosine kinase (ERBB) signaling. To make a more comprehensive analysis of the functional modules, we further evaluated the activity of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (gene sets) (Minoru and Susumu, 2000) within immune subtypes using the single-sample GSVA (ssGSVA). This method quantifies gene set enrichment in individual samples rather than at the group level.

Mutational Status for Oncogenic Signaling Pathways

The mutational status for each signaling pathway is defined through a binary classification: if any gene was mutated in this pathway in one sample, the mutational status of this pathway in this sample was considered as mutated (use 1 to represent mutated status). On the contrary, the mutation status is set to non-mutated (use 0 to represent the non-mutated status). Fisher's exact test was performed on the number of mutated samples and non-mutated samples within immune subtypes for each signaling pathway to compare the level of mutation in different immune subtypes.

Drug Target Selection

To derive mutant genes associated with immune subtypes in breast invasive carcinoma and explore potential drug targets for each subtype, we detected gene mutations associated with each immune subtype using Fisher's exact test (P-value < 0.05). The high-frequency mutant genes (mutation frequency $> 5\%$) associated with each immune subtype were used to find linked therapeutic drugs using OncoKB[2] and Drugbank[3]. Further, we selected the target genes in 11 signaling pathways from Drugbank and then compared whether the expression levels of these target genes were statistically significant among different immune subtypes using Wilcoxon rank-sum test. Target genes that are statistically significant (P-value < 0.05) were used to query DrugBank for drug selection.

[2]https://www.oncokb.org/

[3]https://www.drugbank.ca/

Correlation Analysis

The proliferation scores and leukocyte fractions for breast invasive carcinoma have been previously calculated (Thorsson et al., 2018). The enrichment scores of tumor growth and immune-related pathways in breast invasive carcinoma samples were estimated using GSVA on TCGA gene expression data. We measured the correlation coefficient between the proliferation scores and the enrichment scores of tumor growth-related pathways using Spearman's rank correlation. Similarly, correlations between the enrichment scores of immune-related pathways and the leukocyte fraction were assessed using Spearman's rank correlation. Significant correlations were considered as those pairs with P-value less than 0.05.

RESULTS

Mutation Types and Mutation Signatures Are Associated With Immune Subtypes

To understand the intrinsic tumor cell features that may drive the immune subtypes, we first asked whether there are differences in mutation types among immune subtypes. As mutated genes may produce altered neo-antigens, the mutation load and mutation types may have functional consequences for tumor cells and further drive the formation of immune microenvironments. Using breast invasive carcinoma in the TCGA dataset as an example, five immune subtypes can be detected according to the pan-cancer immune subtyping (Thorsson et al., 2018) (**Figure 1A**). Significant differences in the number of somatic missense mutations are found among different immune subtypes as well as significant difference in the number of somatic frameshift deletions ($P < 10^{-7}$, Kruskal–Wallis test). The frequencies of frameshift deletion and missense mutations in the C1 and C2 immune subtypes were significantly higher than other subtypes ($P < 0.01$, Wilcoxon rank-sum test) (**Figures 1B,C**). In addition, consistent results were observed in LUAD ($P < 0.05$, Wilcoxon rank-sum test) (**Figures 1D,F**). This might hint that these types of mutations were important factors in generating the C1 and C2 immune subtypes in breast invasive carcinoma and LUAD. Both somatic missense mutation and frameshift deletion can introduce abnormal peptides, which may play a key role in recruiting immune cells.

Somatic mutations can be the consequence of multiple mutational processes, such as the deficiency in the DNA replication machinery and DNA repair system, abnormal enzymatic modification of DNA, or exposure to exogenous or endogenous mutagens (Alexandrov et al., 2013). From a large cohort of tumors, somatic mutation spectra have been categorized into 30 mutation signatures, which are associated with different biological processes (Alexandrov et al., 2013; Forbes et al., 2016). Using this concept, we measured the weight of different mutational signatures for each breast invasive carcinoma sample in TCGA and compared the difference of each mutation signature among immune subtypes (Alexandrov et al., 2013, 2015; Mayakonda et al., 2018). Results show that mutation signature 3 (MS3) showed significant differences among immune

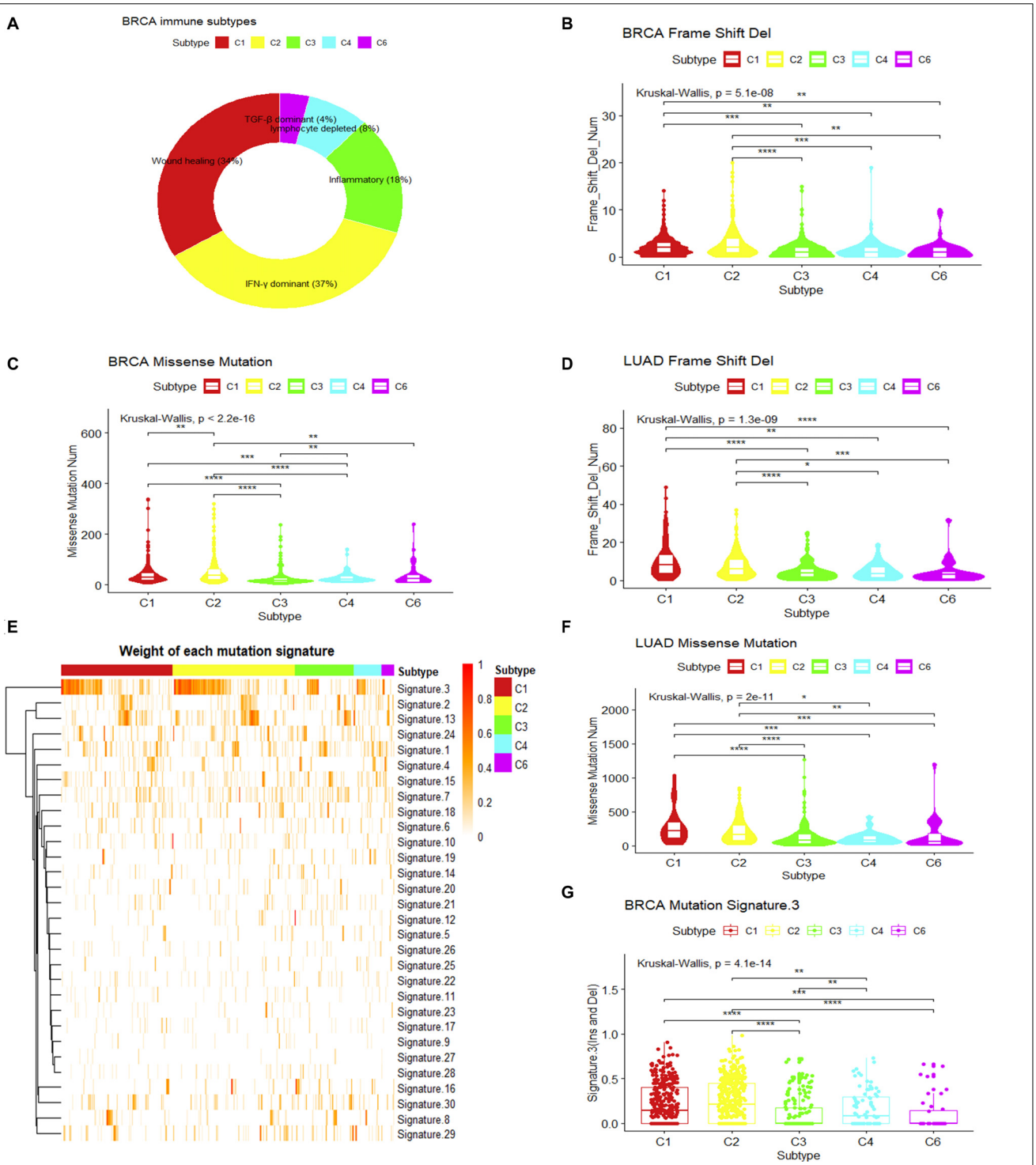


FIGURE 1 | Mutation loads and mutation signatures in immune subtypes. **(A)** Sample proportions of different immune subtypes in breast invasive carcinoma. **(B,C)** Comparison of the frequency of frameshift deletion or missense mutation among different immune subtypes of breast invasive carcinoma (Wilcoxon rank-sum test was used. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$). **(D,F)** Comparison in the frequency of frameshift deletion or missense mutation among different immune subtypes of lung adenocarcinoma (LUAD) (Wilcoxon rank-sum test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$). **(E)** The weight of each mutation signature of breast invasive carcinoma immune subtype. **(G)** The boxplot of the weight of mutation signature 3 for each immune subtype in breast invasive carcinoma (Wilcoxon rank-sum test was used. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

subtypes (*P*.adjust < 0.05, Kruskal–Wallis test, Benjamini and Hochberg adjustment) (**Figure 1E**). It is associated with the failure of DNA double-strand break repair by HR (Alexandrov et al., 2013; Forbes et al., 2016). The higher MS3 weights in C1 and C2 immune subtypes ($P < 0.05$, Wilcoxon rank-sum test) (**Figure 1G**) and the higher mutational load for somatic mutations and frameshift indels suggest that the formation of these two immune subtypes may result from the failure of a DNA double-strand break repair.

Genetic Alteration and Expression Levels of DNA Damage Repair Shape the Immune Subtypes

We then ask whether differences in the DDR system exist among immune subtypes. Loss of DDR function is an important determinant of cancer risk, progression, and therapeutic response (Jeggo et al., 2015). The proportions of samples with mutated DDR genes are shown in **Figure 2A**. The result shows that the proportion of mutated DDR genes in C1 and C2 subtypes of breast invasive carcinoma is significantly higher compared to that in the C3 subtype and slightly higher than those in C4 and C6. Similarly, the result was also observed in LUAD. However, these results do not fully explain how the DDR system interacts with the immune microenvironment.

We further compared the expression level of DDR genes among immune subtypes. ssGSVA (Hanzelmann et al., 2013) was performed for DDR-related pathways in breast invasive carcinoma (**Figure 2B**). BER, FA, and HR genes show higher expression in C1 and C2 subtypes and lower expression in C3 and C6 subtypes ($P < 0.01$, Wilcoxon rank-sum test) (**Figures 2C–E**). This indicates that the C1 and C2 subtypes may be more active in DDR and suggest genomic instability in these subtypes. Additionally, a similar result was found in LUAD ($P < 0.01$, Wilcoxon rank-sum test) (**Supplementary Figures S1C–E**). A promising way forward might be to select optimal drugs targeting the DDR pathway based on specific types of DDR mutations (O'Connor, 2015). The recent approval of olaparib, a PARP inhibitor for treating tumors harboring BRCA1 or BRCA2 mutations, provides a good example. The association of DDR features in tumor cells and the immune subtypes may provide new clues for selecting drug combinations for cancer treatment.

Subtype-Specific Alterations in Signaling Pathways Provide Opportunities for Targeted Therapy and Immune Therapy

Current breast invasive carcinoma drugs are mostly targeted to signaling or cell cycle-related pathways, such as HER2 antibodies, PI3K inhibitors. To bridge the gap between targeted therapy and immune subtypes, we further investigated how signaling pathways are associated with tumor immune subtypes. Oncogenic signaling pathways in the TCGA have been reported to represent the individual and co-occurring actionable alterations which also suggest opportunities for targeted and combination therapies (Sanchezvega et al., 2018). The reported oncogenic pathways as well as well-known drug targetable signaling pathways that include (1) cell cycle, (2) Hippo signaling, (3) Myc signaling, (4) Notch signaling, (5) oxidative stress response/Nrf2, (6) PI3K signaling, (7) RTK/RAS/MAP kinase signaling, (8) TGF-β signaling, (9) TP53 signaling pathway, (10) b-catenin/Wnt signaling, and (11) ERBB signaling were further analyzed to understand the association of signaling pathways and immune subtypes (Sanchezvega et al., 2018). Oncogenic genes in each pathway which show genetic alterations are shown in **Supplementary Table S1** (Sanchezvega et al., 2018).

Results show that the alterations of genes in the TP53 signaling pathway were significantly overrepresented in C2 subtypes (**Figure 3A**). Specifically, the proportion of samples with TP53 mutations is significantly higher in the C2 subtype (**Figure 3B**). The alterations of genes in the RTK-RAS pathways were significantly overrepresented in the C1 subtype (Fisher's exact test) (**Figure 3A**). Among these pathways, several potential target genes show a difference in mutation frequency among immune subtypes. PIK3CA and MAP kinase kinase kinase 1 (MAP3K1) were significantly high frequently mutated in C3 subtype, GATA3 was significantly high frequently mutated in C4 subtype (Fisher's exact test) (**Figures 3B,C**), and BRCA1 or BRCA2 was mutated in a higher percentage of samples in C1 and C2, although not significantly different (**Supplementary Figure S3A**). PIK3CA is a key player in the ERBB signaling pathway which can be targeted by PI3K inhibitors (Wullenkord et al., 2019). So the immune subtype C3, which shows a higher mutation frequency in PIK3CA, may have a better response for PI3K inhibitors (**Supplementary Figures S1A,B**). MAP3K1 mutations are also reported to be associated with sensitivity to MAP kinase kinase (MEK) inhibitors in multiple cancer models (Zheng et al., 2018).

To consider the impact of molecular subtypes for our result, we performed enrichment analysis with the molecular subtypes in immune subtypes and the alteration of carcinogenic signaling pathways in molecular subtypes. Results show that HER2 subtype is associated with a significantly higher proportion of samples with mutations in the RTK–RAS signaling pathway (**Supplementary Figure S2F**). Meanwhile, HER2 subtype is significantly enriched in the C1 immune subtype (**Supplementary Figure S2B**), which is consistent with the significant mutation results of C1 subtype in the RTK–RAS signaling pathway (**Figure 3A**). We observed that HER2 subtype is also associated with a significantly higher proportion of samples with mutations in the ERBB signaling pathway; however, immune subtypes show a difference to that. Basal subtype is associated with a significantly higher proportion of samples with mutations in the p53 signaling pathway (**Supplementary Figure S2F**), while Basal is significantly enriched in the C2 immune subtype (**Supplementary Figure S2A**), which is consistent with the significant mutation results of C2 subtype in the TP53 signaling pathway (**Figure 3A**). TP53 gene is significantly mutated in Basal and HER2 subtypes (**Supplementary Figure S2H**), while Basal and HER2 subtypes are significantly enriched in C2 immune subtypes (**Supplementary Figures S2A,B**), which is consistent with the result that TP53 gene mutated significantly in C2 immune subtypes. PIK3CA and MAP3K1 gene is significantly mutated in Luminal A subtypes (**Supplementary Figure S2H**), while

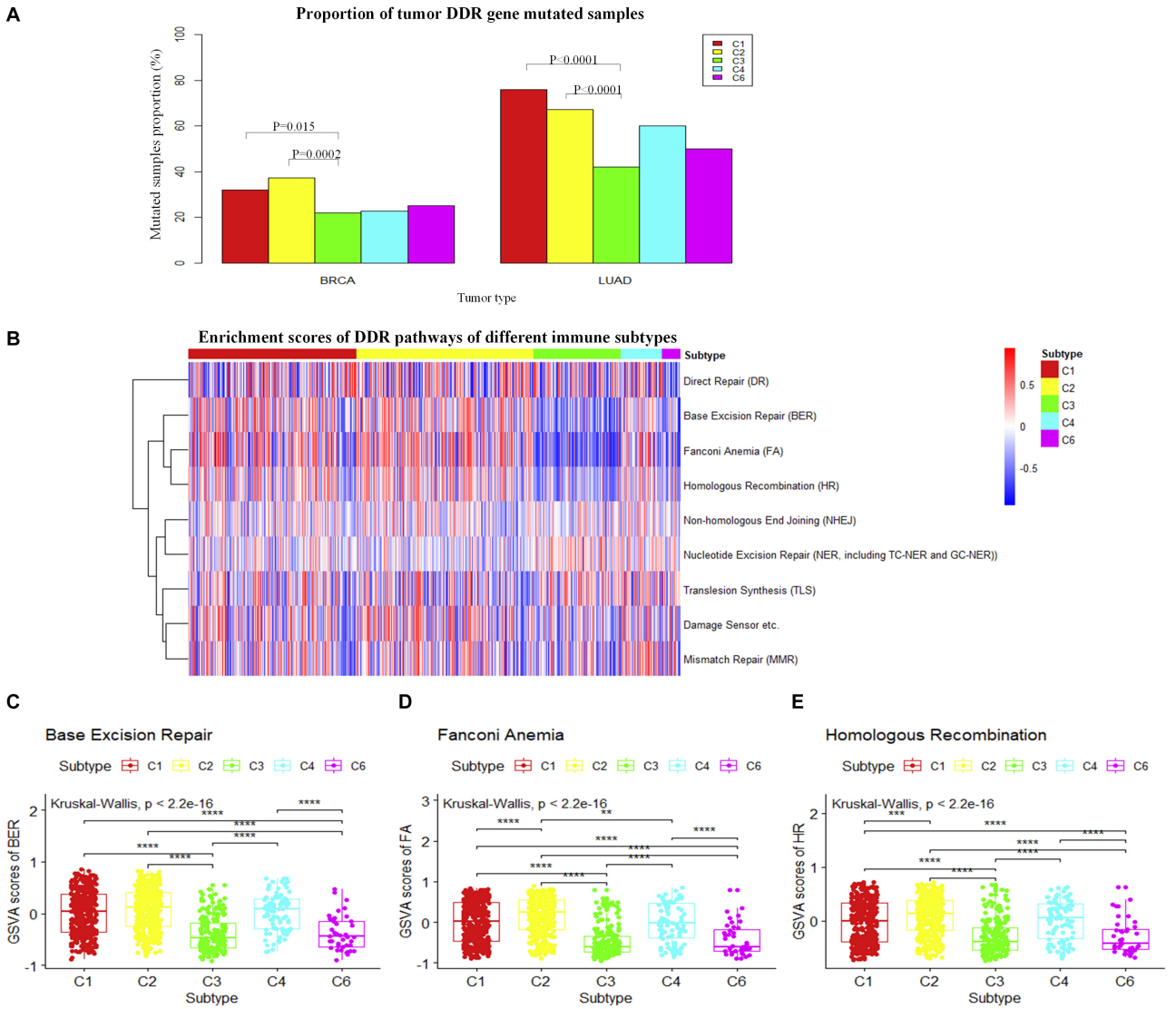


FIGURE 2 | Relationship between immune subtypes and DNA damage repair (DDR) in breast invasive carcinoma. **(A)** The histogram shows the proportion of DDR gene mutations in different immune subtypes. **(B)** The heatmap shows the single-sample gene set variation analysis (ssGSVA) enrichment scores for DDR pathways from immune subtypes. **(C–E)** The boxplot shows the enrichment scores for breast invasive carcinoma immune subtypes in three DDR pathways (Wilcoxon rank-sum test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

Luminal A subtypes are significantly enriched in C3 immune subtype (**Supplementary Figure S2C**), which is consistent with the results that PIK3CA and MAP3K1 gene mutated significantly in C3 immune subtypes. Previous study also suggested that PIK3CA and MAP3K1 alterations imply luminal A status in breast cancer and are associated with clinical benefits from PI3K inhibitors (Nixon et al., 2019). GATA3 gene is significantly mutated in Luminal B subtypes (**Supplementary Figure S2H**), while Luminal B subtypes are significantly enriched in C4 immune subtypes (**Supplementary Figure S2D**), which is consistent with the result that GATA3 gene mutated significantly in C4 immune subtypes. The enrichment analysis between the immune subtypes and the classical molecular subtypes suggest that, for different types of molecular subtypes, their immune environment also show different preference.

These results suggest that MAP3K1 and PIK3CA may be drug targets for patients in C3 subtype. GATA3 may be a potential therapeutic target for patients with the C4 subtype, and TP53 may be a potential therapeutic target for patients with the C2 subtype.

To further explore the differences among immune subtypes of breast invasive carcinoma in the transcriptomic level, from the perspective of signaling pathways, we also performed single-sample gene set enrichment analysis for the same 11 signaling pathways using breast invasive carcinoma samples (**Figure 3D**). Across immune subtypes, the 11 signaling pathways show statistical significance (*P*.adjust $<$ 0.05,

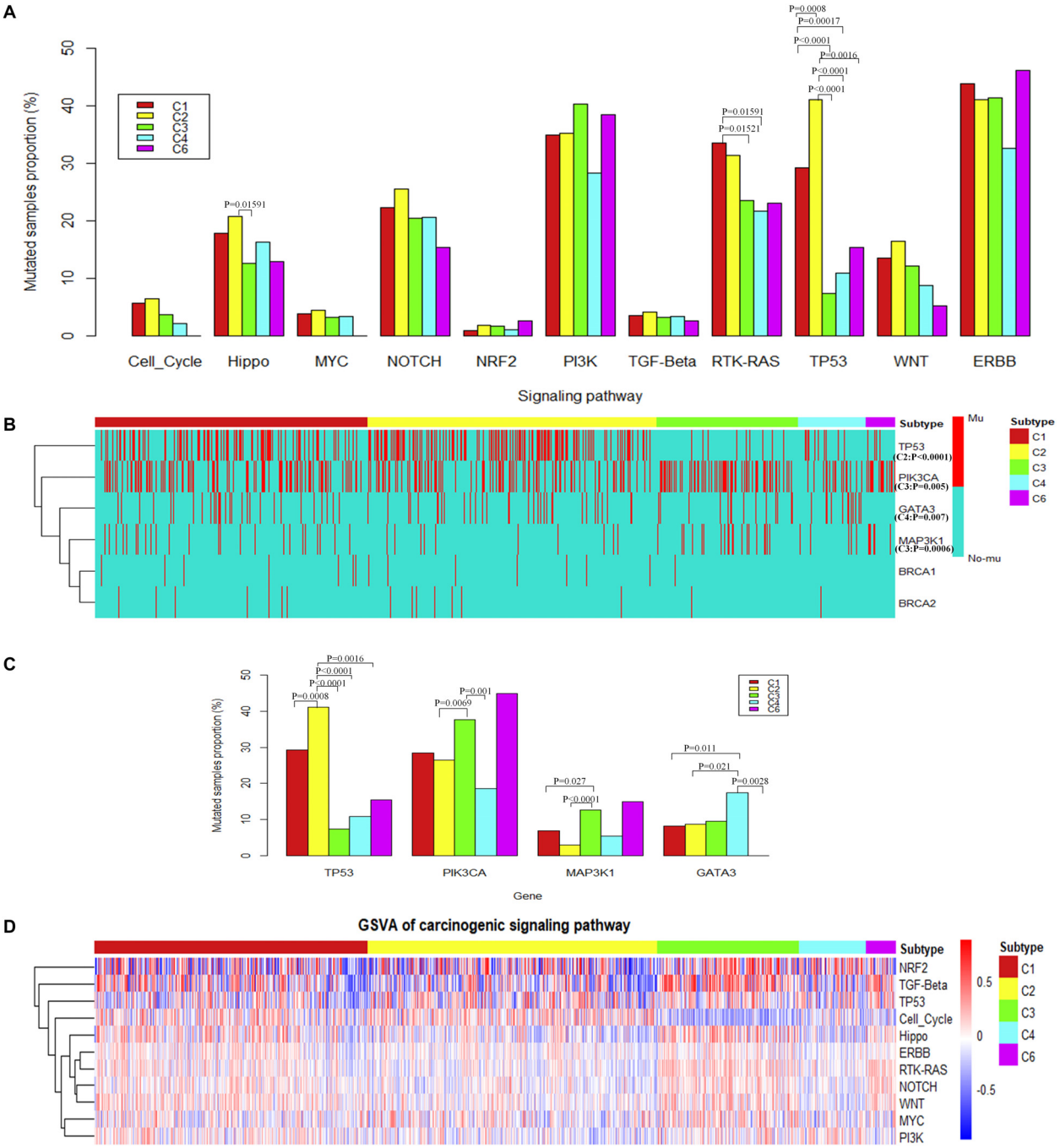


FIGURE 3 | The alteration of carcinogenic signaling pathways in genetic mutations and transcriptional process. **(A)** Proportion of mutated samples for canonical signaling pathways in different immune subtypes (Fisher's exact test, "Mutated samples proportion" is measured as the ratio of the number of samples with mutation in the pathway among the total number of samples in each immune subtype). **(B)** Genes with significant difference of mutations among different immune subtypes (Fisher's exact test) or potential target genes for different immune subtypes. **(C)** The histogram shows the proportion of mutated samples for potential target genes in breast invasive carcinoma immune subtypes. **(D)** The heatmap shows enrichment scores of breast invasive carcinoma immune subtypes in canonical signaling pathways.

Kruskal–Wallis test, Benjamini and Hochberg adjustment) across immune subtypes. Since many drugs can target signaling pathways, understanding subtype-specific target gene expression will provide clues for drug selection. The target genes which show significant differences among immune subtypes are shown in **Supplementary Figure S3B** ($P < 10^{-6}$, Kruskal–Wallis test). In total, our results showed 12 target genes, namely, ataxia telangiectasia mutated (ATM),

TABLE 1 | Therapeutic drugs corresponding to target genes in signaling pathways and which immune subtypes are associated.

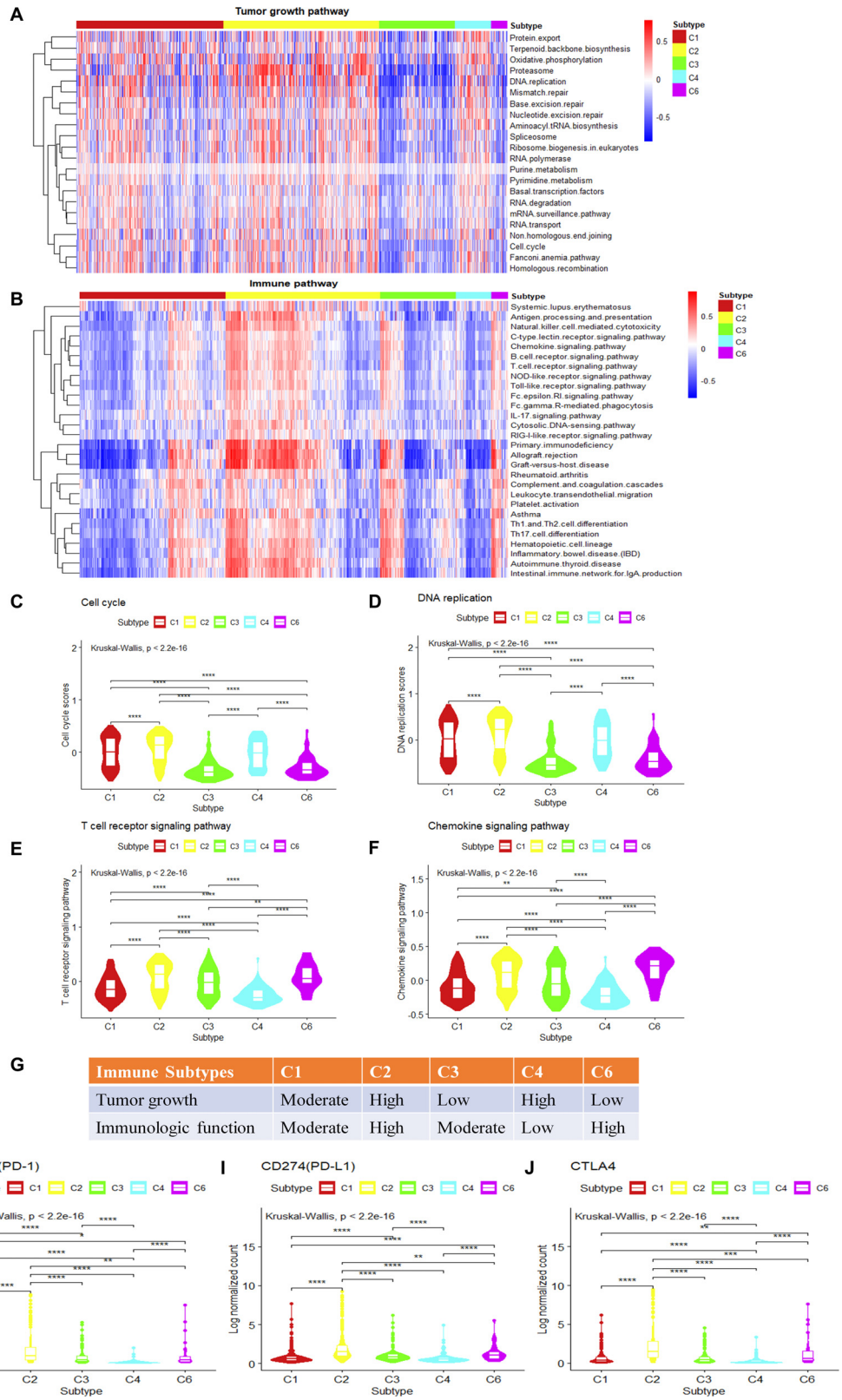
Target	Signaling pathways	Drug	Subtypes
ATM	TP53	Caffeine	C3, C6
BRAF	RTK-RAS, ERBB	Sorafenib	C3
		Vemurafenib	C3
		Regorafenib	C3
		Fostamatinib	C3
		Encorafenib	C3
		Dabrafenib	C3
CDK4	Cell Cycle	Palbociclib	C1, C2
		Abemaciclib	C1, C2
		Fostamatinib	C1, C2
		Ribociclib	C1, C2
CDK6	Cell Cycle	Palbociclib	C2, C6
		Abemaciclib	C2, C6
		Ribociclib	C2, C6
EGFR	RTK-RAS, ERBB	Cetuximab	C1, C2, C3, C6
		Gefitinib	C1,C2,C3,C6
		Erlotinib	C1,C2,C3,C6
		Lapatinib	C1,C2,C3,C6
		Lidocaine	C1,C2,C3,C6
		Necitumumab	C1,C2,C3,C6
		Zalutumumab	C1,C2,C3,C6
		Icotinib	C1,C2,C3,C6
		Vandetanib	C1,C2,C3,C6
		Afatinib	C1,C2,C3,C6
		Osimertinib	C1,C2,C3,C6
		Olmutinib	C1,C2,C3,C6
		Neratinib	C1,C2,C3,C6
		Brigatinib	C1,C2,C3,C6
		Dacomitinib	C1,C2,C3,C6
		Fostamatinib	C1,C2,C3,C6
		Panitumumab	C1,C2,C3,C6
		Zanubrutinib	C1,C2,C3,C6
ERBB2	RTK-RAS, ERBB	Afatinib	C3
		Brigatinib	C3
		Fostamatinib	C3
		Zanubrutinib	C3
		Lapatinib	C3
		Trastuzumab	C3
		Trastuzumab emtansine	C3
		Pertuzumab	C3
FGFR1	RTK-RAS	Regorafenib	C1,C3,C6
		Ponatinib	C1,C3,C6
		Sorafenib	C1,C3,C6
		Lenvatinib	C1,C3,C6
		Nintedanib	C1,C3,C6
		Fostamatinib	C1,C3,C6
		Erdafitinib	C1,C3,C6
FGFR2	RTK-RAS	Thalidomide	C3
		Regorafenib	C3
		Ponatinib	C3
		Nintedanib	C3
		Fostamatinib	C3
		Erdafitinib	C3
		Lenvatinib	C3
NTRK1	RTK-RAS	Imatinib	C3,C6
		Regorafenib	C3,C6
		Fostamatinib	C3,C6
		Larotrectinib	C3,C6
		Entrectinib	C3,C6
NTRK2	RTK-RAS	Larotrectinib	C3,C6
		Entrectinib	C3,C6
		Fostamatinib	C3,C6
NTRK3	RTK-RAS	Fostamatinib	C3,C6
		Larotrectinib	C3,C6
		Entrectinib	C3,C6
PRKCA	ERBB	Ellagic acid	C3,C6
		Midostaurin	C3,C6

ATM, ataxia telangiectasia mutated; CDK, cyclin-dependent kinase; EGFR, epidermal growth factor receptor; ERBB, Erb-B2 receptor tyrosine kinase; FGFR, fibroblast growth factor receptor; NTRK, neurotrophic receptor tyrosine kinase; PRKCA, protein kinase C alpha; RTK, receptor tyrosine kinase; TP53, tumor protein 53. The mechanism of action is given as inhibitors or antagonists (Law et al., 2014).

B-Raf Proto-Oncogene, Serine/Threonine Kinase (BRAF), cyclin-dependent kinase (CDK)4, CDK6, EGF receptor (EGFR), ERBB2, fibroblast growth factor receptor (FGFR)1, FGFR2, neurotrophic receptor tyrosine kinase (NTRK)1, NTRK2, and protein kinase C alpha (PRKCA). The subtype-specific targets and drugs are shown in **Table 1**. Specifically, the high expression levels of CDK4 in C1 and C2 suggest the potential usage of CDK4 inhibitors such as palbociclib and related drugs. The high expression levels of ERBB2 in C3 suggest the potential usage of trastuzumab or lapatinib (or associated drugs). The association of signaling pathway alteration with expression and immune subtypes similarly may provide new ideas in combination drug therapy.

Functional Behaviors in Immune Subtypes

The observation of cell growth potential and immune activities may help to explain prognosis and predict therapeutic opportunities in different subtypes. The tumor proliferation score represents the tumor growth activity, while the leukocyte fraction, to some degree, represents the level of immune activity. Although tumor proliferation and leukocyte fractions have been reported to be statistically significant in different immune subtypes (Wilcoxon rank-sum test) (Thorsson et al., 2018) (**Supplementary Figures S4A,B**), it is not known which functional gene modules cause the observed differences in tumor proliferation and immune microenvironment content. To make a more comprehensive analysis of the functional modules, we further expand our analysis from DNA damage processes and signaling pathways to a more comprehensive pathway set. Single sample gene set enrichment analysis was performed using KEGG pathways with the expression data of breast invasive carcinoma samples (**Supplementary Figure S4C**).



Immune Subtypes	C1	C2	C3	C4	C6
Tumor growth	Moderate	High	Low	High	Low
Immunologic function	Moderate	High	Moderate	Low	High

FIGURE 4 | Single-sample gene set variation analysis (ssGSVA) of immune-related pathways and tumor growth-related pathways for different immune subtypes in breast invasive carcinoma. **(A,B)** The heatmap shows the ssGSVA enrichment scores of breast invasive carcinoma immune subtypes in tumor growth-related

(Continued)

FIGURE 4 | Continued
pathways and immune-related pathways. **(C,D)** The violin plot shows the ssGSVA enrichment scores of breast invasive carcinoma immune subtypes in cell cycle pathway and DNA replication pathway (Wilcoxon rank-sum test was used. $^{****}P < 0.0001$). **(E,F)** The violin plot shows the GSVA enrichment scores of breast invasive carcinoma immune subtypes in T cell receptor (TCR) signaling pathway and chemokine signaling pathway play (Wilcoxon rank-sum test was used. $^{**}P < 0.01$, $^{****}P < 0.0001$). **(G)** Key characteristics of breast invasive carcinoma immune subtypes. **(H–J)** Differences in the expression levels of immune drug targets among breast invasive carcinoma immune subtypes. (Wilcoxon rank-sum test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

The tumor growth-related pathways such as energy metabolism, transcription, translation, replication and repair, folding, sorting and degradation, cell growth and death, nucleotide metabolism show statistical significance among the gene set enrichment scores (P.adjust < 0.05, Kruskal–Wallis test, Benjamini and Hochberg adjustment) among different immune subtypes. The C2 and C4 subtypes present higher enrichment scores; on the contrary, the C3 and C6 subtypes show lower enrichment scores for the tumor growth-related gene sets (P.adjust < 0.05, Kruskal–Wallis test, Benjamini and Hochberg adjustment) (**Figure 4A**). It suggests that the C2 and C4 subtypes might be more active in tumor growth. Among these tumor growth-related pathways, cell cycle and DNA replication were positively correlated with tumor proliferation fraction ($P < 0.05$, Spearman correlation analysis) (**Supplementary Figure S4D**). Comparing gene expression in the cell cycle pathway or the DNA replication pathway among immune subtypes, the C2 and C4 subtypes show significantly higher enrichment scores in tumor growth-related pathways, followed by C1 subtype, C3 and C6 subtypes which have the lowest scores (**Figures 4C,D**). The low tumor growth enrichment scores for C3 and C6 subtypes indicate slow tumor growth.

We also performed GSVA using the immune-related pathways. Results show that 28 immune-related pathways were significantly different among immune subtypes (**Figure 4B**) (P.adjust < 0.05, Kruskal–Wallis test, Benjamini and Hochberg adjustment). Among these immune-related pathways, T cell receptor (TCR) signaling and chemokine signaling pathways were positively correlated with the leukocyte fraction ($P < 0.05$, Spearman correlation) (**Supplementary Figure S4E**). T cell development, differentiation, and maintenance are associated with the antigen-specific TCR and cytokine-mediated signals (Huang and August, 2015). Therefore, TCR signaling pathway and chemokine signaling pathway play a key role in regulating tumor immune microenvironment. The comparison of the gene expression in the T cell signaling pathways or chemokine signaling pathways among immune subtypes shows that the C2 and C6 subtypes have significantly higher enrichment scores, followed by C1 and C3 subtypes, and last, the C4 subtype (**Figures 4E,F**). This is consistent with the previous annotation for C4 as the lymphocyte-depleted subtype (Thorsson et al., 2018).

The C2 subtype shows a high expression level of tumor growth pathways as well as immune-related pathways, and the expressions of immune checkpoint genes such as PD-1, PD-L1, and cytotoxic T lymphocyte-associated antigen (CTLA)4 are also higher than other subtypes (**Figures 4H–J**). These results might suggest the potential usage of both anti-proliferation drugs and immune therapeutic drugs for this subtype. The C4 subtype is rapidly growing for tumor cells but without attracting much immune cells, which may result in a poor prognosis (**Supplementary Figure S1A**). It may also suggest that anti-proliferation drugs might work better for this subtype rather than immune therapy. The C6 subtype shows a high level of expression in immune-related pathways, but with low scores for tumor growth. The poor prognosis (**Supplementary Figure S1A**) for C6 might be caused by other factors that stimulate the activity of the immune system. As C6 is annotated as TGF-β dominant (C6), it also suggests a potential metastatic potential. The relatively high expression of immune drug targets (PD-L1, PD-1, CTLA4) in C6 also suggests the potential usage of immune therapy for this subtype. The C3 subtype shows the best prognosis, and with moderate levels of immune activity and slow tumor growth (**Figure 4G**), it may allow the potential drug combination for anti-proliferation drugs and immune therapeutic drugs. In conclusion, analysis of tumor growth and immune functional activity at the transcriptome level makes some progress in explaining the significant differences observed in the survival rates between immune subtypes as well as provides clues for drug combination selection.

DISCUSSION

The analysis of breast invasive carcinoma immune subtypes is hopefully beneficial to the diagnosis and treatment of breast invasive carcinoma. The pan-cancer classification of immune subtypes is based on immune-related gene sets and molecular markers previously reported in the literature (Thorsson et al., 2018). Our results suggest that mutation types, carcinogenic signaling pathways, and DDR machinery is associated with immune subtypes. The integrative analysis from tumor genetic features and immune subtypes may also provide clues for drug or drug combination selection.

In breast invasive carcinoma, two subtypes (C1 and C2) showed a high frequency of somatic missense mutations and frameshift deletions that may be a result of a failure of DNA double-strand break repair. This is supported by pathway enrichment analysis as well as the relatively high mutation frequency of BRCA1/BRCA2 in these two subtypes, suggesting the potential application of PARP inhibitors. Furthermore, C2 shows high expression of cell cycle-related drug targets such as CDK4, CDK6, as well as immune therapy-related drug targets such as PD-1, PD-L1, CTLA4, suggesting the potential combinatory usage of drugs from multiple categories.

The C3 subtype shows the best prognosis. In breast invasive carcinoma, this subtype shows a low mutation load (fewer number of somatic missense mutations and frameshift deletion) and is enriched with mutations in PIK3CA, with

moderate immune activities and slow tumor growth. All these features suggest that the C3 subtype would be a candidate for treatment with drugs including PI3K inhibitors and anti-proliferation drugs. It also shows a potential for immune therapeutic drug response.

The C4 subtype in breast invasive carcinoma is found to have reduced immune activity coupled with active tumor growth indicating that the C4 subtype tumors are rapidly growing but without attracting immune cells, resulting in a poor prognosis (**Supplementary Figure S1A**). This subtype fits with the idea of "cold" tumors, which cannot be easily targeted by immune therapeutic drugs. So, anti-proliferation drugs might instead be considered.

The C6 subtype in breast invasive carcinoma shows high expression in immune-related pathways and low expression in the tumor growth-related pathways. The poor prognosis (**Supplementary Figure S1A**) for C6 might be caused by other factors that stimulate the activity of the immune system. C6 is annotated as TGF-β dominant (C6), increasing the potential for metastasis. High expression of PD-L1 and PD-1 in this subtype suggests the potential usage of immune therapy for this subtype.

The present study has several limitations. Although most of these observations are similar between breast invasive carcinoma and LUAD, there will likely be tissue-specific features. Therefore, there are likely additional factors that participate in immune subtype formation. Strong tissue specificity reflects differences in inflammatory or immune microenvironments of different tissues. The mutagenesis map of carcinogenic signaling suggests a certain relationship between the signaling pathways and the formation of immune subtypes. The degree to which mutations in signaling pathways are the driving force in the formation of the immune microenvironment still needs experimental verification. Furthermore, our results provide clues for finding drug combinations applicable to immune subtypes; however, for clinical practicality, more detailed experiments must be carried out.

CONCLUSION

This study highlighted important factors potentially affecting the formation of immune subtypes in breast invasive carcinoma and elucidated the potential impact of canonical signaling pathways and DDR on immune subtypes. Functional activities from immune- and tumor growth-related pathways help explain the mechanisms by which there is a significant difference in patient survival between immune subtypes. This study also provides new clues for the therapeutic targets of immune subtypes of breast invasive carcinoma.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

LX and GQ designed the research. YX conducted the research and wrote the manuscript. DL, ZL, DG, LX, and GQ revised the manuscript and performed some of the revised analyses. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Shanghai Municipal Science and Technology Commission of China (No. 17ZR1420300), the National Natural Science Foundation of China (No. 31870829), the Shanghai Municipal Health Commission, and the Collaborative Innovation Cluster Project (No. 2019CXJQ02).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.00053/full#supplementary-material

FIGURE S1 | Prognostic analysis of tumor immune subtypes and gene set variation analysis of DNA damage repair (DDR) in lung adenocarcinoma (LUAD) immune subtypes. **(A,B)** Overall survival analysis and recurrence-free survival analysis for all non-hematologic tumors. **(C–E)** The boxplot shows the gene set variation analysis (GSVA) enrichment scores of LUAD immune subtypes in three DDR pathways (Wilcoxon rank-sum test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

FIGURE S2 | Enrichment analysis with the molecular subtypes in immune subtypes and the alteration of carcinogenic signaling pathways in molecular subtypes. **(A–E)** Enrichment analysis comparing BRCA molecular subtype distribution across C1–C6 immune subtypes (Fisher's exact test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$). **(F)** Proportion of mutated samples for canonical signaling pathways in different molecular subtypes (Fisher's exact test, "Mutated samples proportion" is measured as the ratio of the number of samples with mutation in the pathway among the total number of samples in each molecular subtype. **(G)** Genes with significant difference of mutations among different molecular subtypes or potential target genes for different molecular subtypes. **(H)** The histogram shows the proportion of mutated samples for potential target genes in breast invasive carcinoma molecular subtypes.

FIGURE S3 | Mutation differences and expression differences of genes on signaling pathways in different immune subtypes in breast invasive carcinoma. **(A)** The histogram shows proportion of BRCA1 or BRCA2 mutated samples in breast invasive carcinoma immune subtypes. **(B)** Differences in the gene expression of known drug targets in the signaling pathways among breast invasive carcinoma immune subtypes. (Wilcoxon rank-sum test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

FIGURE S4 | Pathways associated with tumor proliferation and immune microenvironment in immune subtypes. **(A,B)** Tumor proliferation and leukocyte fractions were statistically significant among different immune subtypes from breast invasive carcinoma (Wilcoxon rank-sum test was used. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$). **(C)** The heatmap shows enrichment score of breast invasive carcinoma immune subtypes for KEGG pathways that cover a wide range of functionalities. **(D)** Bar plot of Spearman correlation ecoefficiency between the proliferation fraction and tumor growth related pathways enrichment scores in breast invasive carcinoma. **(E)** Bar plot of Spearman correlation ecoefficiency between the leukocyte fraction and immune-related pathways enrichment scores in breast invasive carcinoma.

TABLE S1 | Gene set specification for oncogenic pathways used in the analysis.

REFERENCES

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.* 47, 1402–1407. doi: 10.1038/ng.3441

Alexandrov, L. B., Nikzainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

Forbes, S. A., Beare, D., Bindal, N., Bamford, S., Ward, S., Cole, C. G., et al. (2016). COSMIC: high-resolution cancer genetics using the catalogue of somatic mutations in cancer. *Curr. Protoc. Hum. Genet.* 91:10.11.1-10.11.37. doi: 10.1002/cphg.21

Fridman, W. H., Zitvogel, L., Sautèsfridman, C., and Kroemer, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* 14, 717–734. doi: 10.1038/nrclinonc.2017.101

Gang, C., Huang, A. C., Wei, Z., Zhang, G., Wu, M., Xu, W., et al. (2018). Exosomal PD-L1 contributes to immunosuppression and is associated with anti-PD-1 response. *Nature* 560, 382–386. doi: 10.1038/s41586-018-0392-8

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Huang, W., and August, A. (2015). The signaling symphony: t cell receptor tunes cytokine-mediated T cell differentiation. *J. Leukoc. Biol.* 97, 477–485. doi: 10.1189/jlb.1RI0614-293R

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., et al. (2016). Genomic and transcriptomic features of response to Anti-PD-1 therapy in metastatic melanoma. *Cell* 165, 35–44. doi: 10.1016/j.cell.2016.02.065

Jeggo, P. A., Pearl, L. H., and Carr, A. M. (2015). DNA repair, genome stability and cancer: a historical perspective. *Nat. Rev. Cancer* 16:35. doi: 10.1038/nrc.2015.4

Knijnenburg, T. A., Wang, L., Zimmermann, M. T., Chambwe, N., Gao, G. F., Cherniack, A. D., et al. (2018). Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas. *Cell Rep.* 23, 239–254. doi: 10.1016/j.celrep.2018.03.076

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, 1091–1097. doi: 10.1093/nar/gkt1068

Li, B., Severson, E., Pignon, J. C., Zhoam, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17:174. doi: 10.1186/s13059-016-1028-7

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28, 1747–1756. doi: 10.1101/gr.239244.118

Minoru, K., and Susumu, G. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Munn, D. H., and Bronte, V. (2016). Immune suppressive mechanisms in the tumor microenvironment. *Curr. Opin. Immunol.* 39, 1–6. doi: 10.1016/j.coi.2015.10.009

Naik, A., Monjazeb, A. M., and Decock, J. (2019). The obesity paradox in cancer, tumor immunology, and immunotherapy: potential therapeutic implications in triple negative breast cancer. *Front. Immunol.* 10:1940. doi: 10.3389/fimmu.2019.01940

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

Nixon, M. J., Formisano, L., Mayer, I. A., Estrada, M. V., González-Ericsson, P. I., Isakoff, S. J., et al. (2019). PIK3CA and MAP3K1 alterations imply luminal A status and are associated with clinical benefit from pan-PI3K inhibitor buparlisib and letrozole in ER+ metastatic breast cancer. *NPJ Breast Cancer* 5:31. doi: 10.1038/s41523-019-0126-6

O'Connor, M. (2015). Targeting the DNA damage response in cancer. *Mol. Cell* 60, 547–560. doi: 10.1016/j.molcel.2015.10.040

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

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2018). The immune landscape of cancer. *Immunity* 48:e14. doi: 10.1016/j.immuni.2018.03.023

Turner, N. C., Ro, J., André, F., Loi, S., Verma, S., Iwata, H., et al. (2015). Palbociclib in hormone-receptor–positive advanced breast cancer. *N. Engl. J. Med.* 373, 209–219. doi: 10.1056/NEJMoa1505270

UCSC Xena (2018). Available online at: http://xena.ucsc.edu/ (accessed March 13, 2018).

Wullenkord, R., Friedrichs, B., Erdmann, T., and Lens, G. (2019). Therapeutic potential of PI3K signaling in distinct entities of B-cell lymphoma. *Exp. Rev. Hematol.* 12, 1053–1062. doi: 10.1080/17474086.2019.1676716

Zheng, X., Vis, J. D., Alejandra, B., Sustic, T., van Wageningen, S., Batra, A. S., et al. (2018). MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Res.* 28, 719–729. doi: 10.1038/s41422-018-0044-4

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

Copyright © 2020 Xu, Li, Liu, Gibbs, Xie and Qin. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 24 April 2020
doi: 10.3389/fmolb.2020.00070

XRCC1 Is a Promising Predictive Biomarker and Facilitates Chemo-Resistance in Gallbladder Cancer

Zhengchun Wu[1], Xiongying Miao[1], Yuanfang Zhang[2], Daiqiang Li[3], Qiong Zou[4], Yuan Yuan[4], Rushi Liu[2] and Zhulin Yang[1]**

[1] Hunan Provincial Key Laboratory of Hepatobiliary Disease Research, Department of General Surgery, Second Xiangya Hospital, Central South University, Changsha, China, [2] Immunodiagnostic Reagents Engineering Research Center of Hunan Province, School of medicine, Hunan Normal University, Changsha, China, [3] Department of Pathology, The Second Xiangya Hospital, Central South University, Changsha, China, [4] Department of Pathology, The Third Xiangya Hospital, Central South University, Changsha, China

Gallbladder cancer is a relatively uncommon human malignant tumor with an extremely poor prognosis. Currently, no biomarkers can accurately diagnose gallbladder cancer and predict patients' prognosis. XRCC1 is involved in tumorigenesis, progression, and chemo-resistance of several human cancers, but the role of XRCC1 in gallbladder cancer is never reported. In this study, we investigated the expression of XRCC1 and its clinicopathological and prognostic significance in gallbladder cancer, and explored the biological role of XRCC1 in gallbladder cancer cells. We found that XRCC1 was significantly up-regulated in gallbladder cancer in protein and mRNA levels. Positive XRCC1 expression was correlated with aggressive clinicopathological features and was an independent poor prognostic factor in gallbladder cancer. The ROC curves suggested that XRCC1 expression had potential clinicopathological diagnostic value in gallbladder cancer. *In vitro*, XRCC1 was overexpression in $CD133^{+}$GBC-SD cells compared to GBC-SD cells. In functional experiment, XRCC1 knockdown had a non-significant impact on proliferation, migration, invasion, and apoptosis of $CD133^{+}$GBC-SD cells. But, XRCC1 knockdown could significantly improve the sensitivity of $CD133^{+}$GBC-SD cells to 5-Fluorouracil via promoting cell necrosis and apoptosis. Thus, this study indicates that XRCC1 may be a promising predictive biomarker of gallbladder cancer and a potential therapeutic target for gallbladder cancer.

Keywords: XRCC1, gallbladder cancer, prognosis, clinicopathological significance, chemo-resistance

OPEN ACCESS

Edited by:
Yongjun Wei,
Zhengzhou University, China

Reviewed by:
Chien-Feng Li,
Chi Mei Medical Center, Taiwan
Jinjin Shi,
Zhengzhou University, China

***Correspondence:**
Rushi Liu
liurushi@hunnu.edu.cn
Zhulin Yang
yangzhulin8@csu.edu.cn

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *19 February 2020*
Accepted: *30 March 2020*
Published: *24 April 2020*

Citation:
Wu Z, Miao X, Zhang Y, Li D, Zou Q, Yuan Y, Liu R and Yang Z (2020) XRCC1 Is a Promising Predictive Biomarker and Facilitates Chemo-Resistance in Gallbladder Cancer. Front. Mol. Biosci. 7:70. doi: 10.3389/fmolb.2020.00070

INTRODUCTION

Gallbladder cancer (GBC) is a relatively uncommon human malignant tumor with an extremely poor prognosis. Histologically, GBC mainly consists of gallbladder adenocarcinoma (AC) (about 90%) and squamous cell/adenosquamous carcinoma (SC/ASC) (accounting for 1–12%) (Roa et al., 2011; Samuel et al., 2018). Although various studies about GBC were performed, GBC clinical outcome remains extremely poor. Currently, radical resection remains the only way to cure

GBC, although adjuvant treatments (chemotherapy and radiotherapy) are available. Because GBC patients at early stages present asymptomatic and are difficult to be diagnosed, most patients are diagnosed at late stages when patients lost the chance to receive radical surgery (Reid et al., 2007; Henley et al., 2015). Additionally, GBC is often resistant to chemotherapy and radiotherapy (Horgan et al., 2012). These above reasons make the prognosis of GBC unsatisfactory. Although many biological marks have been studied, no one can accurately diagnose GBC and predict patients' survival (Sicklick et al., 2016; Sharma et al., 2017). Therefore, discovering reliable early diagnostic biomarkers and exploring the mechanism of treatment resistance are critically important to improve the prognosis of GBC.

DNA repair pathways are related to tumorigenesis and treatment resistance, and base excision repair (BER) is one pathway of DNA repair systems (Wood et al., 2001). BER functions an essential role in protecting the genome against chemical carcinogens and ionizing radiations (Tudek, 2007). Chemotherapy and radiotherapy usually kill tumor cells by causing DNA damage which could be repaired via BER. As a BER protein, x-ray repair cross-complementing group 1 (XRCC1), a 70-kDa protein, is encoded by the gene located on chromosome 19q13.2–13.3 (Thompson and West, 2000). XRCC1 functions as a scaffold protein in the BER and its aberrant expression is associated with carcinogenesis of multiply human malignant tumors (Hanssen-Bauer et al., 2012; Meng et al., 2017; Mei et al., 2019). Currently, researches on XRCC1 mainly concentrate on the relationship between gene polymorphism and cancer susceptibility. Recently, several studies have investigated the clinicopathological and prognostic significance of XRCC1 in human cancers including gastric cancer, ovarian cancer, and non-small cell lung cancer (Wang et al., 2012; Abdel-Fatah et al., 2013; Liu et al., 2015). Moreover, XRCC1 affects the effectiveness of chemotherapy and the role of XRCC1 in chemosensitivity varies in different types of cancer. For example, in gastric and ovarian cancer, patients with low XRCC1 expression exhibited favorable response to platinum-based chemotherapy (Wang et al., 2012; Abdel-Fatah et al., 2013). However, bladder cancer patients with high XRCC1 expression had a favorable chemosensitivity to platinum-based chemotherapy (Sakano et al., 2013). As we know, the role of XRCC1 in GBC is never reported.

Therefore, in this study, we investigated the expression of XRCC1 and its clinicopathological and prognostic significance in gallbladder SC/ASC and AC. Furthermore, the biological function of XRCC1 in $CD133^{+}$GBC cells was evaluated.

MATERIALS AND METHODS

Case Selection

This study was approved by the Ethics Committee for Human Research, Central South University and performed in accordance with the Declaration of Helsinki. The included patients were histologically diagnosed by two pathologists. These patients never received chemotherapy or radiotherapy preoperatively and postoperatively. We collected 69 SC/ASC samples from January 2001 to December 2013 (16 from Xiangya Hospital, 31 from Second Xiangya Hospital, 10 from Third Xiangya Hospital, 5 from Hunan Provincial People Hospital, 5 from Hunan Provincial Tumor Hospital, and 1 each from Changde Central Hospital and Loudi Central Hospital). According to the recommendations of the American Joint Committee on Cancer, tumors with a squamous component $\geq$10% were considered as ASC. The 69 SC/ASCs accounted for 5.5% of 1248 GBCs. We collected 146 AC samples from January 2008 and December 2013 at Second Xiangya Hospital and Third Xiangya Hospital. Survival data for these patients was obtained through letters and telephone calls. The follow-up time was 2 years, and patients who survived over 2 years were considered as censored cases.

EnVision Immunohistochemistry

The rabbit anti-human XRCC1 primary antibody and HRP-conjugated anti-rabbit second antibody were purchased from Santa Cruz Biotechnology (CA, United States). EnVisionTM Detection Kit was purchased from Dako Laboratories (CA, United States). Immunohistochemistry was performed as previously described (Wu et al., 2017). Briefly, four-micrometer-thick sections were cut from routinely paraffin-embedded tissues. The sections were deparaffinized and then incubated with peroxidase inhibitor (3% H_2O_2) in the dark for 15 min, followed by EDTA-trypsin digestion for 15 min. Then, the sections were incubated with primary antibody for 60 min at 37°C. Next, the sections were incubated with the second antibody for 30 min at 37°C after being soaked with PBS for 3 × 5 min. Then, solution A was added to the sections for 30 min, followed by DAB staining and hematoxylin counter-staining. The slides were dehydrated with different concentrations (70%–100%) of alcohol, and soaked in xylene for 3 × 5 min and finally mounted with neutral balsam.

Evaluation of Immunostaining

Ten random fields were examined per section by two independent pathologists. The percent of positively stained cells was determined. Strength of staining was rated on a scale of 1 to 3 (1: little to no positive staining or uncertainly weak staining; 2: weak to moderate staining; 3: moderate to strong staining). A section was determined as positive expression when the percent of positively stained cells was $\geq$10% and staining strength was $\geq$2. The few sections where percent positive staining was 5% to 10% and staining strength was 3 were also regarded as positive.

Western Blot

Total protein was extracted from frozen tissues or cell samples. Protein concentrations were tested via a BCA protein-assay. Protein samples were separated on 10% SDS-PAGE gel. The separated proteins were transferred to Immun-Blot PVDF membrane (Bio-Rad) using a wet transfer system (Bio-Rad). The membrane was blocked with 5% skimmed milk and then incubated with primary antibody (XRCC1, 1:500, proteintech,

China) at 4°C overnight, followed by incubation with HRP-linked anti-rabbit IgG (Merck Millipore) in a dilution of 1: 10000 for 1 h at room temperature.

Real-Time Quantitative PCR (qRT-PCR)

Trizol reagent (Beijing Dingguo Changsheng Biotech, Co., Ltd., China) was applied to extract total RNA. The RNA was reverse-transcribed to cDNA by the PrimeScript RT reagent Kit (Takara Biomedical Tech, Co., Ltd., China). The cDNA was subjected to qRT-PCR using SYBR Premix Ex Taq II (Takara, Co., Ltd., China) and the assay was performed on the CFX connect system (Bio-Rad Co., Ltd., United States). GAPDH was used as an internal control. The primers were synthesized from Tsingke Biological Technology Co., (Changsha, Hunan, China), and sequences of primers were listed as followed: XRCC1: Forward 5′-CCTTTGGCTTGAGTTTTGTACG-3′, Reverse 5′-CCTCCTTCACACGGAACTGG-3′; GAPDH: Forward 5′-ATGACCACAGTCCATGCCATCA-3′, Reverse 5′-TTACTCCTTGGAGGCCATGTAG-3′.

Cell Lines and Culture

The human gallbladder cancer cell line GBC-SD was obtained from the Cell Bank of the Chinese Academy of Sciences (Shanghai, China). Cells were cultured in RPMI-1640 (Hyclone, United States) supplemented with 10% fetal bovine serum (Gibco, Grand Island, NY, United States), Penicillin 100 U/ml and Streptomycin 100 ug/ml (Beyotime, China) in humidified atmosphere at 37°C and 5% CO_2.

Isolation of $CD133^+$cell Population by Magnetic Cell Sorting

For magnetic cell sorting, cells were labeled with CD133 microbeads and sorted using the Miltenyi Biotec CD133 Cell Isolation Kit according to the manufacturer's protocols (Miltenyi Biotec, Germany). Magnetic separation was performed twice to obtain high purity of $CD133^+$cells. The purity of sorted cells was evaluated by flow cytometry with a FACS Calibur machine after labeling with phycoerythrin (PE)-conjugated anti-human CD133 antibody (Biolegend, United States).

Inhibition of XRCC1 Expression by shRNA Transfection

XRCC1 shRNA and negative control shRNA were purchased from GeneChem (Shanghai, China). XRCC1 shRNA or negative control shRNA were mixed with RPIM-1640 (Hyclone, United States) and Lip2000 (Invitrogen, United States), and then incubated at room temperature for 20 min. Approximately 2×10^5 $CD133^+$cells were plated in 6-well plates, followed by treating them with the transfection mixture and incubated at 37°C with 5% CO_2. Cells were harvested at 6 h post-transfection for further studies.

CCK8 Assays

The proliferation of $CD133^+$cells transfected with control shRNA or XRCC1 shRNA was detected by use of Cell Counting Kit-8 (CCK8) (DOJINDO, Japan). Cells were seeded into 96-well culture plates at a density of 1×10^4 cells/100 ul. Four wells of each group were detected every day. At the end of each experiment, CCK-8 solution was added to each well, and the cultures were incubated at 37°C for 4 h. Then, the cultures were detected by use of a microplate reader.

Transwell Assays

Cell migration assays were performed in a 24-well Transwell plate (Corning, United States). Cells in serum-free medium (1×10^5 cells) was added to the upper chamber. Complete medium was added to the bottom wells of the chamber. After 48 h of incubation at 37°C, the cells that did not migrate were removed from the upper face of the filters. The number of cells migrating to the lower face was counted after fixed with 4% formaldehyde and stained with 0.5% crystal violet. The number of cells was counted under a microscope. The cell invasion assay was essentially the same as the migration assays, except that the membrane filters were coated with Matrigel (Becton, Dickinson and Company, United States).

Flow Cytometry Assay for Apoptosis

Cells transfected with control shRNA or XRCC1 shRNA were cultured in a 6-well plate. After 48 h, cells were harvested by trypsinization, washed twice with PBS, and then stained with annexin V-APC (APC) (NanJing KeyGen Biotech, Co., Ltd., China) and propidium iodide (PI) (NanJing KeyGen Biotech, Co., Ltd., China) to detect cell apoptosis. Samples were immediately detected in the flow cytometer. This method can distinguish the cells in early (APC+/PI-) and late (APC+/PI+) apoptosis.

Drug Sensitivity Assay

$CD133^+$cells transfected with XRCC1 shRNA or negative control shRNA were cultured in 96 well plates (1×10^4 cells/well) overnight. On the second day, the cells were treated with 5-Fluorouracil (5-FU, final concentration of 0.1 mg/L) (APExBIO, United States) (Paschall et al., 2016). After 72 h, CCK-8 solution was added to each well, and the cultures were incubated at 37°C for 4 h. Then, the cultures were detected by use of a microplate reader.

$CD133^+$cells transfected with XRCC1 shRNA or negative control shRNA were cultured in 6 well plates (2×10^5 cells/well) overnight. On the second day, the cells were treated with 5-Fluorouracil (5-FU, final concentration of 0.1 mg/L). After 72 h, cells were harvested to assess cell apoptosis by flow cytometry as above.

Statistical Analysis

Data was analyzed using SPSS 13.0. The relationship between XRCC1 expression and clinicopathological factors was analyzed using χ^2 or Fisher's exact test. Kaplan-Meier and Log-rank test were used for univariate survival analysis. Cox proportional hazards model was used for univariate and multivariate analysis. A $P < 0.05$ was considered as statistical significance.

RESULTS

Characteristics of Patients

Among the 69 SC/ASC samples, 44 were collected from female patients and patient ages ranged from 35 to 80 (53.8 ± 10.2) years. Among the 146 AC patients, 85 were female with an age range of 33 to 78 (52.4 ± 9.6) years. The detail clinicopathological information of the 146 SC/ASC patients and the 69 AC patients was presented in **Table 1**. Briefly, among the 69 SC/ASCs, the squamous cell component presented well-differentiated in 19 (27.5%), moderately differentiated in 33 (47.8%), and poorly differentiated in 17 (24.6%). The 146 ACs consisted of 51 well-differentiated types (34.9%), 54 moderately differentiated types (37.0%) and 41 poorly differentiated types (28.1%). Among the SC/ASC patients, invasion to surrounding tissues and organs was observed in 45 patients (65.2%); 42 (60.7%) occurred regional lymph node metastasis; and 38 (55.1%) existed gallstones. Among the 146 AC patients, 74 (50.7%) occurred invasion; 66 (45.2%) presented regional lymph node metastasis; and 68 (46.6%) had gallstones. According to tumor-node-metastasis (TNM) staging, 29 SC/ASCs and 40 SC/ASCs stage I + II and stage III + IV, respectively. Among the 146 ACs, 77 were in a stage of I or II and 69 were in a stage of III or IV. Among all patients, 27 SC/ASC patients and 75 AC patients received radical surgery; 28 SC/ASC patients and 50 AC patients received palliative surgery; 14 SC/ASC patients and 21 AC patients only underwent biops.

TABLE 1 | Comparison of gallbladder SC/ASC and AC clinicopathological characteristics and XRCC1 expression status.

Clinicopathological characteristics	Number of SC/ASC (%)	Number of AC (%)	*P*
Gender			
Male	25 (36.2)	61 (41.8)	0.438
Female	44 (63.8)	85 (58.2)	
Age			
≤45 years	3 (4.3)	20 (13.7)	0.038
>45 years	66 (95.7)	126 (86.3)	
Differentiation			
Well	19 (27.5)	51 (34.9)	0.308
Moderate	33 (47.8)	54 (37.0)	
Poor	17 (24.6)	41 (28.1)	
Maximum tumor diameter			
≤3 cm	39 (56.5)	90 (61.6)	0.474
>3 cm	30 (43.5)	56 (38.4)	
Cholecystolithiasis			
No	31 (44.9)	78 (53.4)	0.245
Yes	38 (55.1)	68 (46.6)	
TNM stages			
I + II	29 (42.0)	77 (52.7)	0.143
III + IV	40 (58.0)	69 (47.3)	
Lymph node metastasis			
No	27 (39.1)	80 (54.8)	0.032
Yes	42 (60.9)	66 (45.2)	
Locoregional invasion			
No	24 (34.8)	72 (49.3)	0.045
Yes	45 (65.2)	74 (50.7)	
Surgical methods			
Radical	27(39.1)	75 (51.4)	0.223
Palliative	28 (40.6)	50 (34.2)	
Without resection	14 (20.3)	21 (14.4)	
XRCC1			
−	28 (40.6)	58 (39.7)	0.905
+	41 (59.4)	88 (60.3)	

−, negative expression; +, positive expression.

XRCC1 Is Significantly Over-Expressed in Gallbladder Cancer Tissues

To evaluate the expression of XRCC1 in GBC tissues and corresponding adjacent non-tumor tissues, qRT-PCR and western blot were performed. The results demonstrated that XRCC1 expression in GBC tissues was significantly higher than adjacent non-tumor tissues both in mRNA and protein levels (**Figures 1A,B**).

We then assessed XRCC1 expression in gallbladder cancer tissues (including 69 SC/ASCs and 146 ACs) and gallbladder epithelium with chronic cholecystitis by immunohistochemistry. The majority of XRCC1 positive-reaction was localized in the cytoplasm of the SC/ASC (**Figure 1C**) and AC (**Figure 1E**). The representative images of XRCC1 negative expression in SC/ASC and AC were seen in **Figure 1D** and **Figure 1F**, respectively. The staining positive rate was significantly higher in SC/ASC (59.4%) and AC (60.3%) than gallbladder epithelium with chronic cholecystitis (6.7%, $P < 0.01$). The epithelium of chronic cholecystitis with high XRCC1 expression showed moderate to severe dysplasia. This suggested that XRCC1 may be a biomarker to evaluate the pre-malignant changes.

Comparison of Gallbladder ASC/SC and AC in Clinicopathological Features Including XRCC1 Expression

As showed in **Table 1**, the percentage of cases with a patient age over 45 years, lymph node metastasis and invasion was significantly higher in SC/ASC compared with AC (all $P < 0.05$). However, there was a non-significant difference between SC/ASC and AC in other clinicopathological features including tumor differentiated degree, tumor size, TNM stages, receiving surgical methods, and XRCC1 positive expression (all $P > 0.05$, **Table1**).

XRCC1 Positive Expression Correlates With Poor Clinicopathological Features of Gallbladder SC/ASC and AC Patients

We further evaluated the clinicopathological significance of XRCC1 expression in SC/ASC and AC patients. We found that XRCC1 positive expression was associated with several poor clinicopathological features of gallbladder cancer. In SC/ASC, XRCC1 positive expression was positively correlated with lymph node metastasis, invasion, and only receiving biopsy (all $P < 0.05$, **Table 2**). Similarly, XRCC1

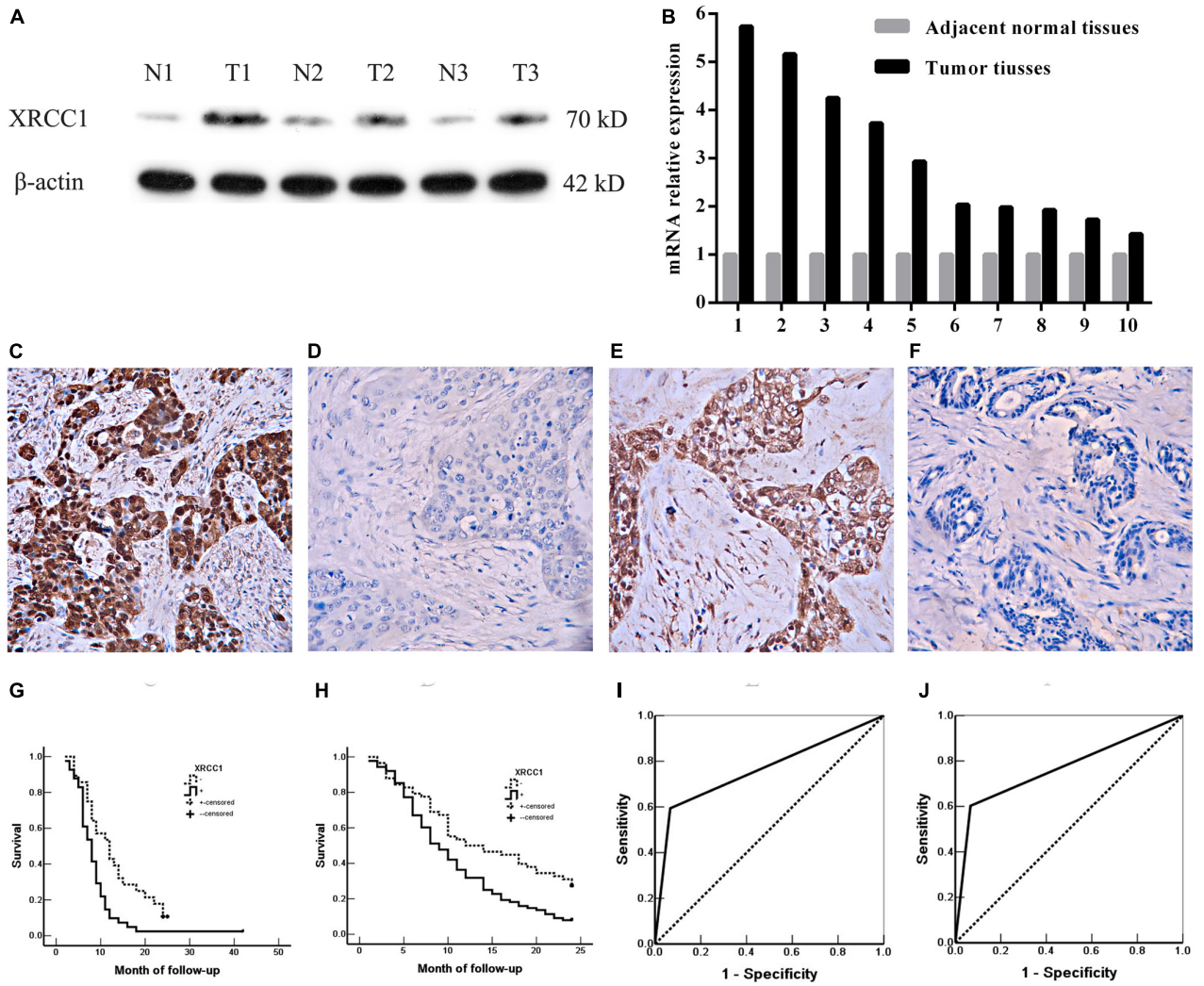


FIGURE 1 | XRCC1 expression is up-regulated in GBC and associated with poor prognosis. (**A**) XRCC1 expression was analyzed by western blot. (**B**) XRCC1 expression was assessed by qRT-PCR. (**C**) Positive expression of XRCC1 in moderately differentiated ASC, ×200. (**D**) Negative expression of XRCC1 in moderately differentiated SC, ×200. (**E**) Positive expression of XRCC1 in moderately differentiated AC, ×200. (**F**) Negative expression of XRCC1 in well differentiated AC, ×200. (**G**) Kaplan-Meier curves for SC/ASC patients with positive and negative XRCC1 expression. (**H**) Kaplan-Meier curves for AC patients with positive and negative XRCC1 expression. (**I**) ROC of diagonal segments was produced by ties of XRCC1 in SC/ASC. (**J**) ROC of diagonal segments was produced by ties of XRCC1 in AC.

positive expression was positively associated with large tumor size (>3 cm), lymph node metastasis, invasion, late TNM stages (III + IV), only receiving biopsy in AC (all $P < 0.05$, **Table 2**).

XRCC1 Positive Expression Is an Independent Risk Factor for the Prognosis of Gallbladder SC/ASC and AC Patients

Gallbladder cancer patients (both AC/ASC and AC) in XRCC1 positive expression group had significantly shorter average survival time than patients in the negative expression group (all $P < 0.01$, **Table 3**). The Kaplan-Meier survival curves demonstrated that patients with XRCC1 positive expression had a poor overall survival than patients with XRCC1 negative expression (**Figures 1G,H**). Moreover, univariate and multivariate analysis showed that XRCC1 positive expression was an independent risk factor for the overall survival of gallbladder SC/ASC and AC patients (**Tables 4**, **5**). Finally, the receiver operating characteristic (ROC) curve was depicted to assess the diagnostic efficacy of XRCC1 expression in SC/ASC and AC. The AUC of XRCC1 expression in SC/ASC and AC was 0.764 (95%CI: 0.669–0.859) and 0.768 (95%CI: 0.689–0.847) respectively (**Figures 1I,J**). These results fully revealed that XRCC1 was closely related to poor survival and might be a novel independent prognosis biomarker for gallbladder SC/ASC and AC patients.

TABLE 2 | Correlations of XRCC1 expression with the clinicopathological characteristics of gallbladder SC/ASC and AC.

Clinicopathological characteristics	SC/ASC			AC		
	Number of patients	Positive Number (%)	*P*	Number of patients	Positive Number (%)	*P*
Differentiation						
Well	19	10 (52.6)	0.738	51	29 (56.9)	0.131
Moderately	33	21 (63.6)		54	29 (53.7)	
Poorly	17	10 (58.8)		41	30 (73.2)	
Tumor size						
≤3cm	39	16 (53.3)	0.366	90	45 (50.0)	0.001
>3cm	30	25 (64.1)		56	43 (76.8)	
Gallstone						
No	31	18 (58.1)	0.836	78	52 (66.7)	0.091
Yes	38	23 (60.5)		68	36 (52.9)	
Lymph node metastasis						
No	27	12 (44.4)	0.037	80	40 (50.0)	0.005
Yes	42	29 (69.1)		66	48 (72.7)	
Invasion						
No	24	10 (41.7)	0.028	72	37 (51.4)	0.030
Yes	45	31 (68.9)		74	51 (68.9)	
TNM stage						
I + II	29	14 (48.3)	0.108	77	38 (49.4)	0.004
III + IV	40	27 (67.5)		69	50 (72.5)	
Surgery						
Radical	27	11 (40.7)	0.031	75	39 (52.0)	0.006
Palliative	28	19 (67.9)		50	30 (60.0)	
Biopsy	14	11 (78.6)		21	19 (90.5)	

XRCC1 Is Significantly Up-Regulated in CD133$^+$GBC-SD Cells Compared With Normal GBC-SD Cells

Previous studies reported that both XRCC1 and CD133$^+$cancer cells are related to tumor drug resistance so that we studied the role of XRCC1 in CD133$^+$GBC-SD cells. CD133$^+$GBC-SD cells were obtained from GBC-SD cells by CD133 magnetic bead sorting. We applied qRT-PCR and western blot to research XRCC1 expression in normal GBC-SD cells and CD133$^+$GBC-SD cells. Compared with GBC-SD cells, XRCC1 mRNA and protein were overexpressed in CD133$^+$GBC-SD cells (**Figure 2**). Based on previous studies, these results indicated that XRCC1 might affect the unique biological features of CD133$^+$GBC-SD cells compared to normal GBC-SD cells, such as chemo-resistance.

Knockdown XRCC1 Has a Non-significant Effect on CD133$^+$GBC-SD Cells Proliferation, Migration, Invasion, and Apoptosis

To further study the function of XRCC1 in CD133$^+$GBC-SD cells, XRCC1 expression in cells was manipulated via short hairpin RNA (shRNA) knockdown. Three shRNAs (shRNA1, shRNA2, and shRNA3) were designed to knockdown XRCC1 expression in CD133$^+$GBC-SD cells. After CD133$^+$GBC-SD cells were infected with XRCC1-shRNA, the expression level of XRCC1 was tested by western blotting to evaluate the efficacy of shRNA knockdown. Among the three XRCC1-shRNAs, shRNA3 was the most effective one (**Figure 3A**) and was selected for further studies. To study the effect of XRCC1 knockdown on the proliferation, migration, invasion, and apoptosis of CD133$^+$GBC-SD cells, CCK8 assay, transwell assay, and flow cytometry were performed. Our results showed that XRCC1 knockdown in CD133$^+$GBC-SD cells had a non-significant impact on the ability of proliferation, migration, invasion, and apoptosis, compared with control-shRNA/CD133$^+$GBC-SD cells (**Figure 3**).

XRCC1 Facilitates CD133$^+$GBC-SD Cells Resistance to 5-FU

To explore the role of XRCC1 in CD133$^+$GBC-SD cells drug resistance, 5-FU was used to treat control-shRNA/CD133$^+$GBC-SD cells and XRCC1-shRNA/CD133$^+$GBC-SD cells. After treated with 5-FU (0.1 mg/L) for 72 h, CCK8 assay was performed to assess the cell totality of every group. Results showed that in CCK8 assay, XRCC1-shRNA/CD133$^+$GBC-SD had a lower absorbance compared to control-shRNA/CD133$^+$GBC-SD, which suggested that XRCC1 could promote CD133$^+$GBC-SD cell resistance to 5-FU (**Figure 4A**). To further validate the results of CCK8 assay, flow cytometry was performed. Under 5-FU treated 72 h, flow cytometry revealed that cell necrosis and apoptosis were significantly increased in XRCC1-shRNA/CD133$^+$GBC-SD compared to control-shRNA/CD133$^+$GBC-SD (**Figure 4B**). Thus, our results

TABLE 3 | Relationship between XRCC1 expression, clinicopathological characteristics and average survival of SC/ASC and AC patients.

Clinicopathological characteristics	SC/ASC			AC		
	Sample (n)	Average survival (month)	*P*	Sample (n)	Average survival (month)	*P*
Differentiation						
Well	19	13.68(5 – 24)	0.000	51	16.69(5 – 24)	0.000
Moderately	33	11.58(4 – 24)		54	12.33(2 – 24)	
Poorly	17	6.12(2 – 14)		41	6.49(1 – 24)	
Tumor size						
≤3cm	30	14.57(6 – 24)	0.000	90	14.60(1 – 24)	0.000
>3cm	39	7.44(2 – 24)		56	8.38(1 – 24)	
Gallstones						
No	31	8.26(3 – 18)	0.008	78	12.19(2 – 24)	0.980
Yes	38	12.90(2 – 24)		68	12.24(1 – 24)	
TNM stage						
I + II	29	16.31(3 – 24)	0.000	77	16.99(3 – 24)	0.000
III + IV	40	6.83(2 – 14)		69	6.88(1 – 24)	
Lymph node metastasis						
No	27	16.04(3 – 24)	0.000	80	16.35(2 – 24)	0.000
Yes	42	7.45(2 – 15)		66	7.20(1 – 24)	
Invasion						
No	24	17.25(3 – 24)	0.000	72	18.08(4 – 24)	0.000
Yes	45	7.38(2 – 20)		74	6.50(1 – 14)	
Surgery						
Radical	27	16.93(5 – 24)	0.000	75	17.84(6 – 24)	0.000
Palliative	28	7.32(2 – 12)		50	6.86(1 – 14)	
Biopsy	14	6.00(4 – 8)		21	4.86(1 – 9)	
XRCC1						
–	28	12.95(4 – 24)	0.002	58	14.47(2 – 24)	0.001
+	41	8.42(2 – 24)		88	10.73(1 – 24)	

–, negative expression; +, positive expression.

TABLE 4 | Univariate Cox regression analysis of survival rate in SC/ASC and AC patients.

Groups	Factors	SC/ASC		AC	
		P	HR (95% CI)	*P*	HR (95% CI)
Differentiated degree	Well/moderately/poorly	0.000	2.040(1.394 – 2.983)	0.000	2.227(1.740 – 2.851)
Tumor size	≤3 cm/>3 cm	0.034	1.765(1.044 – 2.984)	0.000	2.331(1.614 – 3.367)
Gallstone	No/Yes	0.088	1.565(0.935 – 2.261)	0.981	1.004(0.704 – 1.433)
TNM stage	I + II/III + IV	0.000	6.830(3.619 – 12.890)	0.000	5.923(3.898 – 9.002)
Lymph node metastasis	No/Yes	0.000	4.550(2.453 – 8.438)	0.000	5.021(3.312 – 7.612)
Invasion	No/Yes	0.000	5.453(2.942 – 10.104)	0.000	12.808(7.412 – 22.131)
Surgery	Radical/Palliative/Biopsy	0.000	4.240(2.709 – 6.637)	0.000	5.693(4.081 – 7.940)
XRCC1	–/+	0.005	2.125(1.258 – 3.591)	0.002	1.826(1.251 – 2.666)

Abbreviation: HR, hazard risk ratio; CI, confidence interval; –, negative expression; + , positive expression.

indicated that XRCC1 might promote $CD133^{+}$GBC-SD cells resistance to 5-FU through inhibiting cell necrosis and apoptosis.

DISCUSSION

GBC is an aggressive malignant of the biliary tree and consists of several pathological subtypes including AC and SC/ASC. In comparison of AC, the incidence rate of gallbladder SC/ASC is relatively rare and its clinicopathological features remain to be further elucidated. Currently, most reports investigated SC/ASC based on individual cases or small case samples. As far as we know, the 69 SC/ASC cases that we included in this study are relatively large samples in current clinical studies on gallbladder SC/ASC, which could provide more detail clinicopathological knowledge about SC/ASC. In the present study, we found that SC/ASC accounted for 5.5% of GBC and the occurring rate of lymph node metastasis and invasion was significantly higher in

TABLE 5 | Multivariate Cox regression analysis of survival rate in SC/ASC and AC patients.

Groups	Factors	SC/ASC		AC	
		P	HR (95% CI)	*P*	HR (95% CI)
Differentiated degree	Well/moderately/poorly	0.005	1.815(1.198 – 2.750)	0.002	1.514(1.158 – 1.981)
Tumor size	≤3 cm/>3 cm	0.030	1.974(1.067 – 3.653)	0.016	1.772(1.111 – 2.825)
Gallstone	No/Yes	0.461	1.237(0.702 – 2.180)	0.460	1.153(0.791 – 1.679)
TNM stage	I + II/III + IV	0.024	3.662(1.189 – 11.280)	0.002	2.965(1.499 – 5.865)
Lymph node metastasis	No/Yes	0.002	3.823(1.607 – 9.091)	0.000	3.869(2.062 – 7.258)
Invasion	No/Yes	0.016	3.684(1.273 – 10.658)	0.000	6.488(3.287 – 12.809)
Surgery	Radical/Palliative/Biopsy	0.016	1.960(1.132 – 3.393)	0.000	2.284(1.522 – 3.427)
XRCC1	−/+	0.020	1.998(1.116 – 3.576)	0.011	1.721(1.134 – 2.613)

HR, hazard risk ratio; CI, confidence interval; −, negative expression; +, positive expression.

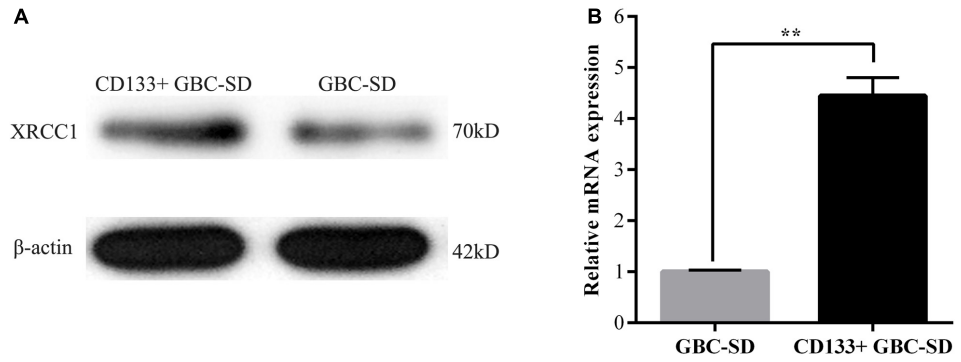


FIGURE 2 | XRCC1 is up-regulated in CD133$^+$ GBC-SD cells. (**A**) XRCC1 expression in GBC-SD cells and CD133$^+$GBC-SD cells was analyzed by western blotting. (**B**) XRCC1 expression in GBC-SD cells and CD133$^+$GBC-SD cells was assessed by qRT-PCR (**$P < 0.01$).

SC/ASC than AC, which was consistent with previous reports (Kim et al., 2011; Roa et al., 2011; Samuel et al., 2018). In agreement with previous researches (Chan et al., 2007; Kim et al., 2011), our results also showed that gallbladder SC/ASC and AC had similar clinicopathological features such as tumor differentiated degree, tumor size, the existence of gallstone, TNM stage, and XRCC1 expression.

Nowadays, the prognosis of GBCs remains extremely poor. In this study, our data revealed that lymph node metastasis, invasion, large tumor size, and advanced TNM stages were independent risk factors for patient's survival, and radical surgery could significantly prolong the mean survival time of patients in SC/ASC and AC. These results suggested that early diagnosis was very important for improving the clinical prognosis of GBC. Thus, it is extremely vital to discover early specific diagnostic biomarkers and explore the reason why GBC resists to chemotherapy. Previous works have demonstrated that XRCC1 is associated with tumor resistance to chemotherapy and radiotherapy, carcinogenesis, and tumor progression (Sak et al., 2005; Hanssen-Bauer et al., 2012; Xu et al., 2014; Li et al., 2018). CD133$^+$cancer cells are a small subgroup of tumor cells and related to tumor resistance to chemotherapy and radiotherapy (Zhang et al., 2010; Desai et al., 2014; Vincent et al., 2014; Kanwal et al., 2018). Thus, we further studied the clinicopathological and prognostic significance of XRCC1 in gallbladder SC/ASC and AC, and evaluated the biological role of XRCC1 in CD133$^+$ GBC-SD cells.

As a DNA repair gene, XRCC1 is involved in tumorigenesis, progression, and poor prognosis of many human cancer types. In this study, we observed that XRCC1 expression was up-regulated in GBC compared with non-tumor tissues, which was consistent with previous studies where XRCC1 was overexpression in ovarian cancer and head and neck squamous cell cancer (Ang et al., 2011; Abdel-Fatah et al., 2013). On the contrary, several reports showed that XRCC1 was down-regulated in glioma, bladder cancer, pancreatic cancer, and gastric cancer (Crnogorac-Jurcevic et al., 2002; Sak et al., 2005; Wang et al., 2012; Mei et al., 2019). This contradiction may be owed to the organ specificity. Furthermore, we found that the epithelium of chronic cholecystitis with high XRCC1 expression showed moderate to severe dysplasia, suggesting that XRCC1 may be involved in the processes that benign lesions evolve into GBC. Thus, we further evaluated the clinicopathological and prognostic significance of XRCC1 in gallbladder SC/ASC and AC. Our data demonstrated that XRCC1 positive expression was significantly related to lymph node metastasis, invasion, and poor prognosis, which was consistent with previous studies (Ang et al., 2011; Abdel-Fatah et al., 2013; Mian et al., 2016). Moreover, cox univariate and multivariate analysis further showed that XRCC1 positive expression was an independent risk factor for the

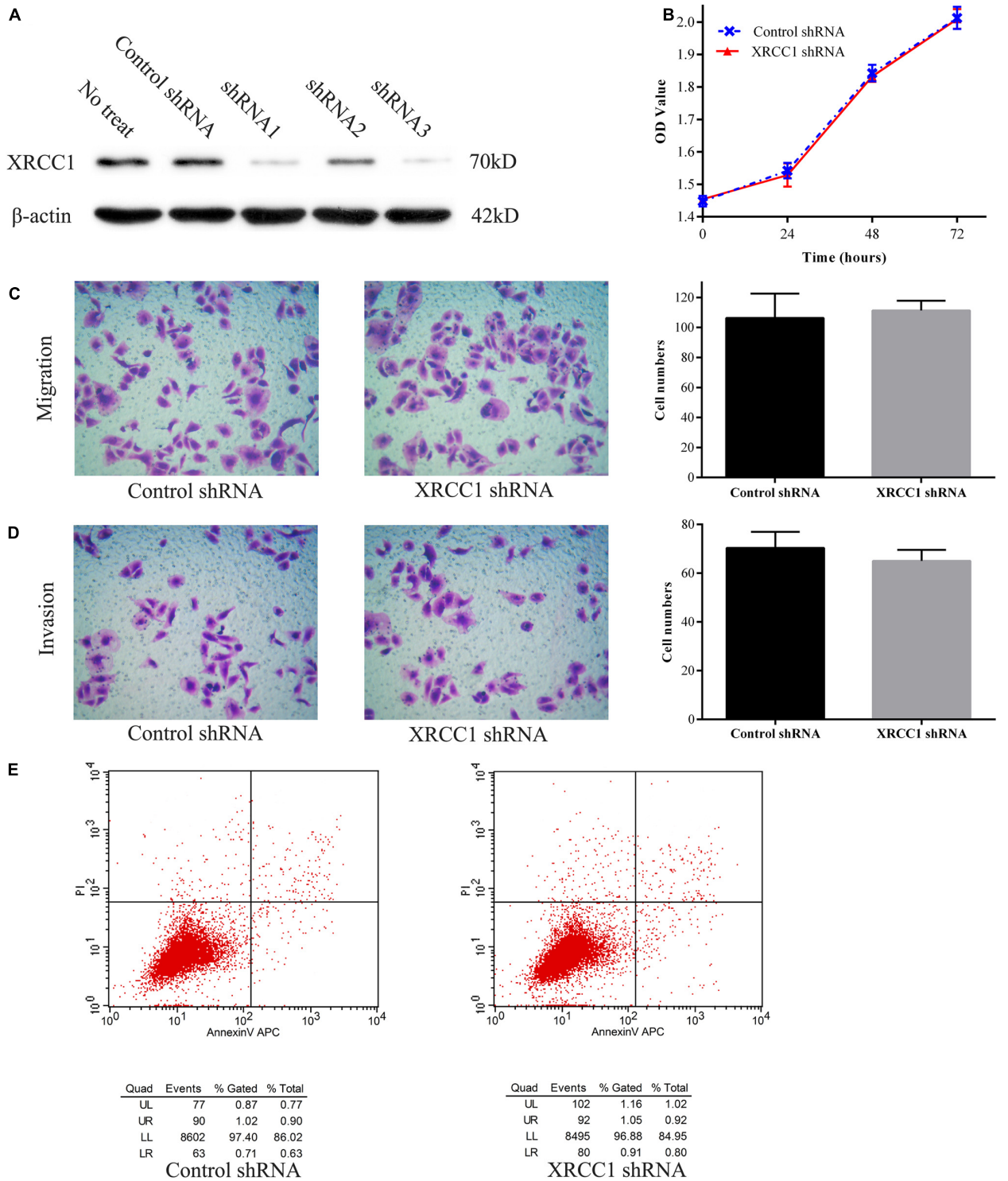


FIGURE 3 | XRCC1 has no effect on proliferation, migration, invasion, and apoptosis of CD133$^+$GBC-SD cells. (**A**) Western blot detected XRCC1 expression in CD133$^+$GBC-SD cells transfected with different shRNAs. (**B**) Proliferation of XRCC1-shRNA/CD133$^+$GBC-SD cells and control-shRNA/CD133$^+$GBC-SD cells was examined by CCK8. (**C**) Transwell migration assay detected the migration capacity of XRCC1-shRNA/CD133$^+$GBC-SD cells and control-shRNA/CD133$^+$GBC-SD cells. (**D**) Transwell invasion assay detected the invasion capacity of XRCC1-shRNA/CD133$^+$GBC-SD cells and control-shRNA/CD133$^+$GBC-SD cells. (**E**) Flow cytometer detected the apoptosis capacity of XRCC1-shRNA/CD133$^+$GBC-SD cells and control-shRNA/CD133$^+$GBC-SD cells; UL: necrosis, UR: late apoptosis, LL: normal cells, LR: early apoptosis.

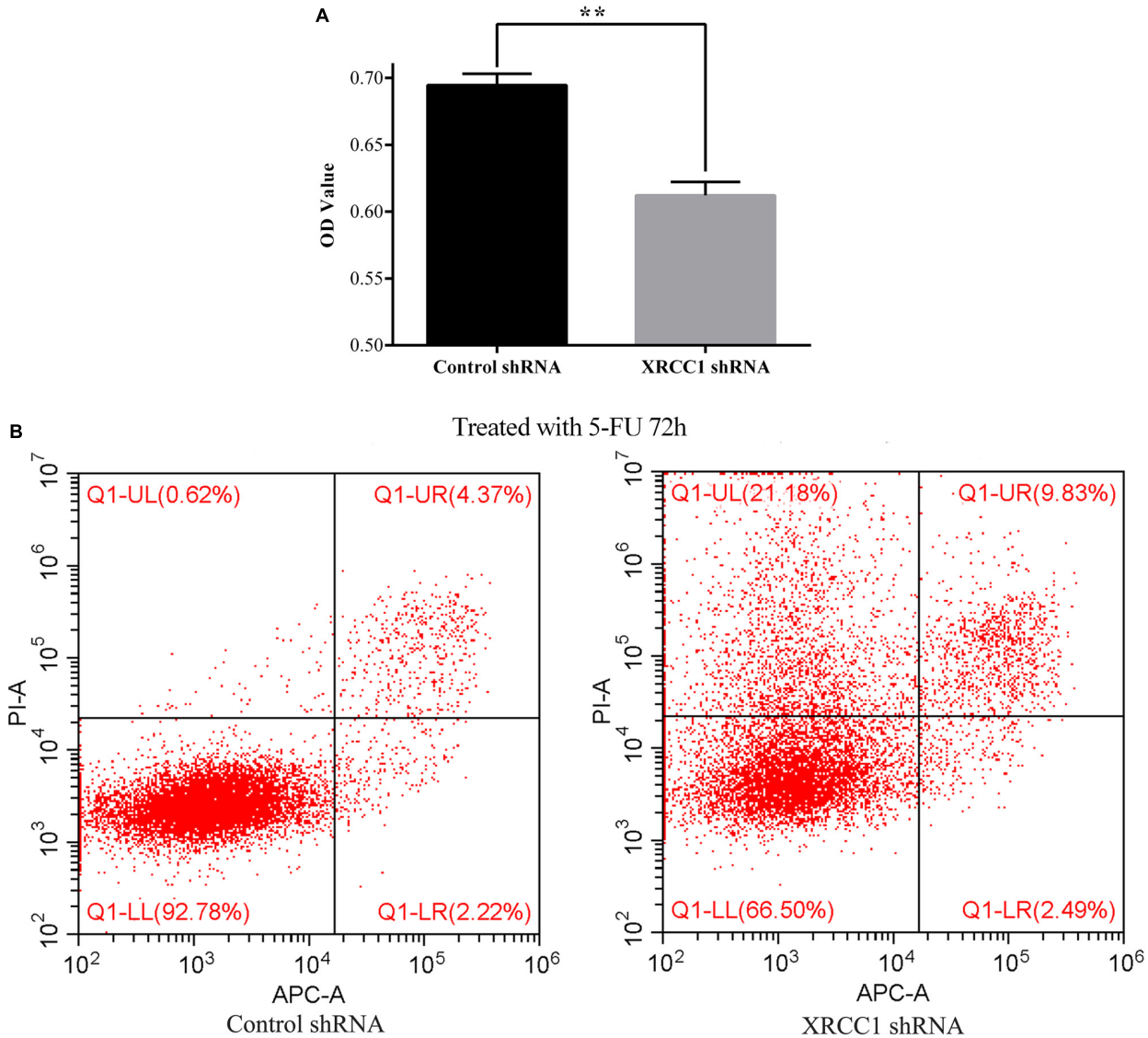


FIGURE 4 | XRCC1 is related to CD133$^+$GBC-SD cells resistance to 5-FU. (**A**) Cell viability of cells treated with 5-FU 72 h was assessed by CCK8 (**$P < 0.01$). (**B**) Apoptosis of cells treated with 5-FU 72 h was examined by Flow cytometer; Q1-UL: necrosis cells, Q1-UR: late apoptosis cells, Q1-LL: normal cells, Q1-LR: early apoptosis cells.

overall survival of SC/ASC and AC. The AUC of XRCC1 indicated that the expression of XRCC1 might have potential clinicopathological diagnostic significance in SC/ASC and AC. These results suggested that XRCC1 might be involved in carcinogenesis and development of GBC.

Previous studies have demonstrated that XRCC1 plays a role in regulating cell biological features such as proliferation, migration, invasion, and drug resistance in several human cancer cell lines (Xu et al., 2014; Meng et al., 2017; Li et al., 2018; Mei et al., 2019). However, there is no study reporting the function of XRCC1 in gallbladder cancer cells. Herein, we firstly studied the biological role of XRCC1 in CD133$^+$GBC-SD cells. Unexpectedly, the functional experiments revealed that knockdown of XRCC1 had no significant effect on the ability of proliferation, migration, invasion, and apoptosis in CD133$^+$GBC-SD cells, which was inconsistent with previous researches (Li et al., 2018; Mei et al., 2019). This inconsistency may be caused by cell specificity. Additionally, we found that XRCC1 was up-regulated in CD133$^+$GBC-SD cells compared with GBC-SD cells, indicating that XRCC1 might be associated with unique biological features of CD133$^+$cancer cells, such as chemo-resistance. Therefore, we further investigated the impact of XRCC1 on CD133$^+$GBC-SD cells resistance to 5-FU. As we suspected, our results showed that XRCC1 were contributed to the resistance of CD133$^+$GBC-SD cells to 5-FU via inhibiting cell necrosis and apoptosis, which was in accordance with previous studies (Abdel-Fatah et al., 2013; Xu et al., 2014). Thus, XRCC1 may promote GBC resistance

to chemotherapy, which needs further studies to validate and explore potential molecular mechanism. Thus, we speculated that XRCC1 might be a promising target to improve the sensitivity of GBC to chemotherapy.

CONCLUSION

In conclusion, this study demonstrated that XRCC1 was overexpression in gallbladder cancer tissues. XRCC1 positive expression was associated with aggressive clinicopathological features and poor prognosis of gallbladder SC/ASC and AC. Moreover, XRCC1 was related to the chemo-resistance of $CD133^{+}$GBC-SD cells to 5-FU. Thus, XRCC1 may be a promising predictive biomarker and a potential therapeutic target for GBC.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

This study was approved by the Ethics Committee for Human Research, Central South University and was carried out in accordance with Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

ZW, RL, and YZ carried out studies and wrote the manuscript. ZY, XM, and RL designed the study and revised the manuscript. ZY and XM performed the statistical analysis. DL, YY, and QZ collected specimens and experimental materials. All authors read and approved the final manuscript.

FUNDING

This work was supported by The National Natural Science Foundation of China (81472738); Natural Science Foundation of Hunan Province, China (2019JJ10002); and Hunan Provincial Key Research and Development Program (2019SK2042).

REFERENCES

Abdel-Fatah, T., Sultana, R., Abbotts, R., Hawkes, C., Seedhouse, C., Chan, S., et al. (2013). Clinicopathological and functional significance of XRCC1 expression in ovarian cancer. *Int. J. Cancer* 132, 2778–2786. doi: 10.1002/ijc.27980

Ang, M. K., Patel, M. R., Yin, X. Y., Sundaram, S., Fritchie, K., Zhao, N., et al. (2011). High XRCC1 protein expression is associated with poorer survival in patients with head and neck squamous cell carcinoma. *Clin. Cancer Res.* 17, 6542–6552.

Chan, K. M., Yu, M. C., Lee, W. C., Jan, Y. Y., and Chen, M. F. (2007). Adenosquamous/squamous cell carcinoma of the gallbladder. *J. Surg. Oncol.* 95, 129–134. doi: 10.1002/jso.20576

Crnogorac-Jurcevic, T., Efthimiou, E., Nielsen, T., Loader, J., Terris, B., Stamp, G., et al. (2002). Expression profiling of microdissected pancreatic adenocarcinomas. *Oncogene* 21, 4587–4594. doi: 10.1038/sj.onc.1205570

Desai, A., Webb, B., and Gerson, S. L. (2014). CD133+ cells contribute to radioresistance via altered regulation of DNA repair genes in human lung cancer cells. *Radiother. Oncol.* 110, 538–545. doi: 10.1016/j.radonc.2013.10.040

Hanssen-Bauer, A., Solvang-Garten, K., Akbari, M., and Otterlei, M. (2012). X-ray repair cross complementing protein 1 in base excision repair. *Int. J. Mol. Sci.* 13, 17210–17229. doi: 10.3390/ijms131217210

Henley, S. J., Weir, H. K., Jim, M. A., Watson, M., and Richardson, L. C. (2015). Gallbladder cancer incidence and mortality, United States 1999-2011. *Cancer Epidemiol. Biomarkers. Prev.* 24, 1319–1326. doi: 10.1158/1055-9965.EPI-15-0199

Horgan, A. M., Amir, E., Walter, T., and Knox, J. J. (2012). Adjuvant therapy in the treatment of biliary tract cancer: a systematic review and meta-analysis. *J. Clin. Oncol.* 30, 1934–1940. doi: 10.1200/JCO.2011.40.5381

Kanwal, R., Shukla, S., Walker, E., and Gupta, S. (2018). Acquisition of tumorigenic potential and therapeutic resistance in CD133+ subpopulation of prostate cancer cells exhibiting stem-cell like characteristics. *Cancer Lett.* 430, 25–33. doi: 10.1016/j.canlet.2018.05.014

Kim, W. S., Jang, K. T., Choi, D. W., Choi, S. H., Heo, J. S., You, D. D., et al. (2011). Clinicopathologic analysis of adenosquamous/squamous cell carcinoma of the gallbladder. *J. Surg. Oncol.* 103, 239–242. doi: 10.1002/jso.21813

Li, Q., Ma, R., and Zhang, M. (2018). XRCC1 rs1799782 (C194T) polymorphism correlated with tumor metastasis and molecular subtypes in breast cancer. *Onco. Targets Ther.* 11, 8435–8444. doi: 10.2147/OTT.S154746

Liu, J. Y., Liu, Q. M., and Li, L. R. (2015). Association of GSTP1 and XRCC1 gene polymorphisms with clinical outcomes of patients with advanced non-small cell lung cancer. *Genet. Mol. Res.* 14, 10331–10337. doi: 10.4238/2015.August.28.19

Mei, P. J., Bai, J., Miao, F. A., Li, Z. L., Chen, C., Zheng, J. N., et al. (2019). Relationship between expression of XRCC1 and tumor proliferation, migration, invasion, and angiogenesis in glioma. *Invest. New Drugs* 37, 646–657. doi: 10.1007/s10637-018-0667-669

Meng, Q., Wang, S., Tang, W., Wu, S., Gao, N., Zhang, C., et al. (2017). XRCC1 mediated the development of cervival cancer through a novel Sp1/Krox-20 swich. *Oncotarget* 8, 86217–86226. doi: 10.18632/oncotarget.21040

Mian, M., McNamara, M. G., Doherty, M., Hedley, D., Knox, J. J., and Serra, S. (2016). Predictive and prognostic values of ERCC1 and XRCC1 in biliary tract cancers. *J. Clin. Pathol.* 69, 695–701. doi: 10.1136/jclinpath-2015-203397

Paschall, A. V., Yang, D., Lu, C., Redd, P. S., Choi, J. H., Heaton, C. M., et al. (2016). CD133+CD24lo defines a 5-Fluorouracil-resistant colon cancer stem cell-like phenotype. *Oncotarget* 7, 78698–78712. doi: 10.18632/oncotarget.12168

Reid, K. M., Ramos-De, L. M. A., and Donohue, J. H. (2007). Diagnosis and surgical management of gallbladder cancer: a review. *J. Gastrointest. Surg.* 11, 671–681. doi: 10.1007/s11605-006-0075-x

Roa, J. C., Tapia, O., Cakir, A., Basturk, O., Dursun, N., Akdemir, D., et al. (2011). Squamous cell and adenosquamous carcinomas of the gallbladder: clinicopathological analysis of 34 cases identified in 606 carcinomas. *Mod. Pathol.* 24, 1069–1078. doi: 10.1038/modpathol.2011.68

Sak, S. C., Harnden, P., Johnston, C. F., Paul, A. B., and Kiltie, A. E. (2005). APE1 and XRCC1 protein expression levels predict cancer-specific survival following radical radiotherapy in bladder cancer. *Clin. Cancer Res.* 11, 6205–6211. doi: 10.1158/1078-0432.CCR-05-0045

Sakano, S., Ogawa, S., Yamamoto, Y., Nishijima, J., Miyachika, Y., Matsumoto, H., et al. (2013). ERCC1 and XRCC1 expression predicts survival in bladder cancer patients receiving combined trimodality therapy. *Mol. Clin. Oncol.* 1, 403–410. doi: 10.3892/mco.2013.85

Samuel, S., Mukherjee, S., Ammannagari, N., Pokuri, V. K., Kuvshinoff, B., Groman, A., et al. (2018). Clinicopathological characteristics and outcomes of rare histologic subtypes of gallbladder cancer over two decades: a population-based study. *PLoS One* 13:e0198809. doi: 10.1371/journal.pone.0198809

Sharma, A., Sharma, K. L., Gupta, A., Yadav, A., and Kumar, A. (2017). Gallbladder cancer epidemiology, pathogenesis and molecular genetics: recent update. *World J. Gastroenterol.* 23, 3978–3998. doi: 10.3748/wjg.v23.i22.3978

Sicklick, J. K., Fanta, P. T., Shimabukuro, K., and Kurzrock, R. (2016). Genomics of gallbladder cancer: the case for biomarker-driven clinical trial design. *Cancer Metastasis Rev.* 35, 263–275. doi: 10.1007/s10555-016-9602-9608

Thompson, L. H., and West, M. G. (2000). XRCC1 keeps DNA from getting stranded. *Mutat. Res.* 459, 1–18. doi: 10.1016/s0921-8777(99)00058-0

Tudek, B. (2007). Base excision repair modulation as a risk factor for human cancers. *Mol. Aspects Med.* 28, 258–275. doi: 10.1016/j.mam.2007.05.003

Vincent, Z., Urakami, K., Maruyama, K., Yamaguchi, K., and Kusuhara, M. (2014). CD133-positive cancer stem cells from Colo205 human colon adenocarcinoma cell line show resistance to chemotherapy and display a specific metabolomic profile. *Genes Cancer* 5, 250–260. doi: 10.18632/genesandcancer.23

Wang, S., Wu, X., Chen, Y., Zhang, J., Ding, J., Zhou, Y., et al. (2012). Prognostic and predictive role of JWA and XRCC1 expressions in gastric cancer. *Clin. Cancer Res.* 18, 2987–2996. doi: 10.1158/1078-0432.CCR-11-2863

Wood, R. D., Mitchell, M., Sgouros, J., and Lindahl, T. (2001). Human DNA repair genes. *Science* 291, 1284–1289. doi: 10.1126/science.1056154

Wu, Z. C., Xiong, L., Wang, L. X., Miao, X. Y., Liu, Z. R., Li, D. Q., et al. (2017). Comparative study of ROR2 and WNT5a expression in squamous/adenosquamous carcinoma and adenocarcinoma of the gallbladder. *World J. Gastroenterol.* 23, 2601–2612. doi: 10.3748/wjg.v23.i14.2601

Xu, W., Wang, S., Chen, Q., Zhang, Y., Ni, P., Wu, X., et al. (2014). TXNL1-XRCC1 pathway regulates cisplatin-induced cell death and contributes to resistance in human gastric cancer. *Cell Death Dis.* 5:e1055. doi: 10.1038/cddis.2014.27

Zhang, Q., Shi, S., Yen, Y., Brown, J., Ta, J. Q., and Le, A. D. (2010). A subpopulation of CD133(+) cancer stem-like cells characterized in human oral squamous cell carcinoma confer resistance to chemotherapy. *Cancer Lett.* 289, 151–160. doi: 10.1016/j.canlet.2009.08.010

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

Copyright © 2020 Wu, Miao, Zhang, Li, Zou, Yuan, Liu and Yang. 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 Molecular Biosciences

MINI REVIEW
published: 26 August 2020
doi: 10.3389/fmolb.2020.00203

Systems Biology of Gastric Cancer: Perspectives on the Omics-Based Diagnosis and Treatment

Xiao-Jing Shi[1], Yongjun Wei[2] and Boyang Ji[3,4]**

[1] Laboratory Animal Center, State Key Laboratory of Esophageal Cancer Prevention and Treatment, Academy of Medical Science, Zhengzhou University, Zhengzhou, China, [2] School of Pharmaceutical Sciences, Key Laboratory of Advanced Drug Preparation Technologies, Ministry of Education, Zhengzhou University, Zhengzhou, China, [3] Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden, [4] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark

Gastric cancer is the fifth most diagnosed cancer in the world, affecting more than a million people and causing nearly 783,000 deaths each year. The prognosis of advanced gastric cancer remains extremely poor despite the use of surgery and adjuvant therapy. Therefore, understanding the mechanism of gastric cancer development, and the discovery of novel diagnostic biomarkers and therapeutics are major goals in gastric cancer research. Here, we review recent progress in application of omics technologies in gastric cancer research, with special focus on the utilization of systems biology approaches to integrate multi-omics data. In addition, the association between gastrointestinal microbiota and gastric cancer are discussed, which may offer insights in exploring the novel microbiota-targeted therapeutics. Finally, the application of data-driven systems biology and machine learning approaches could provide a predictive understanding of gastric cancer, and pave the way to the development of novel biomarkers and rational design of cancer therapeutics.

Keywords: gastric cancer, omics, systems biology, data integration, personalized medicine

OPEN ACCESS

Edited by:
Paula Soares, Universidade do Porto, Portugal

Reviewed by:
Juntaro Matsuzaki, University of California, San Francisco, United States
Nikolay Mikhaylovich Borisov, Moscow Institute of Physics and Technology, Russia

****Correspondence:***
Yongjun Wei
yongjunwei@zzu.edu.cn
Boyang Ji
boyangji@gmail.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *17 April 2020*
Accepted: *27 July 2020*
Published: *26 August 2020*

Citation:
Shi X-J, Wei Y and Ji B (2020) Systems Biology of Gastric Cancer: Perspectives on the Omics-Based Diagnosis and Treatment. Front. Mol. Biosci. 7:203. doi: 10.3389/fmolb.2020.00203

INTRODUCTION

Although the incidences and deaths of gastric cancer are declining in Northern America and Western European, gastric cancer still remains as the fifth most common diagnosed cancer worldwide, and is second compared to lung cancer in terms of worldwide cancer deaths (Bray et al., 2018). Gastric cancer is responsible for over one million new cases and an estimated 783,000 deaths in 2018 (Bray et al., 2018). In Eastern Asia, gastric cancer accounts for ∼31% of all cancer incidences in men and for ∼22% in women. In estimation, most of gastric cancer patients at advanced stages have a 5-year survival rate of <30% (Parkin, 2001). Therefore, early detection and targeted treatment of gastric cancer will be potential therapeutic strategies for increasing the 5-year survival rate of gastric cancer patients.

The vast majority of gastric cancer are adenocarcinomas, which can be classified based on their histological and etiological characteristics. Traditionally, gastric cancer can be divided into two major subtypes: intestinal- and diffuse- types of adenocarcinomas according to the Lauren's criteria (Lauren, 1965). Additionally, the alternative World Health Organization (WHO) classification system differentiates gastric cancer into tubular, papillary, mucinous, and poorly cohesive carcinomas, respectively (Bosman et al., 2010). Both classifications enable a better

understanding of the pathology of gastric cancer. However, these classifications have quite limited success in promoting the development of subtype-specific treatment approaches due to the heterogeneity of gastric cancer and their disability to identify potential molecular targets. With the development of next-generation sequencing (NGS), omics technologies have provided valuable tools to study gastric cancer at the molecular level. Omics based data integration have been extensively applied in gastric cancer research. These studies have successfully identified numerous mutations, gene expression differences, protein abundance differences, epigenetic mutations, and metabolite concentrations to be linked with gastric cancer heterogeneity and staging, which significantly improve our understanding of gastric cancer.

Systems biology approaches aim to the transcendence of individual genes/proteins and to the integration of biological system that taking account into the intrinsic interactions. With more and more available omics data, systems biology approaches have developed many new methods and applications in gastric cancer research. In this review, we will briefly summarize the recent progress in "omics" technologies and their applications in gastric cancer research. We will then highlight the use of omics data integration to classify gastric cancer, and the application of systems approaches and machine learning methods to discover novel biomarkers and potential therapies. Furthermore, how the gastric cancer research shift from human omics to human-microbiota omics for current and future applications will be discussed.

GENOMICS, TRANSCRIPTOMICS, AND EPIGENOMICS IN GASTRIC CANCER

Next-generation sequencing technologies are mainly based on the massively parallel sequencing of short DNA/RNA fragments, which have been extensively reviewed elsewhere (Metzker, 2010). The advances of NGS enable a variety of applications in both DNA and RNA sequencing, including whole-genome, whole-exome, and targeted sequencing of DNA, and total RNA, mRNA, and small RNA. In addition, methylation and ChIP sequencing with NGS are also commonly applied, which remove the biases and limitations generated by previous microarray-based systems (Hurd and Nelson, 2009).

Comprehensive characterization at the genomic, transcriptomic, and epigenomic levels have been applied to define the molecular subgroups of almost all types of cancers. In early studies, the heterogeneity of gastric cancer had been characterized by the expression of a large panel of genes (Cho et al., 2011; Tan et al., 2011). Recently, the genomic landscapes of gastric cancer have been extensively investigated and reviewed elsewhere (Lin et al., 2015; Chia and Tan, 2016; Katona and Rustgi, 2017; Wang et al., 2019). The use of whole genomic data including TCGA (Bass et al., 2014) and ACRG (Cristescu et al., 2015) cohort, have enabled the development of novel and robust molecular classifiers that can guide clinical therapeutics against gastric cancer (**Figure 1**). With unsupervised clustering of molecular data including array-based somatic copy number analysis, array-based DNA methylation profiling, whole-exome sequencing, mRNA sequencing, miRNA sequencing, and reverse-phase protein array (Bass et al., 2014), the gastric cancer can be classified into four subtypes: (1) Epstein–Barr virus (EBV) positive (9%), (2) microsatellite instability (MSI, 22%), (3) genomically stable (GS, 20%), and (4) chromosomal instability (CIN, 50%). Further evaluation of the clinical and histological characteristics of these molecular subtypes revealed the enrichment of the diffuse histological subtype in the GS subtype (Bass et al., 2014). While the ACRG study developed a distinct 4-subtype classification system with gene expression microarray, genome-wide copy number microarrays and targeted gene re-sequencing (Cristescu et al., 2015). As observed in TCGA cohort, gene mutation profiles (e.g., *TP53*) and structural variations are frequently identified in gastric cancer (Zang et al., 2012; Wang et al., 2014; Cristescu et al., 2015; Hu et al., 2016), and these four subtypes show strong associations with clinical phenotypes. Taken together, the accumulation of multiple omics dataset increases the complexity of gastric cancer classification, and the treatment of gastric cancer will be benefit from the clinical-pathological-omics combined subtyping with an individualized way.

Transcriptomics describes the expression levels of RNA transcripts. Gene expression had been shown to dramatically change according to the clinical information of patients, which led to the identification of novel expression biomarkers in patients' group (Tan et al., 2011; Lei et al., 2013). The expression signatures of gastric tumors derived from microarray or NGS had been used to improve the early diagnosis and prognosis prediction (Chia and Tan, 2016). Using 973- and 1024-gene expression signatures, gastric tumors can be distinguished from the normal gastric tissues with high precision in early gastric cancer (Vecchi et al., 2007; Nam et al., 2012). As previous described, gene expression had also been applied for stratification of gastric cancer (Shah et al., 2011; Tan et al., 2011), which reveal distinct transcriptomic subtypes. Moreover, recent advent of single-cell DNA/RNA sequencing provides an opportunity enabling the identification of cell types and state. For instance, the recent study (Zhang et al., 2019) reconstructed single-cell expression atlas underlying the gastric premalignant lesions and early gastric cancer. With expression profiles at the single-cell level, the expression signatures of multiple cell types were identified across different lesions. Furthermore, the single-cell atlas revealed a panel of six high-confidence markers related to early gastric cancer, which could be used as specific biomarkers for early diagnosis targets to recognize the onset of gastric cancer (Zhang et al., 2019). Interestingly, the single-cell RNA sequencing had also been applied to explore the tumor microenvironment of gastric cancer recently (Sathe et al., 2020), which showed distinct expression changes in tumor samples compared with paired normal tissue. The stromal cells, macrophages and cytotoxic T cells were significantly enriched in tumor samples with expression of multiple immune checkpoint and costimulatory molecules (Sathe et al., 2020). Altogether, gene expression profiling at both the population and the single-cell level elucidate the heterogeneity of gastric cancer and the complex relationship between the immune

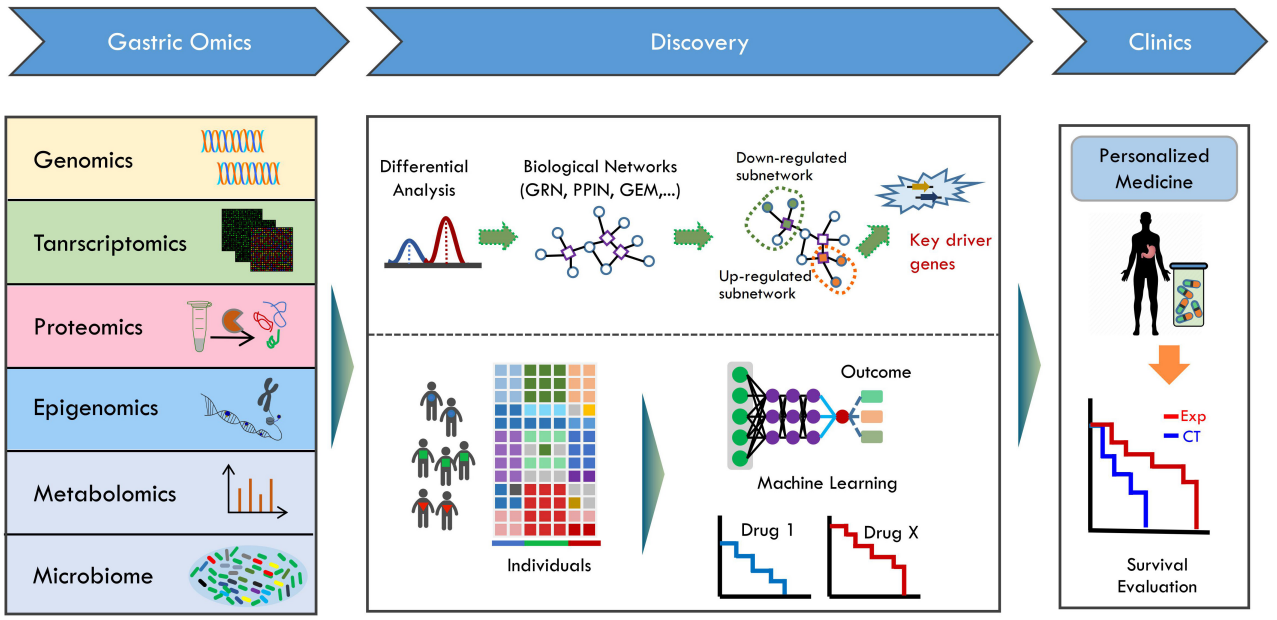


FIGURE 1 | The systems biology approach for gastric cancer research. The different types of omics data including genomics, transcriptomics, proteomics, epigenomics, and metabolomics are obtained using according omics technologies from gastric cancer cohort. Omics data (transcriptome, proteome, or metabolome) is measured for two or more group that differ in clinical information. Generally, differentially expressed genes (DEGs) based methodology is applied. DEGs are identified by comparative analysis of measured omics data. With integrative network-based approach, the biological networks, such as gene regulatory network (GRN), protein-protein interaction network (PPIN), and genome-scale metabolic network (GEM) are used together with omics data in an integrative way. Then, the up-regulated or down-regulated subnetworks are identified by integrating data into network models. Using network modeling tools, the key driver gene linked to clinical information can be identified. Identified key driver genes can be further applied into clinical studies. In addition, the multi-omics data are used to stratify patient into subtypes. Machine learning algorithms utilize omics features to predict potential treatment outcome. The machine learning methodology can be applied to chemotherapy, immunotherapy or their combinations. Finally, the key driver gene information and the predictive models can be used to design the personalized treatment strategy, and applied in clinics. The clinical survival outcome then can be evaluated after personalized medicine treatment.

microenvironment and gastric cancer, which may provide valuable clues to develop rational diagnosis and personalized therapeutic approaches.

Epigenomics describes the modifications of DNA or histones that influence gene expression without altering DNA sequence (Jones and Baylin, 2007). By analyzing the global CpG methylation profiling of gastric cancer and normal tissues, cancer-specific epigenetic alternations were observed in 44% of CpGs in the form of both tumor hyper- and hypomethylation (Toyota et al., 1999; Zouridis et al., 2012). Interestingly, the regions of long-range tumor hypomethylation were strongly associated with increased chromosomal instability (Zouridis et al., 2012). Besides DNA methylation, other types of epigenetic changes, such as histone methylation and acetylation, had been found to be associated with the prognosis of gastric cancer treatment (Calcagno et al., 2019; Li et al., 2019).

PROTEOMICS AND METABOLOMICS IN GASTRIC CANCER

Proteomics complements the genomic and transcriptomics approaches, providing additional information about the protein expression and post-translational modifications. Most of proteomics studies in this field so far focused on the discovery of gastric cancer associated biomarkers from plasma samples (Uen et al., 2013; Abramowicz et al., 2015; Gao et al., 2015; Yoo et al., 2017). An early study (Uen et al., 2013) investigated the glycoprotein profiles of serum samples from gastric cancer patients and healthy subjects. Seventeen significant differentially expressed Con A-bound glycoproteins were identified. Validations using Con A-bound LRG1 glycoprotein revealed an AUC value of 0.65. Another comparative proteomics analysis (Yoo et al., 2017) with serum samples was performed among early gastric cancer, advanced gastric cancer and normal control groups, leading to the identification of hundred protein biomarkers. Using clusterin isoform 1, the highest AUC values to distinguish the advanced or early gastric cancer from normal controls are 0.94 and 0.88, respectively (Yoo et al., 2017). In addition, the comprehensive proteomics studies had also been employed to classify gastric cancer subtypes as genomics data (Ge et al., 2018; Wippel et al., 2018; Mun et al., 2019). The diffuse-type gastric cancer can be further classified into three or four distinct subtypes according to proteome profiling, respectively (Ge et al., 2018; Mun et al., 2019). Moreover, integration of phosphoproteome data with other types of omics data elucidated the signaling pathways associated with somatic mutations (Mun et al., 2019). Most of the metabolomics studies in this field so far focused on the discovery of biomarkers associated with gastric cancer from plasma samples (Abbassi-Ghadi et al., 2013;

Jayavelu and Bar, 2014). Numerous metabolic changes in plasma, urine, gastric juice, and carcinoma tissues had been identified by using targeted or untargeted metabolomics analyses. It provides efficient ways for diagnosis, prognosis, and drug evaluation of gastric cancer, which serves as a potential strategy to develop personalized gastric cancer therapeutics.

GASTROINTESTINAL MICROBIOME IN GASTRIC CANCER

Human microbiome has been confirmed to play critical roles in human health and disease (Knight et al., 2017). The intrinsically heterogeneity of gastric cancer had been extensively explored in decades based on the omics information from human host. However, little is known about how the human microbiota linked to gastric cancer at the function level. Thus, exploring the gastric microbiota at DNA, RNA, and protein level using meta-omics technologies will be helpful for us to understand the potential roles of gastric microbes in cancer development and stage (**Figure 1**).

Helicobacter pylori is one of the gastric pathogen that colonizes in more than 50% persons in the world, and 1% of persons with *H. pylori* infections develop into gastric cancer (Wroblewski et al., 2010; Noto and Peek, 2017; Ferreira et al., 2018). While *H. pylori* was not the dominant bacterial species in some gastric cancer patients, implying other microbes might account for the gastric cancer development (Noto and Peek, 2017). The gastrointestinal microbiota directly interacted with gastric tissue, and affected gastric cancer development (Brawner et al., 2014; Nardone and Compare, 2015). Recent studies indicated that gastric microbiota was strongly associated with gastric cancer (Dias-Jácome et al., 2016). The gastric microbiota of cancer subjects have reduced microbial diversity, decreased *Helicobacter* abundance and the enrichment of other bacterial genera mainly from the intestinal commensals (Ferreira et al., 2018). In addition, significant changes of gut microbiota including microbial richness and diversity were observed in *H. pylori* positive subjects compared to *H. pylori* negative subjects (Guo et al., 2019). Altogether, metagenomics analyses had provided insights into the scenario of gastric microbiota and their interaction with human host. Recently, the drug-microbiota interaction have been extensively investigated (Maier et al., 2018; Vila et al., 2020). However, the influence of gastric cancer treatment, especially the adjunct chemotherapy, on gastric and gut microbiota is still unknown. Therefore, exploring of the gastrointestinal microbiota and gastric cancer associations may provide us novel views in gastric cancer progress and development of microbiota targeted nutrient supplementations or drugs.

DATA-DRIVEN INTEGRATION APPROACHES IN GASTRIC CANCER RESEARCH

Most of gastric studies concentrated on the differential analysis between gastric cancer samples and normal controls using one type of omics data. The comprehensive multi-omics studies of gastric cancer (Bass et al., 2014; Cristescu et al., 2015; Mun et al., 2019) had create a molecular landscape spanning the genome, transcriptome, proteome, and even phosphoproteome. However, there are strong interdependence among different types of omics data. In order to comprehensively understand the gastric cancer and develop efficient diagnosis and treatment approaches, it is critical not only to analyze these omics data as separate layers, but also to dissect how they interact with each another by integrating them together (**Figure 1**).

Cellular processes are represented with networks, whose structures involve in both the species that participate in the biological processes and the interactions between these species (Chiappino-Pepe et al., 2017). The network based multi-omics data integration thus provides us the opportunity to incorporate information across multiple biological layers and describe the gastric cancer (**Figure 1**). For the transcriptome, proteome, and metabolome data, network inference, pathway enrichment analysis and network module identification are three principal steps in network based integration (Borisov et al., 2017; Chiappino-Pepe et al., 2017; Yan et al., 2017). Both the top-down approaches using available experimental data and the bottom-up approaches using reconstructed networks from related organisms as a scaffold to assemble new biological networks with published data are main strategies to infer biological networks (Chiappino-Pepe et al., 2017).

Pathway and network analysis are the two common procedures to explore the functional dynamics linked to cancer. As shown in **Figure 1**, the differentially expressed genes (DEG) are firstly identified using available computational workflows, which are generally performed between gastric cancer samples and normal controls. With the over-expression or under-expression profiles of the DEGs, the related biological pathways are associated with cancer status or stage by pathway enrichment analysis approaches such as gene set enrichment analysis (Subramanian et al., 2005; Buzdin et al., 2017). The DEG-based pathway analysis approach had been successfully applied to identify potential biomarkers distinguishing gastric cancer with normal controls samples using the transcriptomics, proteomics or metabolomics data (Anvar et al., 2018). Nevertheless, DEG-based approach still has a number of limitations, restricting its use in clinics. Firstly, the number of DEGs identified usually exceeds the number that can be experimentally validated. Thus, only parts of DEGs selected according to literature or knowledge are experimentally tested in most of studies. Secondly, not all of DEGs identified are the driver genes for gastric cancer. In fact, it is not easy to discover key driver genes from DEGs, and DEG-based approach cannot always guarantee the successful discovery of key gastric cancer driver genes. Considering such limitations, integrative network-based approach may be useful to intercept omics data and discover cancer driver genes in the context of biological network.

With the predefined biological networks [e.g., protein-protein interaction network (PPIN), gene regulatory network, gene interaction network, and metabolic network], the omics data can be mapped into the biological networks to identify potential functional subnetworks (**Figure 1**). The activity of

subnetwork or modules can be inferred by searching the alternations in predefined networks, providing related regulatory or interaction information linked to clinical information. Furthermore, network-based modeling approaches can be applied to relate the activities of subnetwork components with their influences and consequences on other network components (Creixell et al., 2015). Integrative network analysis utilizing gene expression data identified seven candidates for gastric carcinogenesis with increased levels as disease progression (Takeno et al., 2008; Mansouri et al., 2018). Recent investigations of miRNA and mRNA expression with the human PPIN also reveal a novel miRNA that may function in decreasing gastric tumor proliferation and metastasis through its regulated protein interaction network (Tseng et al., 2011). In summary, transforming the gene-level information to network-level information may provide network biomarkers for understanding the cancer biology (Takeno et al., 2008; Tseng et al., 2011; Mansouri et al., 2018).

MACHINE LEARNING IN GASTRIC CANCER RESEARCH

The applications of machine learning methods, which learn functional relationships from data, had been largely increased in cancer research and drug discovery (Angermueller et al., 2016; Borisov and Buzdin, 2019; Vamathevan et al., 2019; Cuocolo et al., 2020). One important application of machine learning is medical images, and image-based recognition with machine learning had been increasingly applied to diagnosis in various medical fields (Cuocolo et al., 2020). Esophagogastroduodenoscopy (EGD) is the standard procedure for gastric cancer diagnosis. However, the false-negative rate for EGD detection is about 4.6–25.8% (Yalamarthi et al., 2004; Hirasawa et al., 2018). Using convolutional neural networks (CNNs), the machine learning diagnostic system had been trained with >10,000 endoscopic images of gastric cancer (Hirasawa et al., 2018; Yoon and Kim, 2020). The resulting CNN correctly diagnosed 71 of 77 gastric cancer lesions with a overall sensitivity of 92.2% (Hirasawa et al., 2018). Moreover; endoscopic images were used to stratify gastric cancer risk by CNNs, which can diagnose patients as low, moderate, and high risk, respectively (Nakahira et al., 2020).

Not only cancer diagnostics, machine leaning also brings personalized treatment to clinics (Borisov and Buzdin, 2019; Cuocolo et al., 2020). Surgery is the primary treatment for gastric cancer, while the high incidence of distant metastases and the local recurrence of most gastric cancer patients, especially those with advanced gastric cancer, have paved the way for adjuvant therapy (Janunger et al., 2001; Sitarz et al., 2018). The adjuvant treatment may include chemotherapy, targeted drug therapy or immunotherapy, either alone or in combinations (Cunningham et al., 2006). In addition, an emerging chemotherapy method named as neoadjuvant chemotherapy refers to preoperative chemotherapy is recommended for the treatment of patients with resectable advanced-stage gastric cancer (Sitarz et al., 2018). With increased number of omics data linked to gastric cancer treatment, it provided us the opportunities to explore the individual responses to chemotherapy or other types of treatment, and to predict the possible outcome using machine learning and mathematical modeling methods (**Figure 1**). With the gene expression data from TCGA cohort and KUGH cohort, gene expression signatures specific to each of the four molecular subtypes was used to develop predictive models for patients stratification, and the model was tested in other large independent cohorts (Sohn et al., 2017; Oh et al., 2018). Interestingly, these results showed that the subtypes could be as predictors for survival and response to adjuvant chemotherapy (Sohn et al., 2017). Moreover, a recent study characterized key mutational features, copy number alternations and gene expression changes associated with responses to neoadjuvant chemotherapy with multi-omics data of tumor samples from patients responding to neoadjuvant chemotherapy or not (Li et al., 2020). Compared the responders with non-responders tumors and pre- with post-treatment samples, the C10orf71 mutations were found to be associated with treatment resistance by statistical models (Li et al., 2020). Taken together, such machine learning based approach integrates multi-omics data, providing efficient ways to predict the treatment outcome based on the host genetic information.

Immunotherapy has revolutionized both the cancer research and treatment landscape by targeting the host immune system (Coutzac et al., 2019; Szeto and Finley, 2019). Antibodies targeting to blocking immune checkpoints such as programmed cell death-1 (PD-1), programmed death ligand-1 (PD-L1), and cytotoxic T lymphocyte-associated antigen-4 (CTLA-4) have proven efficacies in diverse solid cancers. Several studies had showed the strong correlations between intra-tumoral immune cells and gastric cancer prognosis (Kang B. W. et al., 2017), and the efficiency of checkpoint inhibitors (e.g., nivolumab, pembrolizumab) and their combinations with chemotherapy had been evaluated in clinical trials (Kang Y.-K. et al., 2017; Boku et al., 2019). These results suggest that immunotherapy may be a potential option for patients with advanced gastric cancer. Machine learning has been used to build predictors of drug response and immunotherapy outcomes (Borisov and Buzdin, 2019; Leiserson et al., 2019). However, there is a lack of mechanistic understanding of the effects of gastric cancer immunotherapy in both human host and gastrointestinal microbiota. With the availability of immunotherapy or chemotherapy related multi-omics data, data-driven integration approach and machine learning method will integrate data with known gastric cancer subtyping knowledge in the tumor-specific and patient-specific ways, which can help in stratifying patients before the treatment. In addition, data-driven machine learning or mathematical modeling method may also be useful to learn knowledge and develop predictive models to provide insight into the rational design of cancer therapy in personalized way.

CONCLUSION AND PERSPECTIVES

The advances of omics technologies in decades are enabling the parallel measurement of millions of biomolecules at the same

time. Omics-wide association studies have been widely applied in gastric cancer research, which revealed strong associations between omics features and the gastric cancer development. With the omics data from genome, transcriptome, proteome, and epigenome levels, gastric cancer have been extensively stratified, and the resulting subtypes show strong correlations with the therapeutic outcomes. Both the TCGA and ACRG classifications revealed four distinct gastric cancer subtypes, and the comparison between these two classification systems showed similarities such as tumors with MSI in both data sets, and the TCGA GS, EBV+, and CIN subtypes were enriched in ACRG dataset (Cristescu et al., 2015). However, strong inconsistencies between these two subtype systems were also observed, which covered most of the patient population. The wide variation in study designs, heterogeneity in study cohorts, together with the variations in data analysis strategy, especially in data processing and analysis methods, make the findings of gastric cancer subtyping difficult to applied in clinics (van den Boorn et al., 2018). Therefore, applying robust statistical methods and performing meta-analyses pooling estimates from multiple multi-omics studies may provide a powerful way to investigate gastric cancer across multiple cohorts.

With the proteomics and metabolomics data, numerous gastric cancer-specific biomarkers had been identified, which pave ways for the diagnosis of gastric cancer at the early stages. Systems biology based integration of multi-omics data have provided lot of insights into the cancer diagnosis and therapeutics. However, the application of such methods in gastric cancer still lags behind. Moreover, the application of big data and machine learning approach in gastric cancer studies are still limited. With increased omics data generating from the gastric cancer research field, the application of systems biology approach would provide a systematic scenario of gastric cancer in the future.

AUTHOR CONTRIBUTIONS

YW and BJ conceived the study. X-JS, YW, and BJ wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 31800079 and 31801957) and the Scientific Program of Henan Province (No. 19A350012 for X-JS).

ACKNOWLEDGMENTS

The authors would like to thank Dr. Xin Chen for insightful discussions and the help in preparing the manuscript.

REFERENCES

Abbassi-Ghadi, N., Kumar, S., Huang, J., Goldin, R., Takats, Z., and Hanna, G. B. (2013). Metabolomic profiling of oesophago-gastric cancer: a systematic review. *Eur. J. Cancer* 49, 3625–3637. doi: 10.1016/j.ejca.2013.07.004

Abramowicz, A., Wojakowska, A., Gdowicz-Klosok, A., Polanska, J., Rodziewicz, P., Polanowski, P., et al. (2015). Identification of serum proteome signatures of locally advanced and metastatic gastric cancer: a pilot study. *J. Transl. Med.* 13:304. doi: 10.1186/s12967-015-0668-9

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

Anvar, M. S., Minuchehr, Z., Shahlaei, M., and Kheitan, S. (2018). Gastric cancer biomarkers; a systems biology approach. *Biochem. Biophys. Rep.* 13, 141–146.

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Boku, N., Ryu, M.-H., Kato, K., Chung, H. C., Minashi, K., Lee, K.-W., et al. (2019). Safety and efficacy of nivolumab in combination with S-1/capecitabine plus oxaliplatin in patients with previously untreated, unresectable, advanced, or recurrent gastric/gastroesophageal junction cancer: interim results of a randomized, phase II trial (ATTRACTION-4). *Ann. Oncol.* 30, 250–258. doi: 10.1093/annonc/mdy540

Borisov, N., and Buzdin, A. (2019). New paradigm of machine learning (ML) in personalized oncology: data trimming for squeezing more biomarkers from clinical datasets. *Front. Oncol.* 9:658. doi: 10.3389/fonc.2019.00658

Borisov, N., Suntsova, M., Sorokin, M., Garazha, A., Kovalchuk, O., Aliper, A., et al. (2017). Data aggregation at the level of molecular pathways improves stability of experimental transcriptomic and proteomic data. *Cell Cycle* 16, 1810–1823. doi: 10.1080/15384101.2017.1361068

Bosman, F. T., Carneiro, F., Hruban, R. H., and Theise, N. D. (2010). *WHO Classification Of Tumours Of The Digestive System*. Geneva: World Health Organization.

Brawner, K. M., Morrow, C. D., and Smith, P. D. (2014). Gastric microbiome and gastric cancer. *Cancer J.* 20, 211–216. doi: 10.1097/PPO.0000000000000043

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

Buzdin, A. A., Prassolov, V., Zhavoronkov, A. A., and Borisov, N. M. (2017). "Bioinformatics meets biomedicine: oncofinder, a quantitative approach for interrogating molecular pathways using gene expression data," in *Biological Networks and Pathway Analysis*, eds T. V. Tatarinova and Y. Nikolsky (New York, NY: Springer), 53–83.

Calcagno, D. Q., Wisnieski, F., Mota, E. R. D. S., Maia de Sousa, S. B., Costa da Silva, J. M., Leal, M. F., et al. (2019). Role of histone acetylation in gastric cancer: implications of dietetic compounds and clinical perspectives. *Epigenomics* 11, 349–362. doi: 10.2217/epi-2018-0081

Chia, N.-Y., and Tan, P. (2016). Molecular classification of gastric cancer. *Ann. Oncol.* 27, 763–769. doi: 10.1093/annonc/mdw040

Chiappino-Pepe, A., Pandey, V., Ataman, M., and Hatzimanikatis, V. (2017). Integrating of metabolic, regulatory and signaling networks towards analysis of perturbation and dynamic responses. *Curr. Opin. Syst. Biol.* 2, 59–66. doi: 10.1016/j.coisb.2017.01.007

Cho, J. Y., Lim, J. Y., Cheong, J. H., Park, Y.-Y., Yoon, S.-L., Kim, S. M., et al. (2011). Gene expression signature-based prognostic risk score in gastric cancer. *Clin. Cancer Res.* 17, 1850–1857. doi: 10.1158/1078-0432.CCR-10-2180

Coutzac, C., Pernot, S., Chaput, N., and Zaanan, A. (2019). Immunotherapy in advanced gastric cancer, is it the future? *Crit. Rev. Oncol. Hematol.* 133, 25–32. doi: 10.1016/j.critrevonc.2018.10.007

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

Cristescu, R., Lee, J., Nebozhyn, M., Kim, K.-M., Ting, J. C., Wong, S. S., et al. (2015). Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. Med.* 21, 449–456. doi: 10.1038/nm.3850

Cunningham, D., Allum, W. H., Stenning, S. P., Thompson, J. N., Van de Velde, C. J. H., Nicolson, M., et al. (2006). Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N. Engl. J. Med.* 355, 11–20.

Cuocolo, R., Caruso, M., Perillo, T., Ugga, L., and Petretta, M. (2020). Machine learning in oncology: a clinical appraisal. *Cancer Lett.* 481, 55–62. doi: 10.1016/j.canlet.2020.03.032

Dias-Jácome, E., Libânio, D., Borges-Canha, M., Galaghar, A., and Pimentel-Nunes, P. (2016). Gastric microbiota and carcinogenesis: the role of non-*Helicobacter* pylori bacteria: a systematic review. *Rev. Española Enfermedades. Dig.* 108, 530–540.

Ferreira, R. M., Pereira-Marques, J., Pinto-Ribeiro, I., Costa, J. L., Carneiro, F., Machado, J. C., et al. (2018). Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. *Gut* 67, 226–236. doi: 10.1136/gutjnl-2017-314205

Gao, W., Xu, J., Wang, F., Zhang, L., Peng, R., Shu, Y., et al. (2015). Plasma membrane proteomic analysis of human Gastric Cancer tissues: revealing flotillin 1 as a marker for Gastric cancer. *BMC Cancer* 15:367. doi: 10.1186/s12885-015-1343-5

Ge, S., Xia, X., Ding, C., Zhen, B., Zhou, Q., Feng, J., et al. (2018). A proteomic landscape of diffuse-type gastric cancer. *Nat. Commun.* 2018, 1–16. doi: 10.1038/s41467-018-03121-2

Guo, Y., Zhang, Y., Gerhard, M., Gao, J.-J., Mejias-Luque, R., Zhang, L., et al. (2019). Effect of *Helicobacter pylori* on gastrointestinal microbiota: a population-based study in Linqu, a high-risk area of gastric cancer. *Gut* doi: 10.1136/gutjnl-2019-319696

Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., et al. (2018). Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gast. Cancer* 21, 653–660. doi: 10.1007/s10120-018-0793-2

Hu, N., Kadota, M., Liu, H., Abnet, C. C., Su, H., Wu, H., et al. (2016). Genomic landscape of somatic alterations in esophageal squamous cell carcinoma and gastric cancer. *Cancer Res.* 76, 1714–1723. doi: 10.1158/0008-5472.can-15-0338

Hurd, P. J., and Nelson, C. J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings Funct. Genom. Proteom.* 8, 174–183. doi: 10.1093/bfgp/elp013

Janunger, K. G., Hafström, L., Nygren, P., and Glimelius, B. (2001). A systematic overview of chemotherapy effects in gastric cancer. *Acta Oncol.* 40, 309–326. doi: 10.1080/02841860151116385

Jayavelu, N. D., and Bar, N. S. (2014). Metabolomic studies of human gastric cancer: review. *World J. Gastroenterol.* 20, 8092–8101. doi: 10.3748/wjg.v20.i25.8092

Jones, P. A., and Baylin, S. B. (2007). The epigenomics of cancer. *Cell* 128, 683–692.

Kang, B. W., Kim, J. G., Lee, I. H., Bae, H. I., and Seo, A. N. (2017). Clinical significance of tumor-infiltrating lymphocytes for gastric cancer in the era of immunology. *World J. Gastrointest. Oncol.* 9:293. doi: 10.4251/wjgo.v9.i7.293

Kang, Y.-K., Boku, N., Satoh, T., Ryu, M.-H., Chao, Y., Kato, K., et al. (2017). Nivolumab in patients with advanced gastric or gastro-oesophageal junction cancer refractory to, or intolerant of, at least two previous chemotherapy regimens (ONO-4538-12, ATTRACTION-2): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet* 390, 2461–2471. doi: 10.1016/s0140-6736(17)31827-5

Katona, B. W., and Rustgi, A. K. (2017). Gastric cancer genomics: advances and future directions. *Cell Mol. Gastroenterol. Hepatol.* 3, 211–217. doi: 10.1016/j.jcmgh.2017.01.003

Knight, R., Callewaert, C., Marotz, C., Hyde, E. R., Debelius, J. W., McDonald, D., et al. (2017). The microbiome and human biology. *Annu. Rev. Genomics Hum. Genet.* 18, 65–86.

Lauren, P. (1965). The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma: an attempt at a histo-clinical classification. *Acta Pathol. Microbiol. Scand.* 64, 31–49. doi: 10.1111/apm.1965.64.1.31

Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., et al. (2013). Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* 145, 554–565. doi: 10.1053/j.gastro.2013.05.010

Leiserson, M. D. M., Syrgkanis, V., Gilson, A., Dudik, M., Gillett, S., Chayes, J., et al. (2019). A multifactorial model of T cell expansion and durable clinical benefit in response to a PD-L1 inhibitor. *PLoS One* 13:e0208422. doi: 10.1371/journal.pone.0208422

Li, Y., Guo, D., Sun, R., Chen, P., Qian, Q., and Fan, H. (2019). Methylation patterns of Lys9 and Lys27 on Histone H3 correlate with patient outcome in gastric cancer. *Dig. Dis. Sci.* 64, 439–446. doi: 10.1007/s10620-018-5341-8

Li, Z., Gao, X., Peng, X., May Chen, M.-J., Li, Z., Wei, B., et al. (2020). Multi-omics characterization of molecular features of gastric cancer correlated with response to neoadjuvant chemotherapy. *Sci. Adv.* 6:eaay4211. doi: 10.1126/sciadv.aay4211

Lin, X., Zhao, Y., Song, W., and Zhang, B. (2015). Molecular classification and prediction in gastric cancer. *Comput. Struct. Biotechnol. J.* 13, 448–458. doi: 10.1016/j.csbj.2015.08.001

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* 555, 623–628. doi: 10.1038/nature25979

Mansouri, V., Tavirani, S. R., Zadeh-Esmaeel, M.-M., Rostami-Nejad, M., and Rezaei-Tavirani, M. (2018). Comparative study of gastric cancer and chronic gastritis via network analysis. *Gastroenterol. Hepatol. Bed Bench* 11:343.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Mun, D.-G., Bhin, J., Kim, S., Kim, H., Jung, J. H., Jung, Y., et al. (2019). Proteogenomic characterization of human early-onset gastric cancer. *Cancer Cell* 35, 111–124.

Nakahira, H., Ishihara, R., Aoyama, K., Kono, M., Fukuda, H., Shimamoto, Y., et al. (2020). Stratification of gastric cancer risk using a deep neural network. *JGH Open* 4, 466–471. doi: 10.1002/jgh3.12281

Nam, S., Lee, J., Goh, S.-H., Hong, S.-H., Song, N., Jang, S.-G., et al. (2012). Differential gene expression pattern in early gastric cancer by an integrative systematic approach. *Int. J. Oncol.* 41, 1675–1682. doi: 10.3892/ijo.2012.1621

Nardone, G., and Compare, D. (2015). The human gastric microbiota: is it time to rethink the pathogenesis of stomach diseases? *Unit. Eur. Gastroenterol. J.* 3, 255–260. doi: 10.1177/2050640614566846

Noto, J. M., and Peek, R. M. Jr. (2017). The gastric microbiome, its interaction with *Helicobacter* pylori, and its potential role in the progression to stomach cancer. *PLoS Pathog.* 13:e1006573. doi: 10.1371/journal.pone.1006573

Oh, S. C., Sohn, B. H., Cheong, J., Kim, S., Lee, J. E., Park, K. C., et al. (2018). Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. *Nat. Commun.* 9:1777. doi: 10.1038/s41467-018-04179-8

Parkin, D. M. (2001). Global cancer statistics in the year 2000. *Lancet Oncol.* 2, 533–543. doi: 10.1016/s1470-2045(01)00486-7

Sathe, A., Grimes, S. M., Lau, B. T., Chen, J., Suarez, C., Huang, R. J., et al. (2020). Single-Cell genomic characterization reveals the cellular reprogramming of the gastric tumor microenvironment. *Clin. Cancer Res.* doi: 10.1158/1078-0432.CCR-19-3231

Shah, M. A., Khanin, R., Tang, L., Janjigian, Y. Y., Klimstra, D. S., Gerdes, H., et al. (2011). Molecular classification of gastric cancer: a new paradigm. *Clin. Cancer Res.* 17, 2693–2701. doi: 10.1158/1078-0432.CCR-10-2203

Sitarz, R., Skierucha, M., Mielko, J., Offerhaus, G. J. A., Maciejewski, R., and Polkowski, W. P. (2018). Gastric cancer: epidemiology, prevention, classification, and treatment. *Cancer Manag. Res.* 10:239. doi: 10.2147/cmar.s149619

Sohn, B. H., Hwang, J.-E., Jang, H.-J., Lee, H.-S., Oh, S. C., Shim, J.-J., et al. (2017). Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome atlas project. *Clin. Cancer Res.* 23, 4441–4449. doi: 10.1158/1078-0432.ccr-16-2211

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

Szeto, G. L., and Finley, S. D. (2019). Integrative approaches to cancer immunotherapy. *Trends Cancer* 5, 400–410. doi: 10.1016/j.trecan.2019.05.010

Takeno, A., Takemasa, I., Doki, Y., Yamasaki, M., Miyata, H., Takiguchi, S., et al. (2008). Integrative approach for differentially overexpressed genes in gastric

cancer by combining large-scale gene expression profiling and network analysis. *Br. J. Cancer* 99, 1307–1315. doi: 10.1038/sj.bjc.6604682

Tan, I. B., Ivanova, T., Lim, K. H., Ong, C. W., Deng, N., Lee, J., et al. (2011). Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. *Gastroenterology* 141, 476–485. doi: 10.1053/j.gastro.2011.04.042

Toyota, M., Ahuja, N., Suzuki, H., Itoh, F., Ohe-Toyota, M., Imai, K., et al. (1999). Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype. *Cancer Res.* 59, 5438–5442.

Tseng, C.-W., Lin, C.-C., Chen, C.-N., Huang, H.-C., and Juan, H.-F. (2011). Integrative network analysis reveals active microRNAs and their functions in gastric cancer. *BMC Syst. Biol.* 5:99. doi: 10.1186/1752-0509-5-99

Uen, Y.-H., Lin, K.-Y., Sun, D.-P., Liao, C.-C., Hsieh, M.-S., Huang, Y.-K., et al. (2013). Comparative proteomics, network analysis and post-translational modification identification reveal differential profiles of plasma Con A-bound glycoprotein biomarkers in gastric cancer. *J. Proteom.* 83, 197–213. doi: 10.1016/j.jprot.2013.03.007

Vamathevan, J., Clark, D., Czodrowski, P., and Cleveland, L. S. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477.

van den Boorn, H. G., Engelhardt, E. G., van Kleef, J., Sprangers, M. A. G., van Oijen, M. G. H., Abu-Hanna, A., et al. (2018). Prediction models for patients with esophageal or gastric cancer: a systematic review and meta-analysis. *PLoS One* 13:e0192310. doi: 10.1371/journal.pone.0192310

Vecchi, M., Nuciforo, P., Romagnoli, S., Confalonieri, S., Pellegrini, C., Serio, G., et al. (2007). Gene expression analysis of early and advanced gastric cancers. *Oncogene* 26, 4284–4294. doi: 10.1038/sj.onc.1210208

Vila, A. V., Collij, V., Sanna, S., Sinha, T., Imhann, F., Bourgonje, A. R., et al. (2020). Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nat. Commun.* 11, 1–11.

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H. N., Shi, S. T., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* 46, 573–582. doi: 10.1038/ng.2983

Wang, Q., Liu, G., and Hu, C. (2019). Molecular classification of gastric adenocarcinoma. *Gastroenterol. Res.* 12, 275–282. doi: 10.14740/gr1187

Wippel, H. H., Santos, M. D. M., Clasen, M. A., Kurt, L. U., Nogueira, F. C. S., Carvalho, C. E., et al. (2018). Comparing intestinal versus diffuse gastric cancer using a PEFF-oriented proteomic pipeline. *J. Proteom.* 171, 63–72. doi: 10.1016/j.jprot.2017.10.005

Wroblewski, L. E., Peek, R. M. Jr., and Wilson, K. T. (2010). *Helicobacter* pylori and gastric cancer: factors that modulate disease risk. *Clin. Microbiol. Rev.* 23, 713–739. doi: 10.1128/CMR.00011-10

Yalamarthi, S., Witherspoon, P., McCole, D., and Auld, C. D. (2004). Missed diagnoses in patients with upper gastrointestinal cancers. *Endoscopy* 36, 874–879. doi: 10.1055/s-2004-825853

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2017). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief. Bioinform.* 19, 1370–1381.

Yoo, M., Park, J., Han, H., Yun, Y., Kang, J. W., Choi, D., et al. (2017). Discovery of gastric cancer specific biomarkers by the application of serum proteomics. *Proteomics* 17:1600332. doi: 10.1002/pmic.201600332

Yoon, H. J., and Kim, J. H. (2020). Lesion-based convolutional neural network in diagnosis of early gastric cancer. *Clin. Endosc.* 53, 127–131. doi: 10.5946/ce.2020.046

Zang, Z. J., Cutcutache, I., Poon, S. L., Zhang, S. L., McPherson, J. R., Tao, J., et al. (2012). Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat. Genet.* 44, 570–574. doi: 10.1038/ng.2246

Zhang, P., Yang, M., Zhang, Y., Xiao, S., Lai, X., Tan, A., et al. (2019). Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. *Cell Rep.* 27, 1934–1947. doi: 10.1016/j.celrep.2019.04.052

Zouridis, H., Deng, N., Ivanova, T., Zhu, Y., Wong, B., Huang, D., et al. (2012). Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Sci. Transl. Med.* 4:156ra140. doi: 10.1126/scitranslmed.3004504

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

Copyright © 2020 Shi, Wei and Ji. 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 Molecular Biosciences

REVIEW
published: 08 September 2020
doi: 10.3389/fmolb.2020.562798

Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy

Pei Zhang[1†], Qin Xia[1†], Liqun Liu[1], Shouwei Li[2] and Lei Dong[1]*

[1] School of Life Sciences, Beijing Institute of Technology, Beijing, China, [2] Department of Neurosurgery, Sanbo Brain Hospital, Capital Medical University, Beijing, China

Glioblastoma (GBM) is highly invasive and the deadliest brain tumor in adults. It is characterized by inter-tumor and intra-tumor heterogeneity, short patient survival, and lack of effective treatment. Prognosis and therapy selection is driven by molecular data from gene transcription, genetic alterations and DNA methylation. The four GBM molecular subtypes are proneural, neural, classical, and mesenchymal. More effective personalized therapy heavily depends on higher resolution molecular subtype signatures, combined with gene therapy, immunotherapy and organoid technology. In this review, we summarize the principal GBM molecular classifications that guide diagnosis, prognosis, and therapeutic recommendations.

Keywords: glioblastoma, molecular heterogeneity, transcription-based subtype, genetic alteration-based subtype, DNA methylation-based subtype, subtype-specific therapy

OPEN ACCESS

Edited by:
Cheng Zhang,
Royal Institute of Technology, Sweden

Reviewed by:
Sahin Hanalioglu,
Hacettepe University, Turkey
Jorge Lima,
University of Porto, Portugal

****Correspondence:***
Lei Dong
ldong@bit.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *18 May 2020*
Accepted: *18 August 2020*
Published: *08 September 2020*

Citation:
Zhang P, Xia Q, Liu L, Li S and Dong L (2020) Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy. Front. Mol. Biosci. 7:562798. doi: 10.3389/fmolb.2020.562798

INTRODUCTION

The World Health Organization (WHO) defines adult diffuse gliomas into grade II and grade III astrocytic tumors, grade II and III oligodendrogliomas, and grade IV glioblastomas (Louis et al., 2016). Glioblastoma (GBM) is grade IV, the most invasive and deadly glioma (Brennan et al., 2009; Szopa et al., 2017; Lee et al., 2018; Ghosh et al., 2018; Shergalis et al., 2018; Paolillo et al., 2018). It invades adjacent areas of the brain but rarely spreads outside the brain (Phillips et al., 2006). Clinical data show GBM has a poor prognosis, with less than 5% of patients surviving 5 years after diagnosis (Verhaak et al., 2010). Based on clinicopathologic features, GBM is defined as primary or secondary GBM (Ohgaki and Kleihues, 2013). Primary GBM starts as grade IV, with no evidence of lower grades, and is more aggressive and more likely to affect elderly patients. Secondary GBM develops from astrocytoma (Grade II or III glioma), grows slowly initially then gradually becomes aggressive (Ohgaki and Kleihues, 2007). The mechanism of GBM tumorigenesis is still unclear, many patients relapse due to ineffective treatment options. Notably, recurrent GBM is frequently accompanied by molecular alterations compared with the initial diagnosis (Li et al., 2015; van den Bent et al., 2015; Cioca et al., 2016; Neilsen et al., 2019; Schafer et al., 2019).

Histomorphology ambiguity and tumor heterogeneity pose challenges to GBM diagnosis, prognosis and treatment. Histologic diagnosis often varies among clinicians and limits diagnostic reproducibility. GBM histologically and genetically show significant inter-tumoral and intra-tumoral heterogeneity, differing mutations, and indistinct phenotypic and epigenetic states reflect genomic instability that leads to varying therapy choices and clinical outcomes

(Homma et al., 2006; Marusyk and Polyak, 2010; Szerlip et al., 2012; Brennan et al., 2013). Molecular classification of GBM is a newer tool and a complement to the traditional pathology-based description (Verhaak et al., 2010; Brennan et al., 2013; Ceccarelli et al., 2016).

Molecular-based diagnosis, patient stratification, and personalized treatment are increasingly important. The ISN Haarlem recommends "hierarchical diagnosis with histological classification, WHO classification, and molecular information for comprehensive diagnosis" (Louis et al., 2014). In 2016, the WHO updated guidelines combining morphology and genetic variation, leading to a significant reorganization of the classification of several brain tumor entities, especially in gliomas (Louis et al., 2016). Two significant entities of 2016 WHO classification based on IDH (Isocitrate dehydrogenase) gene mutant status are IDH wild-type and IDH mutated GBM; patients whose full IDH evaluation cannot be assessed are classified as GBM NOS (not otherwise specified) (Louis et al., 2016).

Multi-omics studies from the landscape of GBM in the Cancer Genome Atlas Research Network (TCGA), the Chinese Glioma Genome Atlas (CGGA), and other databases, together reveal the complicated genetic profile of GBM (Cancer Genome Atlas Research Network, 2008; Brennan et al., 2013; Zhao Z. et al., 2020). These aberrant molecules, including 1p and 19q co-deletions (oligodendroglioma-specific), IDH gene mutations, PTEN (Phosphatase and tensin homolog) gene mutations, TP53 mutations, TERT (Telomerase reverse transcriptase) gene promoter mutations, ATRX (Alpha thalassemia/mental retardation syndrome X-linked) gene mutations, and EGFR (Epithelial growth factor receptor) gene amplification, are forcing clinicians to reconsider traditional GBM treatment (Mclendon et al., 2008; Brennan et al., 2013). GBM classification based on aberrant molecules shortens the time from diagnosis to treatment, and significantly improves accuracy and targeting.

In this paper, we summarize the process of GBM classification based on transcription levels, genetic alterations, and DNA methylation. We also describe the molecular characteristics of each category, and the relationship between different classification methods. Finally, we provide the current guiding strategy for diagnosis and treatment.

GBM HETEROGENEITY IDENTIFIED BY TRANSCRIPTION, GENETIC ALTERATION, AND DNA METHYLATION

Deciphering GBM heterogeneity and complexity is the key to understanding it's progression and creating effective therapies. Some important and aberrant molecular events drive GBM malignant transformation, highlighting the importance of molecular classification. First, GBM has a wide variety of chromosomal changes, including amplification in chromosome 4 (Chr.4, PDGFRA), Chr.7 (EGFR; MET, hepatocyte growth factor receptor; CDK6, Cyclin-dependent kinase 6), Chr.12 (CDK4, Cyclin-dependent kinase 6; MDM2, Mouse double minute 2 homolog), and deletion in Chr.10 (PTEN). Notably, some GBM patients have simultaneous gain of Chr.19 and 20 (Brennan et al., 2013).

Second, the TCGA GBM project describes somatic genome changes based on multidimensional and comprehensive features that show significant mutations in GBM, including TP53 (34.4%), EGFR (32.6%), PTEN (32%), NF1 (Neurofibromin 1, 13.7%), PIK3CA (Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform, 12%), PIK3R1 (Phosphatidylinositol 3-kinase regulatory subunit alpha, 11.7%), RB1 (Retinoblastoma-associated protein 1, 9.3%), SPTA1 (Spectrin alpha chain, erythrocytic 1, 9%), ATRX (6%), IDH1 (5.2%), KEL (Kell blood group glycoprotein, 5%), PDGFRA (Platelet-derived growth factor receptor A, 4.5%), and GABRA6 (Gamma-aminobutyric acid receptor subunit alpha-6, 4%) (Cancer Genome Atlas Research Network, 2008; Parsons et al., 2008; Verhaak et al., 2010; Brennan et al., 2013).

Lastly, DNA methylation is a key factor when measuring heterogeneity and stratification of GBM patients. Epigenetic modifications of GBM is related to biological characteristics and are considered therapeutic targets (Hegi et al., 2005; Etcheverry et al., 2010; Romani et al., 2018; Carella et al., 2020). DNA methylation states in GBM are correlated with survival, which has been extensively explored in recent years (Lofton-Day and Lesche, 2003; Hegi et al., 2005; Etcheverry et al., 2010; Christensen et al., 2011). GBM genome-wide methylation data show biologically distinct subtypes (Brennan et al., 2013). For example, DNA methylation of the MGMT (O6-Methyl guanine DNA methyltransferase) gene promoter occurs in 48.5% of GBM patients (174/359); MGMT is a known marker for treatment strategy (Parsons et al., 2008). Additionally, GBM patient data show other methylated genes, including GATA6 (GATA binding protein 6) (68.4%), CD81 (CD81 antigen) (46.1%), DR4 (Death receptor 4) (41.3%) and CASP8 (Caspase-8) (56.8%) (Skiriute et al., 2012). Interestingly, H. Noushmehr et al. found CpG island hypermethylation in a distinct subgroup of gliomas (G-CIMP), however only a small number of GBM patients with a positive prognosis belong to G-CIMP phenotype (Noushmehr et al., 2010).

MOLECULAR-BASED GBM CLASSIFICATION IN DIAGNOSIS AND PROGNOSIS PREDICTION

With the recent development of technology and classification algorithms, GBM is divided into different subtypes based on transcription profiles, genetic alterations, and DNA methylation. This allows targeted therapy based on molecular characteristics of subclasses. For example, clinicians can target the mesenchymal subtype from transcription subtypes in GBM via inhibition of diacylglycerol kinase alpha. In doing so, patients with MGMT methylation had a more robust response to temozolomide (Hegi et al., 2005; Taylor and Schiff, 2015; Olmez et al., 2017). The TCGA GBM project used a multi-platform analysis and comprehensively determined the genomic landscape to better understand the pathogenic and drug-resistant mechanism of GBM (Brennan et al., 2013). Here, we describe the

classical classification, and analyze the differences among various GBM subtypes.

Transcription-Based Subtypes

GBM classification based on gene expression profiles initially used microarray technology, then large-scale high-throughput next-generation sequencing technology. The molecular map of GBM is shown in **Figure 1A**. The classification method proposed by Verhaak et al. has been widely used, includes four subtypes: Proneural, Neural, Classical and Mesenchymal (Verhaak et al., 2010).

Initial Exploration on the Transcription-Based Classification

In the 1990s, scientists acquired data from techniques like PCR, allele analysis, and first-generation sequencing to analyze gliomas. They found a variety of molecular markers of different types and grades, but the landscape was not clear (Sehgal, 1998). Indeed, tumor development is highly complex, involving multiple genetic and epigenetic changes. Through microarray investigations, genes associated with GBM were identified and used as biomarkers in early diagnosis, leading many researchers to begin exploration of molecular diagnosis, classification, and treatment (**Figure 1A**; Schena et al., 1995; Velculescu et al., 1995; Derisi et al., 1996; Dudoit et al., 2002; Irizarry et al., 2003; Hu et al., 2006). Rickman et al. found 360 distinct genes in GBM from pilocytic astrocytomas, including MDM2, IGFBP2 (Insulin-like growth factor-binding protein 2), CD44 (CD44 antigen), and CDK4 (Cyclin-dependent Kinase 4) (Rickman et al., 2001). Sallinen et al. found more than 200 gene expression alterations in GBM and demonstrated a strategy for high-throughput molecular genetic profiling of brain tumors (Sallinen et al., 2000). In addition, Nutt et al. found 14 GBMs and 7 anaplastic oligodendroglioma, diagnosed by pathology, were predicted using gene markers that accurately classified 18 samples (Nutt et al., 2003). The classification prediction model objectively and reliably classifies high-grade non-classical glial tumors (Nutt et al., 2003). Compared with pathological classification, this model reliably predicts the prognosis of atypical lesions more accurately.

In a groundbreaking study, Phillips et al. classified three GBM subtypes: Proneural, Proliferative and Mesenchymal (**Figure 1** and **Table 1**; Phillips et al., 2006). Proneural subtypes are more common in young patients, less pathological compared with proliferative or interstitial GBM and have a better prognosis (Phillips et al., 2006). NCAM (Neural cell adhesion molecule), GABBR1 (Gamma-aminobutyric acid type B receptor subunit 1), and SNAP91 (Clathrin coat assembly protein AP180) are associated with neurons and are more similar to normal brain tissue and expression in proneural subtype (Phillips et al., 2006). The Proliferative

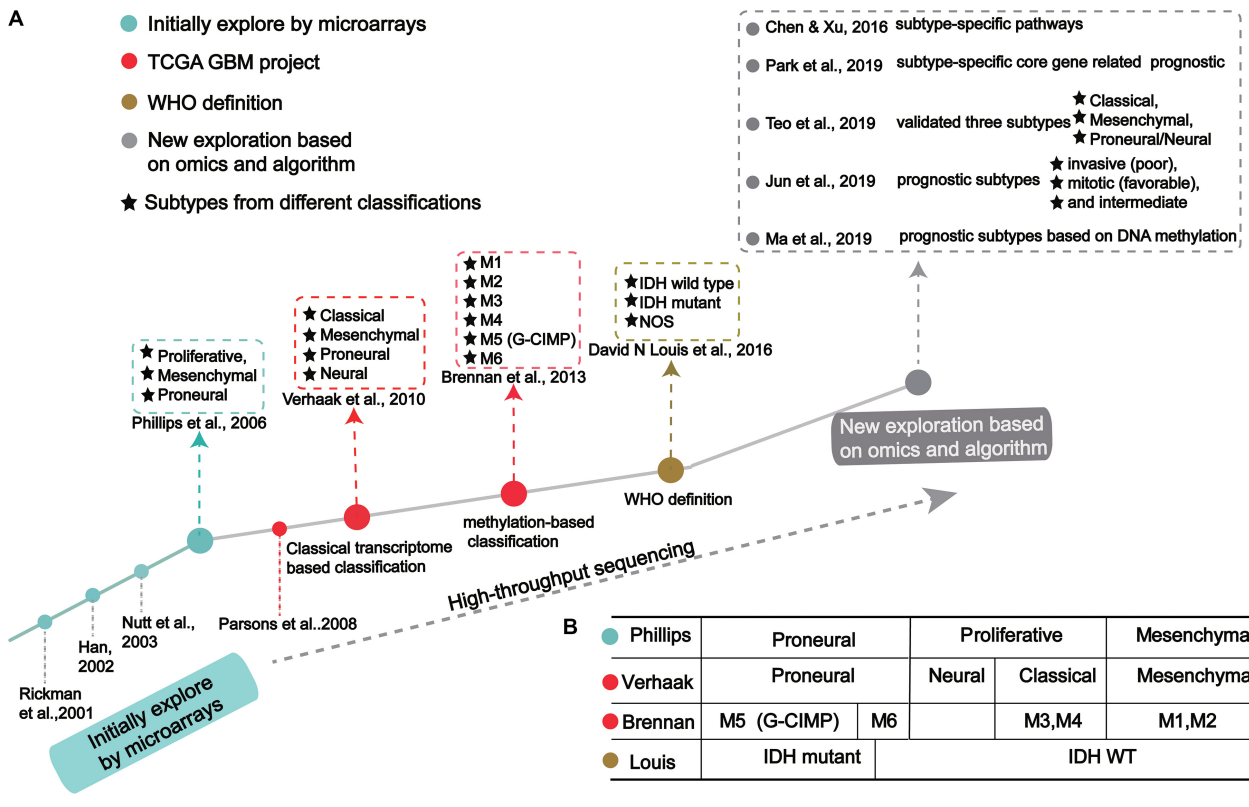


FIGURE 1 | The process of molecular-based GBM classification: **(A)** GBM classification timeline and classical subtypes. **(B)** The relative overlap between subtypes from different classification methods.

subtype is similar to stem cells with significantly up-regulated markers of proliferation, including TOP2A (DNA topoisomerase II alpha) and PCNA (Proliferating cell nuclear antigen) (Phillips et al., 2006). In contrast, the Mesenchymal subtype displays overexpression of angiogenesis markers, including the endothelial marker PECAM1 (Platelet endothelial cell adhesion molecule) gene, VEGF (Vascular endothelial growth factor) gene, VEGFR1 (Vascular endothelial growth factor receptor 1) gene and VEGFR2 (Vascular endothelial growth factor receptor 2) gene, which shows mesenchymal and angiogenic characteristics (Phillips et al., 2006). Proliferative and Mesenchymal subtypes are characterized by activation of PI3K/AKT (Phosphoinositide 3-kinase/Protein kinase B) signaling, loss on Chr.10 (location of PTEN), gain on Chr.7 (location of EGFR), and poor prognosis with invasive growth and angiogenic pathways (Phillips et al., 2006). These three subtypes are reminiscent of the various stages of developmental neurogenesis, which provides the basis and perspective for the molecular classification of GBM.

Deep Analysis of Transcription-Based Classification

The above studies demonstrated tumors often cluster in groups that display heterogeneity, highlighting the weaknesses of conventional diagnosis. With the advent of large-scale, high-throughput, next-generation sequencing methods, and with algorithms in machine learning, complex tumor data is becoming more precise.

Verhaak et al. (2010) offered more in-depth research and treatment possibilities for GBM (**Figure 1A**) based on the four subtypes Proneural, Neural, Classical and Mesenchymal (**Table 1**). The Proneural subtype is found primarily in younger patients, characterized by high PDGFRA gene expression and frequent IDH1 mutation. Compared with the other three subtypes, the Proneural subtypes may have better survival rates. However, Proneural subtypes showed no significant difference from other subtypes in response to chemotherapy and radiotherapy (Colman et al., 2010). The Neural subtype has similar gene expression patterns compared with normal brain tissue and tends to be more responsive to radiation and chemotherapy. GBMs with neural markers like SYT1 (Synaptotagmin 1), SLC12A5 (Solute carrier family 12 members 5), GABRA1 (Gamma-aminobutyric acid type A receptor alpha1) and NEFL (Neurofilament light polypeptide), are classified as the Neural subtype. The Classical subtype shows aberrant changes, including Chr.7 amplification, Chr.10 loss, inactivation of the RB (Retinoblastoma-associated protein) pathway, and focal 9p21.3 homozygous deletion. In addition, Sonic hedgehog pathways (SMO, Smoothened homolog; GAS1, Growth arrest-specific protein 1; GLI2, Growth arrest-specific protein 2), Notch signaling pathways (NOTCH3, Neurogenic locus notch homolog protein 3; JAG1, Jagged1; LFNG, Lunatic fringe) and the neural precursor and stem cell marker NES are highly expressed in the Classical subtype. Importantly, patients with Classical subtype show a significant reduction in mortality with aggressive radiotherapy and chemotherapy. The Mesenchymal subtype is characterized by extensive necrosis and inflammation, upregulation of interstitial and angiogenesis genes, deletion of tumor suppressor genes P53, PTEN, and NF1, and high expression of genes in the tumor necrosis factor superfamily and the NF-κB pathway. Although responsive to aggressive radiotherapy and chemotherapy, the prognosis of Mesenchymal subtypes is the worst among all subtypes (Colman et al., 2010). Recently, Sharma et al. found that VEGF-A (Vascular endothelial growth factor A), VEGF-B (Vascular endothelial growth factor B), ANG1 (Angiopoietin 1) and ANG24 (Angiopoietin 24) genes are highly expressed in the Mesenchymal subtype (Sharma et al., 2017).

TABLE 1 | The classification by Phillips, Verhaak and Wang.

Phillips et al. (2006)		**Proneural**	**Proliferative**		**Mesenchymal**
	Signature	NCAM, GABBR1, SNAP91	PCNA, TOP2A, EGFR		VEGF,VEGFR1, VEGFR2, PECAM1
	Chromosome Gain/loss	None	Gain on Chr.7, loss on Chr.10		Gain on Chr.7, loss on Chr.10
	Biological process	Neurogenesis	Proliferation		Angiogenesis
Verhaak et al. (2010)		**Proneural**	**Neural**	**Classical**	**Mesenchymal**
	Signature	PDGFRA, OLIG2, DDL3,SOX2, NKX2-2	MBP/MAL, NEFL, SLC12A5, SYT1, GABRA1	EGFR, AKT2, SMO, GAS1, GLI2, NOTCH3, JAG1, LFNG	YKL40, MET, CD44, MERTYK, TRADD, RELB, TNFRSF1A
	Mutated genes	TP53, PI3K, IDH1, PDGFRA		PTEN, CHKN2, PDGFRA	NF-κB, NF1
Wang et al. (2017)		**Proneural**	**Neural**	**Classical**	**Mesenchymal**
	Cell source	Tumor cells	Tumor cells	Tumor cells	Non-tumor cell

In 2017, Wang et al. (2017) proved that GBM tumor cells include Classical, Proneural, and Mesenschymal, and Neural subtype is non-tumor cells in the tumor microenvironment. They found the median survival of Mesenchymal, Classical, or Proneural are 11.5, 14.7, and 17.0 months, respectively. Wang's classification is based on tumor cells rather than microenvironmental/non-malignant tumor cells in tumor entities.

Using cancer genome data from the TCGA GBM project and classification from Verhaak et al. (2010) and Park A. K. et al., 2019 identified subtype-specific prognostic core genes and further examined prognostic chromosome changes and mutations (**Figure 1A**). Specific prognostic core genes in Classical subtype exist in DNA repair, cell cycle, Janus kinase, and transcription activation factor (JAK-STAT) pathway. And, specific prognosis genes in Mesenchymal subtype are related to mesenchymal cell movement, PI3K/AKT pathway, Mitogen-activated protein kinase (MAPK) pathways, extracellular signal-regulated kinase (ERK) pathways, and Wnt pathways (Park A. K. et al., 2019). Notably, patients with Mesenchymal subtypes with PIK3R1 or PCLO (Protein piccolo) mutations show a poorer prognosis (Park A. K. et al., 2019). These results demonstrate specific molecular targets and biomarkers for each subtype of GBM.

Recent studies offer new insights into GBM classification based on transcription. Teo et al. validated three robust GBM-subtypes: Proneural/Neural, Classical, and Mesenchymal across six different datasets (**Figure 1A**; Verhaak et al., 2010; Teo et al., 2019). This was validated in subtype-specific patient-derived orthotopic xenograft (PDOX) mice; the Classical subtype showed no survival difference between radiotherapy and temozolomide monotherapy. A Proneural/Neural specific-PDOX model showed temozolomide significantly improved survival compared to radiotherapy. This points to better predictive clinical outcomes based on more precise patient selection in clinical trials.

Park J. et al. (2019) identified three subtypes related to prognosis prediction: Mitotic (favorable), Intermediate, and Invasive (poor) by analyzing and verifying four large-scale gene expression profiles (**Figure 1A**). These new GBM subtypes have different multi-omics features and biological phenotypes. Among GBM prognostic subtypes, the invasiveness in the Invasive subtype is significantly higher than the Mitotic subtype. Interestingly, the methylated MGMT gene promoter is correlated with the Mitotic subtype, indicating Mitotic subtype patients are more likely to respond to temozolomide (Park J. et al., 2019). This study suggests that treatment strategies should be based on prognostic subtypes. For example, patients in the Mitotic subtype can be treated with temozolomide, while patients in Invasive subtypes require therapeutic intervention for the aggressiveness of the GBM. Although the prognostic subtype is based only on transcription and survival time, genomic features such as pathogenic somatic variations of IDH1 and ATRX and DNA methylation are only present in Mitotic subtypes. Since these three subtypes suggest a prognosis for GBM, inhibition of target genes in different subtypes may improve patient survival. Further, these genes may have clinical value as prognostic biomarkers and new drug targets, while also leading to new pathological and etiological factors for the oncogenesis and development for GBM.

Genetic Alteration-Based Subtypes

In recent years, large-scale genomic studies have revealed many mutations in tumor suppressor genes and oncogenes, and significantly improved our understanding of GBM. Specifically, mutated IDH, PTEN and EGFR are related to patient survival and can be used as indicators of patient classification.

IDH-Wild Type and IDH-Mutation Type

The identification of the IDH mutation is an important contribution to the molecular pathology of GBM. In 2008, Parsons et al. (2008) found the IDH1 gene had a point mutation in a small number of glioblastoma samples. Subsequently, Yan et al. (2009) found that GBM patients with IDH1/IDH2 mutations had a higher survival rate than those without these mutations (**Table 2**). Many studies have shown that patients with IDH mutations are significantly different from those without IDH mutations in molecular and clinical characteristics, including prognosis (Ichimura et al., 2009; Nobusawa et al., 2009; Watanabe et al., 2009; Yan et al., 2009; Lu et al., 2012; Songtao et al., 2012; Stancheva et al., 2014; Mondesir et al., 2016). There are three IDH enzymes: IDH1, IDH2, and IDH3 (Yan et al., 2009). IDH1 is mainly cytoplasmic, while IDH2 and IDH3 are mostly present in the mitochondrial matrix. IDH is the central enzyme in the citric acid cycle and plays a vital role in oxidative stress resistance (Marko and Weil, 2013). The most common IDH1 mutation observed in gliomas is the point mutation at position 132 (R132H), which is regarded as a typical IDH1 mutation (Parsons et al., 2008).

In 2016, the WHO divided it into two: IDH mutation and IDH wild type (**Figure 1A**; Louis et al., 2016). IDH wild type GBM with poor survival is dominated by stellate cell differentiation, characterized by nuclear atypia, cell polymorphism, typical diffuse growth patterns, mitotic activity and microvascular proliferation and/or necrosis. There are three variants of IDH wild-type, including giant cell GBM, gliosarcoma and epithelial-like GBM (Ep-GBM) (Louis et al., 2016). Genetically,

TABLE 2 | The characteristics of IDH WT subtype and IDH mutant subtype.

	IDH WT	IDH mutant	References
Corresponds to	Primary GBM	Secondary GBM	Louis et al., 2016
Proportion	90%	~10%	Louis et al., 2016
Age	Usually > 60	Younger adults	Louis et al., 2016
CpG methylator	Less frequent	More frequent	Brennan et al., 2013
TERT promoter mutation	~95%	51%	Yan et al., 2009
homologous deletion of CDKN2A/CDKN2B	~45%	Less	Yan et al., 2009
EGFR alterations	~41%	0%	Yan et al., 2009
PTEN mutation/deletion	~25%	0%	Yan et al., 2009
TP53 mutations	~20%	81%	Yan et al., 2009

giant cell GBM lacks EGFR amplification and homozygous CDKN2A deletion and contains PTEN mutation and TP53 mutation (Meyer-Puttlitz et al., 1997). Patients with the giant cell GBM have outcomes similar to classical GBM. In gliosarcoma, TP53 mutations are rare, and EGFR amplification is also uncommon, and contains CDKN2A deletion (Lowder et al., 2019). The clinical outcome of gliosarcoma differs from classical GBM, but there are still conflicting and uncertain results from various studies. Ep-GBM, as a new variant of GBM, is more prevalent in children and young people, manifesting as superficial brain or mesencephalic masses, and often carries BRAF (Serine/threonine-protein kinase B-Raf) V600E mutations (Chapman et al., 2011; Kleinschmidtdemasters et al., 2013; Broniscer et al., 2014). Ep-GBM is based on the absence of INI1 expression, distinguishing it from similar epithelioid counterparts (Kleinschmidt-DeMasters et al., 2010). Additionally, Ep-GBM often lacks EGFR amplification and PTEN loss, but ODZ3 usually has hemizygous deletions (Alexandrescu et al., 2016).

Multiple studies have confirmed that IDH mutations have prognosis and predictive value (Yan et al., 2009; Beiko et al., 2014; Stancheva et al., 2014). Compared to GBM patients with wild-type IDH, IDH-mutant GBM patients had higher overall survival and were more responsive to temozolomide (Songtao et al., 2012). The inhibitor of IDH mutation, which has been applied in preclinical models, shows activity to retard glioma cell growth (Rohle et al., 2013).

Other Genetic Mutations

In IDH1 wild type GBM, the median survival rate of patients with CDK4/MDM2 co-amplification is 6.6 months after diagnosis, while the median survival rate of patients without an CDK4/MDM2 co-amplification is 12.7 months (Abedalthagafi et al., 2018). The TERT promoter mutation was recently identified as a sign of poor prognosis. It is enriched in elderly patients, with approximately 40% having grade II/III glioma, suggesting TERT's correlation with shorter overall survival as a key pathological player and therapeutic target (Chamberlain and Sanson, 2015; Mosrati et al., 2015; Spiegl-Kreinecker et al., 2015; Yang et al., 2016; Yuan et al., 2016).

EGFR amplification is usually accompanied by EGFR mutation, the most frequent being EGFRvIII (Gan et al., 2009). Under normal physiological conditions, EGFR plays a central role in cell proliferation, differentiation and development. EGFR is located on the short arm of Chr.7 (7p12) and encodes a cell surface receptor tyrosine kinase (Hatanpaa et al., 2010). EGFRvIII is characterized by the absence of 267 amino acids in the extracellular domain, resulting in the inability of the receptor to bind to the ligand but with substitutive activity (Hatanpaa et al., 2010). EGFRvIII enhances the tumorigenic potential of GBM by activating and maintaining mitotic and anti-apoptotic signaling pathways, along with their impaired internalization and degradation (Gan et al., 2009). Some studies have found that EGFRvIII overexpression and EGFR amplification are associated with poor prognosis in young patients, and other data show EGFR overexpression is associated with poor prognosis in elderly patients (Shinojima et al., 2003; Srividya et al., 2010). But recently, Felsberg et al. found EGFRvIII and EGFR SNVs are not prognostic; Chen et al. showed that there is insufficient evidence for the presence of either EGFR amplification or EGFRvIII mutation has prognostic value in patients with GBM using meta-analysis (Chen et al., 2015; Felsberg et al., 2017). These results may be biased by the inherent variability in subtypes, therefore, the exploration of the relationship between EGFR and prognosis needs to be carried out in different subtypes. Notably, compared with patients with both TERT and EGFR gene mutations, the overall survival of TERT/EGFR wild-type patients (EGFR not amplified) is almost twice that of the former (Chamberlain and Sanson, 2015; Yang et al., 2016).

The PTEN protein catalyzes the dephosphorylation of 3' phosphorylation of the inositol ring in PIP3 (phosphatidylinositol-3,4,5-trisphosphate) to produce PIP2 (phosphatidylinositol-4,5-bisphosphate). The dephosphorylation is critical because it inhibits the AKT signaling pathway. The PI3K/AKT pathway is normally dormant in differentiated and quiescent cells, but when activated, the cell cycle modulation leads to cancer. The deficiency of PTEN mainly plays the role of lipid phosphatase through the PI3K/AKT pathway (Endersby and Baker, 2008). Therefore, the loss of PTEN is associated with a more aggressive phenotype.

In addition to the genes above, other genetic mutations also drive GBM development. However, the mutation or deletion of a single gene may not serve to classify GBM independently. The combination of aberrant events related to survival may be a more effective classifier. Data suggest combination of two or three genes provide a robust classifier to diagnostic analysis for clinical applications (Kim et al., 2002). Classification driven by genetic mutation (single or consistent) is the basis for exploring GBM classification, and critical genetic targets can be used as the key for diagnosis, prognosis and treatment.

DNA Methylation-Based Subtypes

Epigenetic changes are common markers of human cancers, including GBM (Kim and Kim, 2008; Romani et al., 2018). DNA methylation is a core element of epigenetic alteration, an essential signaling tool for regulating genomic functions, and a key feature mediating tumorigenesis (Koch et al., 2018; Muhammad et al., 2018; Yamashita et al., 2018). DNA methylation can provide biomarkers for the early diagnosis and prognosis of cancer and provide a new method for further clinical applications (Lofton-Day and Lesche, 2003; Gustafsson et al., 2018; Kim et al., 2018; Li et al., 2018, 2019; Pérez et al., 2018. We know the methylation status of single genes corresponds to expression levels in GBM (Bell et al., 2018; Johannessen et al., 2018). MGMT promoter methylation is a prognostic factor for glioblastoma patients and has a significant correlation with worse survival rates (16.9 months vs. 12.7 months) (Brennan et al., 2013). Due to the diversity of GBM, a broader genome and expression profile is needed to gain insight into the potential response of treatment methods.

Brennan et al. used large-scale methylated sequencing data to classify GBM, divided into six categories based on the expression level of DNA methylation, including Cluster M1 to Cluster M6, in which Cluster M5 was G-CIMP subtype (**Figure 1A**;

Brennan et al., 2013). Cluster M6 is relatively hypomethylated and has the majority of IDH1 wild type patients than the G-CIMP subtype. Cases of missense mutations or deletions in MLL (Histone-lysine N-methyltransferase 2A) genes or HDAC (Histone deacetylase) family genes were concentrated in Cluster M2 (Brennan et al., 2013). These results indicate that the classification based on DNA methylation makes GBM classification clearer.

Recently, Ma et al. (2019) identified specific prognostic subtypes based on DNA methylation status and identified 3 GBM methylation clusters (Cluster 1, Cluster 2, and Cluster 3), which have significantly different survival curves (**Figure 1A**). Among all clusters, Cluster 2 has the best prognosis. The methylation levels in each cluster are related to specific molecular characteristics. Compared with Cluster1 and Cluster2, Cluster 3 showed more TP53 mutations and deletion of wildtype IDH1 and 1p/19q. The genes corresponding to the promoter region of the CpG site annotation are related to the survival and biological processes in GBM. By focusing on the level of DNA methylations in patients with GBM, researchers eventually developed a new prediction panel for 10 CpGs. They are superior to other molecular indicators because these 10 CpG signals reflect the relationship between GBM intrinsic tumor subtypes (Kloosterhof et al., 2013; Paul et al., 2017; Yin et al., 2018). The study also found the enriched CpG sites in genes involved in neuronal differentiation and brain development, including KIFC3 (Kinesin-like protein), OC90 (Otoconin-90), CRB2 (DNA repair protein crb2), IGSF22 (Immunoglobulin superfamily member 22) and NR0B2 (Nuclear receptor subfamily 0 group B member 2) (Wu et al., 2010; Yu et al., 2013).

DNA methylation provides a framework for understanding GBM and guiding a therapeutic strategy. It has offered more molecular biomarkers for each subtype and suggested more targets for treatment. Methylation is a powerful complement to classification based on genetic alterations and transcription, making GBM classification more comprehensive.

The Relationship Among Transcription, Genetic Alterations and DNA Methylation Classifications

Early attempts to identify specific tumor subtypes generally focused only on gene expression patterns. But biological processes are not so simply regulated. Omics data have helped identify clusters of tumors with similar characteristics, including genotypic and epigenetic regulation. Many studies have found that molecular subtypes classified at different levels are related and overlapped, as exampled from the four transcription-based subtypes from Verhaak et al. (2010), six DNA methylation-based subtypes from Brennan et al. (2013) and IDH mutation-based subtypes (**Figure 1B**). Combined analysis with four transcriptome-based subtypes of TCGA, Cluster M1 and M2 are enriched in the Mesenchymal subtypes, Cluster M3 and M4 in the Classical subtype, Cluster G-CIMP in the Proneural subtype, and Cluster M6 is relatively hypomethylated, which belongs to the Proneural subtype (Verhaak et al., 2010). Notably, Cluster G-CIMP increases the likelihood of DNA methylation of MGMT (79% of patients with DNA methylation of MGMT in Cluster G-CIMP and 46% in non-G-CIMP). Interestingly, MGMT DNA methylation is a predicted biomarker of classical subtypes, but not other subtypes. In addition, C-CIMP is a unique and almost invariable feature of IDH1/2 mutant GBMs, and studies have shown that patients with this GBM subtype have a better prognosis (Noushmehr et al., 2010; Baysan et al., 2012). According to the characteristic of DNA methylation pattern causally related to IDH1/2 mutation status and better prognosis, the Proneural subtype is further subdivided into G-CIMP positive and negative groups (Noushmehr et al., 2010).

MOLECULAR SUBTYPE MIGRATION IN RECURRENT GBM

Recently, some studies have shown that subtype migration and molecular changes occur in recurrent GBM, highlighting the need for further research (Wang et al., 2016). The recurrence of GBM is inevitable, although current standards of care for GBM patients include chemotherapy after surgical resection (Stupp et al., 2005). However, when GBM occurs, the tumor always recurs, and treatment options are limited. There is no standard care for patients with relapsed GBM because pathological and molecular features are lacking (Weller et al., 2012; Lukas and Mrugala, 2016). The progression-free survival of recurrent GBM is 2–4 months, and the survival of conventional chemotherapy after progression is 6–8 months (Gorlia et al., 2012).

Transcription-based molecular subtypes are also associated with tumor recurrence. For example, Wang et al. found that two-thirds of patients with primary GBM switched transcriptional subtype after recurrence. Importantly, the Mesenchymal subtype was the most stable primary GBM subtype (Wang et al., 2016). Therefore, further analysis of the molecular changes of recurrent GBM poses significant value in guiding treatment. van den Bent et al. showed that half of recurrent GBM patients lost EGFRvIII compared with the molecular expression of GBM at initial diagnosis (**Table 3**; van den Bent et al., 2015). Cioca et al. found recurrent GBM had lower EGFR expression than primary GBM in 10 cases, and only one case had increased expression

TABLE 3 | The molecular changes in recurrent GBM.

Event	Primary vs. Recurrent		References
EGFRvIII	About half of recurrent GBM patients lose EGFRvIII (7/15)		van den Bent et al., 2015
EGFR	Lower EGFR expression at recurrent GBM		Cioca et al., 2016
	Initial	**Recurrent**	
CDKN2A deletions	86%	53%	Neilsen et al., 2019
CDKN2B deletions	86%	54%	
EGFR mutation	52%	10%	
EGFR amplification	81%	45%	
TERT mutation	95%	51%	

on recurrence (**Table 3**; Cioca et al., 2016). The discrepancies of EGFR expression between primary GBM and recurrence suggest heterogeneity of GBMs is actively fluid. Neilsen et al. analyzed 10 pairs of matched primary and recurrent GBM through genomic changes, and the results indicate all matched tumor pairs showed differences. This study showed that EGFR mutation increased significantly in 3 cases, and the other three genes were generally changed in primary GBM and recurrent GBM, namely CDKN2A and CDKN2B deletion, and TERT mutation. Mutations that cause activation of the PI3K pathway are also common (**Table 3**; Neilsen et al., 2019). Kim et al. (2015) found that recurrent GBM had a hypermutant phenotype that initially occurred in the IDH1 mutant, suggesting IDH1 is associated with a hypermethylated phenotype, resulting in MGMT inhibition, making tumors more susceptible to mutagenesis by temozolomide.

Studies focused on recurrent GBM have shown molecular composition and molecular subtypes of tumors evolve in response to radiotherapy and targeted therapy, therefore molecular signatures guiding treatment protocols may improve patient survival (Campos et al., 2016). However, it is still challenging to develop new molecular therapies for recurrent GBM patients and personalized treatment.

MOLECULAR SUBTYPES AND SIGNATURES GUIDING CLINICAL TREATMENT

Subtype-Specific Molecular Guidance for the Selection of Targeted Drugs

Treatment corresponding to tumor subtypes is an effective strategy to avoid the obstacles caused by molecular heterogeneity (Collisson et al., 2011; Linnekamp et al., 2015; Zhao S. et al., 2020). Chen et al. analyzed the relationship between four subtypes distinguished by Verhaak (**Figure 1A**; Verhaak et al., 2010; Chen and Xu, 2016). The gene signatures in the Mesenchymal subtype is highly enriched in pathways associated with immune response, such as Hepatic Fibrosis/Hepatic Stellate Cell Activation, Coagulation System and IL-10 Signaling. The gene signatures in Proneural subtype are significantly enriched in pathways associated with cellular processes, such as Wnt/β-catenin Signaling and Cyclins and Cell Cycle Regulation. Signatures in the Neural subtype are significantly enriched in pathways associated with nervous system pathways and environmental information processing, such as nNOS Signaling in Skeletal Muscle Cells and cAMP-mediated signaling. Finally, the gene signatures in the Classical subtype are significantly enriched in pathways associated with the metabolism pathways, the nervous system and immune system, such as Fatty Acid Activation, CREB Signaling in Neurons, and PI3K Signaling in B Lymphocytes. They found the response to temozolomide in Classical and Mesenchymal subtypes was higher than that of neurotypes, and the Proneural subtype was lower than these three subtypes. They also developed a computational drug repurposing approach to predict GBM drugs based on the molecular subtypes. Protein kinase inhibitors, antipsychotics, and antidepressants have been identified as the most common drugs for all four subtypes. But in different subtypes, the ranking of drugs is different. In the Proneural subtype, antidepressants and antipsychotics were more effective. Anti-globulin inhibitors of the Mesenchymal subtype are involved in many immune system pathways and phenotypes. These results indicate that different molecular subtypes respond differently to drugs, and GBM subtype-specific therapies should be used.

Further evidence of molecularly guided treatment comes from Sandmann et al. that showed a 4.3 month increase in median survival with the addition of bevacizumab for IDH1 wild-type GBM in the proneural subgroup (Sandmann et al., 2015). The IDH1 R132H vaccine has been developed and shown promising results in animal models of IDH mutant glioblastomas (Schumacher et al., 2014; Dimitrov et al., 2015). These results demonstrate the necessity of diagnosing and developing personalized treatment plans according to IDH status.

Temozolomide is an oral alkylation agent. The main mechanism of temozolomide arrests the cell cycle at G2/M checkpoint, which leads to apoptosis of cancer cells (Alonso et al., 2007). The study showed the median survival was 12 months for patients receiving both temozolomide and radiation therapy and only 8 months for patients receiving radiation therapy alone (Alonso et al., 2007). However, in GBM, due to individual differences, the lack of MGMT methylation in some patients leads to the formation of temozolomide resistance. Herrlinger et al. found that for patients newly diagnosed, without MGMT methylation and with irinotecan/bevacizumab/radiation combination therapy had significantly prolonged mPFS (median progression – free survival) of 9.7 months. Temozolomide/radiation had significant mPFS of 5.9 months, an encouraging result that supports further investigation with this combination (Herrlinger et al., 2014). Therefore, before temozolomide treatment, it is advised to determine the methylation status of MGMT for most effective strategy.

Due to the heterogeneity of GBM, individualized treatment based on specific tumor subtype is clearly a more effective clinical strategy. Gene mutations in TP53, IDH-1 and PDGFR-A in Proneural subtype; mutations or amplification of EGFR gene in Classical subtype; NF-1 gene mutations in Mesenchymal subtype; and the expression of neural markers in Neural subtype are promising therapeutic targets. Several studies have shown that targeting these molecules improves treatment. For example, Sang et al. found the efficacy of SHP099, a potent, selective, and oral SHP-2 inhibitor for treating GBM with activated PDGFR-A signaling; and Liu et al. proved that the third-generation EGFR inhibitor osimertinib overcomes primary resistance by continuously blocking ERK signaling in GBM (Liu et al., 2019; Sang et al., 2019).

Due to subtype migration and molecular changes after recurrence, molecular evaluation of patients must be performed prior to chemotherapy. Patients undergoing surgical resection must undergo immunohistochemical studies to determine various predictors, such as MGMT methylation, to assist in treatment planning.

Individualized Treatment

Gene Therapy

Gene therapy aims to introduce genetic material into cells to compensate for abnormal genes or to make beneficial proteins. If a mutated gene causes a necessary protein loss, gene therapy can introduce a normal gene to supplement the protein's function. Gene therapy is the delivery of a gene through a vector to a cell. Viruses are often used as vectors because they can deliver new genes by infecting cells. These viruses are modified so that when they are used in humans, they do not cause disease. Adenovirus (AAV) vectors have been used to inject directly into GBM cells in the brain to express tumor-killing genes. Crommentujin et al. demonstrated the AAV9 vector, which produces the anticancer agent sTRAIL, killed up to 60% of GBM cells in mouse models and transfected cell lines (Gray et al., 2011; Crommentuijn et al., 2016). AAV9 virus vector is an excellent choice because its serotype can cross the blood-brain barrier during intravenous administration (Gray et al., 2011). CRISPR gene editing also belongs to gene therapy. By combining Cas9 nuclease with synthetic guide RNA and introducing it into the cell, the cell genome can be accurately trimmed, allowing existing genes to be removed or new ones added (Hendel et al., 2015). Using gene therapy technology to repair and compensate the tumor suppressor gene mutation in each subtype of GBM patients, such as PTEN mutation in the Classical subtype, may improve the survival time of patients.

Immunotherapy

Immunotherapy offers the promise of a sustained antitumor immunity that is pathway independent and has the potential to amplify antigens to boost immune responses. Peptide vaccines, such as EGFRvIII found in Classical GBM subtypes, can trigger immunity to GBM tumor cells expressing EGFRvIII. In a phase II trial involving 18 patients, EGFRvIII patients showed an overall survival of 26 months, compared with only 15 months for the control group (Heimberger, 2005). The vaccine has a promising future in immunotherapy for GBM. Tumor-specific antigen vaccines require confirmation that the tumor expresses the targeted antigen. Thus, immunotherapy limits the scope of these vaccines and the population in which they can be used, so specific vaccines can be designed according to the expression of molecules in different subtypes.

Organoid Model

Glioblastoma organs can be an effective model for rapid testing of personalized treatment strategies. The models allow researchers to reconstruct key features of a patient's diseased brain to help paint a clearer picture of the cancer and then allow researchers to explore the best ways to treat it. Researchers have successfully transplanted eight glioblastoma organoids (GMOs) samples into brains of adult mice, administering standard care and targeted therapy to GBOs, including clinical trial drugs and chimeric antigen receptor T (CAR-T) cell immunotherapy (Jacob et al., 2020). For each treatment, the researchers showed that organ-like responses were different, and the effect was linked to genetic mutations in the patient's tumor. The model opens the possibility of future clinical trials that can personalize treatment based on how a patient's tumor responds to different drugs. Notably, the researchers have observed the benefits of treating organ-like organs with CAR-T therapy in clinical trials for EGFRvIII mutations, a driver of the disease. In 6 cases of GBOs, the EGFRvIII mutation was shown to have a specific effect on patient GBOs, with increased CART cells and decreased EGFRvIII expression cells (Jacob et al., 2020). These results highlight the potential of using personalized approaches to detect and treat glioblastoma.

CONCLUSION

The unique and highly reproducible molecular changes discovered in recent years have begun to elucidate the diversity of GBM and contribute to the more effective classification of tumors. These studies provide insights into how to improve current treatment strategies. GBM genomics, transcription, and epigenetic features reveal critical molecular changes that may lead to pathologic disease progression. Large-scale analysis, like the TCGA project, confirm that GBM is a heterogenetic tumor at the molecular level that can be subdivided into different subtypes according to the molecular pathogenesis and biological entities of "driving factor" lesions. Although these comprehensive studies provide useful insights into the characteristics and classification of tumors, their limitations need to be considered when drawing conclusions. Some prognostic markers have appeared in these studies, and there is still a great need to identify true predictive markers to improve the treatment process of personalized care. In addition, the intra-tumoral heterogeneity of GBM needs to be further classified by single-cell sequencing technology to obtain a more complete and more precise inter-tumoral and intra-tumoral classification. We still need large-scale animal experiments and human clinical verification to improve treatment response and survival time among the different subtypes.

Through these molecular-level studies, we can further improve the molecular detection methods, guide the targeted therapy based on molecular classification, and form a set of accurate GBM molecular therapy manuals that can improve patient outcome.

AUTHOR CONTRIBUTIONS

QX and PZ contributed to the conception and design of manuscript. PZ drafted the manuscript. QX provided useful comments and suggestions. LD and QX revised the manuscript. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported by grants from the Beijing Natural Science Foundation (Z190018), the National Natural Science Foundation of China (81870123), the China Postdoctoral Science Foundation Grant (2018M641206), and the National Science Foundation for Young Scientists of China (81902545).

REFERENCES

Abedalthagafi, M., Barakeh, D., and Foshay, K. M. (2018). Immunogenetics of glioblastoma: the future of personalized patient management. *NPJ Precis. Oncol.* 2:27. doi: 10.1038/s41698-018-0070-1

Alexandrescu, S., Korshunov, A., Lai, S. H., Dabiri, S., Patil, S., Li, R., et al. (2016). Epithelioid glioblastomas and anaplastic epithelioid pleomorphic xanthoastrocytomas-same entity or first cousins? *Brain Pathol.* 26, 215–223. doi: 10.1111/bpa.12295

Alonso, M. M., Gomez-Manzano, C., Bekele, B. N., Yung, W. K. A., and Fueyo, J. (2007). Adenovirus-based strategies overcome temozolomide resistance by silencing the O6-Methylguanine-DNA methyltransferase promoter. *Cancer Res.* 67, 11499–11504. doi: 10.1158/0008-5472.can-07-5312

Baysan, M., Bozdag, S., Cam, M. C., Kotliarova, S., Ahn, S., Walling, J., et al. (2012). G-cimp status prediction of glioblastoma samples using mRNA expression data. *PLoS One* 7:e47839. doi: 10.1371/journal.pone.0047839

Beiko, J., Suki, D., Hess, K. R., Fox, B. D., Cheung, V. J., Cabral, M., et al. (2014). IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. *Neuro Oncol.* 16, 81–91. doi: 10.1093/neuonc/not159

Bell, E. H., Zhang, P., Fisher, B. J., Macdonald, D. R., McElroy, J. P., Lesser, G. J., et al. (2018). Association of MGMT promoter methylation status with survival outcomes in patients with high-risk glioma treated with radiotherapy and temozolomide. *Jama Oncol.* 4:1405. doi: 10.1001/jamaoncol.2018.1977

Brennan, C., Momota, H., Hambardzumyan, D., Ozawa, T., Tandon, A., Pedraza, A., et al. (2009). Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations. *PLoS One* 4:e7752. doi: 10.1371/journal.pone.0007752

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

Broniscer, A., Tatevossian, R. G., Sabin, N. D., Klimo, P., Dalton, J., Lee, R., et al. (2014). Clinical, radiological, histological and molecular characteristics of paediatric epithelioid glioblastoma. *Neuropathol. Appl. Neurobiol.* 40, 327–336. doi: 10.1111/nan.12093

Campos, B., Olsen, L. R., Urup, T., and Poulsen, H. S. (2016). A comprehensive profile of recurrent glioblastoma. *Oncogene* 35, 5819–5825. doi: 10.1038/onc.2016.85

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature07385

Carella, A., Tejedor, J. R., Garcia, M. G., Urdinguio, R. G., Bayon, G. F., Sierra, M., et al. (2020). Epigenetic downregulation of TET3 reduces genome-wide 5hmC levels and promotes glioblastoma tumorigenesis. *Int. J. Cancer* 146, 373–387. doi: 10.1002/ijc.32520

Ceccarelli, M., Barthel Floris, P., Malta Tathiane, M., Sabedot Thais, S., Salama Sofie, R., Murray Bradley, A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* 164, 550–563. doi: 10.1016/j.cell.2015.12.028

Chamberlain, M. C., and Sanson, M. (2015). Combined analysis of TERT, EGFR, and IDH status defines distinct prognostic glioblastoma classes. *Neurology* 84:2007. doi: 10.1212/wnl.0000000000001625

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

Chen, J. R., Xu, H. Z., Yao, Y., and Qin, Z. Y. (2015). Prognostic value of epidermal growth factor receptor amplification and EGFRvIII in glioblastoma: meta-analysis. *Acta Neurol. Scand.* 132, 310–322. doi: 10.1111/ane.12401

Chen, Y., and Xu, R. (2016). Drug repurposing for glioblastoma based on molecular subtypes. *J. Biomed. Inform.* 64, 131–138. doi: 10.1016/j.jbi.2016.09.019

Christensen, B. C., Smith, A. A., Zheng, S., Koestler, D. C., Houseman, E. A., Marsit, C. J., et al. (2011). DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma. *J. Natl. Cancer Instit.* 103, 143–153. doi: 10.1093/jnci/djq497

Cioca, A., Gheorghe-Emilian, O., Gisca, M., Morosanu, C., Marin, I., and Florian, S. (2016). Expression of EGFR in paired new and recurrent glioblastomas. *Asian Pac. J. Cancer Prevent. APJCP* 17, 4205–4208.

Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., et al. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* 17, 500–503. doi: 10.1038/nm.2344

Colman, H., Zhang, L., Sulman, E. P., Mcdonald, J. M., Shooshtari, N. L., Rivera, A. L., et al. (2010). A multigene predictor of outcome in glioblastoma. *Neuro Oncol.* 12, 49–57. doi: 10.1093/neuonc/nop007

Crommentuijn, M. H. W., Kantar, R., Noske, D. P., Vandertop, W. P., Badr, C. E., Würdinger, T., et al. (2016). Systemically administered AAV9-sTRAIL combats invasive glioblastoma in a patient-derived orthotopic xenograft model. *Mol. Ther Oncolyt.* 3:16017. doi: 10.1038/mto.2016.17

Derisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14, 457–460. doi: 10.1038/ng1296-457

Dimitrov, L., Hong, C. S., Yang, C., Zhuang, Z., and Heiss, J. D. (2015). New developments in the pathogenesis and therapeutic targeting of the IDH1 mutation in glioma. *Int. J. Med. Sci.* 12, 201–213. doi: 10.7150/ijms.11047

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97, 77–87. doi: 10.1198/016214502753479248

Endersby, R., and Baker, S. J. (2008). PTEN signaling in brain: neuropathology and tumorigenesis PTEN signaling in brain: neuropathology and tumorigenesis. *Oncogene* 27, 5416–5430. doi: 10.1038/onc.2008.239

Etcheverry, A., Aubry, M., de Tayrac, M., Vauleon, E., Boniface, R., Guenot, F., et al. (2010). DNA methylation in glioblastoma: impact on gene expression and clinical outcome. *BMC Genomics* 11:701. doi: 10.1186/1471-2164-11-701

Felsberg, J., Hentschel, B., Kaulich, K., Gramatzki, D., Zacher, A., Malzkorn, B., et al. (2017). Epidermal growth factor receptor variant III (EGFRvIII) positivity in EGFR-amplified glioblastomas: prognostic role and comparison between primary and recurrent tumors. *Clin. Cancer Res.* 23, 6846–6855. doi: 10.1158/1078-0432.CCR-17-0890

Gan, H. K., Kaye, A. H., and Luwor, R. B. (2009). The EGFRvIII variant in glioblastoma multiforme. *J. Clin. Neurosci.* 16, 748–754. doi: 10.1016/j.jocn.2008.12.005

Ghosh, D., Nandi, S., and Bhattacharjee, S. (2018). Combination therapy to checkmate glioblastoma: clinical challenges and advances. *Clin. Transl. Med.* 7:33. doi: 10.1186/s40169-018-0211-8

Gorlia, T., Stupp, R., Brandes, A. A., Rampling, R., Fumoleau, P., Dittrich, C., et al. (2012). New prognostic factors and calculators for outcome prediction in patients with recurrent glioblastoma: a pooled analysis of EORTC Brain Tumour Group phase I and II clinical trials. *Eur. J. Cancer* 48, 1176–1184. doi: 10.1016/j.ejca.2012.02.004

Gray, S. J., Matagne, V., Bachaboina, L., Yadav, S., Ojeda, S. R., and Samulski, R. J. (2011). Preclinical differences of intravascular AAV9 delivery to neurons and glia: a comparative study of adult mice and nonhuman primates. *Mol. Ther.* 19, 1058–1069. doi: 10.1038/mt.2011.72

Gustafsson, J. R., Katsioudi, G., Degn, M., Ejlerskov, P., Issazadeh-Navikas, S., and Kornum, B. R. (2018). DNMT1 regulates expression of MHC class I in post-mitotic neurons. *Mol. Brain* 11:36. doi: 10.1186/s13041-018-0380-9

Hatanpaa, K. J., Burma, S., Zhao, D., and Habib, A. A. (2010). Epidermal growth factor receptor in glioma: signal transduction, neuropathology, imaging, and radioresistance. *Neoplasia* 12, 675–684. doi: 10.1593/neo.10688

Hegi, M. E., Diserens, A., Gorlia, T., Hamou, M., De Tribolet, N., Weller, M., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003.

Heimberger, A. B. (2005). Prognostic effect of epidermal growth factor receptor and EGFRvIII in glioblastoma multiforme patients. *Clin. Cancer Res.* 11, 1462–1466. doi: 10.1158/1078-0432.ccr-04-1737

Hendel, A., Bak, R. O., Clark, J. T., Kennedy, A. B., Ryan, D. E., Roy, S., et al. (2015). Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* 33, 985–989. doi: 10.1038/nbt.3290

Herrlinger, U., Schaefer, N., Steinbach, J. P., Weyerbrock, A., Hau, P., Goldbrunner, R., et al. (2014). Survival and quality of life in the randomized, multicenter GLARIUS trial investigating bevacizumab/irinotecan versus standard temozolomide in newly diagnosed, MGMT-non-methylated glioblastoma patients. *J. Clin. Oncol.* 32:2042. doi: 10.1200/jco.2014.32.15_suppl.2042

Homma, T., Fukushima, T., Vaccarella, S., Yonekawa, Y., Patre, P. L. D., Franceschi, S., et al. (2006). Correlation among pathology, genotype, and patient outcomes

in glioblastoma. *J. Neuropathol. Exp. Neurol.* 65, 846–854. doi: 10.1097/01.jnen.0000235118.75182.94

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96. doi: 10.1186/1471-2164-7-96

Ichimura, K., Pearson, D. M., Kocialkowski, S., Backlund, L. M., Chan, R., Jones, D. T., et al. (2009). IDH1 mutations are present in the majority of common adult gliomas but rare in primary glioblastomas. *Neuro Oncol.* 11, 341–347. doi: 10.1215/15228517-2009-025

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix geneChip probe level data. *Nucleic Acids Res.* 31:e15.

Jacob, F., Salinas, R. D., Zhang, D. Y., Nguyen, P. T. T., Schnoll, J. G., Wong, S. Z. H., et al. (2020). A patient-derived glioblastoma organoid model and biobank recapitulates inter- and intra-tumoral heterogeneity. *Cell* 180, 188.e22–204.e22. doi: 10.1016/j.cell.2019.11.036

Johannessen, L. E., Brandal, P., Myklebust, T. O. R. Å, Heim, S., Micci, F., and Panagopoulos, I. (2018). MGMT gene promoter methylation status – assessment of two pyrosequencing kits and three methylation-specific PCR methods for their predictive capacity in glioblastomas. *Cancer Genomics Proteomics* 15, 437–446. doi: 10.21873/cgp.20102

Kim, D., Lee, W., and Park, J. (2018). Promoter methylation of Wrap53α, an antisense transcript of p53, is associated with the poor prognosis of patients with non-small cell lung cancer. *Oncol. Lett.* 16, 5823–5828. doi: 10.3892/ol.2018.9404

Kim, J., Lee, I.-H., Cho Hee, J., Park, C.-K., Jung, Y.-S., Kim, Y., et al. (2015). Spatiotemporal evolution of the primary glioblastoma genome. *Cancer Cell* 28, 318–328. doi: 10.1016/j.ccell.2015.07.013

Kim, S., Dougherty, E. R., Shmulevich, I., Hess, K. R., Hamilton, S. R., Trent, J. M., et al. (2002). Identification of combination gene sets for glioma classification. *Mol. Cancer Ther.* 1, 1229–1236.

Kim, Y. K., and Kim, W.-J. (2008). Epigenetic markers as promising prognosticators for bladder cancer. *Int. J. Urol.* 16, 17–22. doi: 10.1111/j.1442-2042.2008.02143.x

Kleinschmidtdemasters, B. K., Aisner, D. L., Birks, D. K., and Foreman, N. K. (2013). Epithelioid GBMs show a high percentage of BRAF V600E mutation. *Am. J. Surg. Pathol.* 37, 685–698. doi: 10.1097/pas.0b013e31827f9c5e

Kleinschmidt-DeMasters, B. K., Alassiri, A. H., Birks, D. K., Newell, K. L., Moore, W., and Lillehei, K. O. (2010). Epithelioid versus rhabdoid glioblastomas are distinguished by monosomy 22 and immunohistochemical expression of INI-1 but not Claudin 6. *Am. J. Surg. Pathol.* 34, 341–354. doi: 10.1097/PAS.0b013e3181ce107b

Kloosterhof, N. K., de Rooi, J. J., Kros, M., Eilers, P. H. C., Smitt, P. A. E. S., van den Bent, M. J., et al. (2013). Molecular subtypes of glioma identified by genome-wide methylation profiling. *Genes Chromosomes Cancer* 52, 665–674. doi: 10.1002/gcc.22062

Koch, A., Joosten, S. C., Feng, Z., de Ruijter, T. C., Draht, M. X., Melotte, V., et al. (2018). Author correction: analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* 15:467. doi: 10.1038/s41571-018-0028-9

Lee, E., Yong, R. L., Paddison, P., and Zhu, J. (2018). Comparison of glioblastoma (GBM) molecular classification methods. *Semin. Cancer Biol.* 53, 201–211. doi: 10.1016/j.semcancer.2018.07.006

Li, R., Chen, X., You, Y., Wang, X., Liu, Y., Hu, Q., et al. (2015). Comprehensive portrait of recurrent glioblastoma multiforme in molecular and clinical characteristics. *Oncotarget* 6, 30968–30974. doi: 10.18632/oncotarget.5038

Li, Y., Gong, Y., He, S., Li, X., and Zhou, L. (2019). Downregulation of CLDN7 due to promoter hypermethylation is associated with human clear cell renal cell carcinoma progression and poor prognosis. *Eur. Urol. Suppl.* 18:e91. doi: 10.1016/s1569-9056(19)30069-7

Li, Z., Takenobu, H., Setyawati, A. N., Akita, N., Haruta, M., Satoh, S., et al. (2018). EZH2 regulates neuroblastoma cell differentiation via NTRK1 promoter epigenetic modifications. *Oncogene* 37, 2714–2727. doi: 10.1038/s41388-018-0133-3

Linnekamp, J. F., Wang, X., Medema, J. P., and Vermeulen, L. (2015). Colorectal cancer heterogeneity and targeted therapy: a case for molecular disease subtypes. *Cancer Res.* 75, 245–249. doi: 10.1158/0008-5472.can-14-2240

Liu, X., Chen, X., Shi, L., Shan, Q., Cao, Q., Yue, C., et al. (2019). The third-generation EGFR inhibitor AZD9291 overcomes primary resistance by continuously blocking ERK signaling in glioblastoma. *J. Exp. Clin. Cancer Res.* 38:219. doi: 10.1186/s13046-019-1235-7

Lofton-Day, C., and Lesche, R. (2003). DNA methylation markers in patients with gastrointestinal cancers. *Digest. Dis.* 21, 299–308. doi: 10.1159/000075352

Louis, D. N., Perry, A., Burger, P., Ellison, D. W., Reifenberger, G., von Deimling, A., et al. (2014). International society of neuropathology–haarlem consensus guidelines for nervous system tumor classification and grading. *Brain Pathol.* 24, 429–435. doi: 10.1111/bpa.12171

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarellabranger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Lowder, L., Hauenstein, J., Woods, A., Chen, H. R., Rupji, M., Kowalski, J., et al. (2019). Gliosarcoma: distinct molecular pathways and genomic alterations identified by DNA copy number/SNP microarray analysis. *J. Neuro Oncol.* 143, 381–392. doi: 10.1007/s11060-019-03184-1

Lu, C., Ward, P. S., Kapoor, G. S., Rohle, D., Turcan, S., Abdelwahab, O., et al. (2012). IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature* 483, 474–478. doi: 10.1038/nature10860

Lukas, R. V., and Mrugala, M. M. (2016). Pivotal therapeutic trials for infiltrating gliomas and how they affect clinical practice. *Neuro Oncol. Pract.* 4, 209–219. doi: 10.1093/nop/npw016

Ma, H., Zhao, C., Zhao, Z., Hu, L., Ye, F., Wang, H., et al. (2019). Specific glioblastoma multiforme prognostic-subtype distinctions based on DNA methylation patterns. *Cancer Gene Therap* 1–13. doi: 10.1038/s41417-019-0142-6

Marko, N. F., and Weil, R. J. (2013). The molecular biology of WHO Grade II gliomas. *Neurosurg. Focus* 34:E1. doi: 10.1155/2011/372509

Marusyk, A., and Polyak, K. (2010). Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* 1805, 105–117. doi: 10.1016/j.bbcan.2009.11.002

Mclendon, R. E., Friedman, A. H., Bigner, D. D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature07385

Meyer-Puttlitz, B., Hayashi, Y., Waha, A., Rollbrocker, B., Boström, J., Wiestler, O. D., et al. (1997). Molecular genetic analysis of giant cell glioblastomas. *Am. J. Pathol.* 151, 853–857.

Mondesir, J., Willekens, C., Touat, M., and de Botton, S. (2016). IDH1 and IDH2 mutations as novel therapeutic targets: current perspectives. *J. Blood Med.* 7, 171–180. doi: 10.2147/JBM.S70716

Mosrati, M., Malmström, A., Lysiak, M., Krysztofiak, A., Hallbeck, M., Milos, P., et al. (2015). TERT promoter mutations and polymorphisms as prognostic factors in primary glioblastoma. *Oncotarget* 6, 16663–16673. doi: 10.18632/oncotarget.4389

Muhammad, J. S., Khan, M. R., and Ghias, K. (2018). DNA methylation as an epigenetic regulator of gallbladder cancer: an overview. *Int. J. Surg.* 53, 178–183. doi: 10.1016/j.ijsu.2018.03.053

Neilsen, B. K., Sleightholm, R., McComb, R., Ramkissoon, S. H., Ross, J. S., Corona, R. J., et al. (2019). Comprehensive genetic alteration profiling in primary and recurrent glioblastoma. *J. Neurooncol.* 142, 111–118. doi: 10.1007/s11060-018-03070-2

Nobusawa, S., Watanabe, T., Kleihues, P., and Ohgaki, H. (2009). IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. *Clin. Cancer Res.* 15, 6002–6007. doi: 10.1158/1078-0432.CCR-09-0715

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17, 510–522.

Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., et al. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* 63, 1602–1607.

Ohgaki, H., and Kleihues, P. (2007). Genetic pathways to primary and secondary glioblastoma. *Am. J. Pathol.* 170, 1445–1453. doi: 10.2353/ajpath.2007.070011

Ohgaki, H., and Kleihues, P. (2013). The definition of primary and secondary glioblastoma. *Clin. Cancer Res.* 19, 764–772. doi: 10.1158/1078-0432.CCR-12-3002

Olmez, I., Love, S., Xiao, A., Manigat, L., Randolph, P., McKenna, B. D., et al. (2017). Targeting the mesenchymal subtype in glioblastoma and other cancers via inhibition of diacylglycerol kinase alpha. *Neuro Oncol.* 20, 192–202. doi: 10.1093/neuonc/nox119

Paolillo, M., Boselli, C., and Schinelli, S. (2018). Glioblastoma under siege: an overview of current therapeutic strategies. *Brain Sci.* 8:15. doi: 10.3390/brainsci8010015

Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y., and Zhao, Z. (2019). Subtype-specific signaling pathways and genomic aberrations associated with prognosis of glioblastoma. *Neuro Oncol.* 21, 59–70. doi: 10.1093/neuonc/noy120

Park, J., Shim, J. K., Yoon, S. J., Kim, S. H., Chang, J. H., and Kang, S. G. (2019). Transcriptome profiling-based identification of prognostic subtypes and multi-omics signatures of glioblastoma. *Sci. Rep.* 9:10555. doi: 10.1038/s41598-019-47066-y

Parsons, D. W., Jones, S., Zhang, X., Lin, J., Leary, R. J., Angenendt, P., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807–1812.

Paul, Y., Mondal, B., Patil, V., and Somasundaram, K. (2017). DNA methylation signatures for 2016 WHO classification subtypes of diffuse gliomas. *Clin. Epigenet.* 9:32. doi: 10.1186/s13148-017-0331-9

Pérez, R. F., Tejedor, J. R., Bayón, G. F., Fernández, A. F., and Fraga, M. F. (2018). Distinct chromatin signatures of DNA hypomethylation in aging and cancer. *Aging Cell* 17:e12744. doi: 10.1111/acel.12744

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., et al. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9, 157–173. doi: 10.1016/j.ccr.2006.02.019

Rickman, D. S., Bobek, M. P., Misek, D. E., Kuick, R., Blaivas, M., Kurnit, D. M., et al. (2001). Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res.* 61, 6885–6891.

Rohle, D., Popovicimuller, J., Palaskas, N., Turcan, S., Grommes, C., Campos, C., et al. (2013). An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. *Science* 340, 626–630. doi: 10.1126/science.1236062

Romani, M., Pistillo, M. P., and Banelli, B. (2018). Epigenetic targeting of glioblastoma. *Front. Oncol.* 8:448. doi: 10.3389/fonc.2018.00448

Sallinen, S.-L., Sallinen, P. K., Haapasalo, H. K., Helin, H. J., Helén, P. T., Schraml, P., et al. (2000). Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. *Cancer Res.* 60, 6617–6622.

Sandmann, T., Bourgon, R., Garcia, J., Li, C., Cloughesy, T., Chinot, O. L., et al. (2015). Patients With proneural glioblastoma may derive overall survival benefit from the addition of bevacizumab to first-line radiotherapy and temozolomide: retrospective analysis of the AVAglio trial. *J. Clin. Oncol.* 33, 2735–2744. doi: 10.1200/JCO.2015.61.5005

Sang, Y., Hou, Y., Cheng, R., Zheng, L., Alvarez, A., Hu, B., et al. (2019). Targeting PDGFRα-activated glioblastoma through specific inhibition of SHP-2–mediated signaling. *Neuro Oncol.* 21, 1423–1435. doi: 10.1093/neuonc/noz107

Schafer, N., Gielen, G. H., Rauschenbach, L., Kebir, S., Till, A., Reinartz, R., et al. (2019). Longitudinal heterogeneity in glioblastoma: moving targets in recurrent versus primary tumors. *J. Transl. Med.* 17:96. doi: 10.1186/s12967-019-1846-y

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470. doi: 10.1126/science.270.5235.467

Schumacher, T., Bunse, L., Pusch, S., Sahm, F., Wiestler, B., Quandt, J., et al. (2014). A vaccine targeting mutant IDH1 induces antitumour immunity. *Nature* 512, 324–327. doi: 10.1038/nature13387

Sehgal, A. (1998). Molecular changes during the genesis of human gliomas. *Semin. Surg. Oncol.* 14, 3–12. doi: 10.1002/(sici)1098-2388(199801/02)14:1<3::aid-ssu2<3.0.co;2-f

Sharma, A., Bendre, A., Mondal, A., Muzumdar, D., Goel, N., and Shiras, A. (2017). Angiogenic gene signature derived from subtype specific cell models segregate proneural and mesenchymal glioblastoma. *Front. Oncol.* 7:146. doi: 10.3389/fonc.2017.00146

Shergalis, A., Bankhead, A. III, Luesakul, U., Muangsin, N., and Neamati, N. (2018). Current challenges and opportunities in treating glioblastoma. *Pharmacol. Rev.* 70, 412–445. doi: 10.1124/pr.117.014944

Shinojima, N., Tada, K., Shiraishi, S., Kamiryo, T., Kochi, M., Nakamura, H., et al. (2003). Prognostic value of epidermal growth factor receptor in patients with glioblastoma multiforme. *Cancer Res.* 63, 6962–6970.

Skiriute, D., Vaitkiene, P., Saferis, V., Asmoniene, V., Skauminas, K., Deltuva, V. P., et al. (2012). MGMT, GATA6, CD81, DR4, and CASP8 gene promoter methylation in glioblastoma. *BMC Cancer* 12:218. doi: 10.1186/1471-2407-12-218

Songtao, Q., Lei, Y., Si, G., Yanqing, D., Huixia, H., Xuelin, Z., et al. (2012). IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci.* 103, 269–273. doi: 10.1111/j.1349-7006.2011.02134.x

Spiegl-Kreinecker, S., Lötsch, D., Ghanim, B., Pirker, C., Mohr, T., Laaber, M., et al. (2015). Prognostic quality of activating TERT promoter mutations in glioblastoma: interaction with the rs2853669 polymorphism and patient age at diagnosis. *Neuro Oncol.* 17, 1231–1240. doi: 10.1093/neuonc/nov010

Srividya, M. R., Thota, B., Arivazhagan, A., Thennarasu, K., Balasubramaniam, A., Chandramouli, B. A., et al. (2010). Age-dependent prognostic effects of EGFR/p53 alterations in glioblastoma: study on a prospective cohort of 140 uniformly treated adult patients. *J. Clin. Pathol.* 63, 687–691. doi: 10.1136/jcp.2009.074898

Stancheva, G., Goranova, T., Laleva, M., Kamenova, M., Mitkova, A., Velinov, N., et al. (2014). IDH1/IDH2 but Not TP53 mutations predict prognosis in bulgarian glioblastoma patients. *BioMed. Res. Int.* 2014:654727.

Stupp, R., Mason, W. P., Den Bent, M. J. V., Weller, M., Fisher, B., Taphoorn, M. J. B., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* 352, 987–996.

Szerlip, N., Pedraza, A., Chakravarty, D., Azim, M., Mcguire, J., Fang, Y., et al. (2012). Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proc. Natl. Acad. Sci. U. S. A.* 109, 3041–3046. doi: 10.1073/pnas.1114033109

Szopa, W., Burley, T. A., Kramer-Marek, G., and Kaspera, W. (2017). Diagnostic and therapeutic biomarkers in glioblastoma: current status and future perspectives. *Biomed. Res. Int.* 2017:8013575. doi: 10.1155/2017/8013575

Taylor, J., and Schiff, D. (2015). Treatment considerations for MGMT-unmethylated glioblastoma. *Curr. Neurol. Neurosci. Rep.* 15:507. doi: 10.1007/s11910-014-0507-z

Teo, W. Y., Sekar, K., Seshachalam, P., Shen, J., Chow, W. Y., Lau, C. C., et al. (2019). Relevance of a TCGA-derived glioblastoma subtype gene-classifier among patient populations. *Sci. Rep.* 9:7442. doi: 10.1038/s41598-019-43173-y

van den Bent, M. J., Gao, Y., Kerkhof, M., Kros, J. M., Gorlia, T., van Zwieten, K., et al. (2015). Changes in the EGFR amplification and EGFRvIII expression between paired primary and recurrent glioblastomas. *Neuro Oncol.* 17, 935–941. doi: 10.1093/neuonc/nov013

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270, 484–487.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D. I. S., Zairis, S., et al. (2016). Clonal evolution of glioblastoma under therapy. *Nat. Genet.* 48, 768–776.

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., et al. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* 32:152. doi: 10.1016/j.ccell.2017.06.003

Watanabe, T., Nobusawa, S., Kleihues, P., and Ohgaki, H. (2009). IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. *Am. J. Pathol.* 174, 1149–1153. doi: 10.2353/ajpath.2009.080958

Weller, M., Cloughesy, T., Perry, J. R., and Wick, W. (2012). Standards of care for treatment of recurrent glioblastoma—are we there yet? *Neuro Oncol.* 15, 4–27. doi: 10.1093/neuonc/nos273

Wu, X., Rauch, T. A., Zhong, X., Bennett, W. P., Latif, F., Krex, D., et al. (2010). CpG island hypermethylation in human astrocytomas. *Cancer Res.* 70, 2718–2727. doi: 10.1158/0008-5472.can-09-3631

Yamashita, K., Hosoda, K., Nishizawa, N., Katoh, H., and Watanabe, M. (2018). Epigenetic biomarkers of promoter DNA methylation in the new era of cancer treatment. *Cancer Sci.* 109, 3695–3706. doi: 10.1111/cas.13812

Yan, H., Parsons, D. W., Jin, G., Mclendon, R. E., Rasheed, B. K. A., Yuan, W., et al. (2009). IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.* 360, 765–773.

Yang, P., Cai, J., Yan, W., Zhang, W., Wang, Y., Chen, B., et al. (2016). Classification based on mutations of TERT promoter and IDH characterizes subtypes in grade II/III gliomas. *Neuro Oncol.* 18, 1099–1108. doi: 10.1093/neuonc/now021

Yin, A. A., Lu, N., Etcheverry, A., Aubry, M., Barnholtz-Sloan, J., Zhang, L. H., et al. (2018). A novel prognostic six-CpG signature in glioblastomas. *CNS Neurosci. Ther.* 24, 167–177. doi: 10.1111/cns.12786

Yu, Z.-Q., Zhang, B.-L., Ren, Q.-X., Wang, J.-C., Yu, R.-T., Qu, D.-W., et al. (2013). Changes in transcriptional factor binding capacity resulting from promoter region methylation induce aberrantly high GDNF expression in human glioma. *Mol. Neurobiol.* 48, 571–580. doi: 10.1007/s12035-013-8443-5

Yuan, Y., Qi, C., Maling, G., Xiang, W., Yanhui, L., Ruofei, L., et al. (2016). TERT mutation in glioma: frequency, prognosis and risk. *J. Clin. Neurosci.* 26, 57–62. doi: 10.1016/j.jocn.2015.05.066

Zhao, Z., Zhang, K., Wang, Q., Li, G., Zeng, F., Zhang, Y., et al. (2020). Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data for chinese glioma patients. *bioRxiv* doi: 10.1101/2020.01.20.911982

Zhao, S., Zuo, W.-J., Shao, Z.-M., and Jiang, Y.-Z. (2020). Molecular subtypes and precision treatment of triple-negative breast cancer. *Ann. Transl. Med.* 8:499. doi: 10.21037/atm.2020.03.194

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

Copyright © 2020 Zhang, Xia, Liu, Li and Dong. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fmolb.2020.582593

A Theoretical Approach for Correlating Proteins to Malignant Diseases

Rasha Elnemr[1], Mohammed M. Nasef[2], Passant Elkafrawy[2,3], Mahmoud Rafea[1] and Amani Tariq Jamal[4]*

[1] Climate Change Information Center & Renewable Energy & Expert Systems, Giza, Egypt, [2] Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, Shibin El Kom, Egypt, [3] Information Technology and Computer Science, Nile University, Giza, Egypt, [4] Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Malignant Tumors are developed over several years due to unknown biological factors. These biological factors induce changes in the body and consequently, they lead to Malignant Tumors. Some habits and behaviors initiate these biological factors. In effect, the immune system cannot recognize a Malignant Tumor as foreign tissue. In order to discover a fascinating pattern of these habits, behaviors, and diseases and to make effective decisions, different machine learning techniques should be used. This research attempts to find the association between normal proteins (environmental factors) and diseases that are difficult to diagnose and propose justifications for those diseases. This paper proposes a technique for medical data mining using association rules. The proposed technique overcomes some of the limitations in current association algorithms such as the *Apriori* algorithm and the Equivalence CLAss Transformation (ECLAT) algorithm. A modification to the *Apriori* algorithm has been proposed to mine Erythrocytes Dynamic Antigens Store (EDAS) data in a more efficient and tractable way. The experiments inferred that there is a relation between normal proteins as environment proteins, food proteins, commensal proteins, tissue proteins, and disease proteins. Also, the experiments show that habits and behaviors are associated with certain diseases. The presented tool can be used in clinical laboratories to discover the biological causes of malignant diseases.

Keywords: malignant tumors, data mining, association rule, *apriori* algorithm, eclat algorithm

OPEN ACCESS

Edited by:
Cheng Zhang,
KTH Royal Institute of Technology, Sweden

Reviewed by:
Jacopo Junio Valerio Branca, University of Florence, Italy
Maria Letizia Urban, University of Florence, Italy

****Correspondence:***
Rasha Elnemr
Rashaelnemr127@gmail.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *12 July 2020*
Accepted: *22 September 2020*
Published: *22 October 2020*

Citation:
Elnemr R, Nasef MM, Elkafrawy P, Rafea M and Jamal AT (2020) A Theoretical Approach for Correlating Proteins to Malignant Diseases. Front. Mol. Biosci. 7:582593. doi: 10.3389/fmolb.2020.582593

INTRODUCTION

Lifestyle habits and behaviors affect human general health, like cigarette smoking, excessive alcohol consumption, excessive sunlight exposure, poor diet, lack of exercise, medical drugs, change of hormones, radiation, viruses, bacteria, and environmental chemicals. Chemical factors might be in the air, water, food, and/or workplace. The genetic makeup is essential so that these mentioned factors can lead to malignant transformation (Fymat, 2017; Iqbal, 2017; Ellberg et al., 2018; Ukawa et al., 2018).

Because of the complicated interplay of many habits and behaviors, it is difficult to predict which combination of these habits and behaviors is accountable for certain cancer. The cause of cancer is

still unknown and the human body's readiness to be diseased is unpredictable. One of the important areas of research today is attempting to identify the association between the habits and behavior of an individual and diseases, specifically, Malignant Tumor.

Rafea and Souchelnytskyi (2012) observed and described phenomena related to the protein content of the Red Blood Cell (RBC). It was noticed that plasma contains antibodies against some of the RBC proteins, which are contained within the cytoplasm of RBC of the same person. The discovery is that RBC has a dynamic store of body antigens [Tissue-Specific Antigens (TSA)], food antigens, environment antigens, bacterial commensals antigens, and disease antigens whether microbial, viral, or tumors. This store is named: Erythrocytes Dynamic Antigens Store (EDAS).

To maximize the utility of the EDAS, computer knowledge processing capability was adopted to increase the profit of this discovery. To this endeavor, a random generation of the EDAS model was described in Rafea et al. (2019). This random generation was based on a mathematical model that simulates reality. The random generation of EDAS consists of a set of normal proteins and a set of disease proteins. The normal proteins are environment proteins, food proteins, commensal proteins, and tissue proteins. The diseases proteins are malignant tumor proteins or pathogens proteins. They developed a biomarker discovery technique to detect a minimum set of biomarkers for each disease. They applied their technique on two categories of diseases; Malignancies (Mi) and Pathogens (Gi). Thus, malignancies have 20 types (M1, M2, . . ., M20) and pathogens have 20 types (G1, G2, . . ., G20). In this work, we will use the EDAS data to find which normal proteins (Tissue-Specific Antigens, food antigens, environment antigens, and bacterial commensals antigens) are related to a particular Malignant Tumor. The main challenge of our research is to find the interesting correlations and associations between the set of normal proteins to the set of malignant tumor proteins.

To forecast the association of those biological data, we should use an association rule mining algorithm. The *Apriori* algorithm was described by Agrawal et al. (1993), is widely used to study the relations and associations between items in an ecosystem. The *Apriori* algorithm is simple, and it is easy to program (Ghosh and Dutta, 2016; Patil and Deshmukh, 2016). It applies the *Apriori* property; a candidate itemset is unnecessary if at least one of its subsets is infrequent (Ingle and Suryavanshi, 2015). Hence, it reduces the number of candidate itemsets. However, the *Apriori* algorithm requires multiple scans over the database for generating the itemsets (Ingle and Suryavanshi, 2015; Patil and Deshmukh, 2016). Since the number of database passes is equal to the max length of the frequent itemset, it takes time to scan the database (Han et al., 2000; Mandave et al., 2013; Kaur and Madan, 2015; Rajeswari, 2015). However, the *Apriori* algorithm has low performance in big datasets (Ghosh and Dutta, 2016). In this study, the primary dataset as typical medical data with a considerable number of cases and a large number of features.

Another algorithm is Equivalence CLAss Transformation (ECLAT) was described by Zaki (2000). ECLAT is also used to study the associations between items in a more efficient manner. It only scans the database once. The ECLAT algorithm uses a depth-first search strategy (Shah and Patel, 2015; Shukla and Solanki, 2015; Giri et al., 2016). Thus, it is fast but the accuracy is not preserved, as it violates the *Apriori* property (Shah and Patel, 2015; Ishita and Rathod, 2016). Because the generation of candidate itemset is operated in an equivalence class, the candidate itemsets are not clipped under the prior knowledge. These candidate itemsets still need to be calculated. Although adopting the technology of equivalence classes, ECLAT needs to judge whether two k-itemsets can be joined to generate a $(k + 1)$-itemsets, a great time is needed if the itemset is very long (Kavitha and Selvi, 2016). A large number of conditional branches are used to merge which are highly predictable. There is a waste of time to calculate the support of infrequent itemsets. It fails to manage the main memory at the time of high candidate itemsets (Giri et al., 2016).

Accordingly, there is a need to propose a technique that preserves the *Apriori* property to ignore a candidate itemset if at least one of its subsets has support less than the threshold and works efficiently to infer association relations. Many researchers have done modifications to improve the efficiency of the *Apriori* algorithm and to overcome some of the limitations of the *Apriori* algorithm. They used some methods to improve *Apriori* efficiency as Intersection (Aqra et al., 2018), Hash-based itemset counting (Park et al., 1995; Vyas and Sherasiya, 2016), Partitioning (Jia et al., 2012), Sampling (Toivonen, 1996; Rajeswari, 2015), etc. The Intersection is a method used to improve memory management, efficiently, by reducing the computation cost of *Apriori* and removing its complexity. The Intersection method is designed for vertical data format, by removing the limitations of the horizontal data format used in *Apriori*. Moreover, the support is calculated by counting the common transactions that contain each element of the candidate set. This takes less time than the original algorithm; however, data must be in a vertical layout (Agrawal et al., 2013; Raval et al., 2013; Aqra et al., 2018).

In this research, we shall enhance the *Apriori* algorithm performance by adopting an intersection mechanism while preserving the *Apriori* property to achieve accuracy. Accordingly, the performance is enhanced by two ways; executing one scan to the database and applying the *Apriori* property that helps to reduce the number of candidate itemsets. Therefore, in this research, the proposed algorithm is given to discover the association of proteins to environmental factors, with higher performance and accuracy.

This paper is structured as follows: Section 2 states the related work. Section 3 states the background of using the *Apriori* algorithm in medical problems. Section 4 includes the proposed model. Section 5 describes the experiment and results. Section 6 describes the evaluation. Section 7 describes the discussion. And Section 8 describes the conclusion.

RELATED WORK

Some researchers attempt to improve the *Apriori* algorithm based on Intersection as in Ganesh et al. (2016), the authors proposed a Vertical Format Frequent Mining (VFFM) algorithm. This

algorithm was used to find frequent items from the database. The transaction database is transformed into a vertical data format. They scan the database only one time. They converted the data into 0 and 1 and calculated the support for them. They used the depth strategy for mining. Thus, they could not apply the *Apriori* property. Also, there is no need for converting the data into 0 and 1 because the researcher did not benefit from it. There is no application for their method.

Aqra et al. (2018) presented an Intermediate Transaction ID *Apriori* (ITD *Apriori*) algorithm where a new itemset format structure is adopted to address the problem of threshold that necessitates rescanning the entire database. This approach creates an intermediate itemset. The intermediate itemset has a new structure called Intermediate Transaction Id itemset ITDM list; this structure increases the efficiency of mining. This can be done by scanning the database and by representing data to a vertical data format. After this process, support can be collected by the intersection TID list. Thus, it improves the overall efficiency as no longer the algorithm needs to rescan the entire database. The algorithm also helps to extract frequent itemsets according to pre-determined minimum support with an independent purpose. Furthermore, the association rule set is extracted with high confidence and weak support. However, they used the depth strategy for mining; thus could not apply the *Apriori* property.

Chen and Xiao (2014) proposed the Intersection Maximum Frequent Pattern (ISMFP) algorithm, which was based on set theory and the idea of a top-down search for mining the maximum frequent patterns. Since the maximum frequent patterns have already implied all frequent patterns. They converted the problem from detecting frequent patterns to discover the maximum frequent patterns, avoiding the production of a large number of candidate sets. They forwarded a kind of association rule mining algorithm depending on the intersection, which decreases the search space and the number of cycles by using the principle of the maximum frequent pattern and intersection. Their experimental results showed that the algorithm ISMFP is efficient in mining frequent patterns; especially there exists a low threshold of support degree or long patterns.

BACKGROUND

Multiple papers reviewed the solution of medical problems as Association Rules. They provided a computational study, based on the *Apriori* algorithm to discover the associations among clinical traits and risk factors of different disease [i.e., asthma (Poorani et al., 2018), chronic diseases (Karthiyayini and Jayaprakash, 2015), and heart diseases (Said et al., 2015)]. The *Apriori* algorithm was used to find the frequent symptoms and related causes of a disease from the dataset that was collected from self-reported patients. It was even used in predicting the possibility of chronic occurrence of diseases. In (Karthiyayini and Jayaprakash, 2015), the percentage of possibility for chronic disease was calculated from each symptom of all considered chronic diseases. The higher number of symptoms leads to higher accuracy of calculating the disease possibility.

Others diagnosed by considering NED (No Evidence of Disease) and ED (Evidence of Disease) studies. Fahrudin et al. (2017) experimented with two Association Rules algorithms: *Apriori* and FP-Growth. They categorized NED and ED to detect the relationship between different factors that influences patients. Another study by Gitanjali et al. (2014) was conducted to generate the frequency of diseases that affect patients in various geographical regions and at various periods. The use of association in medicine has numerous critical applications; where mainly *Apriori* is the exemplar algorithm of all those studies.

The state of the art literature concentrated on discovering the relationship between the diseases and their symptoms. However, we have a different goal to discover the relationship between normal proteins and the disease's proteins.

MATERIALS AND METHODS

From the comparison of Association Rule Mining (ARM) algorithms, we propose an enhancement to the *Apriori* algorithm with the concept of vertical format from the ECLAT algorithm. We merged several scientific concepts of mining into *Apriori* employing the set theory of intersection, lattice theory, and vertical data format. Moreover, we use divide and conquer methodology to divide the problem into clusters.

The Proposed Association Rule Model

The proposed association rule model is used to discover the association between the normal proteins and the diseased proteins. It is composed of two phases. Phase one is the *Disease Sub-Typing (DST)*, we attempt to find the different subtypes of a particular disease. Phase two is the *Association Rule (AR) generation*; we attempt to find the association between each subtype to its normal proteins. Inferring, the association rules of factors causing diseases.

Phase One: Disease Sub-Typing

The main idea is to cluster each malignant tumor, M_i, into its subtypes. For each disease, the set of cases is compared to each other to extract similar cases. The similarity is measured by the number of similar proteins. A certain threshold is used to identify the subtype. The rest of the cases are re-evaluated to each other until no further cases remain. The steps are:

1. We select the cases which have the same malignant tumor Mi.
2. For the first case (record), we compare its set of proteins with sets of proteins in the rest cases.
3. The cases which their proteins have similarity with the first case proteins, equal to or greater than the threshold are considered of the same sub-type.
4. The other cases which are less than the threshold are re-evaluated by repeating the previous steps until no further cases remain.

Interestingly, the set of proteins defining a subtype of the disease can be used as a signature of that specific disease for diagnosis.

Disease sub-typing algorithm

Algorithm 1: detecting the type of each Disease.

```
  #Input: PatientsRecords be the list of
all records
  #Output: typeIndexTable of disease types
and the Records to each disease type
  DisTypeProtein proteins of disease types
  //intersection algorithm will be used to
cluster the records
  Initialize SimilarityFactor//this is a
constant user-defined factor
  for each Disease in PatientsRecords

    set dList = select all diseaserecords
    from PatientsRecords
    //find similar cases
    create typeIndexTable//where each
    protein list of records is stored
    create DisTypeProtein//collection of
    proteins in cases of same type
    k = 1
    foreach r in dList do

      add in typeIndexTable(Tk, r)
      DisTypeProtein [k] = r
      remove r from dList
      foreach nextR in dList
        count = | r ∩ nextR |//number of
        intersecting proteins
        avgRecordLength = (length of
        r + length of nextR)/2
        intersectionPerc = count/avgRecord
        Length
        if (intersectionPerc >  =
        SimilarityFactor) then
        add in typeIndexTable(Tk, nextR)
        remove nextR from dList
        DisTypeProtein [k] ∪ nextR//add
        proteins of nextR
      end foreach//nextR
      increment K

    end foreach r
    returntypeIndexTable
    returnDisTypeProtein

end foreach
```

Phase Two: Association Rule

The proposed association rule mining algorithm is considered as an enhancement to the *Apriori* algorithm. The proposed algorithm is based on converting the structure of the dataset from horizontal to vertical format. This vertical format will facilitate applying the intersection concept for getting the support for each itemset. Therefore, the breadth-first search strategy of the proposed new algorithm preserves the *Apriori* property. It is known that most association rule mining algorithms that apply the vertical format works by depth search strategy which violates the *Apriori* property. However, our proposed association rule mining algorithm works by the breadth search strategy. From this point, the proposed algorithm does only one scan to the

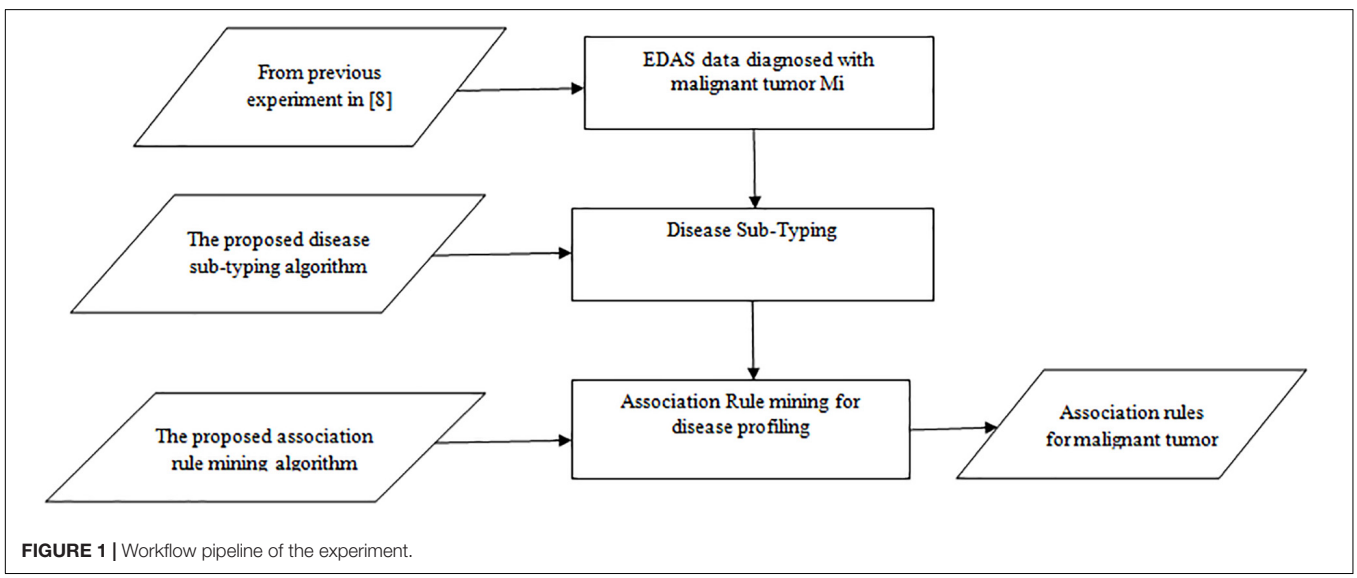


FIGURE 1 | Workflow pipeline of the experiment.

TABLE 1 | Results of the experiment for malignant tumors (Rafea et al., 2019).

Disease	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Number of_records	2063	2109	2083	2053	2035	2094	2062	2135	1982	2096
Disease	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
Number of_records	2040	2084	2076	2149	2130	2115	2059	2080	2116	2181

database, decreasing the candidate sets. Thus, it is fast and guarantees accuracy.

The Proposed Association Rule Mining Algorithm

The main challenge of the model is in its ability to help in finding the relations between the normal proteins and disease conditions that are difficult to diagnose and propose justifications for these diseases through the following steps:

First, we scan all the records to convert the datasets from horizontal to vertical format. This step generates a protein index table containing all 1-itemsets (proteins) without repetition, TID (the records where the protein is located), and its support.

Second, prune the proteins which don't satisfy the minimum support.

Third, self-join the frequent proteins to generate 2-itemsets. Note that the 2-itemsets are unordered sets such that p1p5 is the same as p5p1. Calculate the support for 2-itemsets by the intersection concept where we get the common records between those 2-itemsets by intersecting the record list of each itemset. This eliminated the scans to the database.

Forth, prune the 2-itemsets which don't satisfy the minimum support.

Fifth, self-join the frequent proteins to generate 3-itemsets, however, the ***Apriori* property** has to be applied here. Each subset of the generated 3-itemsets must also be a frequent itemset, i.e. satisfies the minimum support. If at least one 2-itemset of the generated 3-itemsets are infrequent then discard this generated 3-item set.

Sixth, prune the 3-itemsets which don't satisfy the minimum support.

Finally, repeat these steps until no new frequent itemsets are identified.

Proposed Association Rule Mining Algorithm

Algorithm 2: detecting the association between Normal Proteins and Diseases Proteins.

```
  #Input: typeIndexTable be the list of all
records associated with a certain disease
type
  #Output: associationProteins be the list
of proteins associated with each disease
  //Apriori algorithm will be used
  Initialize MIN_SUPPORT//this is a
constant will be changed to search for a
useful result
  for each typeOfDisease in typeIndexTable

    set dList = select all records
    of a certain disease defined by
    typeIndexTable
    set dListLength = no_of_records (dList)
    set proteinList = select all proteins
    in dList//without repetition
    //find frequent proteins
    create proteinIndexTable//where each
    protein list of transactions is stored
    foreachp in proteinList do

      tList = select all transactions, TID,
      in dList that have p
      pCount = no of transactions that have
      P
      pSupport = pCount * 100/dListLength
      if (pSupport >  = MIN_SUPPORT) then
      add in proteinIndexTable(p, pCount,
      pSupport, tList)

    end foreach
```

TABLE 2 | Results of phase two (malignant tumor subtypes).

Disease	Its subtypes	Number of subtypes
Malignant Tumor (M1)	M1T1, M1T2, M1T3, M1T4, and M1T5	5
Malignant Tumor (M7)	M7T1, M7T2, M7T3, M7T4, M7T5, and M7T6	6
Malignant Tumor (M18)	M18T1, M18T2, M18T3, M18T4, M18T5, and M18T6	6
Malignant Tumor (M20)	M20T1, M20T2, M20T3, M20T4, M20T5, M20T6, and M20T7	7

TABLE 3 | Results of phase three (association rule mining).

Disease	Disease Subtype	Associations between proteins	
		Normal proteins	Diseased proteins
Malignant Tumor (M1)	M1T1	P3580	P119913, P119662, P119786, P119535, P119939
Malignant Tumor (M1)	M1T2	P3887	P119640, P119513, P119637, P119515, P119790, P119541, P119917, P119886, P119792, P119668
Malignant Tumor (M7)	M7T5	P3734, P6006	P719501, P719807
Malignant Tumor (M7)	M7T6	P625, P3886	P719785, P719964
Malignant Tumor (M18)	M18T2	P6376	P1819795, P1819919, P1819668, P1819546
Malignant Tumor (M18)	M18T5	P327, P3479	P1819668, P1819696
Malignant Tumor (M20)	M20T3	P777	P2019891, P2019964, P2019765, P2019863, P2019640, P2019643, P2019516, P2019741, P2019614, P2019518, P2019638, P2019889, P2019767, P2019990

TABLE 4 | Rules, confidence, lift, and leverage.

Rule Number	Rule	Confidence	Lift	Leverage
R1	P3479 ^ P327 → P1819668	86.36%	1.2164	0.0711
R2	P3479 ^ P327 → P1819696	84.81%	1.1751	0.0622
R3	P3479 → P1819696	81.48%	1.1111	0.0533
R4	P347 9→ P1819668	81.48%	1.062	0.0433

TABLE 5 | Precision, recall, f-measure, and accuracy of the three algorithms on different support (40%, 50%, 60%, and 70%).

Algorithm	Precision %				Recall %				F-measure %				Accuracy %			
	40%	50%	60%	70%	40%	50%	60%	70%	40%	50%	60%	70%	40%	50%	60%	70%
Apriori Algorithm	42	54	62	64	67	69	84	87	52	61	71	74	91	95	98	98
ITDApriori	42	54	62	64	67	69	84	87	52	61	71	74	91	95	98	98
Proposed Algorithm	**42**	**54**	**62**	**64**	**67**	**69**	**84**	**87**	**52**	**61**	**71**	**74**	**91**	**95**	**98**	**98**

TABLE 6 | The execution time (in second) comparison among *Apriori* Algorithm, ITDApriori, Proposed ARM Algorithm on 500 transactions.

Minimum Support	*Apriori* (in sec.)	ITDApriori (in sec.)	Proposed ARM (in sec.)	Improvement %
40%	12174	9296	**7407**	**44.93**
50%	10979	7312	**5840**	**56.60**
60%	8472	4620	**3448**	**89.85**
70%	3125	1368	**1035**	**117.05**

```
//find 2-itemsets frequent proteins as
an unordered set of pairs
joinedProteins = join pi,pj of proteins
in frequentProteins
foreach pi, pj in joinedProteins do
  tList = get all intersecting
  transactions from proteinIndex table
  pCount = length of tList
  pSupport = pCount * 100/dListLength
  if (pSupport >  = MIN_SUPPORT) then
  add in frequentProteins ({pi, pj},
  pSupport)
end foreach
//find i-itemsets frequent proteins
until no merges can be done
set i = 3//the number of proteins to be
merged
curFreqProteins = frequentProteins
repeat
  joinedProteins = join i number of
  proteins in curFreqProteins
  clear curFreqProteins
  foreach set of proteins, {pi} in
  joinedProteinsdo
    //prune itemsets that have
    non-frequent subsets
    if all subsets of {pi} in
    frequentProteinsthen
    tList = get all intersecting
    transactions of {pi} from
    proteinIndex
    pCount = length of tList
    pSupport = pCount * 100/dListLength
    if (pSupport >  = MIN_SUPPORT) then
    add in curFreqProteins ({pi},
    pSupport)
    endif
  endif
  end foreach
  add curFreqProteins
  tofrequentProteins
  increment i
until no MIN_SUPPORT is satisfied
associationProteins = frequentProteins
return associationProteins
end foreach
```

EXPERIMENT AND RESULTS

The experiment is divided into three phases as shown in **Figure 1**: the random generation of the EDAS data, the Disease Sub-Typing phase, and the association rule mining phase. The aim is to detect the association between Normal Proteins and Malignant Tumor Proteins.

The experiment is performed on MacBook Pro, 2.9 GHz Intel Core i5 and 8 GB of RAM, the database is created in Microsoft SQL Server 2008, the algorithms are implemented in C#.

Phase one: the generation of EDAS data, our data is based on the previous experiment by Rafea et al. (2019), where the total generated cases are 100K record. Malignant tumor patients are 41,742 records from the total 100 K, which was concluded from the experiment by Rafea et al. (2019). Malignant Tumor (Mi) has 20 types (M1, M2, . . ., M20). These patients are divided according to their malignant type as shown in **Table 1**. For example, the number of patients who have a Malignant Tumor (M1) is 2063. The number of patients who have a Malignant Tumor (M7) is 2062. The number of patients who have a Malignant Tumor (M18) is 2080. The number of patients who have a Malignant Tumor (M20) is 2181.

Phase two: apply the Disease Sub-Typing Algorithm to detect the sub-type of each malignant tumor (Mi) as explained in section 4.1.1

Phase three: The aim is to detect the association between Normal Proteins and Malignant Tumor Proteins for each subtype of a malignant tumor. We shall apply the Association Rule Mining Algorithm to the resulted data from phase 2 as defined in section 4.1.2

Results

From phase 2, we concluded that each disease can be divided into several subtypes as shown in **Table 2**. For example,

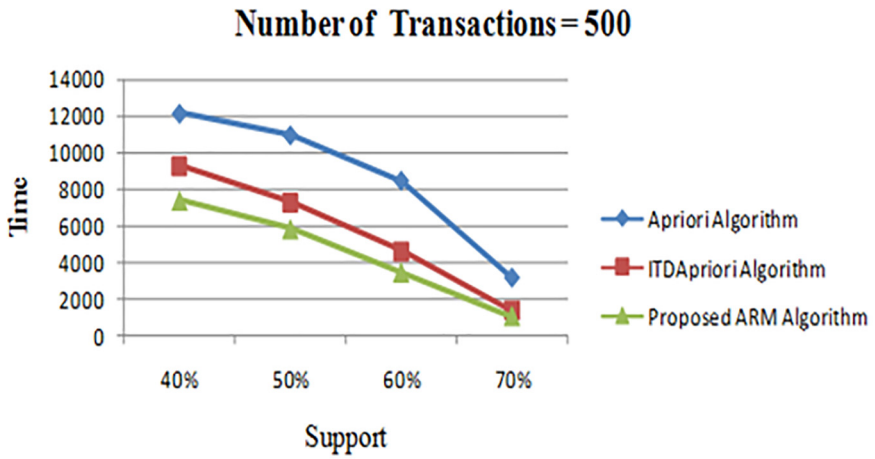


FIGURE 2 | Execution times over different support: *Apriori* Algorithm ITD *Apriori* proposed Algorithm on 500 transactions.

Malignant Tumor (M1) has 5 subtypes which are M1T1, M1T2, M1T3, M1T4, and M1T5. Malignant Tumor (M7) has 6 subtypes which are M7T1, M7T2, M7T3, M7T4, M7T5, and M7T6.

Malignant Tumor (M18) has 6 subtypes which are M18T1, M18T2, M18T3, M18T4, M18T5, and M18T6.

Malignant Tumor (M20) has 7 subtypes which are M20T1, M20T2, M20T3, M20T4, M20T5, M20T6, and M20T7.

From phase three, we concluded that there are interesting associations between malignant tumor proteins and the environmental factors (normal proteins) as shown in **Table 3**. For example,

(1) In Malignant Tumor (M1) which has a subtype (M1T1) and minimum support%60, there is a relation between the food proteins and the disease proteins as the following (P3580, P119913, P119662, P119786, P119535, P119939), where P3580 is a food protein and the rest are malignant tumor proteins.
(2) In Malignant Tumor (M1) which has subtype (M1T2) and minimum support%60, there is a relation between the food proteins and the disease proteins as the following (P3887, P119640, P119513, P119637, P119515, P119790, P119541, P119917, P119886, P119792, P119668), where P3887 is a food protein and the rest are malignant tumor proteins.
(3) In Malignant Tumor (M7) which has subtype (M7T5) and minimum support%60, there is a relation between the food proteins, the commensal proteins, and the disease proteins as the following (P3734, P6006, P719501, P719807), where P3734 is a food protein, P6006 is a commensal protein and the rest are malignant tumor proteins.
(4) In Malignant Tumor (M7) which has subtype (M7T6) and minimum support%60, there is a relation between the environment proteins, the food proteins, and the disease proteins as the following (P625, P3886, P719785, P719964), where P625 is an environment protein, P3886 is a food protein and the rest are malignant tumor proteins.
(5) In Malignant Tumor (M18) which has a subtype (M18T2) and minimum support%60, there is a relation between the commensal proteins and the disease proteins as the following (P6376, P1819795, P1819919, P1819668, P1819546), where P6376 is a commensal protein and the rest are malignant tumor proteins.
(6) In Malignant Tumor (M18) which has a subtype (M18T5) and minimum support%60, there is a relation between the commensal proteins and the disease proteins as the following (P327, P3479, P1819668, P1819696), where P327 is an environmental protein, P3479 is a food protein and the rest are malignant tumor proteins.
(7) In Malignant Tumor (M20) which has subtype (M20T3) and minimum support%60, there is a relation between the environment proteins, and the disease proteins as the following (P777, P2019891, P2019964, P2019765, P2019863, P2019640, P2019643, P2019516, P2019741, P2019614, P2019518, P2019638, P2019889, P2019767, P2019990), where P777 is an environment protein, and the rest are malignant tumor proteins.

Rule Generation

The developed algorithm with a lower minimum support threshold allows for more rules to show up. For example, for the disease M18T5, the generated rules are fifteen rules. Consequently, by applying the statistical measure, like

TABLE 7 | The execution time (in second) comparison among *Apriori* Algorithm, ITDApriori, Proposed ARM Algorithm on 1000 transactions.

Minimum Support	*Apriori* (in sec.)	ITDApriori (in sec.)	Proposed ARM (in sec.)	Improvement (%)
40%	20067	10849	**9195**	**68.11**
50%	16955	9452	**7015**	**88.22**
60%	13433	6028	**4910**	**98.18**
70%	7628	3730	**3275**	**118.71**

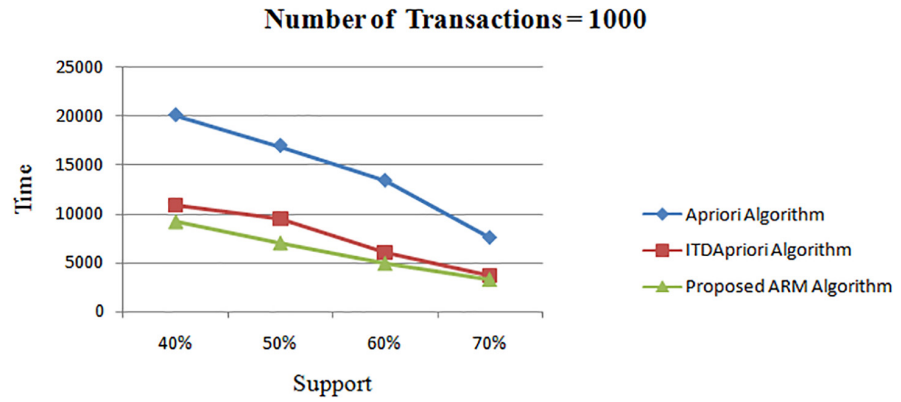


FIGURE 3 | Execution times over different support: *Apriori* Algorithm ITD *Apriori* proposed Algorithm on 1000 transactions.

TABLE 8 | The execution time (in second) comparison among *Apriori* Algorithm, ITDApriori, Proposed ARM Algorithm on 1500 transactions.

Minimum Support	*Apriori* (in sec.)	ITDApriori (in sec.)	Proposed ARM (in sec.)	Improvement %
40%	28176	13880	**11376**	**84.85**
50%	24013	10277	**8949**	**91.58**
60%	17642	8095	**6134**	**109.79**
70%	10384	5260	**3498**	**123.61**

confidence, lift, and leverage, the rules are reduced to four main rules. The generated association rules and their confidence, lift, and leverage of M18T5 are shown in **Table 4**.

Notice that in **Table 4**, the results of R1 show that disease M18T5 profiled by protein P1819668 is affected by proteins

P3479 and P327 which are food and environmental proteins respectively. R2 shows that disease M18T5 profiled by protein P1819696 is affected by proteins P3479 and P327 which are food and environmental proteins respectively. R3 means that disease M18T5 profiled by protein P1819696 is affected by protein P3479 which is food protein. R4 means that disease M18T5 profiled by protein P1819668 is affected by protein P3479 which is food protein.

EVALUATION

The evaluation between the three algorithms (*Apriori* Algorithm, ITDApriori, Proposed Algorithm) consists of two steps; firstly, the common evaluation methodology by calculating precision, recall, F-measure, and accuracy. Secondly, we experiment the performance by calculating the execution time of the three algorithms (*Apriori* Algorithm, ITDApriori, Proposed Algorithm). Thus, the algorithms have been tested over different minimum support and different number of transactions (cases). The minimum support values used are 40%, 50%, 60, and 70%, for the number of transactions 500, 1000, and 1500. For example, we will take (M1T2) to explain the results.

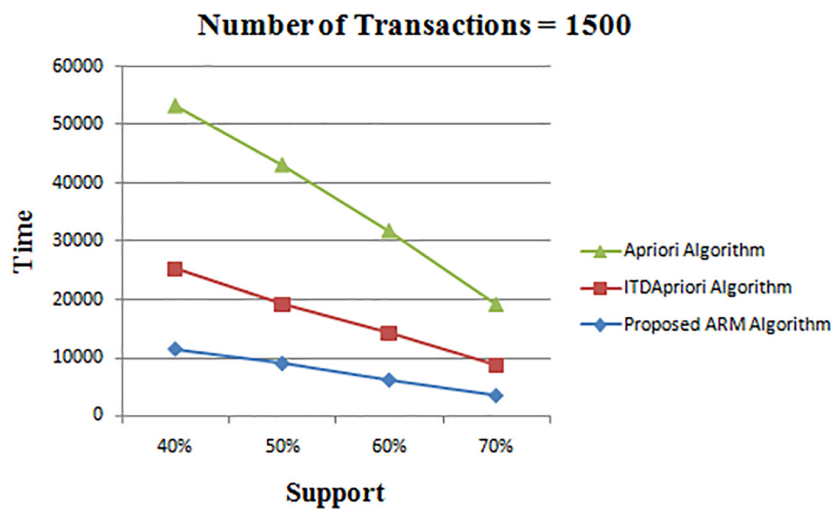


FIGURE 4 | Execution times over different support: *Apriori* Algorithm ITD *Apriori* proposed Algorithm on 1500 transactions.

Table 5 shows the results of the first step of the evaluation methodology on several experiments covering different minimum support (40%, 50%, 60%, and 70%) on the *Apriori* Algorithm, the ITDApriori Algorithm, and the proposed ARM algorithm. As shown in **Table 5**, we noticed that the precision, recall, f-measure, and accuracy are the same for the three algorithms. This is due to the three algorithms gave the same frequent itemset. By which accuracy is not violated by our proposed algorithm.

For the second step of the evaluation, we evaluated the performance of time between the three algorithms. We applied three experiments that are done based on changing the number of transactions and the minimum support values.

The first experiment is carried out for the three algorithms (*Apriori* Algorithm, ITDApriori, and the Proposed Algorithm) on 500 transactions. From **Table 6**, we found that whatever the minimum support value the proposed ARM algorithm is the best. **Table 6** is visualized in **Figure 2**.

The second experiment is conducted for the three algorithms over 1000 transactions. The execution time for the three algorithms is presented in **Table 7**. We noticed that, although changing the value of minimum support, the proposed ARM algorithm is the best. **Table 7** is represented as a graph in **Figure 3**.

Thirdly, the experiment is done for the three algorithms over 1500 transactions. The execution time is presented in **Table 8**. We found that whatever the minimum support value the proposed ARM algorithm is the best. **Table 8** is depicted as a graph in **Figure 4**.

DISCUSSION

From the previous experiments, we concluded that the proposed ARM algorithm is efficient than the ITDApriori and the *Apriori* algorithm. When the number of transactions is increased, the proposed ARM algorithm showed high efficiency, and this is evident from the time of implementation. The proposed ARM algorithm converts the data structure from horizontal to vertical data format. It executes one scan to the database to create an index table.

The proposed ARM algorithm does not do a scan over the entire database for calculating the support of candidate itemsets, it scans only the records of the candidate itemsets in the index table, which causes to reduce the search time. The computation of the support is done by getting the intersection of the Transaction Id sets of the corresponding k-itemsets. The proposed ARM algorithm applies the *Apriori* property which plays a vital role in reducing the search time and space when handling properly. By decreasing the number of candidates itemsets, our algorithm does not waste time for calculating the support for infrequent itemsets. Moreover, it uses the breadth-first search strategy on vertical data layout which guarantees accuracy.

The horizontal format of the data in the *Apriori* algorithm leads to some problems: it does several scans to the entire database which is the main cause of increasing the time. Also, when the number of transactions increases the runtime increases. The *Apriori* algorithm is not efficient in case the number of transactions increases and the number of items increases because more candidate itemsets must be examined during candidate generation and support calculation.

Although the ITDApriori algorithm converts the data structure from horizontal to vertical data format and calculating the support of candidate items by intersection, it does not apply the *Apriori* property which leads to an increase in the number of candidate itemsets. Thus, it takes extra time to calculate the support for infrequent items. Moreover, the larger the number of items the higher the storage space required.

The *Apriori* algorithm scans the database too many times, which reduces the overall performance. Due to this, both the time and space complexity of the *Apriori* algorithm are very high: O (2^D), thus exponential, where D is the width of the transaction (the total number of items) present in the database. While in our proposed Association Rule Mining algorithm the complexity is O $(D + D^2)$. In conclusion, we can say that the proposed ARM Algorithm is faster than the other algorithms while preserving accuracy.

CONCLUSION

This paper is focused on issues related to the design and implementation of an advanced technique. Its main purpose is to help in finding the association between the habits and behavior of the human and causing malignant tumor. The proposed model in this stage is based on hypothetical generated data. Our model is tested by generating databases each with 41742 patients' records who suffering from malignancies. Firstly, the proposed model detects the sub-type of each disease. Secondly, it finds the relation between the normal proteins and each sub-type of malignancies. Lastly, it presents the evaluation of our proposed ARM algorithm against the original *Apriori* Algorithm and, the Intermediate TID *Apriori* (ITDApriori) Algorithm. The evaluation is based on accuracy and time. The results demonstrate that the proposed algorithm achieves superior performance in execution time with preserving accuracy. In the future, the proposed model will be modified by using a parallel method to take in an extremely large database. Improve the proposed model to work on patients that may be infected by more than one disease. Also, it will more interesting to improve the proposed model to predict the relations between these diseases.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

RE is a researcher working on disease diagnosis and this work is a part of her Ph.D. thesis. All the authors worked on reviewing the research.

REFERENCES

Agrawal, P., Kashyap, S., Pandey, V. C., and Keshri, S. P. (2013). A review approach on various form of apriori with association rule mining. *Int. J. Recent Innov. Trends Comput. Commun.* 1, 462–468.

Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Rec.* 22, 207–216. doi: 10.1145/170036.170072

Aqra, I., Herawan, T., Ghani, N. A., Akhunzada, A., Ali, A., Razali, R. B., et al. (2018). A novel association rule mining approach using TID intermediate itemset. *PLoS One* 13:e0179703. doi: 10.1371/journal.pone.0179703

Chen, X., and Xiao, J. (2014). Association rules algorithm based on the intersection. *Open Cybernet. Syst. J.* 8, 1152–1157. doi: 10.2174/1874110X01408011152

Ellberg, C., Olsson, H., and Jernström, H. (2018). Current smoking is associated with a larger waist circumference and a more androgenic profile in young healthy women from high-risk breast cancer families. *Cancer Causes Control* 29, 243–251. doi: 10.1007/s10552-017-0999-3

Fahrudin, T. M., Syarif, I., and Barakbah, A. R. (2017). "Discovering patterns of NED-breast cancer based on association rules using apriori and FP-growth," in *Proceedings of the International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, Bali, 132–139.

Fymat, A. L. (2017). Genetics, epigenetics and cancer. *Cancer Ther. Oncol. Int J.* 4:555634.

Ganesh, C., Sathyabhama, B., and Geetha, D. T. (2016). Fast frequent pattern mining using vertical data format for knowledge discovery. *Int. J. Eng. Res. Manag. Technol.* 5, 141–149.

Ghosh, A., and Dutta, A. (2016). Comparative study of different improvements of apriori algorithm. *Int. J. Recent Innov. Trends Comput. Commun.* 4, 75–78. doi: 10.17762/ijritcc.v4i3.1837

Giri, R., Bhatt, A., and Bhatt, A. (2016). Frequent pattern mining algorithms analysis. *Int. J. Comput. Appl.* 145, 33–36. doi: 10.5120/IJCA2016910763

Gitanjali, J., Ranichandra, C., and Pounambal, M. (2014). Apriori algorithm based medical data mining for frequent disease identification. *IPASJ Int. J. Inform. Technol.* 2, 1–5.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Rec.* 29, 1–12. doi: 10.1145/335191.335372

Ingle, M. G., and Suryavanshi, N. (2015). Association rule mining using improved apriori algorithm. *Int. J. Comput. Appl.* 112, 37–42.

Iqbal, A. (2017). Effect of food on causation and prevention of gastric cancer. *J. Cancer Prev. Curr. Res.* 8:00289. doi: 10.15406/jcpcr.2017.08.00289

Ishita, R., and Rathod, A. (2016). ECLAT with large database parallel algorithm and improve its efficiency. *Int. J. Comput. Appl.* 143, 33–37. doi: 10.5120/ijca2016910462

Jia, Y., Xia, G., Fan, H., Zhang, Q., and Li, X. (2012). "An improved apriori algorithm based on association analysis," in *Proceedings of the 3rd IEEE International Conference (ICNDC)*, Washington, DC, 208–211.

Karthiyayini, R., and Jayaprakash, J. (2015). Association technique on prediction of chronic diseases using apriori algorithm. *Int. J. Innov. Res. Sci. Eng. Technol.* 4, 255–259.

Kaur, J., and Madan, N. (2015). Review of Apriori Algorithm and its Recent Improvements. *Int. J. Emerg. Technol. Comput. Appl. Sci.* 12, 150–152.

Kavitha, M., and Selvi, S. T. (2016). Comparative study on apriori algorithm and Fp growth algorithm with pros and cons. *Int. J. Comput. Sci. Trends Technol.* 4, 161–164.

Mandave, P., Mane, M., and Patil, S. (2013). Data mining using Association rule based on APRIORI algorithm and improved approach with illustration. *Int. J. Latest Trends Eng. Technol.* 3, 107–113.

Park, J. S., Chen, M. S., and Yu, P. S. (1995). An effective hash-based algorithm for mining association rules. *ACM SIGMOD Rec.* 24, 175–186. doi: 10.1145/568271.223813

Patil, S. D., and Deshmukh, R. (2016). Review and analysis of apriori algorithm for association rule mining. *Int. J. Latest Trends Eng. Technol.* 6, 104–112.

Poorani, S., Balasubramanie, P., and Kumar, D. V. (2018). Apriori algorithm for identifying the association rules between clinical traits of asthma. *Int. J. Pure Appl. Math.* 118, 4695–4706.

Rafea, M., ELkafrawy, P., Nasef, M., Elnemr, R., and Jamal, A. T. (2019). Applying machine learning of erythrocytes dynamic antigens store in medicine. *Front. Mol. Biosci.* 6:19. doi: 10.3389/fmolb.2019.00019

Rafea, M., and Souchelnytskyi, S. (2012). *Rediscovering Red Blood Cells: Revealing Their Dynamic Antigens Store and Its Role in Health and Disease, in Blood Cell-an Overview of Studies in Hematology. MoschandreouTE*. London: IntechOpen.

Rajeswari, K. (2015). Improved apriori algorithm–a comparative study using different objective measures. *Int. J. Comput. Sci. Inform. Technol.* 6, 3185–3191.

Raval, M. R., Rajput, I. J., and Gupta, V. (2013). Survey on several improved Apriori algorithms. *IOSR J. Comput. Eng.* 9, 57–61. doi: 10.9790/0661-0945761

Said, I., Haruna, A., and Garko, A. (2015). Association rule mining on medical data to predict heart disease. *Int. J. Sci. Technol. Manag.* 4, 26–35.

Shah, A., and Patel, P. (2015). A Collaborative approach of frequent item set mining: a survey. *Int. J. Comput. Appl.* 107, 34–36. doi: 10.5120/18775-0088

Shukla, R., and Solanki, A. K. (2015). Performance evaluation for frequent pattern mining algorithm. *Int. J. Eng. Res. Gen. Sci.* 3, 910–915.

Toivonen, H. (1996). "September. Sampling large databases for association rules," in *VLDB '96: Proceedings of the 22th International Conference on Very Large Data Bases*, Hong Kong, 134–145.

Ukawa, S., Tamakoshi, A., Mori, M., Ikehara, S., Shirakawa, T., Yatsuya, H., et al. (2018). Association between average daily television viewing time and the incidence of ovarian cancer: findings from the Japan Collaborative Cohort Study. *Cancer Causes Control* 29, 213–219. doi: 10.1007/s10552-018-1001-8

Vyas, K., and Sherasiya, S. (2016). Modified apriori algorithm using hash based technique. *Int. J. Adv. Res. Innov. Ideas Educ.* 2, 1229–1234.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.* 12, 372–390. doi: 10.1109/69.846291

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

Copyright © 2020 Elnemr, Nasef, Elkafrawy, Rafea and Jamal. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 04 December 2020
doi: 10.3389/fmolb.2020.564005

Identification of Tamoxifen-Resistant Breast Cancer Cell Lines and Drug Response Signature

Qingzhou Guan[1†], Xuekun Song[2†], Zhenzhen Zhang[1], Yizhi Zhang[3], Yating Chen[3] and Jing Li[3]*

[1] Co-construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan & Education Ministry of P.R. China, Academy of Chinese Medical Sciences, Henan University of Chinese Medicine, Zhengzhou, China, [2] College of Information Technology, Henan University of Chinese Medicine, Zhengzhou, China, [3] Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China

Breast cancer cell lines are frequently used to elucidate the molecular mechanisms of the disease. However, a large proportion of cell lines are affected by problems such as mislabeling and cross-contamination. Therefore, it is of great clinical significance to select optimal breast cancer cell lines models. Using tamoxifen survival-related genes from breast cancer tissues as the gold standard, we selected the optimal cell line model to represent the characteristics of clinical tissue samples. Moreover, using relative expression orderings of gene pairs, we developed a gene pair signature that could predict tamoxifen therapy outcomes. Based on 235 consistently identified survival-related genes from datasets GSE17705 and GSE6532, we found that only the differentially expressed genes (DEGs) from the cell line dataset GSE26459 were significantly reproducible in tissue samples (binomial test, p = 2.13E-07). Finally, using the consistent DEGs from cell line dataset GSE26459 and tissue samples, we used the transcriptional qualitative feature to develop a two-gene pair (*TOP2A*, *SLC7A5*; *NMU*, *PDSS1*) for predicting clinical tamoxifen resistance in the training data (logrank p = 1.98E-07); this signature was verified using an independent dataset (logrank p = 0.009909). Our results indicate that the cell line model from dataset GSE26459 provides a good representation of the characteristics of clinical tissue samples; thus, it will be a good choice for the selection of drug-resistant and drug-sensitive breast cancer cell lines in the future. Moreover, our signature could predict tamoxifen treatment outcomes in breast cancer patients.

Keywords: breast cancer, tamoxifen, cell line, resistant, sensitive

OPEN ACCESS

Edited by:
Cheng Zhang,
KTH Royal Institute of Technology, Sweden

Reviewed by:
Khyati Shah,
University of California, San Francisco, United States
Ankita Thakkar,
Burke Medical Research Institute, United States

****Correspondence:***
Jing Li
haerbinlisa@hotmail.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *20 May 2020*
Accepted: *15 October 2020*
Published: *04 December 2020*

Citation:
Guan Q, Song X, Zhang Z, Zhang Y, Chen Y and Li J (2020) Identification of Tamoxifen-Resistant Breast Cancer Cell Lines and Drug Response Signature. Front. Mol. Biosci. 7:564005. doi: 10.3389/fmolb.2020.564005

INTRODUCTION

The overall recurrence rate of estrogen receptor positive (ER+) early breast cancer can be reduced by adjuvant treatment with tamoxifen. However, approximately 30–40% of ER + breast cancer patients receiving adjuvant tamoxifen therapy still would relapse or progress to deadly advanced metastatic stages within 15 years follow-up; this is largely attributed to tamoxifen

Abbreviations: DEGs, differentially expressed genes; GEO, Gene Expression Omnibus; ER + , estrogen receptor positive; KEGG, Kyoto Encyclopedia of Genes and Genomes; REO, relative expression ordering; RFS, relapse-free survival; SAM, significance analysis of microarrays.

resistance (Ye et al., 2019). Therefore, it is of great clinical significance to identify the efficacy of tamoxifen in ER + breast cancer patients. Cell lines are a common modeling tool in cancer research (Domcke et al., 2013); they can help us to better understand the biological processes and molecular mechanisms of cancer and aid in the development of anticancer drugs (Kong and Yamori, 2012; Knudsen et al., 2014). However, whether cell line models could adequately reflect the characteristics of clinical tissue samples is controversial (American Type Culture Collection Standards Development Organization Workgroup ASN-0002, 2010; Liedtke et al., 2010; Bayer et al., 2013; Capes-Davis et al., 2019; Wass et al., 2019). It is well known that tumor cell lines might lose some of their tumor-related characteristics owing to the culture environment (Masters, 2000). Cross-contamination (International Cell Line Authentication Committee, 2014) and misidentification (American Type Culture Collection Standards Development Organization Workgroup ASN-0002, 2010) of cell lines exacerbates such issues. Moreover, there is no unified gold standard for the identification of drug-resistant cell lines, which also results in some cell lines poorly reflecting the characteristics of clinical tissue samples (Liedtke et al., 2010). Thus, it is of great value to find resistant/sensitive cell line models that are more representative of clinical tissue samples.

Considering tamoxifen survival-related genes from breast cancer tissue samples as the gold standard, we screened for the optimal cell line model. In the survival-related analysis of tissue samples, we assumed that genes that were positively (negatively) correlated with survival risk in tissue samples were comparable with genes that are upregulated (downregulated) in resistant compared with sensitive cell lines. In this study, through evaluating the consistency of prognosis-related genes in tissue samples from patients undergoing tamoxifen treatment with drug-resistance genes in cell lines, we selected the optimal cell line model to represent the characteristics of clinical tissue samples; the consistent genes between tissues and cell lines were identified as clinical drug-resistance-related genes.

Moreover, the relative expression orderings (REOs) of gene pairs within individual samples, also called qualitative transcriptional characteristics, are robust against experimental batch effects and can be directly applied to samples at an individual level (Eddy et al., 2010; Guan et al., 2019). The robustness property of the qualitative transcriptional characteristics enables integration of multiple datasets from different sources to develop disease signatures or classifiers, which improves the probability of finding robust signatures (Xu et al., 2008; Guan et al., 2019). Thus, based on qualitative transcriptional characteristics and the clinical drug-resistance-related genes that we identified, we developed a tamoxifen-resistance signature for ER + breast cancer and verified it in independent data.

MATERIALS AND METHODS

Data and Preprocessing

Breast cancer gene expression data and corresponding clinical information were downloaded from the GEO database (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/). Relapse-free survival (RFS) time was defined as the interval between the first day of surgery and the date of death from any cause or of recurrence (local and/or distant) (Punt et al., 2007; Merok et al., 2013). Breast cancer tissue samples from ER+ patients who had received post-operative tamoxifen treatment were selected from the seven datasets, as described in **Table 1**. Nine gene expression datasets for breast cancer tamoxifen-resistant/sensitive cell lines were also downloaded from the GEO database, as shown in **Table 1**.

For the array data measured by Affymetrix platform, raw mRNA expression data (.CEL files) were downloaded, and the Robust Multi-array Average algorithm was used for normalization with Affy package in R software (Bolstad et al., 2003; Irizarry et al., 2003). For sequence-based data, the processed data were directly downloaded.

Identification of Survival-related Genes in Tissue

The Cox proportional hazard model was used to study the relationships between gene expression levels and survival (Kreike et al., 2010). For the coefficient β obtained from the Cox model, if $\beta > 0$ for a certain gene, this gene was considered to be positively correlated with survival risk and was comparable with the upregulated gene between resistant and sensitive cell lines. Similarly, if $\beta < 0$, the gene was comparable with the downregulated gene between resistant and sensitive cell lines.

Identification of Differentially Expressed Genes (DEGs) in Cell Lines

In this study, the SAM (significance analysis of microarrays) algorithm (Tusher et al., 2001) was used to identify DEGs between resistant and sensitive cell lines.

Consistency Evaluation Between Tissues and Cell Lines

In this study, we hypothesized that genes positively (negatively) associated with survival in tissues corresponded to those genes upregulated (downregulated) between resistant and sensitive cell lines.

The consistency ratio, which is the number of overlapping and consistent DEGs/number of overlapping DEGs, was used to evaluate the similarity between tissues and cell lines. The significance was evaluated by the binomial distribution test as follows:

$$p = 1 - \sum_{i=0}^{k-1} \binom{n}{i} 0.5^i (1 - 0.5)^{n-i}$$

where n denotes the number of overlapping DEGs between tissue and cell line, and k denotes the number of those overlapping DEGs with the same dysregulation direction.

Then, the p-values were adjusted using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

TABLE 1 | Data used in this study.

Tissue					
GEO Acc	**Platform**	**ER+ Sample**	**Endpoint**		
GSE17705	Affymetrix GPL96	298	RFS		
GSE6532	Affymetrix GPL96	176	RFS		
GSE12093	Affymetrix GPL96	136	RFS		
GSE4922	Affymetrix GPL96	66	RFS		
GSE2990	Affymetrix GPL96	54	RFS		
GSE42568	Affymetrix GPL570	67	RFS		
GSE9195	Affymetrix GPL570	77	RFS		
Cell line					
GEO Acc	**Platform**	**Sensitive**	**Resistant**	**Sample (R vs S)**	**Method**
GSE27473	Affymetrix GPL570	MCF7	MCF7 silenced ER	3:3	RNA silencing
GSE12708	Affymetrix GPL96	SUM44	SUM44/LCCTam	3:3	Drug pressure
GSE26459	Affymetrix GPL570	B7	G11OH-T	3:3	MCF7 subclones
GSE8562	Affymetrix GPL96	MCF7	MCF7/XBP1	3:3	XBP1 overexpression
GSE14986	Affymetrix GPL570	MCF7	T8, T17, T29, T52	4:3	Drug pressure
GSE21618	Affymetrix GPL570	WT	tamR	20:11	Drug pressure
GSE67916	Affymetrix GPL570	MCF7	MCF-7/TAMR	10:8	Drug pressure
#GSE118713	Illumina GPL16791	MCF7	MCF-7/TAMR	3:3	Drug pressure
#GSE125738	HiSeq GPL20795	T47D	T47D-TR	3:3	Drug pressure

RFS: relapse-free survival; ER: estrogen receptor. Sample (R vs S) denotes the number of the resistant and sensitive cell line sample from the corresponding dataset; Method denotes the production process for tamoxifen-resistant breast cancer cell lines. #High-throughput sequencing data.

KEGG Pathway Enrichment

The hypergeometric distribution model was used to determine the significance of KEGG (Kanehisa and Goto, 2000) (Kyoto Encyclopedia of Genes and Genomes) pathways enriched with the genes of interest using the following statistical model:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

where N denotes the number of background genes, n denotes the number of genes of interest, m denotes the number of genes in a given pathway, and k denotes the number of genes of interest in that pathway.

Identification of REO-based Tamoxifen-resistance Signature

Taking the consistent DEGs between tissues and cell lines as candidate genes, we used the Cox model and C-index analysis (Harrell et al., 1984) to develop a tamoxifen-resistance signature. The detailed process was described as follows.

Step 1: Selecting Survival-related Gene Pairs

(1) For the n candidate DEGs, pairwise comparisons were performed for all genes (generating a total of C_n^2 gene pairs), and this gene pair set was defined as Set 1. (2) From all gene pairs (G_i, G_j) in Set 1, the Cox model was used to select those that were significantly correlated with RFS of the tamoxifen-treated breast cancer patients. The set of significantly correlated gene pairs (FDR < 10%) was defined as Set 2.

Step 2: Optimizing the Gene Pair Signature

First, we enumerated all the gene pair combinations in Set 2. For each gene pair combination in a sample, if at least half of the gene pairs in the combination were consistent with tamoxifen sensitivity, the sample was identified as low risk; otherwise, it was considered high risk. Then, we calculated the C-index value for each gene pair combination, and selected the combination with maximum C-index as our tamoxifen-resistance signature (Set 3).

RESULTS

Identification and Evaluation of DEGs in Cell Lines

A flowchart of the analysis procedure is shown in **Figure 1**. We identified the DEGs between tamoxifen-resistant and tamoxifen-sensitive cell line samples within each of the nine datasets using the SAM method (FDR < 20%). We also evaluated the consistency of DEGs among different datasets (a total of $C_9^2 = 36$ combinations). Among the 36 combinations, only 16 showed significant consistency ($p < 0.05$), as described in **Table 2**. These results indicate that there is greater heterogeneity among cell lines from different sources.

Identification of Tamoxifen Survival-related Genes in Tissues

Based on the univariate Cox regression model with FDR < 20%, 893 and 968 tamoxifen survival-related genes were identified

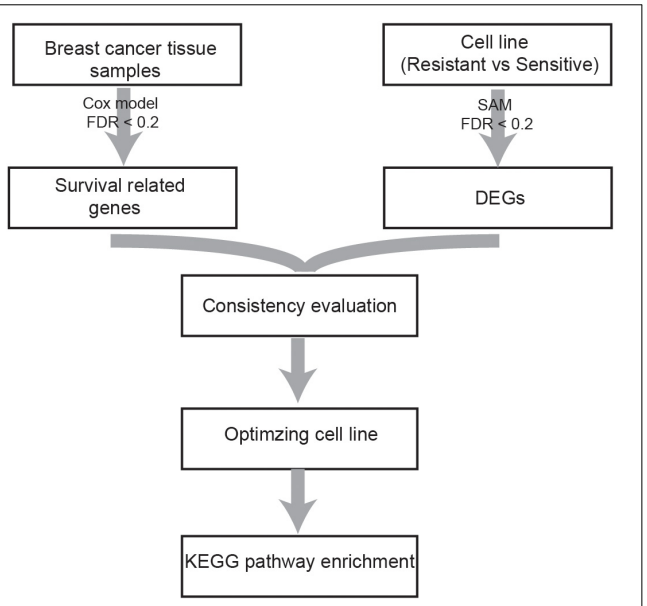


FIGURE 1 | Flowchart of the analysis procedure.

in datasets GSE17705 and GSE6532, respectively; 235 genes were common to the two groups, all of which had the same dysregulation direction (which could not occur by chance; binomial test, $p < 1.0E-16$), further verifying the reliability of the results. These 235 genes were considered to be breast cancer tissue candidate genes.

Owing to the heterogeneity among cell lines, we evaluated the consistency between tissue candidate genes and DEGs from different cell line datasets (resistant vs sensitive) to select an optimal cell line model that could well represent the characteristics of clinical tissue samples. We found that only the DEGs from dataset GSE26459 were well reproduced among tissue candidate genes; the consistency ratio was above 73%, indicating that this did not occur by chance (binomial test, $p = 2.13E-07$). The DEGs from the other cell line datasets were not well reproduced among the tissue candidate genes (**Table 3**). These results demonstrate that the cell line data from dataset GSE26459 could well represent the characteristics of clinical breast cancer tissue samples.

KEGG Pathway Enrichment

KEGG pathway enrichment analysis was performed for the 235 tissue candidate genes from datasets GSE17705 and GSE6532 using a threshold of FDR < 0.2, and for the DEGs from cell line dataset GSE26459 using the same threshold (**Table 4**). There was no pathway commonly enriched between tissues and the cell line, possibly owing to the low statistical power (Zou et al., 2011) or to partial differences between resistant and sensitive cell lines induced by tamoxifen treatment (Dancik et al., 2011). Thus, taking the pathways enriched in tissues as the gold standard, we obtained the p-values of these pathways in dataset GSE26459 (**Table 4**). With $p < 0.2$, the cell cycle, p53 signaling pathway, oocyte meiosis, and progesterone-mediated oocyte maturation were recurring themes in the pathway analysis for both tissues and cell lines. These pathways have been reported to be correlated with tamoxifen resistance.

Studies have shown that tamoxifen could affect the cell cycle of human breast cancer cell lines, the major sensitivity to tamoxifen in terms of both inhibition of cell cycle progression and drug cytotoxicity occurring particularly in the G0-G1 stage (Taylor et al., 1983). Tamoxifen could also affect the mitosis of oocytes and lead to premature centromere separation (London and Mailhes, 2001). The ***PTEN*** protein, encoded by the gene, in the p53 signaling pathway has been shown to be associated with tamoxifen resistance (Shoman et al., 2005). Similarly, the ***PGR*** protein in the progesterone-mediated oocyte maturation signaling pathway has been shown to be associated with tamoxifen response (Elledge et al., 2000). In summary, the pathways found to be enriched in tissues and also in cell line dataset GSE26459 ($p < 0.2$) were correlated with tamoxifen resistance, further demonstrating that the cell line model from dataset GSE26459 could represent the characteristics of clinical tissue samples.

Moreover, with FDR < 20%, the DEGs from cell line dataset GSE26459 were enriched in 31 pathways, compared with only seven pathways for the genes from tissue samples. However, as shown in **Table 4**, many of the pathways enriched for the cell lines from dataset GSE26459 are associated with tamoxifen treatment. For example, the prolactin signaling pathway and neurotrophin signaling pathway are related to side effects of tamoxifen (Lamberts et al., 1982; El-Ashmawy and Khalil, 2014), indicating that some of the differences between resistant and sensitive cell lines were due to tamoxifen treatment.

Identification of Tamoxifen Response Signature

First, we considered the 84 consistent DEGs between tissues and cell line dataset GSE26459 to be clinical tamoxifen-resistance-related genes. In the training dataset GSE12093, pairwise comparisons were performed for all clinical tamoxifen-resistance-related genes, and all the gene pairs were analyzed with a univariate Cox regression model. With FDR < 10%, 20 gene pairs were identified that were significantly associated with RFS. Then, among the 20 gene pairs, we enumerated all the gene pair combinations to calculate their C-index values, and selected the gene combination with the maximum C-index as the tamoxifen response signature. Finally, two gene pairs (***TOP2A***, ***SLC7A5***; ***NMU***, ***PDSS1***) were identified. Based on our signature and the majority vote rule, the training dataset samples could be divided into high- and low-risk samples, which had significantly different RFS (hazard ratio [HR] = 9.509, logrank $p = 1.98E-07$). Our signature was also verified in an independent validation test using combined data from datasets GSE4922 and GSE2990 (HR = 2.191, logrank $p = 0.009909$), as shown in **Figure 2A**. Moreover, we searched public databases again for breast cancer tissue samples treated only with post-operative tamoxifen, for which associated RFS information was available, to further verify the performance of our signature. Finally, two new independent datasets were obtained. For the breast cancer tissue samples

TABLE 2 | Consistency evaluation of DEGs from different cell line datasets.

GEO Acc	Cell line*	Def_gene	Com_gene	Con_gene	Ratio	*P*
GSE27473	si-ER MCF7: MCF7	15937	10795	6147	0.5694	<1.00E-16
GSE14986	T8/17/29/52: MCF7	13391				
GSE27473	si-ER MCF7: MCF7	15937	12580	7427	0.5904	<1.00E-16
GSE21618	TamR: WT	15481				
GSE27473	si-ER MCF7: MCF7	15937	9675	5424	0.5606	<1.00E-16
GSE67916	MCF-7/TAMR:MCF-7	12227				
GSE27473	si-ER MCF7: MCF7	15937	8074	4450	0.5512	<1.00E-16
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE14986	T8/17/29/52: MCF7	13391	10494	7391	0.7043	<1.00E-16
GSE21618	TamR: WT	15481				
GSE14986	T8/17/29/52: MCF7	13391	8125	5396	0.6641	<1.00E-16
GSE67916	MCF-7/TAMR:MCF-7	12227				
GSE14986	T8/17/29/52: MCF7	13391	6534	4139	0.6335	<1.00E-16
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE14986	T8/17/29/52: MCF7	13391	6505	4042	0.6214	<1.00E-16
GSE125738	T47D-TR:T47D	10685				
GSE21618	TamR: WT	15481	9331	5386	0.5772	<1.00E-16
GSE67916	MCF-7/TAMR:MCF-7	12227				
GSE26459	G11OH-T: B7	6375	5525	3192	0.5777	<1.00E-16
GSE27473	si-ER MCF7: MCF7	15937				
GSE21618	TamR: WT	15481	7729	4189	0.5420	8.22E-14
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE118713	MCF-7/TAMR:MCF-7	10023	5808	3161	0.5442	8.16E-12
GSE125738	T47D-TR:T47D	10685				
GSE21618	TamR: WT	15481	7597	4061	0.5346	9.04E-10
GSE125738	T47D-TR:T47D	10685				
GSE67916	MCF-7/TAMR:MCF-7	12227	5824	3212	0.5515	2.00E-15
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE26459	G11OH-T: B7	6375	3767	2044	0.5426	9.10E-08
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE27473	si-ER MCF7: MCF7	15937	7991	4163	0.5210	9.32E-05
GSE125738	T47D-TR:T47D	10685				
GSE26459	G11OH-T: B7	6375	1163	521	0.4480	1.00E + 00
GSE12708	SUM44/LCCTam: SUM44	2538				
GSE26459	G11OH-T: B7	6375	52	21	0.4038	9.37E-01
GSE8562	MCF7/XBP1: MCF7	97				
GSE26459	G11OH-T: B7	6375	4623	2084	0.4508	1.00E + 00
GSE14986	T8/17/29/52: MCF7	13391				
GSE26459	G11OH-T: B7	6375	5262	2643	0.5023	3.76E-01
GSE21618	TamR: WT	15481				
GSE26459	G11OH-T: B7	6375	4090	1946	0.4758	9.99E-01
GSE67916	MCF-7/TAMR:MCF-7	12227				
GSE26459	G11OH-T: B7	6375	3750	1321	0.3523	1.00E + 00
GSE125738	T47D-TR:T47D	10685				
GSE27473	si-ER MCF7: MCF7	15937	2264	1056	0.4664	9.99E-01
GSE12708	SUM44/LCCTam: SUM44	2538				
GSE27473	si-ER MCF7: MCF7	15937	89	33	0.3708	9.95E-01
GSE8562	MCF7/XBP1: MCF7	97				
GSE12708	SUM44/LCCTam: SUM44	2538	23	12	0.5217	5.00E-01
GSE8562	MCF7/XBP1: MCF7	97				
GSE12708	SUM44/LCCTam: SUM44	2538	1885	702	0.3724	1.00E + 00
GSE14986	T8/17/29/52: MCF7	13391				
GSE12708	SUM44/LCCTam: SUM44	2538	2134	920	0.4311	1.00E + 00

(Continued)

TABLE 2 | Continued

GEO Acc	Cell line*	Def_gene	Com_gene	Con_gene	Ratio	*P*
GSE21618	TamR: WT	15481				
GSE12708	SUM44/LCCTam: SUM44	2538	1676	862	0.5143	1.25E-01
GSE67916	MCF-7/TAMR:MCF-7	12227				
GSE12708	SUM44/LCCTam: SUM44	2538	1588	625	0.3936	1.00E + 00
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE12708	SUM44/LCCTam: SUM44	2538	1630	840	0.5153	1.12E-01
GSE125738	T47D-TR:T47D	10685				
GSE8562	MCF7/XBP1: MCF7	97	80	42	0.5250	3.69E-01
GSE14986	T8/17/29/52: MCF7	13391				
GSE8562	MCF7/XBP1: MCF7	97	84	46	0.5476	2.23E-01
GSE21618	TamR: WT	15481				
GSE8562	MCF7/XBP1: MCF7	97	57	30	0.5263	3.96E-01
GSE67916	MCF-7/TAMR:MCF-7	12227				
GSE8562	MCF7/XBP1: MCF7	97	63	25	0.3968	9.62E-01
GSE118713	MCF-7/TAMR:MCF-7	10023				
GSE8562	MCF7/XBP1: MCF7	97	63	25	0.3968	9.62E-01
GSE125738	T47D-TR:T47D	10685				
GSE67916	MCF-7/TAMR:MCF-7	12227	5751	2910	0.5060	1.85E-01
GSE125738	T47D-TR:T47D	10685				

**Resistant and sensitive cell line samples from the corresponding dataset. Taking dataset GSE14986 as an example, among T8/17/29/52: MCF7, T8/17/29/52 denote resistant cell lines, MCF7 denotes sensitive cell line; Def_gene denotes the number of DEGs in the corresponding dataset; Com_gene denotes the number of overlapped DEGs between two datasets; Con_gene denotes the number of overlapping DEGs with the same dysregulation between two datasets; Ratio denotes the consistency ratio of DEGs.*

from dataset GSE42568, 37 samples were identified as high risk, and 30 were identified as low risk (HR = 1.804, logrank p = 0.2), as shown in **Figure 2B**. For the breast cancer tissue samples from dataset GSE9195, 41 samples were identified as high risk and 36 as low risk (HR = 1.516, logrank p = 0.5), as shown in **Figure 2C**. Although the difference between the groups was not significant according to statistical tests, there was a clear trend indicating a difference in RFS between the high- and low-risk groups identified by our signature (**Figure 2B-C**). Moreover, we combined the above two datasets to further verify the performance of our signature. In the combined data from datasets GSE42568 and GSE9195, 78 samples were identified as high risk and 66 samples were identified as low risk (HR = 1.7, logrank p = 0.1), as shown in **Figure 2D**. In summary, the results indicate that our signature (consisting of two gene pairs) can predict drug efficacy to some extent.

DISCUSSION

Cell line models are widely used in various fields of medical research, especially in basic cancer research and drug discovery (Masters, 2000; Mirabelli et al., 2019). Despite the successful application of cell lines in basic research, their use as model systems remains controversial (Masters, 2002; Sandberg and Ernberg, 2005; Peng et al., 2018; Hallas-Potts et al., 2019). Owing to issues such as cross-contamination, mislabeling, or the identification of drug resistance, some cell line models do not adequately represent the characteristics of clinical tissues. In this study, based on evaluation of the consistency of DEGs between tissues and cell lines, we selected the optimal cell line model to represent the characteristics of clinical tissue samples; this was further verified by pathway analysis. Our analysis method is also suitable for other types of cell line modes.

The tamoxifen survival-related genes identified in tissue samples from different datasets were significantly consistent, suggesting that the results were reliable. However, the DEGs found in tamoxifen-resistant and tamoxifen-sensitive cell lines from different sources were less reproducible, indicating that cell line models from different sources show more heterogeneity. Therefore, it will be of great clinical significance to screen

TABLE 3 | Consistency evaluation between tissues and cell lines.

GEO Acc	Def_gene	Com_gene	Con_gene	Ratio	*P*
GSE26459	6375	114	84	0.7368	2.13E-07
GSE27473	15937	211	93	0.4408	9.63E-01
GSE12708	2538	46	15	0.3261	9.94E-01
GSE8562	97	5	3	0.6000	5.00E-01
GSE14986	13391	178	55	0.3090	1.00E + 00
GSE21618	15481	207	82	0.3961	9.99E-01
GSE67916	12227	162	61	0.3765	9.99E-01
GSE118713	10023	159	63	0.3962	9.97E-01
GSE125738	10685	159	32	0.2013	1.00E + 00

Def_gene denotes the number of DEGs in the corresponding dataset; Com_gene denotes the number of overlapping DEGs between the 235 tissue candidate genes and the corresponding cell line dataset; Con_gene denotes the number of overlapping DEGs with the same dysregulation between two datasets; Ratio denotes the consistency ratio of DEGs.

TABLE 4 | KEGG pathway enrichment of tissue and cell line.

Tissue			Cell line		
Pathway num	**Pathway name[a]**	***P****	**Pathway num**	**Pathway name[b]**	**FDR**
hsa04110	Cell cycle	0.0270	hsa03013	RNA transport	4.62E-08
hsa04115	p53 signaling pathway	0.0226	hsa03010	Ribosome	1.14E-05
hsa04114	Oocyte meiosis	0.0726	hsa00970	Aminoacyl-tRNA biosynthesis	1.82E-05
hsa04914	Progesterone-mediated oocyte maturation	0.1176	hsa03008	Ribosome biogenesis in eukaryotes	1.64E-04
hsa03440	Homologous recombination	0.3907	hsa03040	Spliceosome	7.40E-04
hsa04672	Intestinal immune network for IgA production	0.8288	hsa03410	Base excision repair	1.98E-03
hsa04060	Cytokine-cytokine receptor interaction	0.9977	hsa00620	Pyruvate metabolism	9.57E-03
			hsa01230	Biosynthesis of amino acids	0.0119
			hsa01100	Metabolic pathways	0.0194
			hsa01212	Fatty acid metabolism	0.0194
			hsa01200	Carbon metabolism	0.0214
			hsa00510	N-Glycan biosynthesis	0.0244
			hsa00531	Glycosaminoglycan degradation	0.0244
			hsa04360	Axon guidance	0.0244
			hsa04612	Antigen processing and presentation	0.0244
			hsa04917	Prolactin signaling pathway	0.0257
			hsa00511	Other glycan degradation	0.0272
			hsa04144	Endocytosis	0.0272
			hsa03018	RNA degradation	0.0300
			hsa04142	Lysosome	0.0322
			hsa04330	Notch signaling pathway	0.0513
			hsa01040	Biosynthesis of unsaturated fatty acids	0.0573
			hsa04722	Neurotrophin signaling pathway	0.0754
			hsa04910	Insulin signaling pathway	0.0872
			hsa01210	2-Oxocarboxylic acid metabolism	0.0945
			hsa04141	Protein processing in endoplasmic reticulum	0.1101
			hsa00280	Valine, leucine and isoleucine degradation	0.1121
			hsa04120	Ubiquitin mediated proteolysis	0.1121
			hsa00270	Cysteine and methionine metabolism	0.1319
			hsa00020	Citrate cycle (TCA cycle)	0.1527
			hsa03050	Proteasome	0.1848

Tissue: [a]KEGG pathway enriched for survival-related genes in tissues (FDR < 0.2); P denotes the p-value for a KEGG pathway, enriched for tissues, in the cell line dataset GSE26459. Cell line:[b]KEGG pathway enriched by DEGs between resistant and sensitive cell lines in dataset GSE26459 (FDR < 0.2).

for drug-resistant and drug-sensitive cell line models that better represent the characteristics of clinical tissue samples. According to our results, the DEGs from cell line dataset GSE26459 were reproducible in tissue samples, indicating that the cell line model from this dataset was representative of the characteristics of clinical tissue samples. Tissue samples were obtained by surgical resection before tamoxifen therapy. Thus, the survival-related genes obtained from tissues were intrinsic to the patient and not induced by tamoxifen treatment. The resistant and sensitive cell lines from dataset GSE26459 were selected from MCF subclones (Gonzalez-Malerva et al., 2011); this might partly explain why the cell lines from GSE26459 could represent the characteristics of clinical tissue samples. The pathways enriched in tissues and in cell line dataset GSE26459 ($p < 0.2$) have been reported to be associated with tamoxifen resistance (Lamberts et al., 1982; El-Ashmawy and Khalil, 2014). Moreover, the clinical tamoxifen-resistance gene-pair signature we developed was verified in independent validation dataset, which indicates that our signature has some power to predict response to tamoxifen therapy, and further demonstrates that we have selected appropriate tamoxifen-resistant and tamoxifen-sensitive cell line models.

Although the cell line models identified by our analytical method could well reflect the information of clinical tissue samples, there were some limitations. As patients with breast cancer usually have good prognosis, the endpoint of their follow-up is usually survival or recurrence time. Furthermore, as well as the effects of drugs, many factors including mood, marital status, and economic status could affect the survival of patients. The above factors might cause that some of the survival-related genes that we have identified are not involved in tamoxifen resistance. In future work, use of more tissue sample data or an improved algorithm should be considered. Moreover, as DNA methylation patterns, genomic changes, etc., might also predict sensitivity to drugs, the use of other types of data (such as microRNAs,

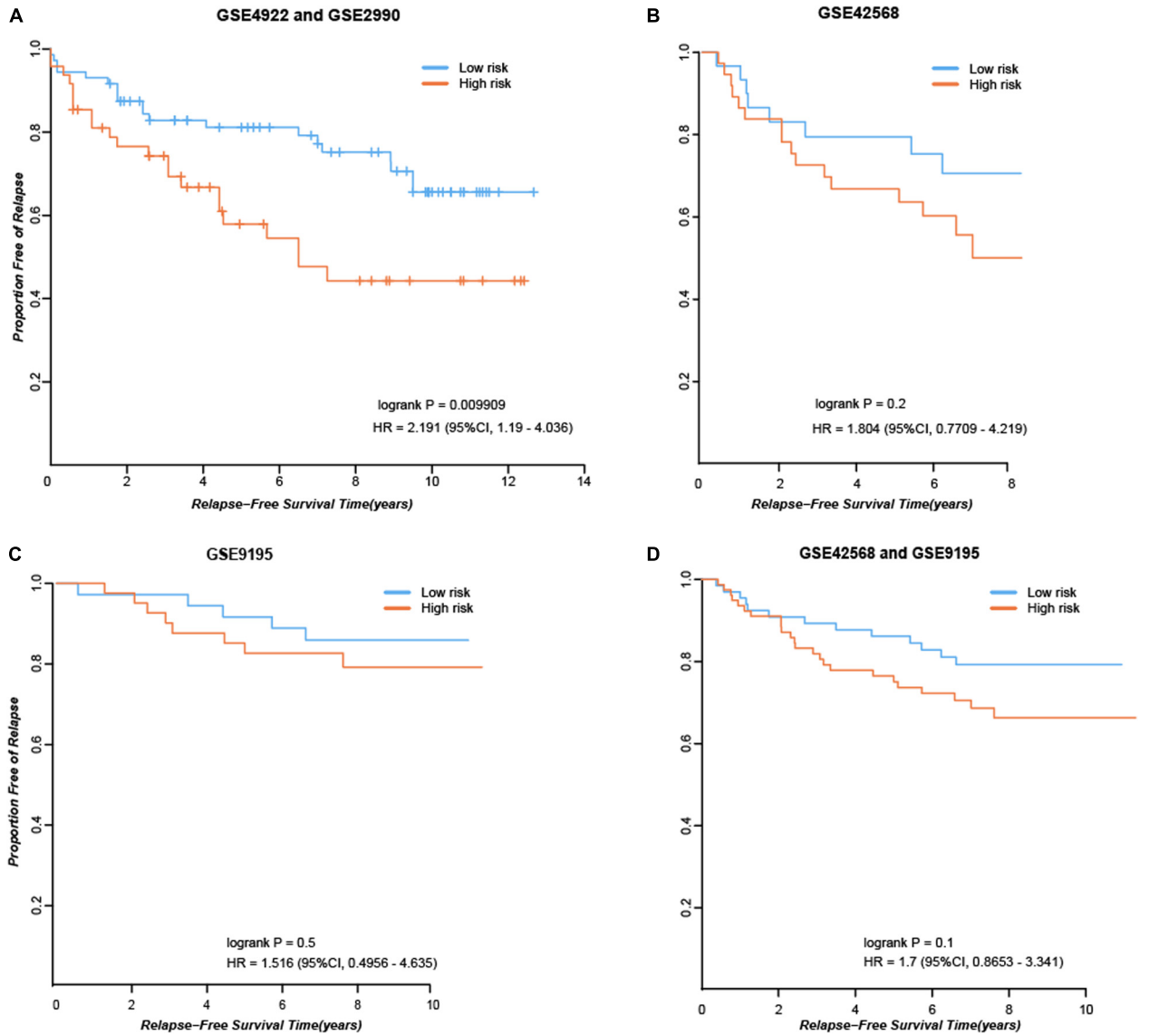


FIGURE 2 | Performance of our signature in independent dataset. **(A)** RFS curves in the combined data from datasets GSE4922 and GSE2990. **(B)** RFS curves in the dataset GSE42568. **(C)** RFS curves in the dataset GSE9195. **(D)** RFS curves in the combined data from datasets GSE42568 and GSE9195.

DNA methylations, and genomic changes) in cell line model optimization deserve consideration in future studies.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

QZG and XKS conceived the study, analyzed the data, produced the figures, performed the statistical analysis, and drafted the manuscript. ZZZ participated in the revision of the manuscript. YZZ and YTC searched the data and participated in the statistical analysis. JL conceived the study and participated in its design and coordination, helped to draft the manuscript, and supervised the work. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the China National Postdoctoral Program for Innovative Talents (BX20200115), National Natural Science Foundation of China (Grant numbers: 61602119 and 61702164), the Joint Technology Innovation Fund of Fujian Province (Grant number: 2017Y9109), Scientific and Technological Project of Henan Province

(Grant numbers: 162102310461 and 172102310535), Natural Science Foundation of Henan Province (Grant number: 162300410184), and Scientific Research Project of Zhengzhou (Grant number: 153PKJGG128).

ACKNOWLEDGMENTS

I would like to especially thank my doctoral mentor Zheng Guo for help in my scientific research and life.

REFERENCES

American Type Culture Collection Standards Development Organization Workgroup ASN-0002 (2010). Cell line misidentification: the beginning of the end. *Nat. Rev. Cancer* 10, 441–448. doi: 10.1038/nrc2852

Bayer, I., Groth, P., and Schneckener, S. (2013). Prediction errors in learning drug response from gene expression data - influence of labeling, sample size, and machine learning algorithm. *PLoS One* 8:e70294. doi: 10.1371/journal.pone.0070294

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery(Rate): a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Capes-Davis, A., Bairoch, A., Barrett, T., Burnett, E. C., Dirks, W. G., Hall, E. M., et al. (2019). Cell lines as biological models: practical steps for more reliable research. *Chem. Res. Toxicol.* 32, 1733–1736. doi: 10.1021/acs.chemrestox.9b00215

Dancik, G. M., Ru, Y., Owens, C. R., and Theodorescu, D. (2011). A framework to select clinically relevant cancer cell lines for investigation by establishing their molecular similarity with primary human cancers. *Cancer Res.* 71, 7398–7409. doi: 10.1158/0008-5472.CAN-11-2427

Domcke, S., Sinha, R., Levine, D. A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* 4:2126. doi: 10.1038/ncomms3126

Eddy, J. A., Sung, J., Geman, D., and Price, N. D. (2010). Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol. Cancer Res. Treat.* 9, 149–159. doi: 10.1177/153303461000900204

El-Ashmawy, N. E., and Khalil, R. M. (2014). A review on the role of L-carnitine in the management of tamoxifen side effects in treated women with breast cancer. *Tumour. Biol.* 35, 2845–2855. doi: 10.1007/s13277-013-1477-1475

Elledge, R. M., Green, S., Pugh, R., Allred, D. C., Clark, G. M., Hill, J., et al. (2000). Estrogen receptor (ER) and progesterone receptor (PgR), by ligand-binding assay compared with ER, PgR and pS2, by immuno-histochemistry in predicting response to tamoxifen in metastatic breast cancer: a Southwest oncology group study. *Int. J. Cancer* 89, 111–117. doi: 10.1002/(sici)1097-0215(20000320)89:2<111::aid-ijc2>3.0.co;2-w

Gonzalez-Malerva, L., Park, J., Zou, L., Hu, Y., Moradpour, Z., Pearlberg, J., et al. (2011). High-throughput ectopic expression screen for tamoxifen resistance identifies an atypical kinase that blocks autophagy. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2058–2063. doi: 10.1073/pnas.1018157108

Guan, Q., Zeng, Q., Yan, H., Xie, J., Cheng, J., Ao, L., et al. (2019). A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci.* 110, 3225–3234. doi: 10.1111/cas.14137

Hallas-Potts, A., Dawson, J. C., and Herrington, C. S. (2019). Ovarian cancer cell lines derived from non-serous carcinomas migrate and invade more aggressively than those derived from high-grade serous carcinomas. *Sci. Rep.* 9:5515. doi: 10.1038/s41598-019-41941-41944

Harrell, F. E. Jr., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Stat. Med.* 3, 143–152. doi: 10.1002/sim.4780030207

International Cell Line Authentication Committee (2014). Cell line cross-contamination: WSU-CLL is a known derivative of REH and is unsuitable as a model for chronic lymphocytic Leukaemia. *Leuk. Res.* 38, 999–1001. doi: 10.1016/j.leukres.2014.05.003

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15. doi: 10.1093/nar/gng015

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

Knudsen, S., Jensen, T., Hansen, A., Mazin, W., Lindemann, J., Kuter, I., et al. (2014). Development and validation of a gene expression score that predicts response to fulvestrant in breast cancer patients. *PLoS One* 9:e87415. doi: 10.1371/journal.pone.0087415

Kong, D., and Yamori, T. (2012). JFCR39, a panel of 39 human cancer cell lines, and its application in the discovery and development of anticancer drugs. *Bioorg. Med. Chem.* 20, 1947–1951. doi: 10.1016/j.bmc.2012.01.017

Kreike, B., Hart, G., Bartelink, H., and van de Vijver, M. J. (2010). Analysis of breast cancer related gene expression using natural splines and the Cox proportional hazard model to identify prognostic associations. *Breast Cancer Res. Treat.* 122, 711–720. doi: 10.1007/s10549-009-0588-586

Lamberts, S. W., Verleun, T., and Oosterom, R. (1982). Effect of tamoxifen administration on prolactin release by invasive prolactin-secreting pituitary adenomas. *Neuroendocrinology* 34, 339–342. doi: 10.1159/000123324

Liedtke, C., Wang, J., Tordai, A., Symmans, W. F., Hortobagyi, G. N., Kiesel, L., et al. (2010). Clinical evaluation of chemotherapy response predictors developed from breast cancer cell lines. *Breast Cancer Res. Treat.* 121, 301–309. doi: 10.1007/s10549-009-0445-447

London, S. N., and Mailhes, J. B. (2001). Tamoxifen-induced alterations in meiotic maturation and cytogenetic abnormalities in mouse oocytes and 1-cell zygotes. *Zygote* 9, 97–104. doi: 10.1017/s0967199401001101

Masters, J. R. (2000). Human cancer cell lines: fact and fantasy. *Nat. Rev. Mol. Cell Biol.* 1, 233–236. doi: 10.1038/35043102

Masters, J. R. (2002). HeLa cells 50 years on: the good, the bad and the ugly. *Nat. Rev. Cancer* 2, 315–319. doi: 10.1038/nrc775

Merok, M. A., Ahlquist, T., Royrvik, E. C., Tufteland, K. F., Hektoen, M., Sjo, O. H., et al. (2013). Microsatellite instability has a positive prognostic impact on stage II colorectal cancer after complete resection: results from a large, consecutive Norwegian series. *Ann. Oncol.* 24, 1274–1282. doi: 10.1093/annonc/mds614

Mirabelli, P., Coppola, L., and Salvatore, M. (2019). Cancer cell lines are useful model systems for medical research. *Cancers* 11:1098. doi: 10.3390/cancers11081098

Peng, A., Xu, X., Wang, C., Ye, L., and Yang, J. (2018). A Bioinformatic profile of gene expression of colorectal carcinoma derived organoids. *Biomed. Res. Int.* 2018:2594076. doi: 10.1155/2018/2594076

Punt, C. J., Buyse, M., Kohne, C. H., Hohenberger, P., Labianca, R., Schmoll, H. J., et al. (2007). Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials. *J. Natl. Cancer Inst.* 99, 998–1003. doi: 10.1093/jnci/djm024

Sandberg, R., and Ernberg, I. (2005). Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *Proc. Natl. Acad. Sci. U.S.A.* 102, 2052–2057. doi: 10.1073/pnas.0408105102

Shoman, N., Klassen, S., McFadden, A., Bickis, M. G., Torlakovic, E., and Chibbar, R. (2005). Reduced PTEN expression predicts relapse in patients with breast carcinoma treated by tamoxifen. *Mod. Pathol.* 18, 250–259. doi: 10.1038/modpathol.3800296

Taylor, I. W., Hodson, P. J., Green, M. D., and Sutherland, R. L. (1983). Effects of tamoxifen on cell cycle progression of synchronous MCF-7 human mammary carcinoma cells. *Cancer Res.* 43, 4007–4010.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5116–5121. doi: 10.1073/pnas.091062498

Wass, M. N., Ray, L., and Michaelis, M. (2019). Understanding of researcher behavior is required to improve data reliability. *Gigascience* 8:giz017. doi: 10.1093/gigascience/giz017

Xu, L., Tan, A. C., Winslow, R. L., and Geman, D. (2008). Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinform.* 9:125. doi: 10.1186/1471-2105-9-125

Ye, L., Lin, C., Wang, X., Li, Q., Li, Y., Wang, M., et al. (2019). Epigenetic silencing of SALL2 confers tamoxifen resistance in breast cancer. *EMBO Mol. Med.* 2019:e10638. doi: 10.15252/emmm.201910638

Zou, J., Hong, G., Guo, X., Zhang, L., Yao, C., Wang, J., et al. (2011). Reproducible cancer biomarker discovery in SELDI-TOF MS using different pre-processing algorithms. *PLoS One* 6:e26294. doi: 10.1371/journal.pone.0026294

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

Copyright © 2020 Guan, Song, Zhang, Zhang, Chen 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 23 December 2020
doi: 10.3389/fmolb.2020.533196

Biased Influences of Low Tumor Purity on Mutation Detection in Cancer

Jun Cheng[1†], Jun He[1†], Shanshan Wang[1], Zhangxiang Zhao[2], Haidan Yan[1], Qingzhou Guan[1], Jing Li[1], Zheng Guo[1] and Lu Ao[1]*

[1] Department of Bioinformatics, Fujian Key Laboratory of Medical Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, The School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China, [2] Department of Systems Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

The non-cancerous components in tumor tissues, e.g., infiltrating stromal cells and immune cells, dilute tumor purity and might confound genomic mutation profile analyses and the identification of pathological biomarkers. It is necessary to systematically evaluate the influence of tumor purity. Here, using public gastric cancer samples from The Cancer Genome Atlas (TCGA), we firstly showed that numbers of mutation, separately called by four algorithms, were significant positively correlated with tumor purities (all $p < 0.05$, Spearman rank correlation). Similar results were also observed in other nine cancers from TCGA. Notably, the result was further confirmed by six in-house samples from two gastric cancer patients and five in-house samples from two colorectal cancer patients with different tumor purities. Furthermore, the metastasis mechanism of gastric cancer may be incorrectly characterized as numbers of mutation and tumor purities of 248 lymph node metastatic (N + M0) samples were both significantly lower than those of 121 non-metastatic (N0M0) samples ($p < 0.05$, Wilcoxon rank-sum test). Similar phenomena were also observed that tumor purities could confound the analysis of histological subtypes of cancer and the identification of microsatellite instability status (MSI) in both gastric and colon cancer. Finally, we suggested that the higher tumor purity, such as above 70%, rather than 60%, could be better to meet the requirement of mutation calling. In conclusion, the influence of tumor purity on the genomic mutation profile and pathological analyses should be fully considered in the further study.

Keywords: tumor purity, gastric cancer, microsatellite instability status, mutation calling algorithms, number of mutation

OPEN ACCESS

Edited by:
Peng Zhang,
University of Maryland, United States

Reviewed by:
Christos K. Kontos,
National and Kapodistrian University of Athens, Greece
Dong Wang,
Southern Medical University, China

***Correspondence:**
Lu Ao
lukey@fjmu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *07 February 2020*
Accepted: *22 October 2020*
Published: *23 December 2020*

Citation:
Cheng J, He J, Wang S, Zhao Z, Yan H, Guan Q, Li J, Guo Z and Ao L (2020) Biased Influences of Low Tumor Purity on Mutation Detection in Cancer. Front. Mol. Biosci. 7:533196. doi: 10.3389/fmolb.2020.533196

INTRODUCTION

Somatic mutation is accumulated during tumor development, which is commonly believed to play an important role in revealing the mechanism of carcinogenesis (Stratton et al., 2009; Stratton, 2011; Nik-Zainal et al., 2012). Recently, through sequencing analysis of cancer genomes, considerable advancements have been made in identifying cancer genes with "driver" mutation, such as TP53 (Moon et al., 2019), KRAS (Polom et al., 2019), BRAF (Yang et al., 2018a), EGFR (Paez et al., 2004), and PIK3CA (Harada et al., 2016). They provide insights into understand cancer development, find targets for therapeutic intervention (Alexandrov et al., 2013a,b) and develop diagnostic biomarkers.

However, it has been reported that the identification of somatic mutation may be influenced by tumor purity (Koboldt et al., 2012; Cibulskis et al., 2013). As is known to all, tumor tissues collected patients contain not only tumor cells, but also non-tumor cells, e.g., infiltrating stromal cells, immune cells, fibroblasts and normal cells (Joyce and Pollard, 2009), which could dilute the purity of tumor cells. Specifically, DNA from tumor samples are inevitably contaminated with non-tumor DNA. Various tumor purities might affect mutation detections through disturbed the numbers of mutated read (Raphael et al., 2014), and consequently affect the biological interpretations of genomic analyses (Aran et al., 2015).

Several approaches have been proposed to reduce the influence of tumor purity on mutation detection. For example, most studies generally require samples with at least 60% of tumor nuclei. However, the threshold of tumor purity might remain to be further evaluated (Aran et al., 2015). Practically, it is often difficult to obtain some cancer samples with sufficient tumor purity, such as diffuse gastric cancer and pancreatic adenocarcinomas. The laser capture microdissection (LCM) is commonly used to isolated pure tumor cells from tumor tissues (Espina et al., 2006), but it is cost and time consuming, which makes it difficult to be widely used in clinical scenes. Meanwhile, other collection technologies have been reported to isolate pure or putative tumor cells from tumor tissues. For example, DEPArray technology could isolate putative tumor cells from cancer samples (Lee et al., 2018), but it is difficult to handle large number of cells from large volume of cancers because of sorting time and the expenses (Lee et al., 2018). Furthermore, several algorithms have been proposed to evaluate tumor purities based on the copy number ploidy variations (Carter et al., 2012), methylation (Zheng et al., 2014), or expression levels of signature genes (Yoshihara et al., 2013). However, these tumor purities commonly reflect the average proportion of various cell types or are biased to a certain cell type. And the measurements of genes are sensitive to experimental batch effects (Leek et al., 2010; Oesper et al., 2014). The evaluation and correction of tumor purity is very hard and the golden standard is still dependent on the pathologists. Therefore, it is necessary to fully evaluate the influence of tumor purity on the analysis of genome mutation profile.

Gastric cancer is one of the common malignant tumors (Siegel et al., 2017). Tumor progression of gastric cancer, e.g., metastasis or post-surgery relapse, is the main death cause, and the tumor-node-metastasis (TNM) staging is an important indicator for tumor progression, which T represents primary tumor, N represents metastasis of regional lymph nodes and M represents distant metastasis of cancer. Based on the TNM system, the absence or presence of lymph node metastasis is identified as N0M0 or N + M0. Meanwhile, according to the Lauren's pathological classification, gastric cancer could be distinguished as intestinal, diffuse, or mixed subtypes (Shah et al., 2011). Compared with intestinal subtype, diffuse subtype has a different pattern of spread and behavior with a worse prognosis (Shah et al., 2011). The TNM staging system and the pathological classification are always used to determine the treatment strategies for gastric cancer patients. Besides, the microsatellite instability (MSI) status is another indicator for determining the treatment regimen in gastric cancer and colon cancer, which patients with high level of MSI (MSI-H) are less likely to benefit from the 5-Fu-based chemotherapy (Ilson, 2018). The MSI status were commonly identified by using immunohistochemistry and polymerase chain reaction (PCR), which measured the expressions of putative genes or the mutations of putative sites. However, molecular analyses between N0M0 and N + M0, or between diffuse and intestinal subtypes, or the identification of MSI status, may be affected by various tumor purities.

In this study, mainly using public gastric cancer samples from The Cancer Genome Atlas (TCGA) for example, the influence of tumor purity on mutation detection, pathological subtypes and the identification of MSI status were evaluated. Moreover, the biased influences were further evaluated in other nine cancers from TCGA and the in-house samples with different tumor purities from the same cancer patients. To obtain the robustly biological interpretations of genomic and pathological analyses, we suggested that the biased influences of various tumor purities should be fully considered.

MATERIALS AND METHODS

Data and Pre-processing

Public Data and Pre-processing

The mutation profiles called by four algorithms (MuSE, MuTect2, SomaticSniper, and VarScan2) and the clinical information of stomach adenocarcinoma (STAD) samples were downloaded from TCGA (**Table 1**).[1] Generally, multiple slides which were sampled from the top to bottom of the same tumor tissue were collected. Each slide was consisted of tumor cells and non-tumor cells. The percent of tumor nuclei in each slide was evaluated by pathologists. According to the report by Yoshihara et al. (2013), the tumor purity of a sample was the arithmetic mean percent of tumor nuclei in all slides. If the information of percent of tumor nuclei in one of the multiple slides was

[1] http://cancergenome.nih.gov/

TABLE 1 | Description of the number of public data/samples used in this study.

Cancer type	Sample size			
	MuSE	MuTect2	SomaticSniper	VarScan2
STAD	432	436	426	432
BRCA	979	982	970	981
CRC	534	534	535	534
GBM	389	389	383	388
LGG	502	504	497	503
LIHC	361	363	360	363
LUAD	504	508	497	502
LUSC	485	487	482	485
PAAD	161	169	140	150
PRAD	472	486	456	475

unavailable or the percent of tumor nuclei of all slides are zeros, the sample is excluded. Moreover, the mutation profiles and corresponding clinical information of other nine cancer types, included breast invasive carcinoma (BRAC), colorectal carcinoma (CRC), glioblastoma multiforme (GBM), brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), and prostate adenocarcinoma (PRAD) were also downloaded, respectively. And 723 cancer genes were downloaded from the COSMIC database (Tate et al., 2019),[2] which were used to analyze the influences of tumor purity on mutation callings of cancer genes.

In-house Data and Measurement

Six surgical resection specimens from two gastric cancer patients were measured by whole-exome sequencing with mean depth of 80–100×. For each patient, three specimens were sampled in three different locations, whose diameters of tumor tissues were at least 50 mm, respectively. The tumor purities of six samples, measured by pathologists, ranged from 26.5 to 92.5%, as shown in **Table 2**. Meanwhile, five surgical resection specimens collected from two colorectal cancer patients in our previous study were used to validate the influence of tumor purity on mutation detection (Yan et al., 2019). The tumor purities of five colorectal cancer samples ranged from 40 to 100% (**Table 2**). This study was approved by the institutional review boards of all participating institutions, and written consent forms were obtained from all participants.

Afterward, according to the manufacture's protocol, total DNA was isolated from the fresh frozen gastric tumor tissues and the generated raw whole-exome sequencing files (.fastq) were preprocessed using Trimmomatic (Bolger et al., 2014), and the reference genome (GRCh37) was used to align reads using Burrows-Wheeler aligner (BWA; Li and Durbin, 2009). Finally, the mutations were called using default parameters. Mutations included single nucleotide variation (SNV), indel (insertion and deletion, less than 50 bp) in this study. And they were filtered to exclude the mutation sites of germline risk based on gnomAD variant dataset file.[3] Only those SNVs which were identified as mutations were further analyzed.

Statistical Analysis

The spearman rank correlation analysis was used to assess the correlation between numbers of mutation and corresponding tumor purities in tumor samples. The wilcoxon rank-sum test was used to assess the difference of tumor purities (or numbers of mutation) between two groups of samples. And the fisher exact test was used to evaluate the significance of mutation frequencies of genes between high-purity and low-purity samples or between N0M0 and N + M0 samples. N0M0 and N + M0 represent non-metastatic samples and lymph node metastatic samples of gastric cancer, respectively. The hypergeometric test and cumulative binomial test were used to assess the impact of sample size on the correlation between numbers of mutation and tumor purities, respectively.

[2]https://cancer.sanger.ac.uk/cosmic/

[3]https://gnomad.broadinstitute.org/downloads

TABLE 2 | The tumor purities of in-house gastric cancer and colorectal cancer samples.

Patient	Position A (%)	Position B (%)	Position C (%)
GC-1	92.50	72.50	26.50
GC-2	88.00	56.50	33.00
CRC-1	100.00	100.00	40.00
CRC-2	70.00	40.00	–

RESULTS

Tumor Purity Confounds Mutation Detection

Taken gastric cancer as an example, we firstly analyzed the associations between numbers of mutation called by four mutation calling algorithms (MuSE, MuTect2, SomaticSniper, and VarScan2) and corresponding tumor purities, respectively. Tumor purities of gastric cancer samples distributed dispersedly, ranging from 5 to 100%. The tumor purity of about 72% gastric cancers were higher than 70%. The results showed that numbers of mutation called by MuSE and SomaticSniper algorithms were significant positively correlated with tumor purities (p = 2.22e-05 for MuSE and p = 1.84e-05 for SomaticSniper). Similar results were also observed in numbers of mutation called by MuTect2 (p = 1.00e-04) and VarScan2 (p = 7.73e-06) algorithms which are implanted the correction parameters of tumor purity. Notably, the significantly positive correlation between numbers of mutation and tumor purities in other nine cancer types could also be observed (**Table 3**). These results suggested that mutation detections might be significantly influenced by various tumor purities.

Then we verified the influence of tumor purity on mutation detection using MuTect2 algorithm in six in-house gastric tumor samples, which were sampled from three different locations with different tumor purities from each gastric cancer patient. The results showed that, for the samples from the same patient, the numbers of mutation decreased as the tumor purities decreased,

TABLE 3 | The p-values of spearman's rank correlations between tumor purities and numbers of mutation in other nine cancer types.

Cancer types	MuSE	MuTect2	SomaticSniper	VarScan2
BRCA	2.00e-04*	1.02e-02*	1.47e-06*	3.00e-04*
CRC	7.23e-02	2.03e-01	1.52e-04*	9.35e-02
GBM	3.42e-02*	1.16e-05*	1.66e-01	4.67e-02*
LGG	7.25e-08*	5.49e-06*	6.75e-10*	1.45e-08*
LIHC	5.65e-02	1.22e-01	5.20e-03*	2.98e-02*
LUAD	4.17e-02	8.84e-01	1.54e-02*	3.03e-01
LUSC	9.35e-05*	5.30e-03*	3.07e-08*	8.69e-05*
PAAD	2.34e-02*	3.42e-02*	7.40e-03*	6.71e-02
PRAD	4.00e-04*	4.00e-04*	3.61e-07*	7.05e-05*

Represented the significance of p-value.

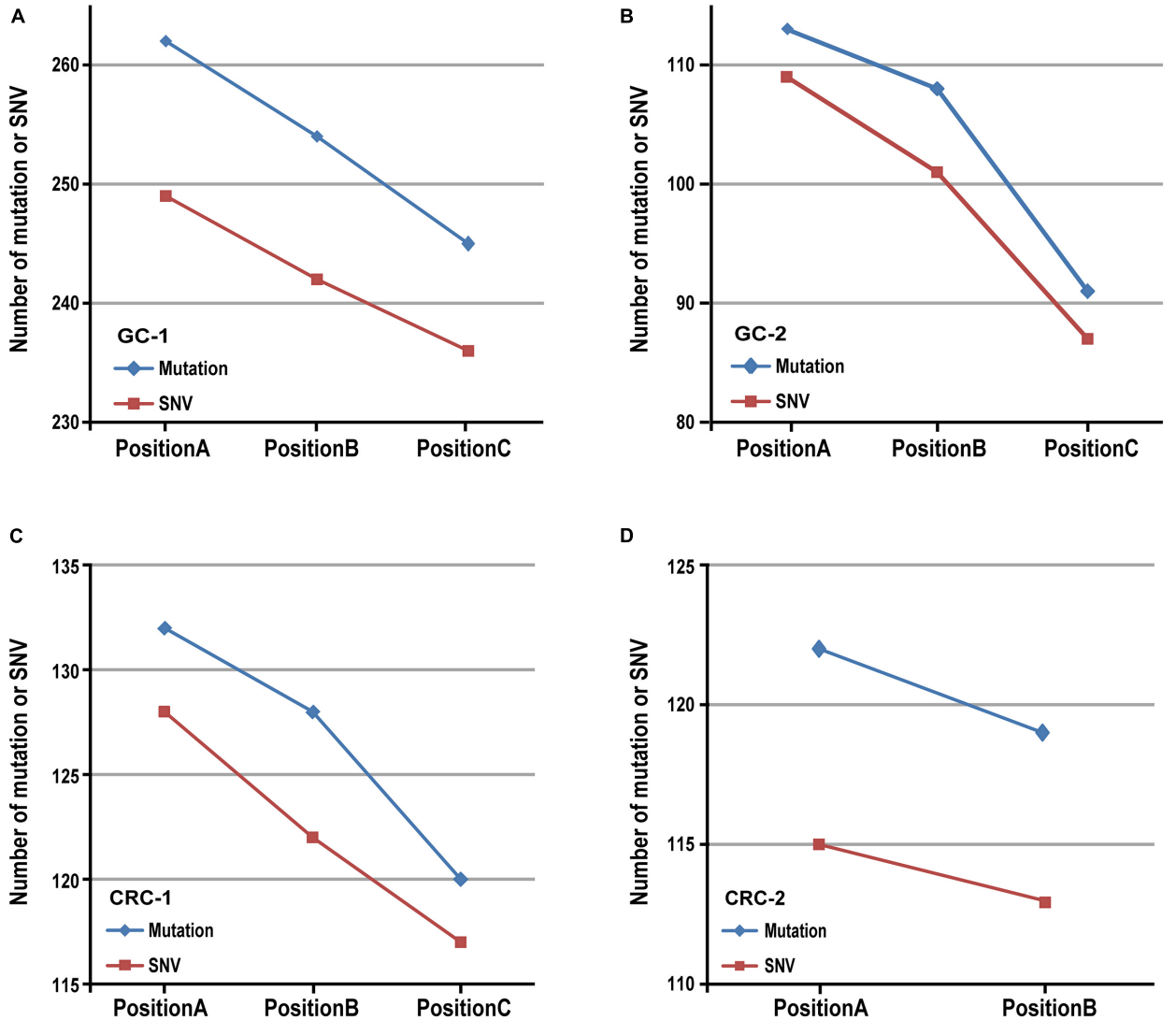


FIGURE 1 | The result of mutation calling in different sampling positions. The number of mutation (or SNV) decreased as tumor purities decreased in two patients with gastric cancer **(A,B)** and two patients with colorectal cancer **(C,D)**.

as shown in **Figures 1A,B**. Similar results were also observed in five in-house colorectal tumor samples collected from two patients, as shown in **Figures 1C,D**. The results further confirmed that various tumor purities might affect numbers of mutation. Moreover, similar results were observed in numbers of mutation detected by the Varscan2, SomaticSniper and MuSE algorithms, respectively, which decreased with the tumor purities, as shown in **Supplementary Table 1**.

Additionally, we further analyzed the numbers of mutated reads aligned to each mutation site in measured gastric cancer samples. For GC-1 patient, among 19 SNVs that were identified in samples with tumor purities of 92.50 and 72.50%, 15 SNVs were not detected in sample with the lowest tumor purity of 26.50%. Nevertheless, they were aligned to several mutated fragments (14 SNVs: 1–4 reads and 1 SNV: 6 reads). Similarly, 14 SNVs were not identified as mutations in the position C with 33% of tumor purity for GC-2 patient, but they were also aligned to several mutated fragments (1–5 reads). Those unidentified mutation sites in the position C of two patients included the genes *FBXO11* and *XPO1*, which were identified as cancer genes in the COSMIC database,[4] shown in **Table 4**. These results indicated that the artificially low mutation burden might result from low tumor purities.

Tumor Purity Confounds the Mutation Differences Between Metastasis and Non-metastasis of Gastric Cancer

Based on the non-synonymous mutation data of primary gastric cancer samples from TCGA database, which were called by MuTect2 algorithm, we found that the numbers of mutation

[4]https://cancer.sanger.ac.uk/cosmic/

TABLE 4 | Mutations of *FBXO11* and *XP01* in different sampling positions.

Patient/ gene	Sites	Mutation type	Mutation site	Mutation reads	Aligned reads
GC-1/ FBXO11	PositionA	disruptive_ inframe_del	c.11_37del	5	57
	PositionB	disruptive_ inframe_ del	c.11_37del	8	51
	PositionC	no	no	1	81
GC-2/ XP01	PositionA	missense	c.1426T > C	7	68
	PositionB	missense	c.1426T > C	11	100
	PositionC	no	no	0	98

in 248 N + M0 samples tended to be significantly less than those in 121 N0M0 samples (p = 5.14e-02, Wilcoxon rank-sum test, **Figure 2A**). Then we compared the differences of multiple clinical factors between two subgroups, including age, gender, tumor purity and grade, and found that only tumor purity was significantly different between two subgroups. The tumor purities in N + M0 samples were significantly lower than those in N0M0 samples (p = 1.77e-02, Wilcoxon rank-sum test, **Figure 2B**). In order to remove the biased influence of sample sizes, we randomly selected 121 samples from 248 N + M0 samples and compared tumor purities and numbers of mutation between 121 N0M0 and 121 N + M0 samples. The random experiment was repeated 1,000 times. The result showed that there were 546 times of significantly different tumor purities between N0M0 and N + M0 samples, 388 times of significantly different numbers of mutation, and 246 times that tumor purity and number of mutation were both significantly different (all $p < 0.05$, Wilcoxon rank-sum test). The results were not happened randomly ($p <$ 1.00e-16, hypergeometric test), which indicated that the biased sample sizes could not be the main cause of mutation differences between N0M0 and N + M0 samples. Removing diffuse gastric tumor samples with high heterogeneity, similar phenomena were also observed in intestinal gastric cancer that numbers of mutation in 115 N + M0 samples were significantly less than those in 46 N0M0 samples ($p <$ 8.40e-03, Wilcoxon rank-sum test, **Figure 2C**), and tumor purities in 115 N + M0 samples were also significantly less than those in 46 N0M0 samples ($p <$ 4.24e-02, Wilcoxon rank-sum test, **Figure 2D**). The results indicated that the difference of numbers of mutation between N0M0 and N + M0 may be mainly caused by the variations of tumor purity. The lower tumor purities of N + M0 samples could lead to the artificially lower mutation burden than that of N0M0 samples.

Meanwhile, we also found the mutation frequency of 1,184 genes were significantly different between N0M0 and N + M0 samples ($p < 0.05$, Fisher's exact one-side test). Subsequently, we divided the primary gastric tumor tissues into two groups according to tumor purities. Totally, 129 samples whose tumor purities were at least 80% were divided into the high-purity group, while 127 samples whose tumor purities were less than 70% were divided into the low-purity group. The information of low- and high-purity samples in different categories was shown

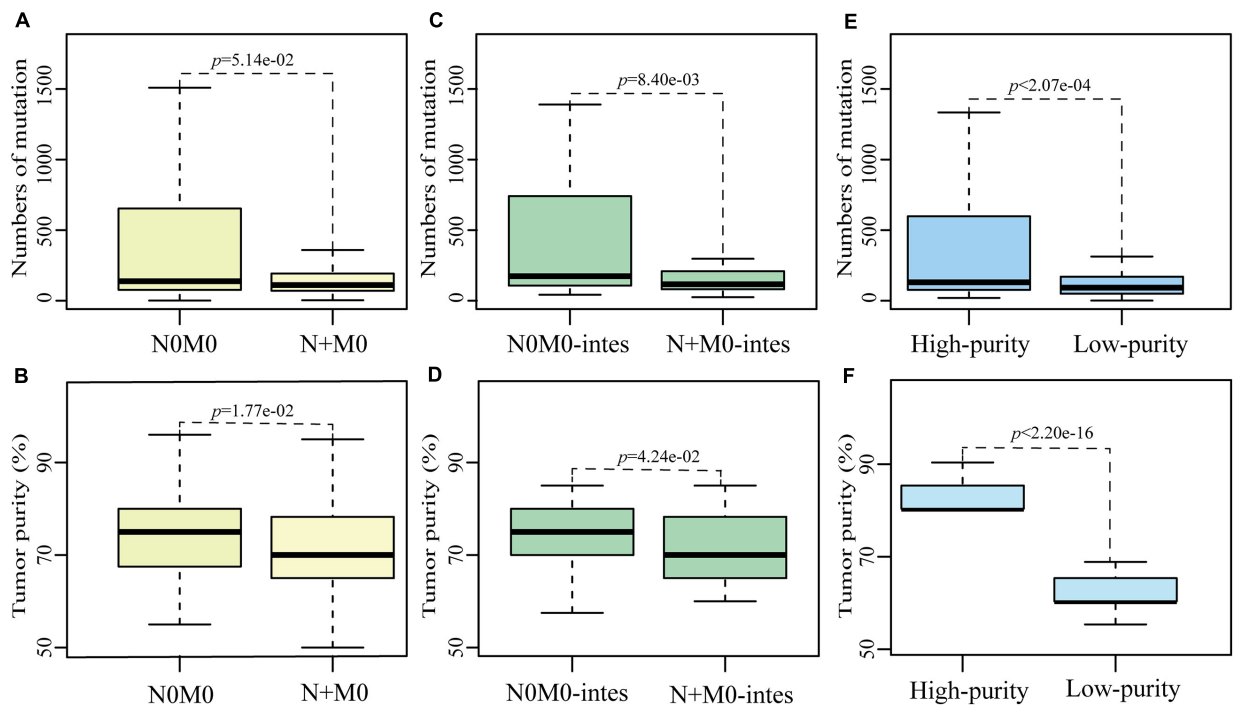


FIGURE 2 | The influence of tumor purity on mutation detection between N0M0 and N + M0 samples. The differences of number of mutation or tumor purity between N0M0 and N + M0 samples **(A,B)**, between N0M0 (N0M0-intes) and N + M0 (N + M0-intes) samples in intestinal gastric cancer **(C,D)**, and between high-purity and low-purity samples **(E,F)**. Outline points were deleted.

TABLE 5 | The number of low- and high-purity samples in different categories.

Sample size	High_purity ≥ 80%	Low_purity < 70%
All(436)	129	127
N0M0(121)	46	31
N + M0(248)	61	80
N + M0-intes(115)	28	40

in **Table 5**. The numbers of mutation in low-purity samples were significantly lower than those in high-purity samples ($p < 0.05$, Wilcoxon rank-sum test, **Figures 2E,F**). Similarly, the mutation frequencies of 1,247 genes were significantly different between high-purity and low-purity groups ($p < 0.05$, Fisher's exact one-side test). There were 184 genes overlapped with the 1,184 genes of differentially mutated frequency between N0M0 and N + M0 samples, of which 182 genes had significantly higher mutation frequency in both N0M0 samples and high-purity samples. Gene *SLC3A2* and *APC*, which were associated with metastasis and neoplasia (Ghatak et al., 2017; Wang et al., 2017), were included. These results indicated that various tumor purities had an impact on mutation differences between N0M0 and N + M0 samples, which might confound the interpretation of metastasis mechanism for gastric cancer.

Tumor Purity Confounds the Molecular Analysis of Gastric Cancer Subtypes

We then evaluated the influence of tumor purity on the mutation analysis between the diffuse and intestinal histological subtypes of gastric cancer. No significant difference of tumor purity was observed between 70 diffuse samples and 190 intestinal samples (p = 1.45e-01, Wilcoxon rank-sum test). However, after excluding five intestinal and four diffuse unrepresentative samples that only had one slide with more than 90% of tumor purity, the tumor purities of 66 diffuse samples tend to be significantly lower than those of 185 intestinal samples (p = 5.04e-02, Wilcoxon rank-sum test), while numbers of mutation in diffuse subtype were significantly less than those in intestinal subtype (p = 9.49e-05, Wilcoxon-rank test), as showed in **Figure 3A**. Furthermore, similar phenomena that the significant differences of tumor purities and numbers of mutation between the histological subtypes of lung cancer (including LUAD and LUSA) or glioma (including GBM and LGG) were also observed, respectively, as shown in **Figure 3B**. The results suggested the various tumor purities might confound the mutation differences between different histological subtypes of cancer.

Tumor Purity Confounds the Identification of MSI Status

We further evaluated the influence of various tumor purities on the identification of a known pathological biomarker, the MSI status, which is commonly used to determine the follow-up treatment regimen for gastric and colon cancer patients. According to the MSI status of gastric cancer, the tumor purities of 241 samples with stable level of MSI were significantly lower than both 72 MSI-H samples and 56 low level of MSI (MSI-L) samples, respectively (all $p < 0.05$, Wilcoxon rank-sum test, **Figure 3C**). Compared with the distribution of tumor purities of gastric cancer samples, the tumor purities of colon cancer samples distributed narrowly, and 86% of the colon cancer samples were with ≥70% of tumor purities. No significant correlation was observed between number of mutation and tumor purity in colon cancer. However, the tumor purities of 83 MSI-H samples were significantly higher than those of 82 MSI-L samples (p = 4.46e-02, Wilcoxon rank-sum test) and tentative significantly higher than those of 291 samples with stable level of MSI (p = 7.20e-02, Wilcoxon rank-sum test), respectively, as shown in **Figure 3C**. The above results suggested that various tumor purities might confound the identification of MSI status.

An Appropriate Threshold of Tumor Purity for Mutation Calling

Finally, we took gastric cancer as an example to identify an appropriate tumor purity for mutation calling. According to the at least 60% of tumor purity required in most researches, we firstly removed the gastric cancer samples with tumor purity less than 60%, and observed that numbers of mutation called by four algorithms were still significant positively correlated with tumor purities ($p < 0.05$, **Table 6**). These results indicated that higher tumor purity may be needed for mutation calling. Then we analyzed samples with higher than 70% of tumor purity. No significant correlation was observed between tumor purity and number of mutation, except for SomaticSniper algorithm. Moreover, similar results that non-significant correlation between tumor purities and numbers of mutation were observed in other nine cancer types, except for LGG (**Table 6**).

In order to remove the influence of sample size, the same size of gastric samples with above 70% of tumor purity were randomly selected from samples with ≥60% of tumor purity and the correlations between tumor purities and numbers of mutation were calculated. The random experiment was repeated 1,000 times. Finally, a cumulative binomial test was used to assess the significance of positive correlation in the 1,000 random experiments. The results showed that 65.50% of 1,000 random experiments were significant correlations in Mutect2 algorithms and more than 80% of 1,000 random experiments were significant correlations in other three algorithms, respectively (all $p < 0.05$, binomial test, **Supplementary Table 2**). Similar results of random experiments were also observed in other multiple cancer types (**Supplementary Table 1**). These results indicated that the sample sizes could not be the major factor of correlation between number of mutation and tumor purity. In a word, above 70% of tumor purity, rather than 60%, might be better to meet the requirement of mutation calling.

DISCUSSION

As showed in this study, numbers of mutation and tumor purities were significantly positive correlation in gastric cancer and other nine cancer types, regardless of calling algorithms. The

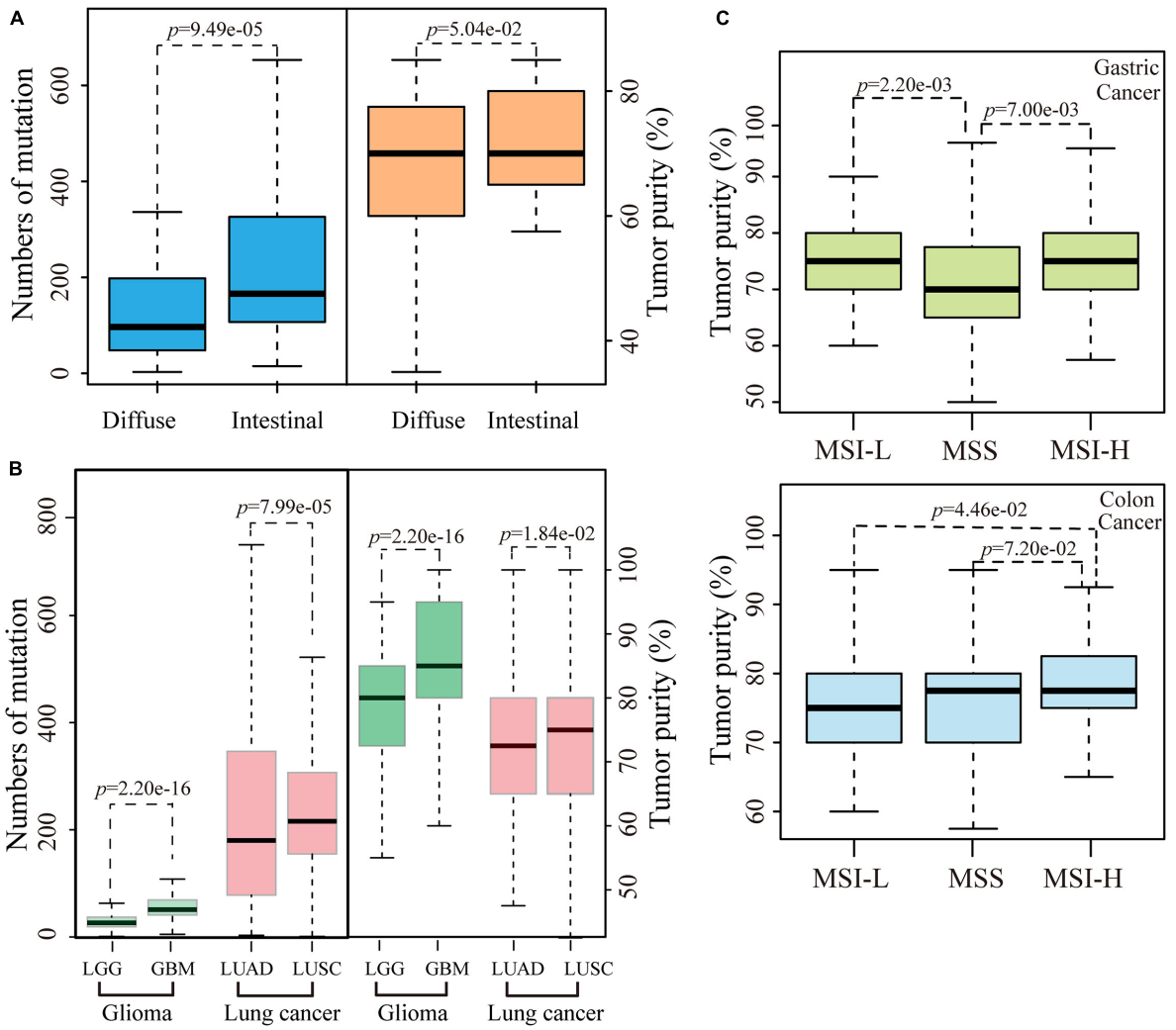


FIGURE 3 | The influence of tumor purity on MSI status and pathological subtypes. **(A)** The differences of number of mutation or tumor purity between the diffuse and intestinal gastric cancer. **(B)** The differences of number of mutation or tumor purity between GBM and LGG or between LUAD and LUSC. **(C)** The differences of tumor purity between different MSI status both in gastric and colon cancer. Outline points were deleted.

lower tumor purities may lead to the artificially lower mutation burden, which may consequently cause the misleading biological interpretation of metastasis mechanism, pathological subtypes, as well as pathological biomarker analyses. Finally, we suggested that above 70% of tumor purity could be better to meet the requirement of mutation callings.

Moreover, gene *FBXO11*, *XPO1*, *SLC3A2*, and *APC*, whose mutation detections may be affected by various tumor purities in gastric cancer, were closely related with cancer occurrence and development. For examples, protein FBXO11 has both the E3 ubiquitin ligase and methyltrasferase activity, which could facilitate epithelial-mesenchymal transition (EMT), promote PI3K/AKT pathway activation, and regulate metastasis and apoptosis in human cancer (Kim et al., 2018, 2020; Sun et al., 2018). Protein XPO1 is positively correlated with cell proliferation and growth transformation, and negatively correlated with poor survival outcomes, which could be a promising molecular target in gastric cancer (Subhash et al., 2018; Gruffaz et al., 2019; Sexton et al., 2019). Protein SLC3A2 is associated with the migration and invasion of tumor cells (Wang et al., 2017), which is a potential biomarker for molecular imaging-based detection of gastric cancer (Yang et al., 2012). Gene *APC*, which is involved in Wnt/β-catenin signaling pathway, has been reported to be associated with tumorigenesis, tumor metastasis and resistance (Yang et al., 2018b).

Currently, many studies have been proposed that tumor mutation burden (TMB) could predict the response to immunotherapy (Goodman et al., 2017; Qin et al., 2018), which patients with high TMB commonly responds better to immunotherapy than patients with low TMB. However, due to the differences in surgical sampling or biopsy sites of tumor tissue, the TMB or the pathologic biomarkers, such as PDL-1 (Anagnostou et al., 2017; Qin et al., 2018), could be affected by various tumor purities. For this problem, some

TABLE 6 | The *p*-values of spearman's rank correlation between tumor purity higher than 60 or 70% and number of mutation.

Cancer types	MuSE	MuTect2	SomaticSniper	VarScan2
Tumor purity ≥ 60%				
STAD	1.24e-04*	1.40e-03*	7.05e-05*	8.44e-05*
BRCA	1.40e-03*	4.68e-02*	5.57e-05*	4.80e-03*
CRC	1.08e-01	2.64e-01	5.48e-04*	1.14e-01
GBM	8.00e-02	5.37e-05*	1.96e-01	1.07e-01
LGG	1.26e-06*	1.88e-05*	2.59e-08*	1.72e-07*
LIHC	5.65e-02	1.22e-01	5.20e-03*	2.98e-02*
LUAD	7.03e-01	8.10e-01	4.01e-02*	5.49e-01
LUSC	6.85e-04*	1.58e-02*	5.63e-07*	4.54e-04*
PAAD	NAN	NAN	NAN	NAN
PRAD	7.61e-06*	3.31e-05*	4.10e-07*	3.29e-05*
Tumor purity > 70%				
STAD	1.43e-01*	4.19e-01*	3.17e-02*	1.29e-01*
BRCA	1.78e-01	5.47e-01	1.76e-02*	7.00e-02
CRC	3.02e-01	5.69e-01	6.30e-03*	3.75e-01
GBM	1.14e-01	9.54e-05#	3.60e-01	1.34e-01
LGG	7.00e-03*	1.87e-02*	7.40e-03*	1.40e-03*
LIHC	1.76e-01	3.71e-01	3.58e-02*	1.42e-01
LUAD	3.74e-01	6.55e-01	4.80e-02*	4.17e-01
LUSC	3.40e-01	3.05e-01	1.84e-01	4.43e-01
PAAD	NAN	NAN	NAN	NAN
PRAD	8.21e-02	2.25e-01	2.61e-02*	7.45e-02

*# and * represented non-significant and significant p-value (< 0.05) calculated by spearman rank correlation, respectively. NAN represented that the p-value was not calculated due to small sample size.*

researches proposed to increase the sequencing depth to reduce the false negatives from low tumor purity, but it might also sharply increase the false positives of mutation detection, work burden and cost.

Additionally, for the threshold of tumor purity, TCGA originally required at least 80% of tumor nuclei (Aran et al., 2015), but it is generally difficult to collect enough amount of samples. Then, this threshold was later reduced to 60% as the RNA-seq technology developed. And most current studies set the threshold as 60%. However, the research by Dvir Aran et.al (Aran et al., 2015) indicated that the impact of 60% of tumor purity on the interpretation of genomic analyses remained to be evaluated. Our results in ten cancer types showed that, above 70% of tumor purity, rather than 60%, might be better to meet the requirement of mutation calling and obtain relatively sufficient and reliable mutation profiles. Certainly, a novel mutation detection algorithm for tumor sample with low purity should be developed as soon as possible.

A major limitation is that the tumor heterogeneity, pathological subtypes, and the colonal selection of mutations do affect mutation callings during the process of tumor occurrence and development (Gerlinger et al., 2012), which could not be excluded in this study. However, our study revealed that there were universal significantly correlations between numbers of mutation and tumor purities in ten cancer types. Although the sample size of in-house data is small in this study, the low tumor purities resulted in less mutations that were further demonstrated in six gastric cancer samples from two patients and five colorectal cancer samples from two patients with different tumor purities. That suggested that numbers of mutation were influenced by tumor purities regardless of tumor types and the influence of tumor purity on number of mutation should be noticed.

In conclusion, the influences of various tumor purities on mutation detection and pathological analyses should be fully considered in further analysis. And we suggested that more than 70% of tumor purity could be better to meet the requirement of mutation calling.

DATA AVAILABILITY STATEMENT

The in-house data used and analyzed during the current study is available from the corresponding authors upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Affiliated Union Hospital of Fujian Medical University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

ZG conceived the idea. LA and JC conceived and designed the experiments and wrote the manuscript. JH designed the experiments and made figures. SSW and ZXZ searched the data and participated in the statistical analysis. QZG and HDY helped in interpreting the results and writing the manuscript. JL helped in writing the manuscript. All authors approved the final version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 81602738 and 61801118), the Joint Scientific and Technology Innovation Fund of Fujian Province (Grant No. 2018Y9065), the Joint research program of health and education in Fujian Province (2019-WJ-32), and the Natural Science Foundation of Fujian Province (Grant No. 2020J01600). The authors declare no conflicts of interest.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.533196/full#supplementary-material

Supplementary Table 1 | Numbers of mutation called by other three algorithms.

Supplementary Table 2 | The number of significant *p*-value in 1000 random experiments.

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013a). Signatures of mutational processes in human cancer. *Nature* 500, 415–421.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 3, 246–259. doi: 10.1016/j.celrep.2012.12.008

Anagnostou, V., Smith, K. N., Forde, P. M., Niknafs, N., Bhattacharya, R., White, J., et al. (2017). Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov.* 7, 264–276.

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* 6:8971.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30, 413–421. doi: 10.1038/nbt.2203

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. doi: 10.1038/nbt.2514

Espina, V., Wulfkuhle, J. D., Calvert, V. S., VanMeter, A., Zhou, W., Coukos, G., et al. (2006). Laser-capture microdissection. *Nat. Protoc.* 1, 586–603.

Gerlinger, M., Rowan, A. J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892.

Ghatak, S., Chakraborty, P., Sarkar, S. R., Chowdhury, B., Bhaumik, A., and Kumar, N. S. (2017). Novel APC gene mutations associated with protein alteration in diffuse type gastric cancer. *BMC Med. Genet.* 18:61.

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

Gruffaz, M., Yuan, H., Meng, W., Liu, H., Bae, S., Kim, J. S., et al. (2019). CRISPR-Cas9 screening of Kaposi's sarcoma-associated herpesvirus-transformed cells identifies XPO1 as a vulnerable target of cancer cells. *mBio* 10:e00866-19.

Harada, K., Baba, Y., Shigaki, H., Ishimoto, T., Miyake, K., Kosumi, K., et al. (2016). Prognostic and clinical impact of PIK3CA mutation in gastric cancer: pyrosequencing technology and literature review. *BMC Cancer* 16: 400.

Ilson, D. H. (2018). Advances in the treatment of gastric cancer. *Curr. Opin. Gastroenterol.* 34, 465–468.

Joyce, J. A., and Pollard, J. W. (2009). Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* 9, 239–252. doi: 10.1038/nrc2618

Kim, Y. J., Hwang, K. C., Kim, S. W., and Lee, Y. C. (2018). Potential miRNA-target interactions for the screening of gastric carcinoma development in gastric adenoma/dysplasia. *Int. J. Med. Sci.* 15, 610–616. doi: 10.7150/ijms.24061

Kim, Y. J., Jeong, S., Jung, W. Y., Choi, J. W., Hwang, K. C., Kim, S. W., et al. (2020). miRNAs as potential biomarkers for the progression of gastric cancer inhibit CREBZF and regulate migration of gastric adenocarcinoma cells. *Int. J. Med. Sci.* 17, 693–701. doi: 10.7150/ijms.42654

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Lee, J. W., Shin, J. Y., and Seo, J. S. (2018). Identification of novel mutations in FFPE lung adenocarcinomas using DEPArray sorting technology and next-generation sequencing. *J. Appl. Genet.* 59, 269–277. doi: 10.1007/s13353-018-0439-4

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Moon, S., Balch, C., Park, S., Lee, J., Sung, J., and Nam, S. (2019). Systematic inspection of the clinical relevance of TP53 missense mutations in gastric cancer. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1693–1701. doi: 10.1109/tcbb.2018.2814049

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell* 149, 979–993.

Oesper, L., Satas, G., and Raphael, B. J. (2014). Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* 30, 3532–3540. doi: 10.1093/bioinformatics/btu651

Paez, J. G., Janne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500. doi: 10.1126/science.1099314

Polom, K., Das, K., Marrelli, D., Roviello, G., Pascale, V., Voglino, C., et al. (2019). KRAS mutation in gastric cancer and prognostication associated with microsatellite instability status. *Pathol. Oncol. Res.* 25, 333–340. doi: 10.1007/s12253-017-0348-6

Qin, B. D., Jiao, X. D., and Zang, Y. S. (2018). Tumor mutation burden to tumor burden ratio and prediction of clinical benefit of anti-PD-1/PD-L1 immunotherapy. *Med. Hypotheses* 116, 111–113. doi: 10.1016/j.mehy.2018.05.005

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* 6:5. doi: 10.1186/gm524

Sexton, R., Mahdi, Z., Chaudhury, R., Beydoun, R., Aboukameel, A., Khan, H. Y., et al. (2019). Targeting nuclear exporter protein XPO1/CRM1 in gastric cancer. *Int. J. Mol. Sci.* 20:4826. doi: 10.3390/ijms20194826

Shah, M. A., Khanin, R., Tang, L., Janjigian, Y. Y., Klimstra, D. S., Gerdes, H., et al. (2011). Molecular classification of gastric cancer: a new paradigm. *Clin. Cancer Res.* 17, 2693–2701. doi: 10.1158/1078-0432.ccr-10-2203

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

Stratton, M. R. (2011). Exploring the genomes of cancer cells: progress and promise. *Science* 331, 1553–1558. doi: 10.1126/science.1204040

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458, 719–724.

Subhash, V. V., Yeo, M. S., Wang, L., Tan, S. H., Wong, F. Y., Thuya, W. L., et al. (2018). Anti-tumor efficacy of selinexor (KPT-330) in gastric cancer is dependent on nuclear accumulation of p53 tumor suppressor. *Sci. Rep.* 8:12248.

Sun, C., Tao, Y., Gao, Y., Xia, Y., Liu, Y., Wang, G., et al. (2018). F-box protein 11 promotes the growth and metastasis of gastric cancer via PI3K/AKT pathway-mediated EMT. *Biomed. Pharmacother.* 98, 416–423. doi: 10.1016/j.biopha.2017.12.088

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947.

Wang, S., Han, H., Hu, Y., Yang, W., Lv, Y., Wang, L., et al. (2017). SLC3A2, antigen of mAb 3G9, promotes migration and invasion by upregulating of mucins in gastric cancer. *Oncotarget* 8, 88586–88598. doi: 10.18632/oncotarget.19529

Yan, H., Deng, X., Chen, H., Cheng, J., He, J., Guan, Q., et al. (2019). Identification of common and subtype-specific mutated sub-pathways for a cancer. *Front. Genet.* 10:1228.

Yang, Q., Huo, S., Sui, Y., Du, Z., Zhao, H., Liu, Y., et al. (2018a). Mutation status and immunohistochemical correlation of KRAS, NRAS, and BRAF in 260 Chinese colorectal and gastric cancers. *Front. Oncol.* 8:487.

Yang, X. Z., Cheng, T. T., He, Q. J., Lei, Z. Y., Chi, J., Tang, Z., et al. (2018b). LINC01133 as ceRNA inhibits gastric cancer progression by sponging miR-106a-3p to regulate APC expression and the Wnt/beta-catenin pathway. *Mol. Cancer* 17:126.

Yang, Y., Toy, W., Choong, L. Y., Hou, P., Ashktorab, H., Smoot, D. T., et al. (2012). Discovery of SLC3A2 cell membrane protein as a potential gastric cancer biomarker: implications in molecular imaging. *J. Proteome Res.* 11, 5736–5747. doi: 10.1021/pr300555y

Yoshihara, K., Shahmoradgoli, M., Martinez, 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.

Zheng, X., Zhao, Q., Wu, H. J., Li, W., Wang, H., Meyer, C. A., et al. (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol.* 15:419.

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

Copyright © 2020 Cheng, He, Wang, Zhao, Yan, Guan, Li, Guo and Ao. 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 Molecular Biosciences

ORIGINAL RESEARCH
published: 07 January 2021
doi: 10.3389/fmolb.2020.587822

A Novel Prognostic Model of Endometrial Carcinoma Based on Clinical Variables and Oncogenomic Gene Signature

Fang Deng[1], *Jing Mu*[1], *Chiwen Qu*[2], *Fang Yang*[1], *Xing Liu*[1], *Xiaomin Zeng*[1]* and *Xiaoning Peng*[2,3,4]*

[1] Department of Epidemiology and Health Statistics, Xiangya School of Public Health, Central South University, Changsha, China, [2] School of Mathematics and Statistics, Hunan Normal University, Changsha, China, [3] Department of Pathology and Pathophysiology, Hunan Normal University School of Medicine, Changsha, China, [4] Department of Pathophysiology, Jishou University School of Medicine, Jishou, China

Due to the difficulty in predicting the prognosis of endometrial carcinoma (EC) patients by clinical variables alone, this study aims to build a new EC prognosis model integrating clinical and molecular information, so as to improve the accuracy of predicting the prognosis of EC. The clinical and gene expression data of 496 EC patients in the TCGA database were used to establish and validate this model. General Cox regression was applied to analyze clinical variables and RNAs. Elastic net-penalized Cox proportional hazard regression was employed to select the best EC prognosis-related RNAs, and ridge regression was used to construct the EC prognostic model. The predictive ability of the prognostic model was evaluated by the Kaplan–Meier curve and the area under the receiver operating characteristic curve (AUC-ROC). A clinical-RNA prognostic model integrating two clinical variables and 28 RNAs was established. The 5-year AUC of the clinical-RNA prognostic model was 0.932, which is higher than that of the clinical-alone (0.897) or RNA-alone prognostic model (0.836). This clinical-RNA prognostic model can better classify the prognosis risk of EC patients. In the training group (396 patients), the overall survival of EC patients was lower in the high-risk group than in the low-risk group [HR = 32.263, (95% CI, 7.707–135.058), P = 8e-14]. The same comparison result was also observed for the validation group. A novel EC prognosis model integrating clinical variables and RNAs was established, which can better predict the prognosis and help to improve the clinical management of EC patients.

Keywords: endometrial carcinoma, cancer genomics, the Cancer Genome Atlas (TCGA), integrative model, prognosis

OPEN ACCESS

Edited by:
Peng Zhang,
University of Maryland, United States

Reviewed by:
Siddharth Shukla,
Howard Hughes Medical Institute (HHMI), United States
Junjiang Fu,
Southwest Medical University, China

***Correspondence:**
Xiaomin Zeng
zxiaomin@csu.edu.cn
Xiaoning Peng
pxiaoning@hunnu.edu.cn

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *27 July 2020*
Accepted: *23 November 2020*
Published: *07 January 2021*

Citation:
Deng F, Mu J, Qu C, Yang F, Liu X, Zeng X and Peng X (2021) A Novel Prognostic Model of Endometrial Carcinoma Based on Clinical Variables and Oncogenomic Gene Signature. Front. Mol. Biosci. 7:587822. doi: 10.3389/fmolb.2020.587822

INTRODUCTION

Endometrial carcinoma (EC) is a common malignant tumor of the female reproductive system, and its incidence is increasing (Chen W. et al., 2016; Siegel et al., 2020). Metastasis or recurrence often occurs in EC patients after surgery, and the median survival time of patients with recurrence or metastasis is generally <12 months (Obel et al., 2006). Chemotherapy and radiation therapy fail to kill tumor cells with high specificity. The 5-year overall survival rate of EC patients without

metastasis is between 74 and 91% (Morice et al., 2016), while the rate is reduced to 68 or 17% for EC patients with local or distant metastases, respectively (Colombo et al., 2016). Therefore, it is urgent to study the factors and mechanisms that affect the prognosis of EC patients and improve the clinical management.

At present, the prognosis prediction of EC patients is mainly based on the age at diagnosis, FIGO stage, pathological classification, treatment method, and other clinical variables. Due to the strong individual differences in the stages of occurrence, development, and metastasis of EC, it is difficult to accurately predict the prognosis of EC patients through clinical variables only (Frederick and Straughn, 2009). Studies have shown that specific genes or molecular changes influence the prognosis of EC patients (Bell and Ellenson, 2019). Molecules such as *ER*, *PR*, *p53*, *HER-2/neu*, and *Ki-67* have been used to predict EC recurrence or prognosis; nevertheless, the results are still controversial (Fanning et al., 2002; Jeon et al., 2006).

In recent years, a class of non-coding RNA (ncRNA), including microRNA (miRNA) and long non-coding RNA (lncRNA), which cannot encode proteins, has been found to play an important role in life regulation (Djebali et al., 2012). More and more studies show that abnormal expression of ncRNA is closely related to the prognosis of EC patients. For example, *miRNA-200c*, *miR-944*, *HOTAIR*, *H19*, and *SRA* are related to prognosis of EC patients or the malignant degree of EC tumors (Smolle et al., 2015; He et al., 2017; Wilczynski et al., 2018). The expression levels of *miR-142-3p*, *miR-142-5p*, and *miR-15a-5p* are higher in EC patients with progression-free survival (PFS)>21 months than in EC patients with PFS<21 months, suggesting that *miR-142* and *miR-15a* may be useful for EC prognosis prediction (Jayaraman et al., 2017). *Hsa-mir-15a.MIMAT0000068*, *hsa-mir-142.MIMAT0000433*, *hsa-mir-142.MIMAT0000434*, *hsa-mir-3170.MIMAT0015045*, *hsa-mir-1976.MIMAT0009451*, and *hsa-mir-146a.MIMAT0000449* are significantly related to EC overall survival (OS), and the six-microRNA signature is an independent prognostic factor of EC (Wang Y. et al., 2019). However, so far, there has been no report on EC prognostic model integrating mRNAs, miRNAs, lncRNAs, and clinical variables.

This study intends to analyze the clinical and genome-wide mRNA, miRNA, and lncRNA expression data of EC patients in The Cancer Genome Atlas (TCGA) database and screen RNAs and clinical variables that are related to EC prognosis, with the expectation of discovering new EC prognostic molecular markers and establishing the integrated clinical-mRNA–miRNA–lncRNA prognostic model, thus providing a theoretical basis for EC prognostic risk assessment and individualized treatment.

MATERIALS AND METHODS

Data Acquisition and Selection

We searched the TCGA database and other open databases, including Gene Expression Omnibus (GEO), International Cancer Genome Consortium (ICGC), ArrayExpress, Oncomine, etc. Only the TCGA database has an EC-related dataset with both gene profiles and clinical survival information. Clinical data, RNAseq-HTSeq FPKM data (including mRNA and lncRNA profiles), and miRNAseq data of EC patients were downloaded from the TCGA database (https://gdc-portal.nci.nih.gov/) in June 2019, and the dataset obtained contains 548 EC patients. Then, 52 of 548 EC patients were excluded. The reasons for exclusion were as follows: (1) 14 EC patients had no clinical data or mRNA, miRNA, and lncRNA gene expression profiles, and (2) 38 EC patients survived <30 days after the first pathological diagnosis. Eventually, 496 EC patients were included in this study, and the data missing rates for each clinical variable and the expression of each gene were <10%. The missing data of the 496 EC patients included in this study were filled by predictive mean matching. Differential expression analysis was performed based on log2 transformation of RNA expression data. From among all the patients involved in this study, 100 EC patients were randomly selected as the validation group, while the remaining 396 EC patients were used as the training group for the construction of the EC prognostic model. Furthermore, 33 out of the 496 EC patients had mRNA, miRNA, and lncRNA gene expression profiles of paracancerous tissues available, which were used as the control group for differential expression analysis.

Construction of the Prognosis Model

The univariate Cox regression (the 2-sided log-rank test) was applied to analyze 12 clinical variables, including age at initial pathologic diagnosis, height, weight, histologic grade, clinical stage, histologic type, initial pathological diagnosis method, time since last menstruation, neoplasm status, race, surgical approach, and tissue indicator (prospective or retrospective). Clinical variables resulting in a univariate Cox regression $P < 0.05$ were initially screened for inclusion multivariate Cox regression analysis ($\alpha_{in} = 0.10$, $\alpha_{out} = 0.15$). Then, the clinical variables selected by the multivariate Cox regression model were identified, and the prognosis clinical model (model 1) of EC patients based on the identified clinical variables was established. Clinical variables that had a hazard ratio (HR) for death>1 were considered to be risk-increasing clinical variables, and those with HR<1 were defined as protective clinical variables.

The RNA genes related to EC prognosis were screened by the following three steps: (1) A fold change (FC) and false discovery rate (FDR) were applied to identify RNAs with differential expression between the EC patient group (396 patients) and the control group (33 controls). mRNAs, miRNAs, and lncRNAs with a FC>2 or<-2 and FDR<0.05 were screened as differentially expressed RNAs. (2) Univariate Cox regression analysis was used to explore the relationship between the differentially expressed RNAs and the prognosis of EC patients, and the differentially expressed RNA with a univariate Cox regression $P < 0.05$ is considered to be a prognosis-related RNA in EC patients. (3) The three types of RNAs (i.e., mRNA, miRNA, and lncRNA) with $P < 0.05$ identified in the univariate Cox regression analysis were further subjected to elastic net-penalized Cox proportional hazards regression analysis with 10,000 iterations and 10 cross-validation (Zou and Hastie, 2005; Pak et al., 2020). Lastly, the mRNAs, miRNAs, and lncRNAs with a non-zero elastic net-penalized Cox proportional hazards regression coefficient were the final selected RNAs considered to be related to the OS of EC. Then, the ridge regression Cox model was used to fit the selected RNAs (mRNAs, miRNAs, and lncRNAs) to construct

the prognosis models of mRNA (model 2), miRNA (model 3), and lncRNA (model 4), respectively. The integrated RNA molecular prognostic model (model 5) was then constructed by fitting the selected mRNAs, miRNAs, and lncRNAs with the ridge regression Cox model. Eventually, the integrated clinical-RNA prognostic model was established by fitting the screened prognosis-related clinical variables and RNAs with the ridge regression Cox model (model 6). RNAs that had a HR for death>1 were considered to be risk-increasing RNAs, and those with HR<1 were defined as protective RNAs.

Evaluation of the Prognostic Model

The prognostic index (PI) is a weighted linear combination of various factors in the prognostic model. In the prognostic model, the PI value reflects the prognosis of the patient. PI is positively proportional to the risk function. A greater PI value indicates worse prognosis, and conversely, a smaller PI values means better prognosis. Standardization was carried out for PI to obtain a weighted prognostic index (WPI). The formulas used for calculating PI and WPI of each patient are as follows:

$$PI = \sum_{i} (\beta_i \times V_i) \quad (1)$$

$$WPI = \frac{PI - mean(PI)}{SD(PI)}, \quad (2)$$

Where β_i is the regression coefficient of the i-th factor in the model, V_i is the value of the i-th factor of EC patients, and mean (PI) and SD (PI) are the mean and standard deviation (SD) of the PI vector in EC patients, respectively. Applying WPI = 0 as the cutoff point, the patients were classified into two groups in terms of the predicted prognosis. Specifically, patients with WPI≤0 were in the low-risk group, whereas those with WPI > 0 were in the high-risk group. The Kaplan–Meier curves of patients in the high-risk group and low-risk group were drawn and subjected to the log-rank test. $P \leq 0.05$ indicates statistically significant difference in the OS between the two groups. The areas under the time-dependent ROC curves (AUC-ROC) of the six prognostic models were calculated. The model with the greatest AUC value was selected as the optimal prognostic model. An AUC value between 0.7 and 0.9 is generally believed to indicate medium predictive ability, while an AUC value greater than 0.9 indicates relatively ideal predictive ability. The larger is the AUC value, the stronger is the predictive ability of the model.

GO and KEGG Enrichment Analysis

The online tool DAVID (The Database for Annotation, Visualization and Integrated Discovery, version 6.8, http://david.abcc.ncifcrf.gov) was used to perform the GO and KEGG enrichment analysis for mRNAs, miRNA-targeted mRNAs (mRNAs with miRDB database-predicted scores higher than 90), and lncRNA-related mRNAs (Spearman's correlation coefficient r_s >0.50 and $P < 0.05$) in the EC prognostic molecular model. Fisher's exact test was employed to select terms with $P < 0.05$ as significant GO and KEGG pathway terms. GO analysis annotates and classifies genes through biological process (BP), molecular function (MF), and cell composition (CC).

Statistical Analysis

Data analyses in this study were conducted by R, version 3.6.1. The missing data were filled by the "mice" R package (version 3.11.0), and differential expression analysis was performed with the "limma" R package (version 3.26.9). The "survival" R package (version 3.2-3) was applied for univariate Cox regression, multivariate Cox regression, and plotting Kaplan–Meier curves. The elastic net-penalized Cox proportional hazards regression model and the ridge regression Cox model were analyzed using the "glmnet" R package (version 3.0-2). The "timeROC" R package (version 0.4) was used to plot the time-dependent ROC curves and calculate the AUC values, and the "ggplot2" R package (version 3.3.1) was used to generate figures of GO and KEGG analysis.

RESULTS

Workflow

Figure 1 shows the process of our Study. RNA expression data and corresponding clinical variable data from TCGA for EC were analyzed. Cox proportional hazards regression was used to analyze clinical variables related to EC prognosis. RNAs related to EC prognosis were screened by differential expression analysis, univariate Cox proportional hazards regression, and elastic net-penalized Cox proportional hazards regression. The EC prognostic model was constructed by using EC prognostic-related RNAs/clinical variables, and the performance of the prognostic model was evaluated.

Clinical Characteristics and Prognosis Model of EC Patients

Among the 396 EC patients in the training group, 33 patients died by the follow-up deadline and 363 patients survived. The minimum and maximum ages of patients at initial pathological diagnosis were 33 years and 89 years, respectively, with the average age being 64.22 years ($SD = 10.86$ years). Univariate Cox regression analysis indicated that histological grade, clinical stage, and neoplasm status were statistically significant ($P < 0.05$) among the 12 clinical variables. Further multivariate Cox regression analysis was performed for the three factors, and the results suggested that histological grade and neoplasm status are independent prognostic clinical variables of EC, and both of them are EC risk factors (HR > 1). Then, a clinical prognostic model of EC was established based on histological grade and neoplasm status (**Table 1**).

Differentially Expressed and OS-Related RNAs of EC

The EC RNA expression data acquired from TCGA database were preprocessed, and a total of 36,844 RNAs (19,754 mRNA, 2,243 miRNA, and 14,847 lncRNA) were included in this study. 1060 differential expression RNAs were screened by differential expression analysis, including 920 mRNAs (353 upregulation and 567 downregulation, **Figure 2A**), 100 miRNAs (83 upregulation and 17 downregulation, **Figure 2B**), and 40 lncRNAs (21 upregulation and 19 downregulation, **Figure 2C**). Univariate Cox regression was performed on these 1,060

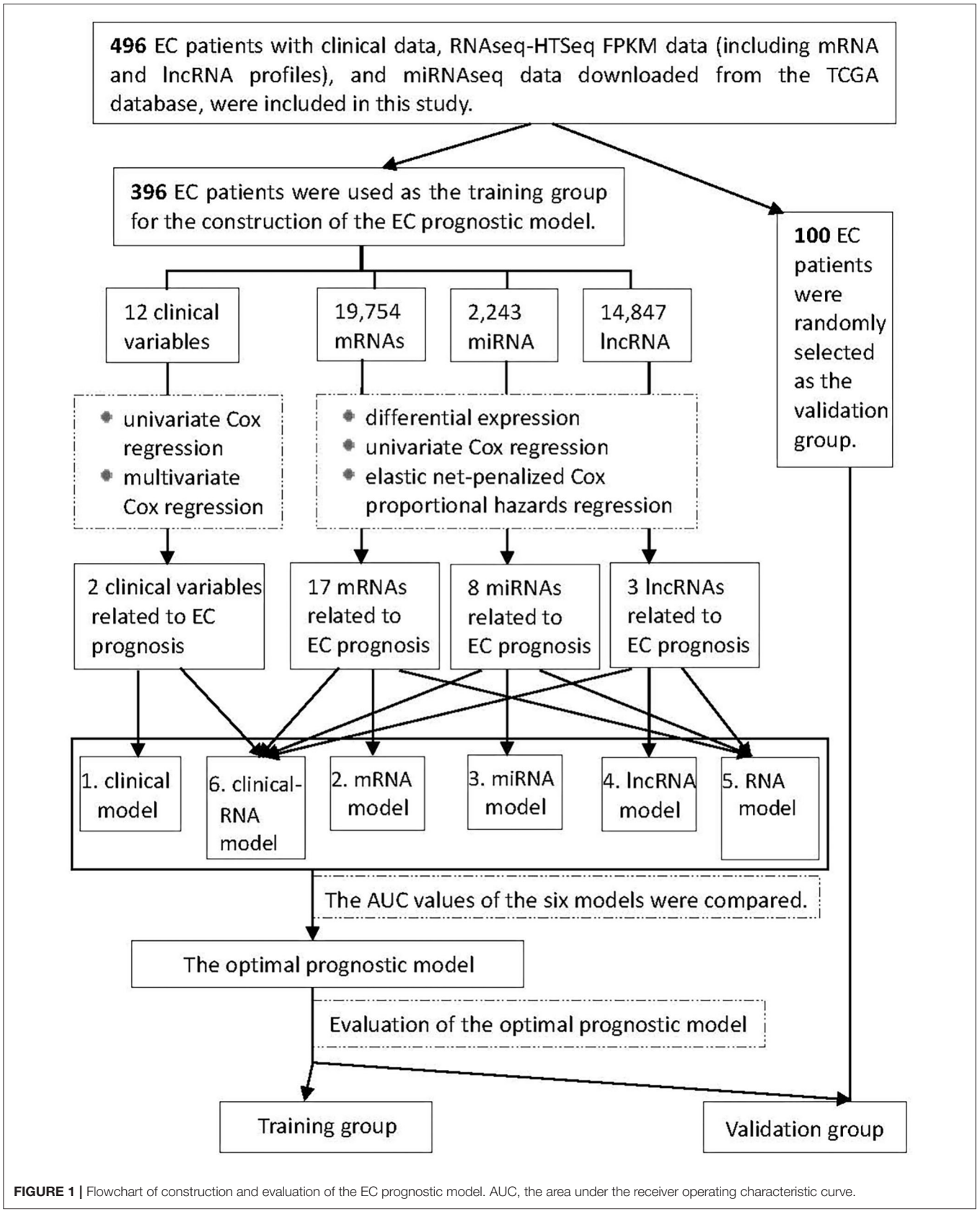


FIGURE 1 | Flowchart of construction and evaluation of the EC prognostic model. AUC, the area under the receiver operating characteristic curve.

TABLE 1 | Survival analysis results of demographic and clinical variables for EC patients in the prognostic model training group (396 patients).

Variables	*n* (%)[a]	Univariate Cox		Multivariate Cox[c]	
		HR[b] (95% CI)	*P*-value	HR[b] (95% CI)	*P*-value
Histological grade		3.27 (1.48–7.23)	0.003	1.89 (0.87–4.10)	0.107
G1	70 (17.68)				
G2	88 (22.22)				
G3	238 (60.10)				
Neoplasm status		22.45 (9.25–54.47)	<0.001	18.24 (7.38–45.08)	<0.001
Yes	319 (84.62)				
No	58 (15.38)				
Data missing	19				
Clinical stage		2.20 (1.61–2.98)	<0.001		
I	241 (60.86)				
II	42 (10.60)				
III	95 (23.99)				
IV	18 (4.55)				
Age at initial pathological diagnosis (years)	1.69 (0.78–3.63)	0.182			
≤60	143 (36.11)				
>60	253 (63.89)				
Height (cm)		1.13 (0.84–1.53)	0.418		
≤155	77 (20.70)				
≤160	102 (27.42)				
≤165	95 (25.54)				
>165	98 (26.34)				
Data missing	24				
Weight (kg)		1.01 (0.73–1.39)	0.969		
≤50	118 (31.13)				
≤60	101 (26.65)				
≤70	85 (22.43)				
>70	75 (19.79)				
Data missing	17				
Histological type[d]					
Serous	88 (22.22)	–	0.111		
Endometrioid	294 (74.24)	1.19 (0.15–9.22)	0.867		
Mixed serous and endometrioid	14 (3.54)	0.56 (0.07–4.17)	0.568		
Initial pathological diagnosis method	1.82 (0.91–3.65)	0.093			
Biopsy	248 (63.42)				
Other	143 (36.58)				
Data missing	5				
Time since last menstruation (months)	0.76 (0.44–1.31)	0.323			
≤6	25 (6.87)				
≤12	9 (2.47)				
>12	330 (90.66)				
Data missing	32				
Race[d]					
Black	87 (23.32)	–	0.649		
White	260 (69.71)	0.39 (0.08–2.00)	0.257		
Asian	17 (4.56)	0.45 (0.11–1.90)	0.277		
Other	9 (2.41)	0.31 (0.04–2.25)	0.249		
Data missing	23				
Surgical approach		1.26 (0.61–2.63)	0.532		
Open surgery	222 (59.04)				

(Continued)

TABLE 1 | Continued

Variables	*n* (%)[a]	Univariate Cox		Multivariate Cox[c]	
		HR[b] (95% CI)	*P*-value	HR[b] (95% CI)	*P*-value
Minimally invasive	154 (40.96)				
Data missing	20				
Tissue collection indicator		0.80 (0.18–3.53)	0.769		
Prospective	79 (19.95)				
Retrospective	317 (80.05)				

[a]*Before missing data is filled.* [b]*Protective RNA had a HR <1 and risky RNA had a HR > 1 in EC patients.* [c]*The multivariate Cox regression analysis ($\alpha_{in} = 0.10$, $\alpha_{out} = 0.15$) was carried out for clinical variables with P <0.05 in the univariate Cox regression analysis.* [d]*Dummy variables were applied.*
EC, endometrial carcinoma; CI, confidence interval; HR, hazard ratio.

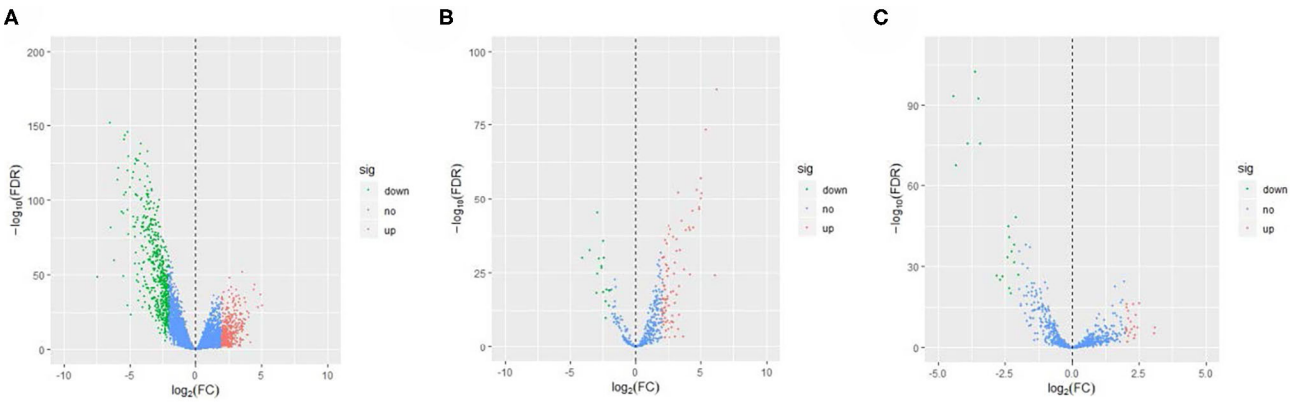


FIGURE 2 | Volcano diagram of the differential expression RNAs in ECs and controls. **(A)** Volcano diagram of mRNA, **(B)** volcano diagram of miRNA, and **(C)** volcano diagram of lncRNA.

differentially expressed RNAs individually, and there were 126 RNAs with $P < 0.05$ (115 mRNAs, 8 miRNAs, and 3 lncRNAs). Subsequently, the whole 126 RNAs were subjected to elastic net-penalized Cox proportional hazard regression analysis, and 17 RNAs related to EC prognosis were selected, including 15 mRNAs (*ANGPTL1*, *ALDH1A1*, *FIBIN*, *GFPT2*, *HIST1H3H*, *HOXD8*, *IGFBP5*, *MAL*, *MMP1*, *PRKAR2B*, *PROM2*, *SCARA3*, *SNAP25*, *TFPI*, and *TSPYL5*), and 2 miRNAs (*has-miR-215-5p* and *has-miR-592*). There is no lncRNA in the 17 RNAs. We do not think it can reflect the whole picture of RNAs associated with EC prognosis. Recent studies have shown that although lncRNA does not encode proteins, lncRNA participates in gene expression regulation at various levels, such as transcriptional regulation and posttranscriptional regulation. The abnormal expression of lncRNA is usually associated with the occurrence, recurrence, and metastasis of tumors (Kopp and Mendell, 2018). Therefore, we used elastic net-penalized Cox proportional hazard regression to screen the three types of RNAs (mRNA, miRNA, and lncRNA), respectively. Eventually, 28 EC OS-related RNAs (17 mRNAs, 8 miRNAs, and 3 lncRNAs) were identified (**Figure 3**, **Table 2**). These 28 EC OS-related RNAs contain all 17 RNAs screened by elastic net-penalized Cox proportional hazard regression using whole genes.

RNA Molecular Prognostic Models of EC

The EC OS-related 17 mRNAs, 8 miRNAs, and 3 lncRNAs were fitted with the ridge regression Cox model respectively to obtain the corresponding mRNA prognostic model, miRNA prognostic model, and lncRNA prognostic model. The 28 EC OS-related RNAs were fitted with the ridge regression Cox model, and an integrated mRNA–miRNA–lncRNA molecular prognostic model was established (**Table 3**).

Integrated Clinical-RNA Prognostic Model of EC

The EC prognosis-related 28 RNAs and 2 clinical variables were fitted with the ridge regression Cox model, and an integrated clinical-RNA prognostic model was established (**Table 3**). As listed in **Table 4**, the AUC (95% CI) values of the lncRNA model based on the selected three lncRNAs at 1, 3, and 5 years were 0.823 (0.719–0.928), 0.646 (0.505–0.788), and 0.737 (0.608–0.870), respectively. These results suggest that the three lncRNAs have a certain predictive effect on the prognosis of EC. The AUC-ROC showed that except for the lncRNA molecular model (model 4) with a 3-year prognosis AUC of 0.646 (<0.7), the minimum AUC value of other models was 0.733. The AUC of the integrated RNA molecular prognostic model (model 5) was ≥0.821, which is greater than the AUC of mRNA, miRNA, and lncRNA models,

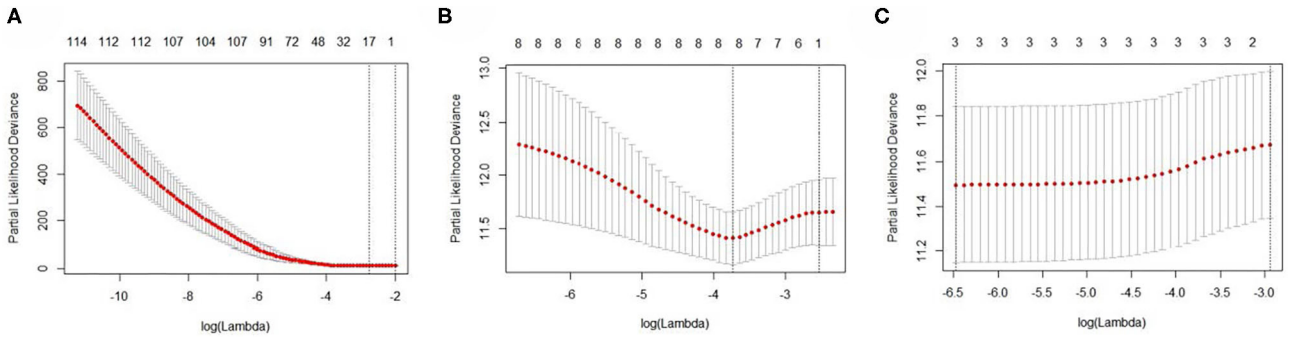


FIGURE 3 | The cross-validation curves of the elastic net-penalized Cox proportional hazard regression. **(A)** Cross-validation curve for mRNA, **(B)** cross-validation curve for miRNA, and **(C)** cross-validation curve for lncRNA.

TABLE 2 | The 28 identified EC prognosis-related RNAs.

Number	Gene symbol	HR[a]	95% CI	*P*-value[b]	Regulation[c]	Coefficient[d]
mRNA						
1	*ALDH1A1*	1.001	1.001–1.002	<0.001	Down	0.0002
2	*ANGPTL1*	1.001	1.001–1.002	<0.001	Down	0.0018
3	*COL4A6*	1.113	1.061–1.169	<0.001	Down	0.0324
4	*FIBIN*	1.028	1.016–1.040	<0.001	Down	0.0024
5	*GFPT2*	1.026	1.014–1.038	<0.001	Down	0.0042
6	*HIST1H3H*	1.011	1.005–1.016	<0.001	Up	0.0059
7	*HOXD8*	1.019	1.009–1.028	<0.001	Down	0.0022
8	*IGFBP5*	1.001	1.0005–1.0014	<0.001	Down	0.0001
9	*MAL*	1.001	1.001–1.002	<0.001	Up	0.0005
10	*MMP1*	1.006	1.003–1.009	<0.001	Up	0.0005
11	*PRKAR2B*	1.036	1.016–1.055	<0.001	Down	0.0034
12	*PROM2*	1.008	1.004–1.012	<0.001	Up	0.0031
13	*RAB26*	1.020	1.008–1.032	0.001	Up	0.0001
14	*SCARA3*	1.003	1.002–1.005	<0.001	Down	0.0011
15	*SNAP25*	1.125	1.075–1.177	<0.001	Down	0.0567
16	*TFPI*	1.057	1.030–1.084	<0.001	Down	0.0114
17	*TSPYL5*	1.015	1.008–1.022	<0.001	Down	0.0078
miRNA						
18	*hsa-miR-141-3p*	1.034	1.001–1.068	0.041	Up	−0.0002
19	*hsa-miR-191-5p*	0.999	0.999–1.000	0.048	Up	−0.00002
20	*hsa-miR-192-5p*	1.000	1.00001–1.00007	0.011	Up	0.00002
21	*hsa-miR-215-5p*	1.003	1.001–1.005	<0.001	Up	0.00205
22	*hsa-miR-3170*	0.875	0.779–0.983	0.024	Up	−0.0465
23	*hsa-miR-3613-5p*	0.963	0.931–0.997	0.034	Up	−0.0143
24	*hsa-miR-592*	1.006	1.003–1.009	0.001	Up	0.00483
25	*hsa-miR-7-5p*	1.034	1.001–1.068	0.041	Up	0.03034
lncRNA						
26	*DNM3OS*	1.057	1.004–1.112	0.036	Down	0.05966
27	*FAM83H-AS1*	1.014	1.001–1.027	0.042	Up	0.01465
28	*RP11-295G20.2.1*	1.010	1.000–1.020	0.049	Up	0.00946

[a]Protective RNA had a HR <1 and risky RNA had a HR > 1 in EC patients. [b]Univariate Cox regression P value <0.05 was considered statistically significant. [c]Type of regulation (upregulated or downregulated) in ECs vs controls. [d]Elastic net-regulated Cox regression coefficient.
EC, endometrial carcinoma; HR, hazard ratio; CI, confidence interval.

TABLE 3 | Ridge regression Cox prognostic model.

Number	Variable	Coefficient[a]	Coefficient[b]	Coefficient[c]	Coefficient[d]	Coefficient[e]
mRNA						
1	*ALDH1A1*	0.000460	-	-	0.0003	0.0003
2	*ANGPTL1*	0.005925	-	-	0.0056	0.0033
3	*COL4A6*	0.033186	-	-	0.0322	0.0343
4	*FIBIN*	0.004171	-	-	0.0053	0.0037
5	*GFPT2*	0.005855	-	-	0.0058	0.0033
6	*HIST1H3H*	0.006833	-	-	0.0068	0.0071
7	*HOXD8*	0.003590	-	-	0.0039	0.0034
8	*IGFBP5*	0.000262	-	-	0.0003	0.0003
9	*MAL*	0.000456	-	-	0.0004	0.0005
10	*MMP1*	0.001393	-	-	0.0011	0.0004
11	*PRKAR2B*	0.009734	-	-	0.0097	0.0043
12	*PROM2*	0.002367	-	-	0.0020	0.0018
13	*RAB26*	0.005232	-	-	0.0051	0.004
14	*SCARA3*	0.001498	-	-	0.0014	0.0019
15	*SNAP25*	0.044875	-	-	0.0377	0.0333
16	*TFPI*	0.015739	-	-	0.0141	0.0173
17	*TSPYL5*	0.008143	-	-	0.0079	0.0066
miRNA						
18	*hsa-miR-141-3p*	-	−0.000121699	-	−0.00006	−0.00007
19	*hsa-miR-191-5p*	-	−0.000123281	-	−0.00004	−0.00003
20	*hsa-miR-192-5p*	-	2.08217E-05	-	0.00001	0.00001
21	*hsa-miR-215-5p*	-	0.001517209	-	0.0009	0.0009
22	*hsa-miR-3170*	-	−0.02585058	-	−0.019	−0.0256
23	*hsa-miR-3613-5p*	-	−0.00941356	-	−0.0058	−0.0063
24	*hsa-miR-592*	-	0.00339596	-	0.0033	0.005
25	*hsa-miR-7-5p*	-	0.01726421	-	0.0144	0.0159
lncRNA						
26	*DNM3OS*	-	-	0.054349555	0.0044	0.0074
27	*FAM83H-AS1*	-	-	0.012946912	0.0023	0.0021
28	*RP11-295G20.2.1*	-	-	0.008863742	0.0034	0.0042
Clinical						
29	Histologic grade	-	-	-	-	0.1432
30	Neoplasm status	-	-	-	-	1.1664

[a]*Coefficient of the mRNA ridge regression Cox model (model 2).* [b]*Coefficient of the miRNA ridge regression Cox model (model 3).* [c]*Coefficient of the lncRNA ridge regression Cox model (model 4).* [d]*Coefficient of the mRNA–miRNA–lncRNA ridge regression Cox model (model 5).* [e]*Coefficient of the clinical-RNA ridge regression Cox model (model 6).*

suggesting that the integrated RNA molecular prognostic model is superior to the mRNA or miRNA or lncRNA model in terms of predictive ability. As for the clinical prognostic model, the AUC value was ≥0.830, implying that the two clinical variables (i.e., histological grade and neoplasm status) screened in this study can predict the prognosis of EC. The 1-, 3-, and 5-year AUC values of the integrated clinical-RNA prognostic model (model 6) were ≥0.919, being greater than the AUC values of other models at the same time point. This indicates that the integrated clinical-RNA prognostic model has the best predictive ability among the six models (**Figure 4**).

The integrated clinical-RNA prognostic model was used to calculate the WPI value of each EC patient in the training group (396 patients). The WPI value of EC patients ranged from −1.957 to 4.831. Taking WPI = 0 as the cutoff point, the EC patients were divided into the high-risk group (147 patients) and low-risk group (249 patients) (**Figure 5A**). The difference between the two groups' Kaplan–Meier curves was statistically significant (P = 8e-14). The prognosis of EC patients was worse in the high-risk group than in the low-risk group [HR = 32.263, (95% CI, 7.707–135.058)], suggesting that the integrated clinical-RNA prognostic model enables accurate prediction of the prognosis of EC patients (**Figure 5C**). Furthermore, this model was used to calculate the WPI value of each EC patient in the validation group (100 patients) to predict the prognosis of EC patients. The WPI value of EC patients in the validation group ranged from −1.455 to 3.822. Taking WPI = 0 as the cutoff point, the EC patients in the validation group were divided into the high-risk

TABLE 4 | 1-, 3-, and 5-year AUC of the EC prognostic model.

Model	1-year		3-year		5-year	
	AUC	95% CI	AUC	95% CI	AUC	95% CI
1. Clinical model	0.830	0.695–0.964	0.872	0.808–0.936	0.897	0.828–0.966
2. mRNA model	0.894	0.764–1.025	0.756	0.634–0.879	0.783	0.664–0.903
3. miRNA model	0.733	0.579–0.888	0.761	0.653–0.869	0.763	0.646–0.879
4. lncRNA model	0.823	0.719–0.928	0.646	0.505–0.788	0.737	0.608–0.870
5. RNA model	0.927	0.813–1.042	0.821	0.717–0.925	0.836	0.733–0.938
6. Clinical-RNA model	0.979	0.949–1.008	0.919	0.860–0.978	0.932	0.875–0.989

AUC, the area under the receiver operating characteristic curve; CI, confidence interval.

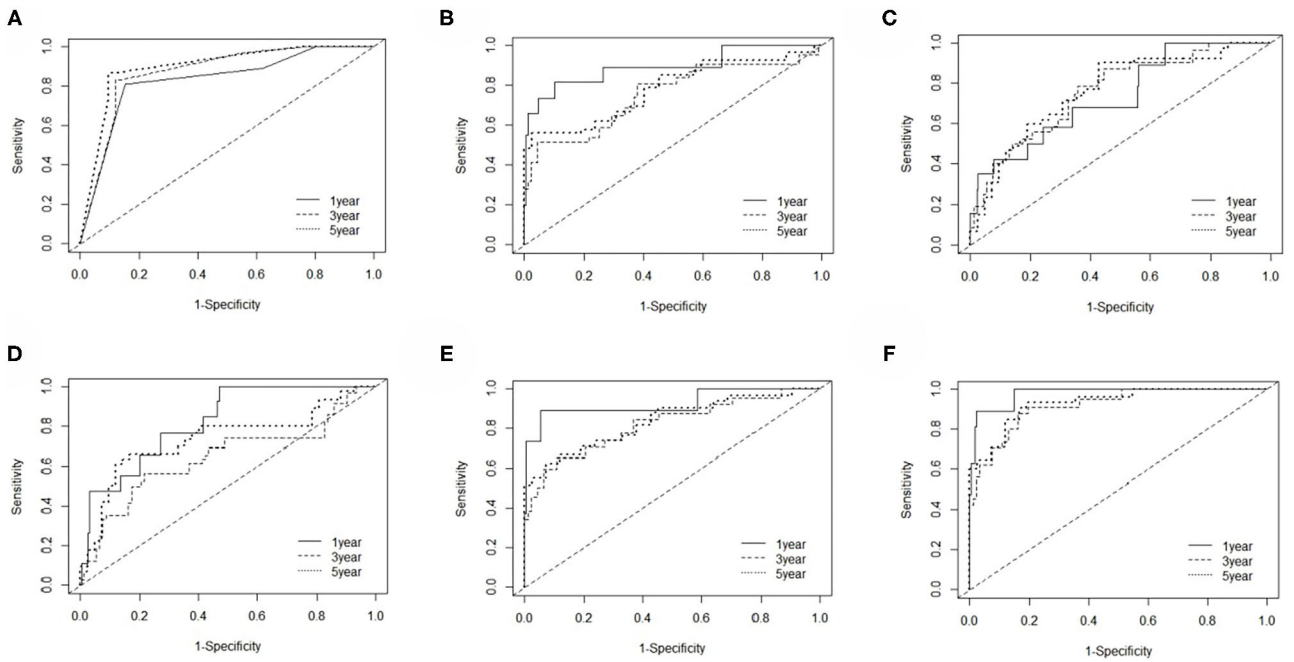


FIGURE 4 | The time-dependent receiver operating characteristic curves. **(A–F)** The 1-, 3-, and 5-year ROC curves of the clinical model, mRNA model, miRNA model, lncRNA model, RNA model, and clinical-RNA model.

group (41 patients) and low-risk group (59 patients) (**Figure 5B**). The difference between the two groups' Kaplan–Meier curves was statistically significant ($P = 0.0052$). The prognosis of EC patients in the high-risk group was worse than that of patients in the low-risk group [HR = 6.674, (95% CI, 1.437–30.995)], showing that the integrated clinical-RNA prognostic model also has satisfactory accuracy in predicting the prognosis of EC patients in the validation group (**Figure 5D**).

Functional Analysis of EC Prognosis-Related Genes

Taking the union of the 17 mRNAs, the 371 target mRNAs of the 8 miRNAs, and the 300 mRNAs related to the 3 lncRNAs in the RNA prognostic model, a gene set with 652 mRNAs was obtained and subjected to GO and KEGG analyses. The GO analysis results show that in biological processes, target genes are mainly enriched in "signaling," "positive regulation of RNA polymerase II promoter transcription," and "negative regulation of RNA polymerase II promoter transcription" (**Figure 6A**). In terms of cellular composition, the target genes are mainly located in the "cytoplasm" and "plasma membrane" (**Figure 6B**). As for the molecular function, "protein binding" is the most important mode (**Figure 6C**). According to the results of KEGG analysis, the top three pathways are pathways in cancer, the PI3K-Akt signaling pathway, and the focal adhesion pathway (**Figure 7**).

DISCUSSION

At present, researchers have been searching for EC prognosis-related biomarkers and establishing EC prognostic prediction model with higher accuracy to provide better clues for formulating reasonable individualized treatment plans, thereby

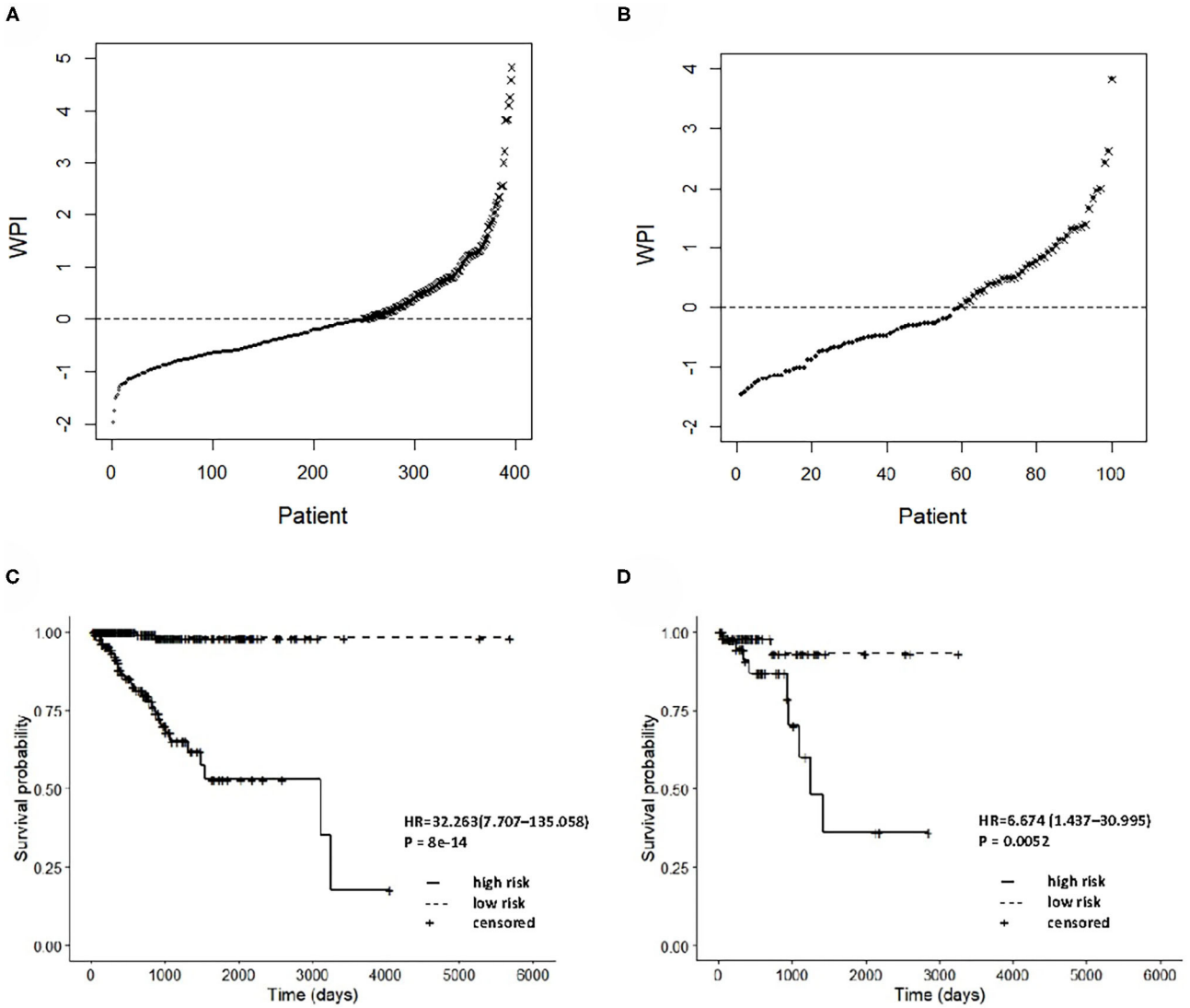


FIGURE 5 | Prognostic performance assessment of integrated clinical-RNA prognostic model. WPI distributions and stratifications of EC patients in training set **(A)** and validating set **(B)**, respectively. Kaplan–Meier curves for stratifications in training set **(C)** and validating set **(D)**, respectively. WPI, weighed prognostic index; HR, hazard ratio.

improving patients' prognostic quality of life. After acquiring the clinical data of EC patients from the TCGA database and the full mRNA, miRNA, and lncRNA genome expression profiles, 30 factors related to EC prognosis, including two clinical variables and 28 RNAs, were identified in this study. Based on the 30 EC prognosis-related factors, 6 EC prognosis models were established: clinical model, mRNA model, miRNA model, lncRNA model, integrated RNA model, and integrated clinical-RNA model. The clinical-RNA model displayed the highest AUC ($\geq$ 0.919), indicating the strongest predictive ability among the six prognostic models.

Previous studies have shown that the clinical variables related to the prognosis of EC include pathological grade, pathological stage, FIGO stage, age at initial pathological diagnosis, degree of muscular invasion, vascular tumor thrombus, and lymph node metastasis (Braun et al., 2016; Morice et al., 2016). Our study suggests that histological grade and neoplasm status are independent prognostic factors for EC overall survival.

There were 17 prognosis-related mRNAs with HR>1, indicating that an increased expression level of these mRNAs will increase the risk of death in EC patients. Most of these genes are reportedly related to the occurrence, development, or prognosis of cancer. The expression of *ALDH1A1* is upregulated in endometrial carcinoma cells (Shiba et al., 2019), and *ALDH1A1* is a confirmed oncogene for lung cancer (Gao et al., 2015). As a member of angiopoietin-like protein genes, *ANGPTL1* acts as a tumor-suppressor gene in various tumors (Chen H. A. et al., 2016). The absence of *COL4A6* may cause familial

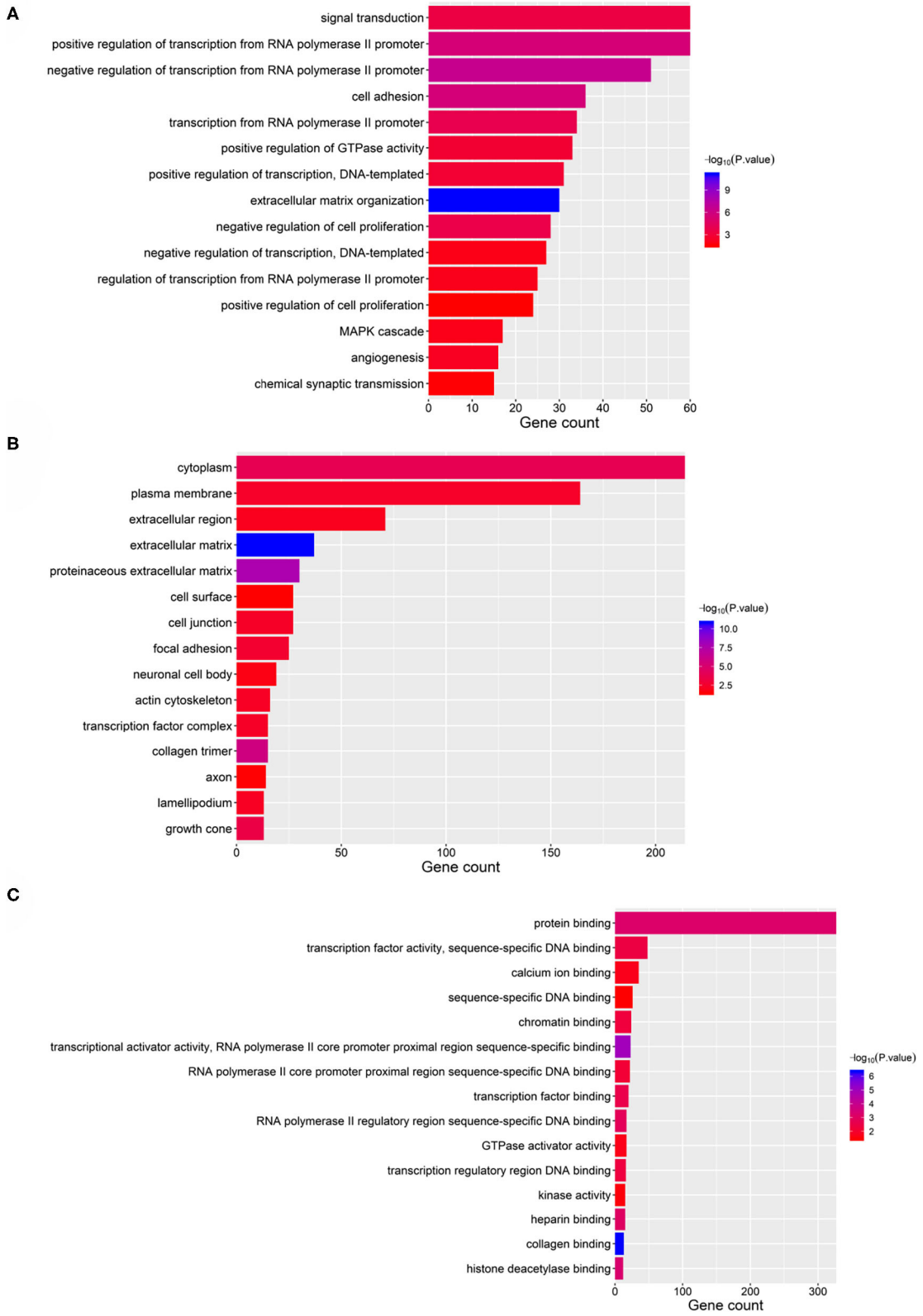


FIGURE 6 | GO analysis of the 28 validated genes. **(A)** Biological process (BP), **(B)** cellular component (CC), and **(C)** molecular function (MF).

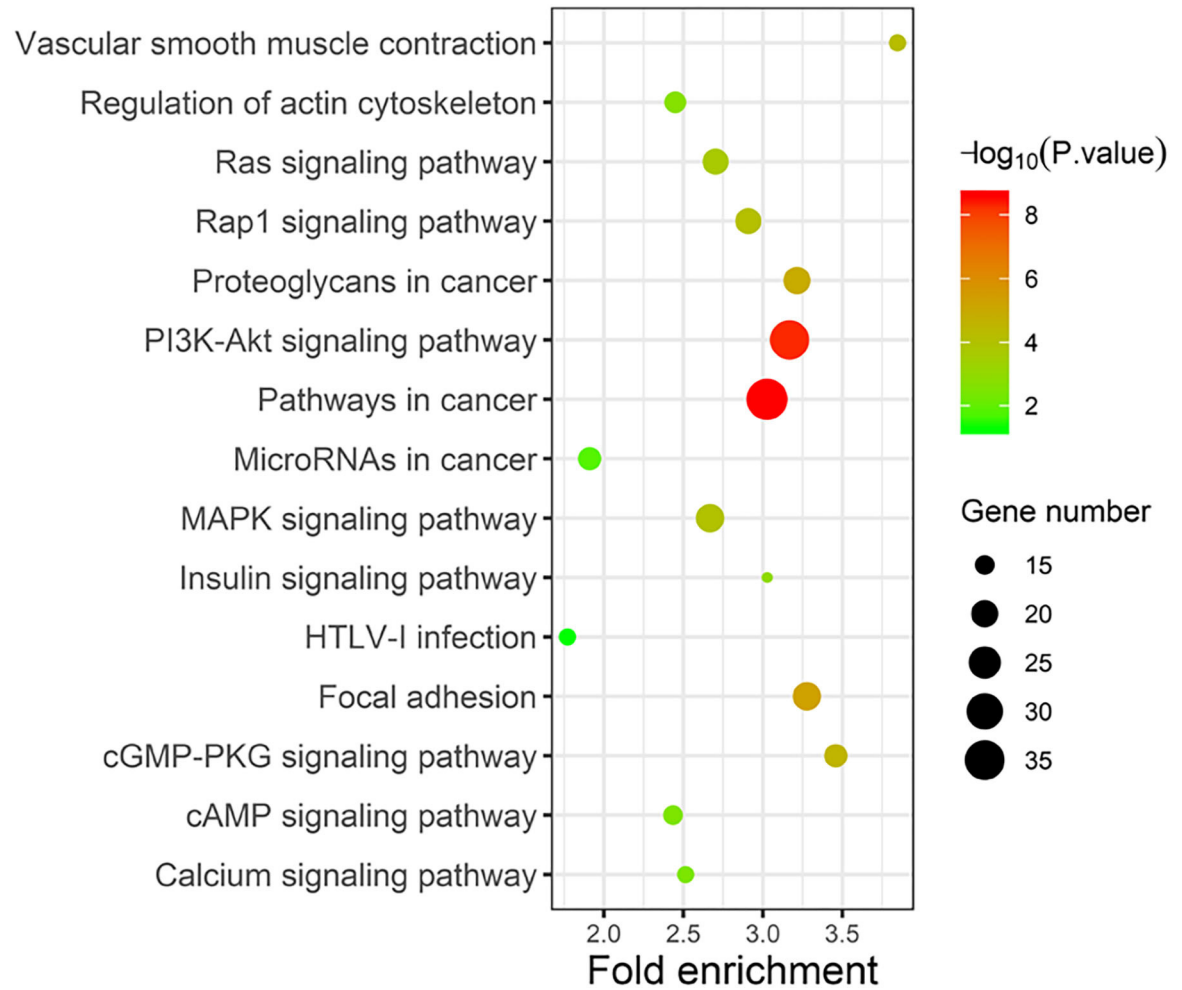


FIGURE 7 | KEGG analysis of the 28 validated genes.

hemorrhagic nephritis (Murata et al., 2016). *GFPT2* is highly expressed in lung cancer (Zhang et al., 2018). *HIST1H3H* is a histone gene, and its high expression is related to the OS, relapse-free survival (RFS), and distant metastasis-free survival (DMFS) of breast cancer patients (Xie et al., 2019). *HOXD8* belongs to a homeobox gene family and is closely related to cell proliferation, apoptosis, and cell cycle. Studies have found that *HOXD8* is a downstream target gene of *miR-5692a*. *MiR-5692a* plays the role of an oncogene in the occurrence and development of liver cancer by regulating the expression of *HOXD8* (Sun et al., 2019). *IGFBP5* is a tumor-suppressor gene for leukemia, osteosarcoma, breast cancer, and pancreatic cancer and participates in cell biological functions, such as cell metastasis and apoptosis (Baxter, 2014). Hypermethylation of *MAL* in cervical intraepithelial neoplasia accelerates cervical lesions (Meršaková et al., 2018). High expression of *MMP1* in cancer tissues leads to accelerated angiogenesis, thus promoting the proliferation and migration of cancer cells (Pahwa et al., 2014). It has been clarified that the *PRKAR2B* gene is overexpressed in castration-resistant prostate cancer (CRPC), which mainly promotes cell-cycle biological processes, accelerates CRPC cell proliferation and invasion, and inhibits CRPC cell apoptosis (Sha et al., 2017). The expression of *PROM2* (prominin 2) is upregulated in kidney cancer and melanoma (Rohan et al., 2006; Winnepenninckx et al., 2006). The function of *RAB26* is mainly related to membrane transport and cell autophagy. Some studies have found that *RAB26* can affect the metastasis and invasion of breast cancer (Schwartz et al., 2007). *SCARA3* inhibits the lethal effect of dexamethasone and bortezomib on myeloma cells (Brown et al., 2013). *SNAP25* is mainly involved in the occurrence and development of mental diseases (González-Giraldo and Forero, 2020). *TFPI* reduces tumor cell-induced coagulation activation and lung metastasis, and it has shown inhibitory effect on primary and metastatic tumors in mice (Hembrough et al., 2003; Amirkhosravi et al., 2007). Some studies suggest that the methylation of the tumor-suppressor gene *TSPYL5* will cause its expression silencing and, thereby, gastric cancer (Jung et al., 2008).

In this study, eight EC prognosis-related miRNAs were identified. The HR values of *miR-141-3p*, *miR-192-5p*, *miR-215-5p*, *miR-592*, and *miR-7-5p* were greater than 1, indicating that the five miRNAs are highly expressed in EC, acting as tumor genes and prognostic risk factors. *MiR-141-3p* acts as a tumor-suppressor gene in colorectal cancer and enhances the sensitivity of colorectal cancer cells to cetuximab by inhibiting *EGFR* (Xing et al., 2020). *MiR-192-5p* plays different roles in different cancers, e.g., it is highly expressed in gastric cancer and pancreatic ductal cancer, while the expression is low in lung cancer (Feng et al., 2011; Zhao et al., 2013; Chen et al., 2014). Overexpression of *miR-215-5p* in colorectal cancer leads to G2/M phase cell-cycle arrest and *p53*-dependent apoptosis induction, thus reducing the proliferation and migration of colorectal cancer cells (Vychytilova-Faltejskova et al., 2017). The biological function of *miR-592* varies according to the cancer type. Its overexpression in liver cancer inhibits the proliferation and metastasis of cancer cells, while the opposite effect is observed in prostate cancer (Wang et al., 2012; Lv et al., 2015). Studies have shown that *miR-7-5p* can inhibit tumor development by regulating the PI3K/Akt pathway and the expression of the target gene *KLF4* (Fang et al., 2012; Okuda et al., 2013). The other three miRNAs (i.e., *miR-191-5p*, *miR-3170*, and *miR-3613-5p*) have HR values lower than 1, indicating that these three genes are protective factors for EC prognosis, i.e., their high expression reduces the risk of death in EC patients. The overexpression of *miR-191-5p* in lung adenocarcinoma downregulates Wnt signaling via the target gene *SATB1*, thus blocking lung cancer cell migration and proliferation (Zhou et al., 2020). Studies have shown that the prognosis is better in EC patients with high expression of *miR-3170* than in those with low expression of *miR-3170* (Wang Y. et al., 2019). The expression level of *miR-3613-5p* in the serum of patients with endometriosis is significantly reduced (Cosar et al., 2019).

The HR values of the three identified EC prognosis-related lncRNAs (e.g., *DNM3OS*, *FAM83H-AS1*, and *RP11-295G20.2.1*) are all greater than 1, indicating that they are prognostic risk factors for EC. Studies have observed that the expression of *DNM3OS* is upregulated in gastric cancer tissues and cell lines. Knocking out *DNM3OS* hinders snail-mediated epithelial-to-mesenchymal transition, thereby inhibiting the proliferation, migration, and invasion of gastric cancer cells (Wang S. et al., 2019). *FAM83H-AS1* promotes radiation resistance and metastasis of ovarian cancer via targeted HuR protein (Dou et al., 2019). At present, the specific biological function of *RP11-295G20.2.1* is not clear, and its relationship with the occurrence, development, and prognosis of EC needs to be confirmed by further experimental research.

CONCLUSIONS

In summary, 28 RNAs that are related to the prognosis of EC patients were identified in this study, and a clinical-mRNA–miRNA–lncRNA prognostic model for EC patients was established. The predictive ability of this clinical-RNA model is significantly better than the clinical-alone model and RNA-alone model in terms of prognosis prediction for EC patients. This study provides a scientific basis for discovering new prognostic markers for EC patients, clarifying the molecular mechanism of EC prognosis, and improving prognosis and clinical management of EC patients.

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/s.

AUTHOR CONTRIBUTIONS

Conceptualization: XZ and XP. Data curation: FD. Formal analysis: FD and JM. Methodology: CQ. Software: FD and CQ. Writing—original draft: FD, JM, CQ, FY, XL, and XZ. Writing—review and editing: XP.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant Number 81472860).

ACKNOWLEDGMENTS

We thank The Cancer Genome Atlas project for contributing the data for analyses of our study.

REFERENCES

Amirkhosravi, A., Meyer, T., Amaya, M., Davila, M., Mousa, S. A., Robson, T., et al. (2007). The role of tissue factor pathway inhibitor in tumor growth and metastasis. *Semin. Thromb Hemost.* 33, 643–652. doi: 10.1055/s-2007-991531

Baxter, R. C. (2014). IGF binding proteins in cancer: mechanistic and clinical insights. *Nat. Rev. Cancer* 14, 329–341. doi: 10.1038/nrc3720

Bell, D. W., and Ellenson, L. H. (2019). Molecular genetics of endometrial carcinoma. *Annu. Rev. Pathol.* 14, 339–367. doi: 10.1146/annurev-pathol-020117-043609

Braun, M. M., Overbeek-Wager, E. A., and Grumbo, R. J. (2016). Diagnosis and management of endometrial cancer. *Am. Fam. Phys.* 93, 468–474.

Brown, C. O., Schibler, J., Fitzgerald, M. P., Singh, N., Salem, K., Zhan, F., et al. (2013). Scavenger receptor class A member 3 (SCARA3) in disease progression and therapy resistance in multiple myeloma. *Leuk. Res.* 37, 963–969. doi: 10.1016/j.leukres.2013.03.004

Chen, H. A., Kuo, T. C., Tseng, C. F., Ma, J. T., Yang, S. T., Yen, C. J., et al. (2016). Angiopoietin-like protein 1 antagonizes MET receptor activity to repress sorafenib resistance and cancer stemness in hepatocellular carcinoma. *Hepatology* 64, 1637–1651. doi: 10.1002/hep.28773

Chen, Q., Ge, X., Zhang, Y., Xia, H., Yuan, D., Tang, Q., et al. (2014). Plasma miR-122 and miR-192 as potential novel biomarkers for the early detection of distant metastasis of gastric cancer. *Oncol. Rep.* 31, 1863–1870. doi: 10.3892/or.2014.3004

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA. Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Colombo, N., Creutzberg, C., Amant, F., Bosse, T., González-Martín, A., Ledermann, J., et al. (2016). ESMO-ESGO-ESTRO Consensus conference on endometrial cancer: diagnosis, treatment and follow-up. *Ann. Oncol.* 27, 16–41. doi: 10.1093/annonc/mdv484

Cosar, E., Mamillapalli, R., Moridi, I., Duleba, A., and Taylor, H. S. (2019). Serum microRNA biomarkers regulated by simvastatin in a primate model of endometriosis. *Reprod. Sci.* 26, 1343–1350. doi: 10.1177/1933719118765971

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233

Dou, Q., Xu, Y., Zhu, Y., Hu, Y., Yan, Y., and Yan, H. (2019). LncRNA FAM83H-AS1 contributes to the radioresistance, proliferation, and metastasis in ovarian cancer through stabilizing HuR protein. *Eur. J. Pharmacol.* 852, 134–141. doi: 10.1016/j.ejphar.2019.03.002

Fang, Y., Xue, J. L., Shen, Q., Chen, J., and Tian, L. (2012). MicroRNA-7 inhibits tumor growth and metastasis by targeting the phosphoinositide 3-kinase/Akt pathway in hepatocellular carcinoma. *Hepatology* 55, 1852–1862. doi: 10.1002/hep.25576

Fanning, J., Brown, S., Phibbs, G., Kramer, T., and Zaher, A. (2002). Immunohistochemical evaluation is not prognostic for recurrence in fully staged high-risk endometrial cancer. *Int. J. Gynecol. Cancer* 12, 286–289. doi: 10.1136/ijgc-00009577-200205000-00008

Feng, S., Cong, S., Zhang, X., Bao, X., Wang, W., Li, H., et al. (2011). MicroRNA-192 targeting retinoblastoma 1 inhibits cell proliferation and induces cell apoptosis in lung cancer cells. *Nucleic Acids Res.* 39, 6669–6678. doi: 10.1093/nar/gkr232

Frederick, P. J., and Straughn, J. M. (2009). The role of comprehensive surgical staging in patients with endometrial cancer. *Cancer Control* 16, 23–29. doi: 10.1177/107327480901600104

Gao, F., Zhou, B., Xu, J. C., Gao, X., Li, S. X., Zhu, G. C., et al. (2015). The role of LGR5 and ALDH1 Al in non-small cell lung cancer: cancer progression and prognosis. *Biochem. Biophys. Res. Commun.* 462, 91–98. doi: 10.1016/j.bbrc.2015.04.029

González-Giraldo, Y., and Forero, D. A. (2020). A functional SNP in the synaptic SNAP25 gene is associated with impulsivity in a Colombian sample. *3 Biotech* 10:134. doi: 10.1007/s13205-020-2110-0

He, Z., Xu, H., Meng, Y., and Kuang, Y. (2017). miR-944 acts as a prognostic marker and promotes the tumor progression in endometrial cancer. *Biomed. Pharmacother.* 88, 902–910. doi: 10.1016/j.biopha.2017.01.117

Hembrough, T. A., Swartz, G. M., Papathanassiu, A., Vlasuk, G. P., Rote, W. E., Green, S. J., et al. (2003). Tissue factor/factor VIIa inhibitors block angiogenesis and tumor growth through a non-hemostatic mechanism. *Cancer Res.* 63, 2997–3000.

Jayaraman, M., Radhakrishnan, R., Mathews, C. A., Yan, M., Husain, S., Moxley, K. M., et al. (2017). Identification of novel diagnostic and prognostic miRNA signatures in endometrial cancer. *Genes Cancer* 8, 566–576. doi: 10.18632/genesandcancer.144

Jeon, Y. T., Park, I. A., Kim, Y. B., Kim, J. W., Park, N. H., Kang, S. B., et al. (2006). Steroid receptor expressions in endometrial cancer: clinical significance and epidemiological implication. *Cancer Lett.* 239, 198–204. doi: 10.1016/j.canlet.2005.08.001

Jung, Y., Park, J., Bang, Y. J., and Kim, T. Y. (2008). Gene silencing of TSPYL5 mediated by aberrant promoter methylation in gastric cancers. *Lab. Invest.* 88, 153–160. doi: 10.1038/labinvest.3700706

Kopp, F., and Mendell, J. T. (2018). Functional classification and experimental dissection of long noncoding RNAs. *Cell* 172, 393–407. doi: 10.1016/j.cell.2018.01.011

Lv, Z. H., Rao, P., and Li, W. (2015). MiR-592 represses FOXO3 expression and promotes the proliferation of prostate cancer cells. *Int. J. Clin. Exp. Med.* 8, 15246–15253.

Meršaková, S., Holubeková, V., Grendár, M., Višnovský, J., and Nachajová, M., Kalman, M., et al. (2018). Methylation of CADM1 and MAL together with HPV status in cytological cervical specimens serves an important role in the progression of cervical intraepithelial neoplasia. *Oncol. Lett.* 16, 7166–7174. doi: 10.3892/ol.2018.9505

Morice, P., Leary, A., Creutzberg, C., Abu-Rustum, N., and Darai, E. (2016). Endometrial cancer. *Lancet* 387, 1094–1108. doi: 10.1016/S0140-6736(15)00130-0

Murata, T., Katayama, K., Oohashi, T., Jahnukainen, T., Yonezawa, T., Sado, Y., et al. (2016). COL4A6 is dispensable for autosomal recessive Alport syndrome. *Sci. Rep.* 6:29450. doi: 10.1038/srep29450

Obel, J. C., Friberg, G., and Fleming, G. F. (2006). Chemotherapy in endometrial cancer. *Clin. Adv. Hematol. Oncol.* 4, 459–468.

Okuda, H., Xing, F., Pandey, P. R., Sharma, S., Watabe, M., Pai, S. K., et al. (2013). miR-7 suppresses brain metastasis of breast cancer stem-like cells by modulating KLF4. *Cancer Res.* 73, 1434–1444. doi: 10.1158/0008-5472.CAN-12-2037

Pahwa, S., Stawikowski, M. J., and Fields, G. B. (2014). Monitoring and inhibiting MT1-MMP during cancer initiation and progression. *Cancers* 6, 416–435. doi: 10.3390/cancers6010416

Pak, K., Oh, S., Goh, T. S., Heo, H. J., Han, M., Jeong, D. C., et al. (2020). A user-friendly, web-based integrative tool (ESurv) for survival analysis: development and validation study. *J. Med. Internet Res.* 22:e16084. doi: 10.2196/16084

Rohan, S., Tu, J. J., Kao, J., Mukherjee, P., Campagne, F., Zhou, X. K., et al. (2006). Gene expression profiling separates chromophobe renal cell carcinoma from oncocytoma and identifies vesicular transport and cell junction proteins as differentially expressed genes. *Clin. Cancer Res.* 12, 6937–6945. doi: 10.1158/1078-0432.CCR-06-1268

Schwartz, S. L., Cao, C., Pylypenko, O., Rak, A., and Wandinger-Ness, A. (2007). RabGTPases at a glance. *J. Cell Sci.* 120, 3905–3910. doi: 10.1242/jcs.015909

Sha, J., Xue, W., Dong, B., Pan, J., Wu, X., Li, D., et al. (2017). PRKAR2B plays an oncogenic role in the castration-resistant prostate cancer. *Oncotarget* 8, 6114–6129. doi: 10.18632/oncotarget.14044

Shiba, S., Ikeda, K., Suzuki, T., Shintani, D., Okamoto, K., Horie-Inoue, K., et al. (2019). Hormonal regulation of patient-derived endometrial cancer stem-like cells generated by Three-dimensional culture. *Endocrinology* 160, 1895–1906. doi: 10.1210/en.2019-00362

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA. Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Smolle, M. A., Bullock, M. D., Ling, H., Pichler, M., and Haybaeck, J. (2015). Long non-coding RNAs in endometrial carcinoma. *Int. J. Mol. Sci.* 16, 26463–26472. doi: 10.3390/ijms161125962

Sun, S., Wang, N., Sun, Z., Wang, X., and Cui, H. (2019). MiR-5692a promotes proliferation and inhibits apoptosis by targeting HOXD8 in hepatocellular carcinoma. *J. BUON.* 24, 178–186.

Vychytilova-Faltejskova, P., Merhautova, J., Machackova, T., Gutierrez-Garcia, I., Garcia-Solano, J., Radova, L., et al. (2017). MiR-215-5p is a tumor suppressor in colorectal cancer targeting EGFR ligand epiregulin and its transcriptional inducer HOXB9. *Oncogenesis* 6, 399–406. doi: 10.1038/s41389-017-0006-6

Wang, S., Ni, B., Zhang, Z., Wang, C., Wo, L., Zhou, C., et al. (2019). Long non-coding RNA DNM3OS promotes tumor progression and EMT in gastric cancer by associating with Snail. *Biochem. Biophys. Res. Commun.* 511, 57–62. doi: 10.1016/j.bbrc.2019.02.030

Wang, W., Zhao, L. J., Tan, Y. X., Ren, H., and Qi, Z. T. (2012). Identification of deregulated miRNAs and their targets in hepatitis B virusassociatedhepato-cellular carcinoma. *World J. Gastroenterol.* 18, 5442–5453. doi: 10.3748/wjg.v18.i38.5442

Wang, Y., Xu, M., and Yang, Q. (2019). A six-microRNA signature predicts survival of patients with uterine corpusendometrial carcinoma. *Curr. Probl. Cancer* 43, 167–176. doi: 10.1016/j.currproblcancer.2018.02.002

Wilczynski, M., Danielska, J., Domanska-Senderowska, D., Dzieniecka, M., Szymanska, B., and Malinowski, A. (2018). Association of microRNA-200c expression levels with clinicopathological factors and prognosis in endometrioid endometrial cancer. *Acta Obstet. Gynecol. Scand.* 97, 560–569. doi: 10.1111/aogs.13306

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., et al. (2006). Gene expression profiling of primary cutaneous melanoma and clinical outcome. *J. Natl. Cancer. Inst.* 98, 472–482. doi: 10.1093/jnci/djj103

Xie, W., Zhang, J., Zhong, P., Qin, S., Zhang, H., and Fan, X. (2019). Expression and potential prognostic value of histone family gene signature in breast cancer. *Exp. Ther. Med.* 18, 4893–4903. doi: 10.3892/etm.2019.8131

Xing, Y., Jing, H., Zhang, Y., Suo, J., and Qian, M. (2020). MicroRNA-141-3p affected proliferation, chemosensitivity, migration and invasion of colorectal cancer cells by targeting EGFR. *Int. J. Biochem. Cell Biol.* 118:105643. doi: 10.1016/j.biocel.2019.105643

Zhang, W., Bouchard, G., Yu, A., Shafiq, M., Jamali, M., Shrager, J. B., et al. (2018). GFPT2-expressing cancer-associated fibroblasts mediate metabolic reprogramming in human lung adenocarcinoma. *Cancer Res.* 78, 3445–3457. doi: 10.1158/0008-5472.CAN-17-2928

Zhao, C., Zhang, J., Zhang, S., Yu, D., Chen, Y., Liu, Q., et al. (2013). Diagnostic and biological significance of microRNA-192 in pancreatic ductal adenocarcinoma. *Oncol. Rep.* 30, 276–284. doi: 10.3892/or.2013.2420

Zhou, L. Y., Zhang, F. W., Tong, J., and Liu, F. (2020). MiR-191-5p inhibits lung adenocarcinoma by repressing SATB1 to inhibit Wnt pathway. *Mol. Genet. Genomic Med.* 8:e1043. doi: 10.1002/mgg3.1043

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

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

Copyright © 2021 Deng, Mu, Qu, Yang, Liu, Zeng and Peng. 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