ANALYSIS OF BIOINFORMATICS TOOLS IN SYSTEMS GENETICS

EDITED BY: Shuai Cheng Li, Sandro Jose De Souza and Bairong Shen
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-182-3
DOI 10.3389/978-2-88974-182-3

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

ANALYSIS OF BIOINFORMATICS TOOLS IN SYSTEMS GENETICS

Topic Editors:
Shuai Cheng Li, City University of Hong Kong, Hong Kong, SAR China
Sandro Jose De Souza, Federal University of Rio Grande do Norte, Brazil
Bairong Shen, Sichuan University, China

Citation: Li, S. C., De Souza, S. J., Shen. B., eds. (2022). Analysis of Bioinformatics Tools in Systems Genetics. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-182-3

Table of Contents

frontiers in Genetics

DATA REPORT
published: 06 October 2020
doi: 10.3389/fgene.2020.00907

CRC-EBD: Epigenetic Biomarker Database for Colorectal Cancer

Xingyun Liu[1,2†], Xueli Zhang[2,3,4†], Jing Chen[5†], Benchen Ye[2], Shumin Ren[1], Yuxin Lin[2], Xiao-Feng Sun[6], Hong Zhang[3] and Bairong Shen[1,2*]*

[1] Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu, China, [2] Center for Systems Biology, Soochow University, Suzhou, China, [3] School of Medicine, Institute of Medical Sciences, Örebro University, Örebro, Sweden, [4] Department of Ophthalmology, Guangdong Academy of Medical Sciences, Guangdong Provincial People's Hospital, Guangzhou, China, [5] School of Science, Kangda College of Nanjing Medical University, Lianyungang, China, [6] Department of Oncology and Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

OPEN ACCESS

Edited by:
Xiaogang Wu,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Lorena Aguilar Arnal,
National Autonomous University of Mexico, Mexico
Georges Nemer,
American University of Beirut, Lebanon

***Correspondence:**
Bairong Shen
bairong.shen@scu.edu.cn
Hong Zhang
hong.zhang@oru.se

[†] These authors have contributed equally to this work

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

Received: 14 May 2020
Accepted: 22 July 2020
Published: 06 October 2020

Citation:
Liu X, Zhang X, Chen J, Ye B, Ren S, Lin Y, Sun X-F, Zhang H and Shen B (2020) CRC-EBD: Epigenetic Biomarker Database for Colorectal Cancer. Front. Genet. 11:907. doi: 10.3389/fgene.2020.00907

Keywords: colorectal cancer, database, epigenetics, DNA methylation, histone modification

INTRODUCTION

Colorectal cancer (CRC) is one of the most common forms of cancer and a major cause of cancer-related death in both men and women worldwide (Lao and Grady, 2011; Siegel et al., 2017). Over 881,000 people globally have died from CRC, while 1.8 million were newly diagnosed with CRC in 2018 (Bray et al., 2018). The death rate of CRC has been steadily declining since 1990 (Siegel et al., 2017), but precise diagnosing and treating CRC remains challenging. Many patients exhibit few symptoms until the tumor has metastasized, making biomarkers for early diagnosis essential. Liquid biopsy is an easy and non-invasive method to detect ctDNA (circulating tumor DNA) in plasma or serum samples for early diagnosis, prognosis, or treatment (Tarazona and Cervantes, 2018). But ctDNA from a liquid biopsy is difficult to process and the lack of accuracy is still a problem (Kolencik et al., 2020); as a result, most previous studies focused on tumor tissue samples. For CRC patients, ctDNA is used to detect not only RAS mutations, but also DNA methylation, such as SEPTIN9 methylation (Song et al., 2018).

Epigenetic modifications play an important role in CRC genesis and progression (Danese and Montagnana, 2017). Epigenetics investigates heritable phenotype changes without alterations in the DNA sequence (Dupont et al., 2009). Epigenetic modifications include DNA methylation, histone modification, and genomic imprinting. DNA methylation is one of the best-characterized epigenetic mechanisms (Li and Zhang, 2014), which adds methyl groups to DNA, often at CpG sequences (Ehrlich et al., 1982). Emerging evidence suggests that some epigenetic modifications, DNA methylation in particular, can be important biomarkers for CRC (Ahmed, 2007). Aberrant DNA methylation is tissue-specific and often appears at early stages of cancer development (Jahn et al., 2011), making it a potentially ideal biomarker for early diagnosis of CRC.

We have previously constructed a biomarker database for colorectal cancer (CBD) (Zhang et al., 2018). Despite numerous reports on this subject so far, to our best knowledge, no database for cancer epigenetic biomarkers has been built yet. To enable the systematic study of epigenetics in CRC, we hereby established the first cancer epigenetic biomarker database, which was named CRC-EBD (Epigenetic Biomarker Database for Colorectal Cancer). CRC-EBD stores the epigenetic biomarkers information on CRC from PubMed literature. As precision medicine is becoming the new scientific paradigm (Morere, 2012), our database is built with more focus on collecting information regarding clinical samples in order to promote future translational researches on CRC.

MATERIALS AND METHODS

Data in CRC-EBD was manually collected from PubMed. We used "*(colon[ti] OR rectosigmoid junction[ti] OR rectal[ti] OR anus[ti] OR bowel[ti] OR colorectum[ti] OR colorectal[ti]) AND (biomarker*[tiab] OR marker*[tiab] OR indicator*[tiab] OR predicator*[tiab] OR (drug target*[tiab]) OR (therapeutic target*[tiab])*" as the term to search the PubMed for the CRC biomarkers. In addition, we used the keyword "*AND methylat*[tiab]*" for methylation biomarker, "*AND histone*[tiab]*" for histone modification, and "*AND epigenetics*[tiab] NOT methylation[tiab] NOT histone*[tiab]*" for other epigenetic biomarkers. In total, 1,444 articles were screened for these biomarkers in PubMed citations until December, 2019.

The following rules were applied to screen articles about CRC epigenetic biomarkers.

1) The article should contain clear statements like "Epigenetic modification (such as DNA methylation, histone modification, or other epigenetics modifications) is a biomarker/marker/indicator of CRC." If the statement includes expressions like "can/may/has potential," the corresponding data is included. This key statement can be searched in our database under "Description".
2) Reviews or meta-analyses are excluded in the screening of CRC biomarkers.
3) If the article includes information about AUC/sensitivity/specificity or other assessment of the accuracy of the biomarker for prediction or classification of CRC, the value should be statistically significant.
4) Biomarkers from different articles have different IDs in our database, even if they share the same name, but with different clinical conditions for CRC, such as biomarker for diagnosis, prognosis, or treatment of CRC.
5) If both single and combinatorial biomarkers are included in one article, all the reported markers are given different IDs in our database.

We eventually selected 355 biomarkers, along with 694 records of sample information and 420 records of epigenetics information from the articles. The various cancer names in the original articles were uniformly changed to colorectal/colon/rectal cancer. A common format, as in "methylation of APC," was adopted for all the biomarker names in CRC-EBD. All gene symbols and miRNA names were annotated as the official gene symbols from NCBI and miRBase. The biomarkers were also labeled with sample resources (blood, stool, and tissue) and clinical applications (diagnosis, prognosis, and treatment). Moreover, sample information of the patients (e.g., nationality, age, and TNM stage,) was collected for further analysis in personalized medicine. The pipeline of data collection, database construction, and functions of CRC-EBD is shown in **Figure 1**.

B/S (Browser/Server) structure and WAMP (Windows Server 2016 + Apache 2.4.39 + MySQL (10.4.6-MariaDB) + PHP 7.3.8) were used to construct the database. Users can access our database using their own browsers without installing other components. HTML and CSS were used to create the web pages and display the information. PHP and JavaScript were applied to connect the database and realize the search function. The data is stored in the MySQL database, which can be easily and

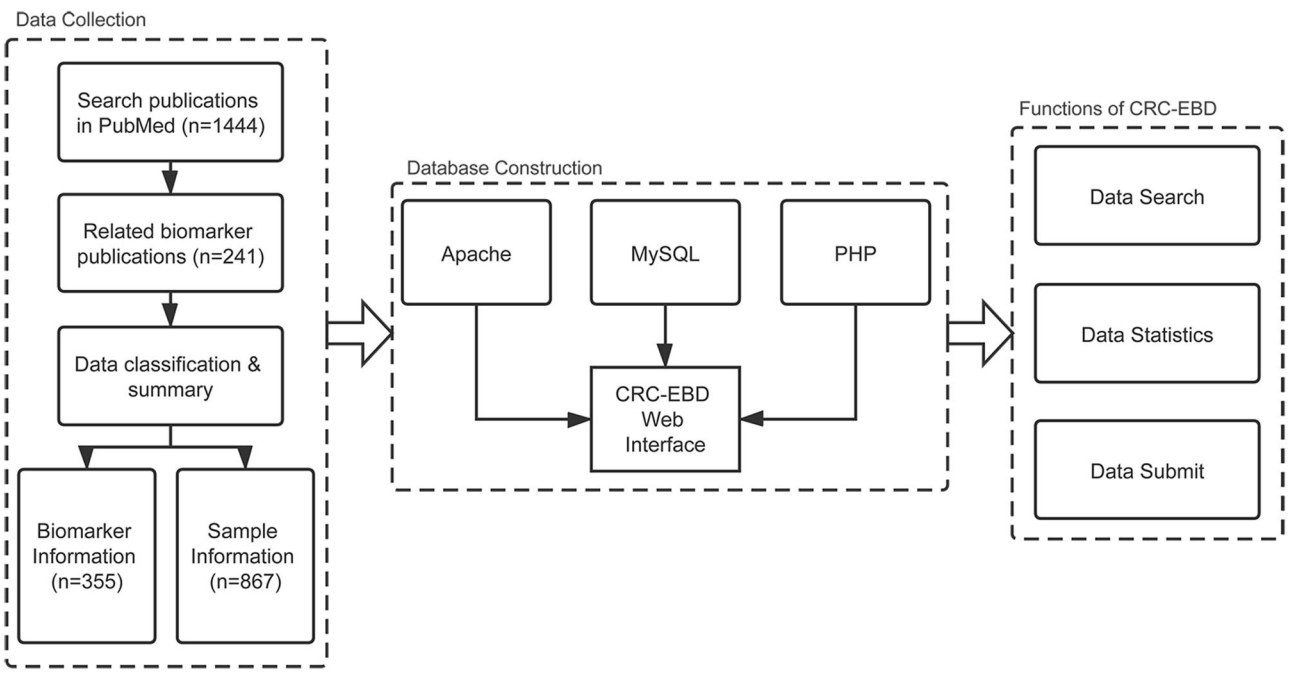


FIGURE 1 | Pipeline of database construction.

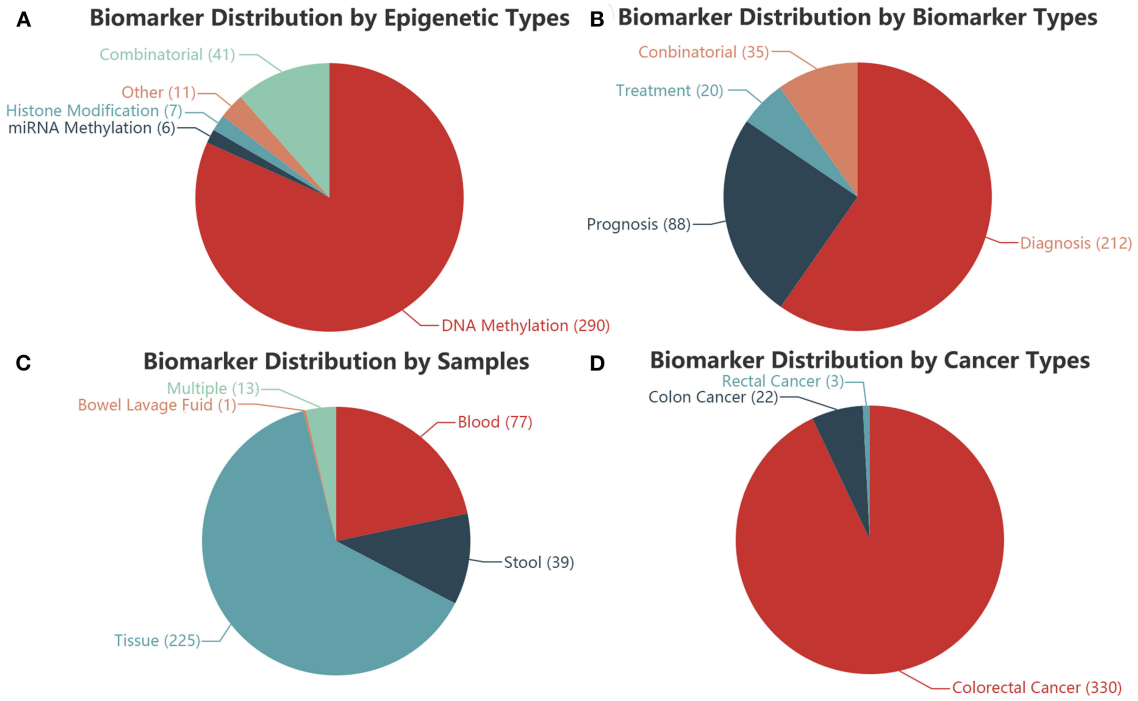


FIGURE 2 | Distributions of the biomarkers in CRC-EBD. **(A)** Epigenetic types. **(B)** Biomarker types. **(C)** Sample types. **(D)** Cancer types.

quickly accessed. The charts in the statistics page were generated dynamically using ECharts (Li et al., 2018).

DISCUSSION

The epigenetic biomarkers in our online database can be searched by epigenetics name, epigenetics type, CRC subtype, biomarker type, and application. Epigenetics name searching mode allows users to enter the name of a gene, miRNA, or histone in a text box. Similarly, under CRC subtype searching mode, users can type in a text box the CRC subtypes or cancer names. Epigenetics types can be searched by DNA methylation, RNA methylation, histone modification, or others. Furthermore, users can select the biomarker type (diagnostic, prognostic, or therapeutic) and the application mode (blood, stool, tissue, or bowel lavage fluid) for their searches. The search result will be shown in a new webpage containing the list of biomarkers, and users can click each item for more detailed information.

Among the 355 epigenetics biomarkers in our CRC-EBD, 81.69% (290) of them are single DNA methylation biomarkers, whereas 11.52% are combinatorial (**Figure 2A**). Based on the clinical applications, 59.72% of the biomarkers are diagnostic, among which 9.86% are combinatorial for diagnosis, prognosis, or treatment (**Figure 2B**). 225 (63.38%) biomarkers are applied for tissue samples, 39 (10.99%) for stool, 77 (21.69%) for blood, and 13 (3.66%) for multiple sample types. A combined biomarker (miR-124-3, ZNF582-AS1, and SFRP1 methylation) is the only one reported for bowel lavage fluid detection (**Figure 2C**). 92.98% of the biomarkers in our database are applied for colorectal cancer research, demonstrating its prominence in the current field of studies (**Figure 2D**).

Six hundred and ninety four groups of samples in total are collected in the CRC-EBD: 457 are tissue samples (tumor samples and healthy samples), 73 are stool samples, 131 are serum/plasma samples or others. Though stool or blood samples are easier and more convenient to acquire, most of the previous studies are based on tissue samples directly connected to cancer genesis and progress.

CRC-EBD is the first online resource for epigenetic biomarkers of cancer. We will expand the database to other cancers in the future. This database will offer the users a systematic perspective on the heterogeneous cancer and promote epigenetics research on cancers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.sysbio.org.cn/EBD/.

AUTHOR CONTRIBUTIONS

XL, XZ, HZ, X-FS, and BS conducted and designed this study. XL, XZ, JC, BY, SR, and YL collected data and implemented the database. XL and SR wrote the manuscript. BS supervised the project. All authors reviewed and approved the paper for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant no. 31670851).

REFERENCES

Ahmed, F. E. (2007). Colorectal cancer epigenetics: the role of environmental factors and the search for molecular biomarkers. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* 25, 101–154. doi: 10.1080/10590500701399184

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

Danese, E., and Montagnana, M. (2017). Epigenetics of colorectal cancer: emerging circulating diagnostic and prognostic biomarkers. *Ann. Transl. Med.* 5:279. doi: 10.21037/atm.2017.04.45

Dupont, C., Armant, D. R., and Brenner, C. A. (2009). Epigenetics: definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* 27, 351–357. doi: 10.1055/s-0029-1237423

Ehrlich, M., Gama-Sosa, M. A., Huang, L. H., Midgett, R. M., Kuo, K. C., McCune, R. A., et al. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.* 10, 2709–2721. doi: 10.1093/nar/10.8.2709

Jahn, K. A., Su, Y., and Braet, F. (2011). Multifaceted nature of membrane microdomains in colorectal cancer. *World J. Gastroenterol.* 17, 681–690. doi: 10.3748/wjg.v17.i6.681

Kolencik, D., Shishido, S. N., Pitule, P., Mason, J., Hicks, J., and Kuhn, P. (2020). Liquid biopsy in colorectal carcinoma: clinical applications and challenges. *Cancers* 12:1376. doi: 10.3390/cancers12061376

Lao, V. V., and Grady, W. M. (2011). Epigenetics and colorectal cancer. *Nat. Rev. Gastroenterol. Hepatol.* 8, 686–700. doi: 10.1038/nrgastro.2011.173

Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., et al. (2018). ECharts: a declarative framework for rapid construction of web-based visualization. *Vis. Inform.* 2, 136–146. doi: 10.1016/j.visinf.2018.04.011

Li, E., and Zhang, Y. (2014). DNA methylation in mammals. *Cold Spring Harb. Perspect. Biol.* 6:a019133. doi: 10.1101/cshperspect.a019133

Morere, J. F. (2012). Oncology in 2012: from personalized medicine to precision medicine. *Target Oncol.* 7, 211–212. doi: 10.1007/s11523-012-0238-5

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., et al. (2017). Colorectal cancer statistics, 2017. *CA Cancer J. Clin.* 67, 177–193. doi: 10.3322/caac.21395

Song, L., Guo, S., Wang, J., Peng, X., Jia, J., Gong, Y., et al. (2018). The blood mSEPT9 is capable of assessing the surgical therapeutic effect and the prognosis of colorectal cancer. *Biomark Med.* 12, 961–973. doi: 10.2217/bmm-2018-0012

Tarazona, N., and Cervantes, A. (2018). Liquid biopsy: another tool towards tailored therapy in colorectal cancer. *Ann. Oncol.* 29, 7–8. doi: 10.1093/annonc/mdx641

Zhang, X., Sun, X. F., Cao, Y., Ye, B., Peng, Q., Liu, X., et al. (2018). CBD: a biomarker database for colorectal cancer. *Database* 2018, 1–12. doi: 10.1093/database/bay046

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.

frontiers in Genetics

REVIEW
published: 25 November 2020
doi: 10.3389/fgene.2020.607812

Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins From Their Amino Acid Sequences

Kenichiro Imai[1] *and Kenta Nakai*[2]*

[1]*Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan,* [2]*The Institute of Medical Science, The University of Tokyo, Tokyo, Japan*

At the time of translation, nascent proteins are thought to be sorted into their final subcellular localization sites, based on the part of their amino acid sequences (i.e., sorting or targeting signals). Thus, it is interesting to computationally recognize these signals from the amino acid sequences of any given proteins and to predict their final subcellular localization with such information, supplemented with additional information (e.g., *k*-mer frequency). This field has a long history and many prediction tools have been released. Even in this era of proteomic atlas at the single-cell level, researchers continue to develop new algorithms, aiming at accessing the impact of disease-causing mutations/cell type-specific alternative splicing, for example. In this article, we overview the entire field and discuss its future direction.

Keywords: protein sorting/targeting, subcellular loalization, sorting/targeting signals, prediction methods, bacteria, archaea, eukarya

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong, Hong Kong

Reviewed by:
Litao Sun,
Sun Yat-sen University, China
Marti Aldea,
Instituto de Biología Molecular de Barcelona (IBMB), Spain

***Correspondence:**
Kenta Nakai
knakai@ims.u-tokyo.ac.jp

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

Received: *18 September 2020*
Accepted: *03 November 2020*
Published: *25 November 2020*

Citation:
Imai K and Nakai K (2020) Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins From Their Amino Acid Sequences. Front. Genet. 11:607812. doi: 10.3389/fgene.2020.607812

INTRODUCTION

Although we should not underestimate the importance of non-coding genes, the main players of the genetic system of living organisms are still regarded as protein-coding genes, which specify amino acid sequence information. Thus, in principle, we should be able to infer the *in vivo* fate of any protein from its amino acid sequence, if its environmental conditions, such as the cell type where it is synthesized, are appropriately given. For example, we should be able to predict the three-dimensional structure of a protein from its sequence or to design novel amino acid sequences that take a desired three-dimensional structure (Baker, 2019), as well as to predict how it binds/interacts with other proteins/small molecule ligands (Vakser, 2020). Another important information to be predicted is which kind of post-translational modifications, if any, it will take [at which residue(s); Audagnotto and Dal Peraro, 2017]. Also, it may be possible to predict the half-life of a given protein/peptide-based on the degradation signals (degrons) and/or other properties (Mathur et al., 2018; Eldeeb et al., 2019). Finally, the prediction of subcellular localization of a protein based on its amino acid sequence is a challenging field in bioinformatics. It is well accepted that the protein sorting for subcellular localization is regulated by so-called protein sorting (or targeting) signals, which are typically represented as a short stretch(es) of its amino acid sequence. Nowadays, many of the protein localization mechanisms/pathways that recognize and utilize such signals have been clarified. Therefore, many predictors

have been developed for the recognition of such sorting signals and attempts have been done to combine such predictors, leading to the comprehensive prediction of the final localization site. However, not all such signals have been clarified. Moreover, not all proteins are equipped with such typical signals and use some alternative (minor/exceptional) pathways. Adding the knowledge of such exceptional cases will make the prediction system gradually more realistic but the objective assessment of its performance, like the ones commonly used in the field of machine learning, will become difficult because the knowledge of exceptional cases are quite unlikely to be generalized (in other words, any sequence features of such exceptional proteins, which are nothing to do with their sorting mechanisms, would work as clues for their prediction). It should be also noted that the practical value of subcellular localization predictors has been degraded because the localization information is being comprehensively determined with subcellular proteomics experiments (Harvey Millar and Taylor, 2014). However, the rise of synthetic biology as well as precision medicine will demand prediction tools that enable the prediction against artificial proteins and/or the prediction of the impact of mutations/polymorphic variations on potential sorting signals.

In this review article, we will introduce the outline of this field, emphasizing its recent progress. The readers are recommended to refer to additional reviews by other authors and ourselves, too (Imai and Nakai, 2010, 2019; Du and Xu, 2013; Nielsen, 2017; Nielsen et al., 2019).

PREDICTION OF SUBCELLULAR LOCALIZATION SITES FOR BACTERIAL/ARCHAEAL PROTEINS

Even in the simplest type of organisms, which are unicellular organisms without any subcellular compartments, proteins can be localized at either the cytoplasmic space, the cellular membrane, or the extracellular space (i.e., secreted). This is basically the case for so-called Gram-positive bacteria and archaea, but, in reality, they also have a cell wall for another localization site. The basic prediction strategy for these proteins is to combine two kinds of predictors: a predictor for N-terminal signal peptides and that for transmembrane segments. Namely, a protein that neither has an N-terminal (and cleavable) signal peptide nor any hydrophobic transmembrane segment(s) is predicted to be localized at the cytoplasmic space; a protein that has any transmembrane segment(s) (including an N-terminal uncleavable segment) is predicted to be localized at the cellular membrane; and finally, a protein that has a cleavable N-terminal signal peptide but does not have any transmembrane segment(s) is predicted to be secreted to the extracellular space or to be localized at the cell wall. In Gram-positive bacteria, proteins that are anchored to the cell wall are characterized with the existence of the LPXTG-motif, followed by a hydrophobic domain and a tail of positively-charged residues (for recent review, see Siegel et al., 2017). On the other hand, Gram-negative bacteria contain one more membrane, the outer membrane, instead of the cell wall. Therefore, their possible localization sites are the cytoplasmic space, the inner membrane (which is equivalent to the membrane of Gram-positive bacteria), the periplasm, the outer membrane, and the extracellular space. Generally speaking, proteins that are localized at the latter three sites (the periplasm, the outer membrane, and the extracellular space) have an N-terminal cleavable signal peptide but do not have any hydrophobic transmembrane segment(s). Proteins that are integrated into the outer membrane are typically β-barrel proteins (Bakelar et al., 2017). To distinguish these three types of proteins, their difference in amino acid composition and/or *k*-mer frequency as well as motif/homology-based methods are often used.

A pioneering work to propose the above formalism is published in 1991 (Nakai and Kanehisa, 1991), where the predictor was named PSORT (I). In 2003, its approach was inherited and elaborated by Fiona Brinkman's group (Gardy et al., 2003); their software is named PSORTb (or PSORT-B). Its latest version is PSORTb 3.0 (Yu et al., 2010). The group published an excellent review of bacterial protein subcellular localization in 2006 (Gardy and Brinkman, 2006). According to the assessment shown in the review, PSORTb was the best predictor at that time. The group also releases PSORTdb, which contains a collection of experimentally-determined information of subcellular localization as well as systematic outputs of PSORTb applied to thousands of bacterial proteomes [its latest reference reports v. 3.0: (Peabody et al., 2016) but its latest version is v. 4.0]. The same group also proposes PSORTm, a variant of PSORTb designed for the prediction of metagenomic data (Peabody et al., 2020). The basic idea of PSORTm is to first identify the taxonomy of each read based on a reference database of microbial proteins. From the estimated taxonomy, the read is automatically classified with cell envelope types and then it is subject to a variant of PSORTb, which uses various types of analyses (such as motif/profile analysis) for its subcellular localization prediction. Although the assessment of its precise accuracy would be difficult, they report an assessment using artificial data and the comparison with the prediction against pre-assembled data. In view of the rapid growth of microbiome analyses, the need of characterizing metagenome data should increase even more and thus the field looks promising. Of course, other groups have developed a variety of predictors for bacterial/archaeal proteins, among which PSO-LocBact (Lertampaiporn et al., 2019), GPos-ECC-mPLoc/Gneg-ECC-mPLoc (Wang et al., 2015), BUSCA (Savojardo et al., 2018b), which will be introduced below, and ClubSub-P (Paramasivam and Linke, 2011) are released relatively recently. Some of them claim that they can deal with proteins with multiple-locations. Although once a database for (eukaryotic) proteins with multiple subcellular localizations is released (Zhang et al., 2008), it still seems difficult to classify multiple localizations objectively and quantitatively because the data come from different sources which rely on different experimental conditions (but see the discussion below).

Beyond the basic scheme described above, there are several issues to be further explored. One is the prediction of several specialized localization sites, such as host-associated, type III

secretion, fimbrial, flagellar, and spore. In PSORTb, they are treated as subcategories. Of course, it is favorable that a predictor can deal with such localization sites but it is questionable if such a predictor can also deal with artificial proteins that are transported to such locations. In other words, it is likely that such predictions are easily done with simple homology transfer from known examples. Another issue is how to deal with the proteins that are transported with minor pathways. For the users' convenience, it is desirable that a predictor can inform users which pathway the input protein will use. For example, it is surely useful if a predictor informs us that the input protein will be transported *via* the twin-arginine translocation pathway (Palmer and Stansfeld, 2020) or the lipoprotein signal peptidase II-dependent pathway (El Arnaout and Soulimane, 2019). This can already be done with several predictors, including SignalP-5.0 (Almagro Armenteros et al., 2019, see below). Hopefully, more knowledge of various protein sorting pathways should be incorporated into predictors, even if the objective assessment of their predictability would become difficult. In this sense, more benchmarking efforts/systematic analysis of subcellular localization from various viewpoints would be valuable (Stekhoven et al., 2014; Orioli and Vihinen, 2019; see below).

PREDICTION OF SUBCELLULAR LOCALIZATION SITES FOR EUKARYOTIC PROTEINS

So far, many prediction methods of eukaryotic protein subcellular localization have been developed. They are mainly based on biological/empirical sequence features related to subcellular localization. In these methods, a variety of machine learning algorithms, such as the *k*-nearest neighbor (*k*-NN) classifier, the Random Forest classifier, the support vector machine (SVM), and the deep learning, have been used. Those methods usually target 10 main localization sites, where subcompartments of localization sites are merged into 10 major sites in order to increase the number of proteins per localization site (see **Table 1**). As further explained below, for the prediction of subcellular localization sites, three types of prediction features are generally used: targeting signal features, sequence-based features, and annotation-based features (**Figure 1**). The features associated with targeting signals are most powerful, when available, and many subcellular localization predictors based on

TABLE 1 | Representative subcellular locations covered by predictors for eukaryotic proteins.

Main location	Representative subcompartments
Nucleus	inner and outer membranes, matrix, chromosome, nucleus speckle, etc.
Mitochondrion	inner and outer membranes, matrix, intermembrane space
Endoplasmic reticulum (ER)	ER membrane and lumen, microsome, rough ER, smooth ER, etc.
Plastid	inner and outer membranes, stroma, thylakoid, etc.
Golgi apparatus	Golgi apparatus membrane, lumen
Lysosome/Vacuole	vacuole lumen and membrane, lysosome lumen and membrane, etc.
Peroxisome	matrix, membrane
Cytoplasm	cytosol, cytoskeleton
Cell membrane	cell membrane, cell projection, apical, basal, etc.
Extracellular	–

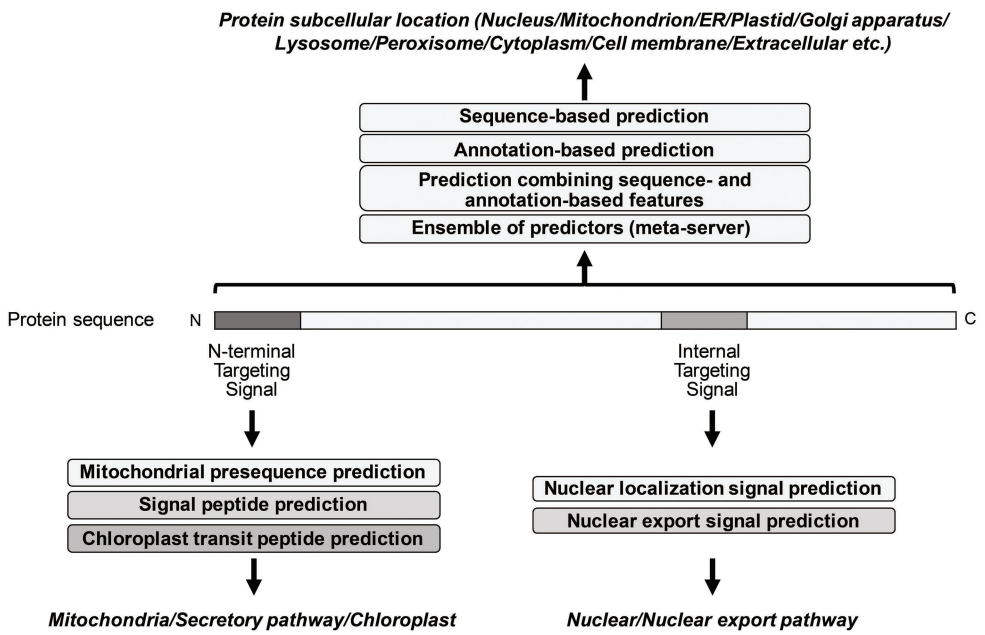


FIGURE 1 | Summary of representative prediction approaches of different subcellular localization.

targeting signal features have been developed. Thus, we first overview the representative targeting-signal predictors and then predictors for localization sites.

Prediction of Targeting Signals

The targeting signals are roughly grouped into two categories: N-terminal targeting signals and non-N-terminal targeting signals. The mitochondrial targeting signal (presequences), the signal sequence for the secretory pathway (signal peptides), and the transit signal for chloroplast (transit peptides) are well-known as N-terminal targeting signals, while the nuclear localization signal (NLS) and the nuclear export signal (NES) are internal signal sequences. Peroxisome matrix proteins contain peroxisomal targeting signal type 1 (PTS1) in the C-terminus.

Prediction of Mitochondrial Targeting Signal

Mitochondria have been estimated to host 1,000 to 1,500 distinct proteins. Approximately, 99% of mitochondrial proteins are encoded in the nuclear genome and are imported by translocases in the mitochondrial outer and inner membranes. Approximately 60% of mitochondrial proteins possess an N-terminal cleavable targeting signal (presequence; Vögtle et al., 2009). These presequences are typically recognized by the translocase of the outer membrane (TOM) receptors, which consist of Tom20 and Tom22, in the TOM complex. Then, they direct the translocation of signal-containing proteins through the main protein translocation channel, Tom40 (Pfanner et al., 2019). Upon translocation across the outer membrane, the presequence-containing proteins are transferred across the inner membrane by the translocase of the inner membrane complex (TIM23) with the presequence translocase-associated motor (PAM). The length of presequences is 20–60 amino acid residues (Calvo et al., 2017). The representative features of presequences are high and low composition of arginine residues and negatively-charged residues, respectively (von Heijne, 1986; Schneider et al., 1998). Positively charged amphiphilicity (amphiphilic α-helical structure with hydrophobic residues on one face and positively-charged residues on the opposite face) is also a well-characterized feature (Chacinska et al., 2009; Fukasawa et al., 2015). Recently, the TOM complex structure was revealed by cryo-electron microscopy and it provided structural insights into the import path of precursor protein containing presequence through the TOM complex (Araiso et al., 2019). Presequence is typically cleaved by three mitochondrial peptidases in the matrix (MPP, Icp55, and Oct1; Mossmann et al., 2012). The cleavage by MPP occurs after the position of two amino acids of C-terminal to an arginine (the R-2 motif). Icp55 and Oct1 subsequently cleave off one amino acid and eight amino acids from the newly-emerged N-terminus, respectively. Therefore, proteins processed by MPP and Icp55 have an arginine at position -3 (the R-3 motif) in the presequence, while proteins processed by MPP and Oct1 have an arginine at position -10 (the R-10 motif).

MitoProtII (Claros, 1995), TargetP (Emanuelsson et al., 2000), Predotar (Small et al., 2004), TPpred3.0 (Savojardo et al., 2015), and MitoFates (Fukasawa et al., 2015) were widely used presequence prediction methods. Those are developed using machine-learning techniques with these features of presequences. Those tools are also capable of predicting the existence of presequence as well as their cleavage site. MitoProtII and MitoFates are specific predictors for (mitochondrial) presequences, while TargetP, Predotar, and TPpred3.0 can also predict other N-terminal targeting signals, such as secretory signal sequence and chloroplastic targeting signal. Recently, TargetP2.0 is developed as a deep learning model, using bidirectional long-short-term memory (LSTM) and a multi-attention mechanism (Armenteros et al., 2019). Among existing tools, three of them (MitoFates, TPpred3.0, and TargetP2.0) perform better in the prediction of both the presequence existence and its cleavage site. MitoFates employs an SVM classifier by combining amino acid composition and physicochemical properties with positively charged amphiphilicity, discovered presequence motifs, and position-weight matrices of cleavage site patterns. TPpred3.0 is a combination of a Grammatical Restrained Hidden Conditional Random Field, N-to-1 Extreme Learning Machines, and SVMs. We compared the performance of those three methods, using recent proteomic data of the N-termini of mouse mitochondrial proteins (we omitted proteins whose length of cleaved N-terminal sequences is shorter than 10 or longer than 100 amino acids in the comparison; Calvo et al., 2017). The recalls of presequence prediction by TPpred3.0, MitoFates, and TargetP2.0 are 63.2, 75.9, and 79.9%, respectively. Whereas the recalls of the cleavage prediction by TPpred3.0, MitoFates, and TargetP2.0 are 27.0, 28.8, and 45.5%, respectively. MitoFates and TargetP2.0 show better performance on the presequence prediction. In the cleavage site prediction, TargetP2.0 far outperformed other methods, though the cleavage site prediction is still a challenging task. About 20% of mouse cleavage site data does not match with the R-2, R-3, and R-10 motifs (Calvo et al., 2017). It will be necessary to better characterize these untypical presequences.

Prediction of Signal Sequence

The targeting signal sequence for the secretory pathway (signal peptides) is located at the N-terminal of protein sequence in both eukaryotes and prokaryotes. The length of signal peptides is 16–30 amino acid residues. It is estimated that about 10–20% of eukaryotic proteome and 10% of bacterial proteome have the signal peptide at N-terminus (Kanapin et al., 2003; Ivankov et al., 2013). In eukaryotic cells, the signal recognition particle (SRP) co-translationally recognizes signal peptides upon their emergence from the ribosome and transfers them to the Sec61 translocon in the endoplasmic reticulum (ER) membrane *via* the SRP receptor (Nilsson et al., 2015). The signal peptidase cleaves off signal peptides and thus mature proteins are generated. Signal peptides share several characteristic features (von Heijne, 1990); they have tripartite architecture: a positively charged N-terminus (n-region), a hydrophobic segment (h-region), and a cleavage site for signal peptidase (c-region). The cleavage site is characterized by the (-1, -3) rule; amino acids with

small, uncharged side chains at the -1 and -3 position relative to the cleavage site.

For predicting signal peptides and their cleavage sites, many prediction methods, such as SignalP 4.0 (Petersen et al., 2011), SPEPlip (Fariselli et al., 2003), Phobius (Krogh et al., 2007), and DeepSig (Savojardo et al., 2018a), have been developed. The discrimination between secretory and non-secretory proteins based on the signal peptide prediction has been most successful in targeting signal predictions because SignalP 3.0 has already achieved the best Matthews' Correlation Coefficient (MCC) of 0.76 in eukaryotic data sets in an assessment study in 2009 (Choo et al., 2009). Recently, SignalP has been further improved as a deep neural network-based method, combining with conditional random field classification and optimized transfer learning (SignalP-5.0; Almagro Armenteros et al., 2019). According to their benchmark results, SignalP-5.0 outperforms other methods in predicting both the signal peptide existence and the cleavage site: the MCC was 0.88 in the signal peptide prediction and the recall of cleavage site detection was 72.9%.

Prediction of Chloroplastic Targeting Signal

The translocons at the outer and the inner membranes of chloroplasts, the TOC and TIC complexes mediate the targeting and import of ~3,500 different nuclear-encoded proteins. Those proteins are imported from the cytoplasm *via* interaction between their cleavable, N-terminal chloroplast targeting signal (transit peptides), and the TOC–TIC import systems (Li and Chiu, 2010; Paila et al., 2015). The transit peptide is removed off by the activity of stroma processing peptidase (SPP), which is related to the mitochondrial peptidase, MPP. SPP does not interact stably with the TOC–TIC import system, thus the cleavage event occurs after protein translocation or upon the emergence of the transit peptide cleavage site into the stroma. Chloroplast transit peptides are mostly unstructured but can form α-helical structures in hydrophobic environments (Bruce, 2001; Jarvis, 2008). In addition, chloroplast transit peptides have a high content of hydroxylated amino acids (e.g., serine residues) and positively charged amino acids and a very low content of negatively charged amino acids (Bhushan et al., 2006). Transit peptides and presequences are therefore similar in several aspects. In spite of the similarities, chloroplast transit peptides direct precursor proteins specifically to chloroplasts. Ge et al. (2014) demonstrated that transit peptides and presequences can be discriminated by their charge properties and hydrophobicity. Also, the analysis of 916 chloroplast proteins revealed an N-terminal domain beginning with Met-Ala and the low composition of arginine in the N-terminal portion (Zybailov et al., 2008). Moreover, Lee et al. (2019) recently showed that mitochondrial or chloroplast targeting specificities are characterized by the N-terminal regions of these targeting signals: an N-terminal multiple-arginine motif was identified as the mitochondrial specificity factor and chloroplast evasion signal. Cleavage sites of transit peptides are characterized by higher content of Ala, Ile, Cys, and Val residues (Gavel and von Heijne, 1990). The three motifs, [V,I][R,A]↓[A,C]AAE, S[V,I][R,S,V]↓[C,A]A, and [A,V]N↓A[A,M]AG[E,D], are derived by a set of 198 cleavage sites (Savojardo et al., 2015).

The existing prediction tools for the chloroplastic targeting signal deal with cleavable N-terminal transit peptides. Widely used prediction methods have been integrated as a part of prediction of N-terminal targeting signals in general: e.g., TargetP (Emanuelsson et al., 2000), iPSORT (Bannai et al., 2002), Predotar (Small et al., 2004), and TPpred3 (Savojardo et al., 2015). Among those tools, TPpred3 achieved better performance for transit peptide prediction (46% precision and 64% recall). As mentioned above, TargetP is recently updated to version 2.0 as a deep learning model (TargetP2.0; Armenteros et al., 2019). In their comparison, the precision and recall of chloroplastic transit peptide identification of TargetP2.0 are 90 and 86%, respectively, while those of TPpred3 are 76 and 69%. In the cleavage site prediction, the recalls of TargetP2.0 and TPpred3 are 49 and 30%, respectively. Like mitochondrial presequence prediction, the cleavage site prediction of chloroplastic targeting signal is a difficult problem. Comparing with the data size of signal peptides, that of transit peptides is quite small and thus the lower performance could have been caused by this reason. Larger-scale N-terminal proteomics data of chloroplastic proteins would be necessary for the improvement of their cleavage site prediction.

Prediction of Nuclear Localization Signals and Nuclear Export Signals

Nuclear proteins are transported into or out of the nuclei through the nuclear pore complex by the importin-β (Impβ) family nucleocytoplasmic transport receptors (Kimura and Imamoto, 2014). The human proteome contains 20 Impβ family proteins: 10 are nuclear import receptors (importin-β, transportin-1, -2, -SR, importin-4, -5 (RanBP5), -7, -8, -9 and -11), seven are export receptors (exportin-1 (CRM1), -2(CAS/CSE1L), -5, -6, -7, -t, and RanBP17), two are bi-directional receptors (imporin-13 and exportin-4), while the function of remaining RanBP6 is undetermined (Kimura and Imamoto, 2014). Those nucleocytoplasmic transport receptors are thought to recognize specific targeting signals on those cargo proteins. Several types of NLSs and NESs have been reported, so far. The most studied NLS is the classical NLS (cNLS) that binds to Impα, which is a cargo-binding adaptor exclusively used for Impβ (Lange et al., 2007). Sequences similar to the Impβ binding (IBB)-domain in Impα act as NLSs that bind directly to Impβ. Other known NLSs/NESs that bind directly to Impβ family are: the PY-NLS for Trn1 and Trn2 (Lee et al., 2006), the Leu-rich NES for CRM1 (Hutten and Kehlenbach, 2007), the SR-domain for TrnSR (Maertens et al., 2014), and the β-like importin binding (BIB)-domain, which binds to several nucleocytoplasmic transport receptors (Jäkel and Görlich, 1998). In addition, the RG/RGG-rich segment for Trn1 and the RSY-rich segment for TrnSR were reported recently (Bourgeois et al., 2020). However, these known NLSs/NESs do not explain all of the cargo recognition sites. Moreover, recent proteomic analysis for the identification of cargo proteins of 12 nucleocytoplasmic transport receptors (10 nuclear import

receptors and 2 bi-directional receptors; Kimura et al., 2017) also pointed out that about 30% of identified cargos are shared by multiple receptors. The degree of multiplicity and diversity of cargo recognition by nucleocytoplasmic transport receptors are still controversial.

Among known nuclear targeting signals, cNLS and NES of CRM1 are well characterized. Thus, existing prediction methods of NLSs and NESs mainly target these two types. cNLSs are grouped into monopartite and bipartite NLSs. Monopartite NLS is characterized with a single stretch of basic residues (e.g., KR[K/R]R and K[K/R]RK), while bipartite NLS has two clusters of basic residues, separated by a spacer region of 10–12 amino acids (e.g., $KRX_{10-12}K[K/R][K/R]$; Kosugi et al., 2009). Lisitsyna et al. (2017) assessed the prediction performance of widely used methods, Nucpred (Brameier et al., 2007), cNLSmapper (Kosugi et al., 2008a), NLStradamus (Ba et al., 2009), NucImport (Mehdi et al., 2011), and SeqNLS (Lin and Hu, 2013), using a human NLS dataset (Lisitsyna et al., 2017). NucPred, seqNLS, and NLStradamus showed better MCCs (~0.3); however, the recalls of those methods were still ~45%. Recently, Guo et al. (2020) reported INSP, which is a NLS predictor based on a multivariate regression model integrating PSSM-based conservation score, protein language-based SVM learning score, disorder-based structural score, and amino acid physical chemistry property-based score. On their test dataset, INSP showed 50.6% precision at 67.0% recall, whereas seqNLS, NLStradamus, and cNLSmapper obtained 60.6% precision at 36.4% recall, 53.9% precision at 35.6% recall, and 50.9% precision at 50.9% recall, respectively. INSP showed a favorable balance between the prediction precision and recall, but NLS prediction seems to be still difficult because the cNLS sequence patterns are often observed in non-nuclear protein sequences (i.e., false positives).

Nuclear export signals function as essential regulators for the export of hundreds of distinct cargo proteins by interacting with CRM1. So far, 11 consensus patterns of NES have been proposed by a peptide-library study and structure analyses of CRM1-NES (Kosugi et al., 2008b; Fung et al., 2015, 2017). In general, NESs are represented by $\Phi 0\text{-}x_{1\text{-}2}\text{-}\Phi 1\text{-}(x)_{2\text{-}3}\text{-}\Phi 2\text{-}(x)_{2\text{-}3}\text{-}\Phi 3\text{-}x\text{-}\Phi 4$ (Φ1-4 denote Leu, Val, Ile, Phe, and Met while x is any amino acid. Φ0 is not restricted to the hydrophobic amino acids). Those hydrophobic residues in Φ0–Φ4 are bound to the corresponding hydrophobic pockets in CRM1. Based on the pattern of these Φ's and spacing sequences, the NES motifs are classified into seven classes and four additional reverse classes, representing binding in the opposite direction. Several prediction tools for NESs, such as NetNES (La Cour et al., 2004), NESsential (Fu et al., 2011), NESmapper (Kosugi et al., 2014), Wregex (Prieto et al., 2014), LocNES (Xu et al., 2015), and NoLogo (Liku et al., 2018) have been developed, representing the consensus sequences with regular expressions or PSSMs as well as biophysical properties (disorder propensity, solvent accessibility, and secondary structure information). Among those tools, LocNES outperformed other prediction tools; however, the precision is ~50% at 20% recall. The low performance is caused by high false-positive rates. As mentioned above, the NES consensus patterns are simple and commonly observed in other protein sequences. Thus, it seems to be difficult to improve the prediction performance by only using the sequence information. Recently, Lee et al. (2019) provided a comprehensive table for cargo proteins, containing the location of the NES motifs with the disordered propensity, the predicted secondary structures, and the conserved domain information. They also proposed a structure modeling-based prediction which predicts the binding energy of the NES peptide bound to the binding groove of CRM1, using multiple structures of CRM1-NES peptide complex as templates (Lee et al., 2019). The structure-based methods performed at the same level as LocNES in recall rate but outperformed LocNES in specificity and false-positive rate. Thus, combining sequence-based and structure-based predictions seems promising in significantly improving the NES prediction. Moreover, NLSdb, which is a database containing NLSs and NESs, has been recently updated (Bernhofer et al., 2018). In this update, the potential set of novel NLSs and NESs has been generated by an *in silico* mutagenesis protocol. Then, the potential NLSs and NESs match at least one nuclear protein but do not match any non-nuclear proteins. The updated NLSdb contains 2,253 NLSs (1,614 are potential NLSs) and 398 NESs (192 are potential NESs). The data would be useful to further improve the NLS and NES prediction performances.

Prediction of Subcellular Localization Site of Protein in a Cell

Existing methods for predicting subcellular localization sites can be grouped into four categories. The first category of prediction methods uses only sequence-based features. Some sequence-based features are used in localization site prediction because their differences are empirically known to be correlated with the differences between localization sites. Such empirical features include the frequency of dipeptides, *n*-grams, and *k*-mers as well as the pseudo amino acid composition of the entire amino acid sequence (or that of predicted mature sequence). Pseudo amino acid composition is more informative in terms of incorporating sequence-order information of a protein sequence (Chou, 2001). These empirical sequence-based features have also been popular in various amino acid sequence-based predictions. Besides these systematically defined features, sequence features of various known targeting signals are more or less useful, as mentioned above. Functional motifs are also used in the prediction because sequence motifs associated with the function of a protein are closely related to its localization site (for example, a protein containing a DNA-binding motif is likely to be localized in the nucleus). The first sequence-based method was PSORT (I) (Nakai and Kanehisa, 1992), which was developed about 30 years ago, and later many other methods, such as WoLF PSORT (Horton et al., 2007), CELLO2.5 (Yu et al., 2006), and DeepLoc (Almagro Armenteros et al., 2017), have been developed. WoLF PSORT is an update of PSORT II (Horton and Nakai, 1997), which converts the input amino acid sequences into a numerical vector consisting of amino acid composition and PSORT/iPSORT (Nakai and Kanehisa, 1992; Bannai et al., 2002)

localization features, and then classifies proteins into subcellular locations with a weighted *k*-NN classifier. CELLO2.5 is a two-level SVM classifier system: the first level comprises a number of SVM classifiers, each based on distinctive sets of feature vectors generated from amino acid sequence data, and the second level SVM classifier functions as the jury machine to generate the probability distribution of decisions for possible localizations. Recently, several deep learning-based predictors are developed. DeepLoc is their representative. DeepLoc uses recurrent neural networks (RNNs) with long short-term memory (LSTM) cells that process the entire amino acid sequence and an attention mechanism identifying sequence regions important for the subcellular localization.

The second category of predictors uses annotation-based features obtained from experimental evidence. GO terms, localization annotation in UniProt, functional domain, protein-protein interaction, and literature information from PubMed abstracts are categorized into this type of features. mGOASVM (Wan et al., 2012) is a predictor for the subcellular localization of multi-location proteins based on GO-terms. In mGOASVM, multi-label GO vectors, which are the occurrences of GO terms of homologous proteins, are constructed, and then GO vectors are recognized by SVM classifiers equipped with a decision strategy that can produce multiple-class labels for a query protein. pLoc-mEuk (Cheng et al., 2018) is recently developed by extracting the key GO information into "Chou's general Pseudo Amino Acid Composition." pLoc-mEuk can also deal with proteins with multiple locations. Generally speaking, however, compared with those features, the transfer of localization annotation from homologous protein seems to be simpler and more useful. We previously pointed out that a simple homology-based inference outperforms methods based on machine learning if a homologous protein with localization annotation is available (Imai and Nakai, 2010).

The third category is the predictors combining sequence-based and annotation-based features, such as MultiLoc2 (Blum et al., 2009), SherLoc2 (Briesemeister et al., 2009), YLoc (Briesemeister et al., 2010), and LocTree3 (Goldberg et al., 2014). MultiLoc2 utilizes an SVM predictor, MultiLoc (Höglund et al., 2006), which is based on overall amino acids and the presence of known sorting signals, combined with phylogenetic profiles and GO terms. SherLoc2 combines MultiLoc2 and EpiLoc (Brady and Shatkay, 2008), a prediction system based on features derived from PubMed abstracts. YLoc is based on a simple naive Bayes classifier, which combines various features ranging from simple amino acid composition to annotation information, like PROSITE domains, and GO terms from close homologs. LocTree3 improves over a machine learning-based predictor, LocTree2 (Goldberg et al., 2012), by the combination of the machine learning-based method with a homology-based inference transfer through PSI-BLAST.

The fourth category is the ensemble of several prediction methods (meta-servers), which collects prediction scores of several predictors, and then they are trained by a machine learning technique, such as the Random Forest classifier and SVM. SubCons (Salvatore et al., 2017) is a recent ensemble method, which combines four predictors (CELLO2.5, LocTree2, MultiLoc2, and SherLoc2) using a Random Forest classifier. BUSCA also integrates different prediction methods. Prediction pipeline of BUSCA consists of predictors for targeting signals [DeepSig (Savojardo et al., 2018a) and TPpred3 (Savojardo et al., 2015)], for GPI-anchors [PredGPI (Pierleoni et al., 2008)], for transmembrane domains [ENSEMBLE3.0 (Martelli et al., 2003) and BetAware (Savojardo et al., 2013)], and for discriminators of subcellular localization of both globular and membrane proteins [BaCelLo (Pierleoni et al., 2006), MemLoci (Pierleoni et al., 2011), and SChloro (Savojardo et al., 2017)].

Recent Benchmarks for Subcellular Localization Prediction

Evaluation of prediction performance of subcellular localization prediction is often difficult due to the following reasons: (i) There are often overlaps between their own training data and the test data of different methods. In those cases, the performances could be overestimated. (ii) Comparison of sequence-based methods with annotation-based methods or methods combining sequence- and annotation-based methods tends to be unfair. For example, the measured accuracy of annotation-based methods would become apparently higher if the majority of test data used for sequence-based methods are included in the databases used for the prediction by annotation-based methods.

To evaluate the prediction performance with less bias, Salvatore et al. recently made a benchmark dataset which consists of proteins containing identical subcellular annotations in at least two out of the three resources (Salvatore et al., 2017): two large-scale study data on subcellular localization of human proteins (Uhlen et al., 2010; Fagerberg et al., 2011; Breckels et al., 2013; Christoforou et al., 2014) and proteins with "manually curated" annotation of subcellular localization in UniProt (UniProt Consortium, 2019). Then, they examined the performance of six state-of-the-art methods [CELLO2.5 (Yu et al., 2006), LocTree2 (Goldberg et al., 2012), MultiLoc2 (Blum et al., 2009), SherLoc2 (Briesemeister et al., 2009), WoLF PSORT (Horton et al., 2007), and YLoc (Briesemeister et al., 2010)] as well as SubCons (Salvatore et al., 2017) for eight localization sites (nucleus, mitochondria, ER, Golgi apparatus, lysosome, peroxisome, plasma membrane, and cytoplasm). They used the Generalized Squared Correlation (GC^2; Baldi et al., 2000) for performance evaluation. GC^2 is a subtype of Gorodkin measure (Gorodkin, 2004), which can be seen as a generalization of MCC that applies to *K*-categories. The Gorodkin measure is more informative than the accuracy measure when there is an imbalance of classes. For $K = 2$, the Gorodkin measure squared is GC^2. In this assessment, SubCons showed the best overall prediction performance, $GC^2 = 0.32$, and the second best was SherLoc2 ($GC^2 = 0.27$). On the other hand, during the development of DeepLoc (Almagro Armenteros et al., 2017), the authors made an independent test set by performing a stringent homology partitioning against experimentally annotated protein data in UniProt. Homologous proteins that fulfill a certain threshold of similarity were clustered, and then each

cluster of homologous proteins was assigned to one of the five folds, ensuring that similar proteins were not mixed between the different folds. Four were used for the training and validation while the remaining one for testing. Using the test set, they compared the prediction performance of DeepLoc with the above six methods (CELLO2.5, LocTree2, MultiLoc2, SherLoc2, WoLF PSORT, and YLoc) and iLoc-Euk (Chou et al., 2011) in 10 localization sites (extracellular and plastid are added into the above eight localization sites). DeepLoc showed the best Gorodkin measure of 0.735, and the second and third best were achieved by iLoc-Euk at 0.682 and YLoc at 0.533, respectively.

Although efforts to evaluate the prediction performance with less bias have been made, more efforts seem to be necessary. According to recent benchmarking reports based on human data sets and membrane proteins (Orioli and Vihinen, 2019; Shen et al., 2020), sequence-based methods tend to show lower performance than annotation-based methods, including meta methods. However, a certain number of proteins (or their highly homologous ones) in the benchmark test data seem to be included in the database used in annotation-based methods. In addition, methods trained and tested with newly constructed data tend to show better performance because older data tend to include more mislabeled or questionable examples. Indeed, Almagro Armenteros et al. (2017) pointed out a considerable decrease of experimentally confirmed proteins in UniProt after a major change in the annotation standards on release 2014_09. The prediction performances of machine learning algorithms significantly depend on the datasets used. Some of the previously developed methods may outperform newer methods when they are trained and tested with the latest datasets. For fair assessments, performance comparison should therefore be done in each category with standardized benchmark data sets, ensuring independence between training and test data sets. Unfortunately, to the best of our knowledge, such standardized benchmark data sets have not been constructed so far. The data sets used in previous studies are often used in the development of novel methods. The standardization of prediction performance comparison is a big challenge but this is essential and important in this field. Recent progress in proteome-wide subcellular protein mapping (see below) would provide substantial information on the subcellular localization of unverified or unseen proteins as well as the information for correcting mislabeled proteins, which should be helpful in constructing standardized benchmark data sets, obviously.

PROTEIN LOCALIZATION RESOURCES OBTAINED FROM RECENT SPATIAL PROTEOMICS APPROACHES

Proteomics data for capturing the spatial distribution of proteins at the subcellular level (subcellular protein mapping) are useful resources for their predictive studies. Recent advances in high-throughput microscopy, quantitative mass spectrometry (MS), interactome mapping, and machine learning applications for data analysis have enabled proteome-wide subcellular protein mapping (Lundberg and Borner, 2019; Borner, 2020). Three experimental approaches are generally used for spatial proteomics: proteome-wide imaging of protein localization, protein–protein interaction network analysis, and MS-based organelle profiling. All of these approaches have produced numerous available data of human protein subcellular localization. The Human Cell Atlas provides an invaluable resource of imaging data at a single-cell level (localization of 12,003 proteins; Thul et al., 2017). The global organellar map based on biotin identification (BioID) data is now available as a resource of protein–protein interaction network analysis (4,145 proteins; Go et al., 2019). Several organelle profiling resources are obtained from fibroblasts (2,533 proteins; Jean Beltran et al., 2016) and cell lines: HeLa (8,710 proteins; Itzhak et al., 2016), five different cancer cell lines (12,418 proteins; Orre et al., 2019), and U-2 OS (2,412 proteins; Geladaki et al., 2019). In addition, organelle profiling resources of mouse primary neurons (Itzhak et al., 2017), mouse liver (Krahmer et al., 2018), mouse pluripotent stem cell (Christoforou et al., 2016), rat liver (Jadot et al., 2017), and *Saccharomyces cerevisiae* (Nightingale et al., 2019) are also available.

Each of these approaches has its own merits for the identification of protein localization: the imaging approach provides multiple localizations and has a single-cell resolution while the MS-based approach can provide peptide-level resolution and reveal the differential localization of splicing isoforms, proteolytically processed forms, and the isoforms *via* differential post-translational modifications. A recent imaging-based large-scale study reports that about a half of all proteins are localized at multiple compartments, suggesting that there is a shared pool of proteins even among functionally unrelated organelles (Thul et al., 2017). Prediction of proteins that exist in two or more subcellular location sites is an important issue for understanding the biological process in a cell. A recent review summarizes the prediction methods that can deal with proteins with multiple locations (Chou, 2019).

A number of differentially localized isoform pairs were found by MS-based approaches (Christoforou et al., 2016; Geladaki et al., 2019). Such localization change at the isoform level is an interesting issue in terms of targeting signal usage. Protein isoforms seem to be generated by a stress response or in a tissue-specific manner. Thus, a number of localization changes at the isoform level may have been unidentified still. For mitochondrial proteins, we previously applied MitoFates to search for differentially-localized candidates of isoforms and obtained 517 genes, which were 44% of the predicted mitochondrial genes (Fukasawa et al., 2015), suggesting that the major localization changes of mitochondrial protein isoforms are regulated by the changes in their N-terminal targeting signal. Recently, relative protein levels of more than 12,000 genes across 32 normal human tissues were quantified and tissue-specific or tissue-enriched proteins were identified (Jiang et al., 2020). Also, they identified a total of 2,436 tissue-enriched protein isoforms. Those isoforms may be useful for the investigation of tissue-specific localization changes at the isoform level.

Multiple localization proteins and localization changes among isoforms imply potential "moonlighting" activity. Comprehensive analyses of these proteins should boost our further understanding in cell biology.

CONCLUSION

A number of computational tools for the analyses of protein subcellular localization are introduced in this review. Although many of the localization sites of a given protein would be able to be predicted through a mere homology transfer nowadays, we would like to emphasize that the subcellular localization prediction problem is not a pedantic one at all. The authors believe that the *in silico* accumulation of various knowledge on protein sorting/targeting processes is important. Prediction methods can be used for assessing how much we understand these processes quantitatively. The future methods should be useful for various purposes, such as for the evaluation of artificial proteins, for understanding why some proteins are localized at multiple positions and for inferring how tissue-specific and/or condition-specific isoforms can change their localization sites. Therefore, in our opinion, the knowledge-based approach would be most important in the future of this field and such knowledge should be integrated into the wider knowledge on the *in vivo* fate of proteins since all of the processes are interrelated with each other (Nakai, 2001).

AUTHOR CONTRIBUTIONS

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

FUNDING

KI acknowledges support from JSPS KAKENHI (grant number 18K11543).

REFERENCES

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387–3395. doi: 10.1093/bioinformatics/btx431

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Araiso, Y., Tsutsumi, A., Qiu, J., Imai, K., Shiota, T., Song, J., et al. (2019). Structure of the mitochondrial import gate reveals distinct preprotein paths. *Nature* 575, 395–401. doi: 10.1038/s41586-019-1680-7

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., et al. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* 2:e201900429. doi: 10.26508/lsa.201900429

Audagnotto, M., and Dal Peraro, M. (2017). Protein post-translational modifications: in silico prediction tools and molecular modeling. *Comput. Struct. Biotechnol. J.* 15, 307–319. doi: 10.1016/j.csbj.2017.03.004

Ba, A. N. N., Pogoutse, A., Provart, N., and Moses, A. M. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10:202. doi: 10.1186/1471-2105-10-202

Bakelar, J., Buchanan, S. K., and Noinaj, N. (2017). Structural snapshots of the β-barrel assembly machinery. *FEBS J.* 284, 1778–1786. doi: 10.1111/febs.13960

Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* 28, 678–683. doi: 10.1002/pro.3588

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412–424. doi: 10.1093/bioinformatics/16.5.412

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18, 298–305. doi: 10.1093/bioinformatics/18.2.298

Bernhofer, M., Goldberg, T., Wolf, S., Ahmed, M., Zaugg, J., Boden, M., et al. (2018). NLSdb-major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Res.* 46, D503–D508. doi: 10.1093/nar/gkx1021

Bhushan, S., Kuhn, C., Berglund, A. K., Roth, C., and Glaser, E. (2006). The role of the N-terminal domain of chloroplast targeting peptides in organellar protein import and miss-sorting. *FEBS Lett.* 580, 3966–3972. doi: 10.1016/j.febslet.2006.06.018

Blum, T., Briesemeister, S., and Kohlbacher, O. (2009). MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 10:274. doi: 10.1186/1471-2105-10-274

Borner, G. H. H. (2020). Organellar maps through proteomic profiling - a conceptual guide. *Mol. Cell. Proteomics* 19, 1076–1087. doi: 10.1074/mcp.R120.001971

Bourgeois, B., Hutten, S., Gottschalk, B., Hofweber, M., Richter, G., and Sternat, J. (2020). Nonclassical nuclear localization signals mediate nuclear import of CIRBP. *Proc. Natl. Acad. Sci. U. S. A.* 117, 8503–8514. doi: 10.1073/pnas.1918944117

Brady, S., and Shatkay, H. (2008). EPILOC: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.* 13, 604–615. doi: 10.1142/9789812776136_0058

Brameier, M., Krings, A., and MacCallum, R. M. (2007). NucPred — predicting nuclear localization of proteins. *Bioinformatics* 23, 1159–1160. doi: 10.1093/bioinformatics/btm066

Breckels, L. M., Gatto, L., Christoforou, A., Groen, A. J., Lilley, K. S., and Trotter, M. W. B. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *J. Proteomics* 88, 129–140. doi: 10.1016/j.jprot.2013.02.019

Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., and Shatkay, H. (2009). SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.* 8, 5363–5366. doi: 10.1021/pr900665y

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). Going from where to why-interpretable prediction of protein subcellular localization. *Bioinformatics* 26, 1232–1238. doi: 10.1093/bioinformatics/btq115

Bruce, B. D. (2001). The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta* 1541, 2–21. doi: 10.1016/s0167-4889(01)00149-5

Calvo, S. E., Julien, O., Clauser, K. R., Shen, H., Kamer, K. J., Wells, J. A., et al. (2017). Comparative analysis of mitochondrial N-termini from mouse, human, and yeast. *Mol. Cell. Proteomics* 16, 512–523. doi: 10.1074/mcp.M116.063818

Chacinska, A., Koehler, C. M., Milenkovic, D., Lithgow, T., and Pfanner, N. (2009). Importing mitochondrial proteins: machineries and mechanisms. *Cell* 138, 628–644. doi: 10.1016/j.cell.2009.08.005

Cheng, X., Xiao, X., and Chou, K. C. (2018). pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. *Genomics* 110, 50–58. doi: 10.1016/j.ygeno.2017.08.005

Choo, K. H., Tan, T. W., and Ranganathan, S. (2009). A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics* 10:S2. doi: 10.1186/1471-2105-10-S15-S2

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Genet.* 43, 246–255. doi: 10.1002/prot.1035

Chou, K. C. (2019). Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.* 26, 4918–4943. doi: 10.2174/0929867326666190507082559

Chou, K. C., Wu, Z. C., and Xiao, X. (2011). iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One* 6:e18258. doi: 10.1371/journal.pone.0018258

Christoforou, A., Arias, A. M., and Lilley, K. S. (2014). Determining protein subcellular localization in mammalian cell culture with biochemical fractionation and iTRAQ 8-plex quantification. *Methods Mol. Biol.* 1156, 157–174. doi: 10.1007/978-1-4939-0685-7_10

Christoforou, A., Mulvey, C. M., Breckels, L. M., Geladaki, A., Hurrell, T., Hayward, P. C., et al. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* 7:8992. doi: 10.1038/ncomms9992

Claros, M. G. (1995). MitoProt, a macintosh application for studying mitochondrial proteins. *Bioinformatics* 11, 441–447. doi: 10.1093/bioinformatics/11.4.441

Du, P., and Xu, C. (2013). Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev. Proteomics* 10, 227–237. doi: 10.1586/epr.13.16

El Arnaout, T., and Soulimane, T. (2019). Targeting lipoprotein biogenesis: considerations towards antimicrobials. *Trends Biochem. Sci.* 44, 701–715. doi: 10.1016/j.tibs.2019.03.007

Eldeeb, M. A., Siva-Piragasam, R., Ragheb, M. A., Esmaili, M., Salla, M., and Fahlman, R. P. (2019). A molecular toolbox for studying protein degradation in mammalian cells. *J. Neurochem.* 151, 520–533. doi: 10.1111/jnc.14838

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016. doi: 10.1006/jmbi.2000.3903

Fagerberg, L., Stadler, C., Skogs, M., Hjelmare, M., Jonasson, K., Wiking, M., et al. (2011). Mapping the subcellular protein distribution in three human cell lines. *J. Proteome Res.* 10, 3766–3777. doi: 10.1021/pr200379a

Fariselli, P., Finocchiaro, G., and Casadio, R. (2003). SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* 19, 2498–2499. doi: 10.1093/bioinformatics/btg360

Fu, S. C., Imai, K., and Horton, P. (2011). Prediction of leucine-rich nuclear export signal containing proteins with NESsential. *Nucleic Acids Res.* 39:e111. doi: 10.1093/nar/gkr493

Fukasawa, Y., Tsuji, J., Fu, S. C., Tomii, K., Horton, P., and Imai, K. (2015). MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteomics* 14, 1113–1126. doi: 10.1074/mcp.M114.043083

Fung, H. Y. J., Fu, S. C., Brautigam, C. A., and Chook, Y. M. (2015). Structural determinants of nuclear export signal orientation in binding to exportin CRM1. *eLife* 4:e10034. doi: 10.7554/eLife.10034

Fung, H. Y. J., Fu, S. C., and Chook, Y. M. (2017). Nuclear export receptor CRM1 recognizes diverse conformations in nuclear export signals. *eLife* 6:e23961. doi: 10.7554/eLife.23961

Gardy, J. L., and Brinkman, F. S. L. (2006). Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4, 741–751. doi: 10.1038/nrmicro1494

Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnády, G. E., Simon, I., et al. (2003). PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Res.* 31, 3613–3617. doi: 10.1093/nar/gkg602

Gavel, Y., and von Heijne, G. (1990). A conserved cleavage-site motif in chloroplast transit peptides. *FEBS Lett.* 261, 455–458. doi: 10.1016/0014-5793(90)80614-O

Ge, C., Spånning, E., Glaser, E., and Wieslander, Å. (2014). Import determinants of organelle-specific and dual targeting peptides of mitochondria and chloroplasts in Arabidopsis thaliana. *Mol. Plant* 7, 121–136. doi: 10.1093/mp/sst148

Geladaki, A., Kočevar Britovšek, N., Breckels, L. M., Smith, T. S., Vennard, O. L., Mulvey, C. M., et al. (2019). Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* 10:331. doi: 10.1038/s41467-018-08191-w

Go, C., Knight, J., Rajasekharan, A., Rathod, B., Hesketh, G., Abe, K., et al. (2019). A proximity biotinylation map of a human cell. bioRxiv [Preprint]. doi: 10.1101/796391

Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics* 28, i458–i465. doi: 10.1093/bioinformatics/bts390

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., et al. (2014). LocTree3 prediction of localization. *Nucleic Acids Res.* 42, W350–W355. doi: 10.1093/nar/gku396

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Guo, Y., Yang, Y., Huang, Y., and Shen, H. B. (2020). Discovering nuclear targeting signal sequence through protein language learning and multivariate analysis. *Anal. Biochem.* 591:113565. doi: 10.1016/j.ab.2019.113565

Harvey Millar, A., and Taylor, N. L. (2014). Subcellular proteomics-where cell biology meets protein chemistry. *Front. Plant Sci.* 5:55. doi: 10.3389/fpls.2014.00055

Höglund, A., Dönnes, P., Blum, T., Adolph, H. W., and Kohlbacher, O. (2006). MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* 22, 1158–1165. doi: 10.1093/bioinformatics/btl002

Horton, P., and Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5, 147–152.

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Hutten, S., and Kehlenbach, R. H. (2007). CRM1-mediated nuclear export: to the pore and beyond. *Trends Cell Biol.* 17, 193–201. doi: 10.1016/j.tcb.2007.02.003

Imai, K., and Nakai, K. (2010). Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10, 3970–3983. doi: 10.1002/pmic.201000274

Imai, K., and Nakai, K. (2019). "Prediction of protein localization" in *Encyclopedia of Bioinformatics and Computational Biology*, *Vol. 2*. Elsevier, 53–59.

Itzhak, D. N., Davies, C., Tyanova, S., Mishra, A., Williamson, J., Antrobus, R., et al. (2017). A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell Rep.* 20, 2706–2718. doi: 10.1016/j.celrep.2017.08.063

Itzhak, D. N., Tyanova, S., Cox, J., and Borner, G. H. H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* 5:e16950. doi: 10.7554/eLife.16950

Ivankov, D. N., Payne, S. H., Galperin, M. Y., Bonissone, S., Pevzner, P. A., and Frishman, D. (2013). How many signal peptides are there in bacteria? *Environ. Microbiol.* 15, 983–990. doi: 10.1111/1462-2920.12105

Jadot, M., Boonen, M., Thirion, J., Wang, N., Xing, J., Zhao, C., et al. (2017). Accounting for protein subcellular localization: a compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* 16, 194–212. doi: 10.1074/mcp.M116.064527

Jäkel, S., and Görlich, D. (1998). Importin β, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells. *EMBO J.* 17, 4491–4502. doi: 10.1093/emboj/17.15.4491

Jarvis, P. (2008). Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* 179, 257–285. doi: 10.1111/j.1469-8137.2008.02452.x

Jean Beltran, P. M., Mathias, R. A., and Cristea, I. M. (2016). A portrait of the human organelle proteome in space and time during cytomegalovirus infection. *Cell Syst.* 3, 361–373.e6. doi: 10.1016/j.cels.2016.08.012

Jiang, L., Wang, M., Lin, S., Jian, R., Li, X., Chan, J., et al. (2020). A quantitative proteome map of the human body. *Cell* 183, 269–283.e19. doi: 10.1016/j.cell.2020.08.036

Kanapin, A., Batalov, S., Davis, M. J., Gough, J., Grimmond, S., Kawaji, H., et al. (2003). Mouse proteome analysis. *Genome Res.* 13, 1335–1344. doi: 10.1101/gr.978703

Kimura, M., and Imamoto, N. (2014). Biological significance of the importin-β family-dependent nucleocytoplasmic transport. *Traffic* 15, 727–748. doi: 10.1111/tra.12174

Kimura, M., Morinaka, Y., Imai, K., and Kose, S. (2017). Extensive cargo identification reveals distinct biological roles of the 12 importin pathways. *eLife* 6:e21184. doi: 10.7554/eLife.21184

Kosugi, S., Hasebe, M., Entani, T., Takayama, S., Tomita, M., and Yanagawa, H. (2008a). Article design of peptide inhibitors for the importin α/β nuclear import pathway by activity-based profiling. *Chem. Biol.* 15, 940–949. doi: 10.1016/j.chembiol.2008.07.019

Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-sato, E., Tomita, M., et al. (2009). Six classes of nuclear localization signals specific

to different binding grooves of importin α. *J. Biol. Chem.* 284, 478–485. doi: 10.1074/jbc.M807017200

Kosugi, S., Hasebe, M., Tomita, M., and Yanagawa, H. (2008b). Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic* 9, 2053–2062. doi: 10.1111/j.1600-0854.2008.00825.x

Kosugi, S., Yanagawa, H., Terauchi, R., and Tabata, S. (2014). NESmapper: accurate prediction of leucine-rich nuclear export signals using activity-based profiles. *PLoS Comput. Biol.* 10:e1003841. doi: 10.1371/journal.pcbi.1003841

Krahmer, N., Najafi, B., Schueder, F., Quagliarini, F., Steger, M., Seitz, S., et al. (2018). Organellar proteomics and Phospho-proteomics reveal subcellular reorganization in diet-induced hepatic steatosis. *Dev. Cell* 47, 205–221.e7. doi: 10.1016/j.devcel.2018.09.017

Krogh, A., Sonnhammer, E. L. L., and Ka, L. (2007). Advantages of combined transmembrane topology and signal peptide prediction — the Phobius web server. *Nucleic Acids Res.* 35, W429–W432. doi: 10.1093/nar/gkm256

La Cour, T., Kiemer, L., Mølgaard, A., Gupta, R., Skriver, K., and Brunak, S. (2004). Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng. Des. Sel.* 17, 527–536. doi: 10.1093/protein/gzh062

Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E., and Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin α. *J. Biol. Chem.* 282, 5101–5105. doi: 10.1074/jbc.R600026200

Lee, B. J., Cansizoglu, A. E., Su, K. E., Louis, T. H., Zhang, Z., and Chook, Y. M. (2006). Rules for nuclear localization sequence recognition by karyopherin β 2. *Cell* 126, 543–558. doi: 10.1016/j.cell.2006.05.049

Lee, D. W., Lee, S., Lee, J., Woo, S., Razzak, M. A., Vitale, A., et al. (2019). Molecular mechanism of the specificity of protein import into chloroplasts and mitochondria in plant cells. *Mol. Plant* 12, 951–966. doi: 10.1016/j.molp.2019.03.003

Lertampaiporn, S., Nuannimnoi, S., Vorapreeda, T., Chokesajjawatee, N., Visessanguan, W., and Thammarongtham, C. (2019). PSO-LocBact: a consensus method for optimizing multiple classifier results for predicting the subcellular localization of bacterial proteins. *Biomed. Res. Int.* 2019:5617153. doi: 10.1155/2019/5617153

Li, H. M., and Chiu, C. C. (2010). Protein transport into chloroplasts. *Annu. Rev. Plant Biol.* 61, 157–180. doi: 10.1146/annurev-arplant-042809-112222

Liku, M. E., Legere, E. A., and Moses, A. M. (2018). NoLogo: a new statistical model highlights the diversity and suggests new classes of Crm1-dependent nuclear export signals. *BMC Bioinformatics* 19:65. doi: 10.1186/s12859-018-2076-7

Lin, J., and Hu, J. (2013). SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. *PLoS One* 8:e76864. doi: 10.1371/journal.pone.0076864

Lisitsyna, O. M., Seplyarskiy, V. B., and Sheval, E. V. (2017). Comparative analysis of nuclear localization signal (NLS) prediction methods. *Biopolym. Cell* 33, 147–154. doi: 10.7124/bc.00094C

Lundberg, E., and Borner, G. H. H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* 20, 285–302. doi: 10.1038/s41580-018-0094-y

Maertens, G. N., Cook, N. J., Wang, W., Hare, S., Shree, S., and Öztop, I. (2014). Structural basis for nuclear import of splicing factors by human transportin 3. *Proc. Natl. Acad. Sci. U. S. A.* 111, 2728–2733. doi: 10.1073/pnas.1320755111

Martelli, P. L., Fariselli, P., and Casadio, R. (2003). An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* 19, i205–i211. doi: 10.1093/bioinformatics/btg1027

Mathur, D., Singh, S., Mehta, A., Agrawal, P., and Raghava, G. P. S. (2018). In silico approaches for predicting the half-life of natural and modified peptides in blood. *PLoS One* 13:e0196829. doi: 10.1371/journal.pone.0196829

Mehdi, A. M., Sehgal, M. S. B., Kobe, B., Bailey, T. L., and Bodén, M. (2011). A probabilistic model of nuclear import of proteins. *Bioinformatics* 27, 1239–1246. doi: 10.1093/bioinformatics/btr121

Mossmann, D., Meisinger, C., and Vögtle, F. N. (2012). Processing of mitochondrial presequences. *Biochim. Biophys. Acta Gene Regul. Mech.* 1819, 1098–1106. doi: 10.1016/j.bbagrm.2011.11.007

Nakai, K. (2001). Review: prediction of *in vivo* fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* 134, 103–116. doi: 10.1006/jsbi.2001.4378

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins Struct. Funct. Bioinforma.* 11, 95–110. doi: 10.1002/prot.340110203

Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14, 897–911. doi: 10.1016/S0888-7543(05)80111-9

Nielsen, H. (2017). Protein sorting prediction. *Methods Mol. Biol.* 1615, 23–57. doi: 10.1007/978-1-4939-7033-9_2

Nielsen, H., Tsirigos, K. D., Brunak, S., and von Heijne, G. (2019). A brief history of protein sorting prediction. *Protein J.* 38, 200–216. doi: 10.1007/s10930-019-09838-3

Nightingale, D. J. H., Oliver, S. G., and Lilley, K. S. (2019). Mapping the *Saccharomyces cerevisiae* spatial proteome with high resolution using hyperLOPIT. *Methods Mol. Biol.* 2049, 165–190. doi: 10.1007/978-1-4939-9736-7_10

Nilsson, I., Lara, P., Hessa, T., Johnson, A. E., von Heijne, G.V., and Karamyshev, A. L. (2015). The code for directing proteins for translocation across ER membrane: SRP cotranslationally recognizes specific features of a signal sequence. *J. Mol. Biol.* 427, 1191–1201. doi:10.1016/j.jmb.2014.06.014

Orioli, T., and Vihinen, M. (2019). Benchmarking subcellular localization and variant tolerance predictors on membrane proteins. *BMC Genomics* 20:547. doi: 10.1186/s12864-019-5865-0

Orre, L. M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., Fernandez Woodbridge, A., et al. (2019). SubCellBarCode: proteome-wide mapping of protein localization and relocalization. *Mol. Cell* 73, 166–182.e7. doi: 10.1016/j.molcel.2018.11.035

Paila, Y. D., Richardson, L. G. L., and Schnell, D. J. (2015). New insights into the mechanism of chloroplast protein import and its integration with protein quality control, organelle biogenesis and development. *J. Mol. Biol.* 427, 1038–1060. doi: 10.1016/j.jmb.2014.08.016

Palmer, T., and Stansfeld, P. J. (2020). Targeting of proteins to the twin-arginine translocation pathway. *Mol. Microbiol.* 113, 861–871. doi: 10.1111/mmi.14461

Paramasivam, N., and Linke, D. (2011). Clubsub-P: cluster-based subcellular localization prediction for gram-negative bacteria and archaea. *Front. Microbiol.* 2:218. doi: 10.3389/fmicb.2011.00218

Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R., and Brinkman, F. S. L. (2016). PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.* 44, D663–D668. doi: 10.1093/nar/gkv1271

Peabody, M. A., Lau, W. Y. V., Hoad, G. R., Jia, B., Maguire, F., Gray, K. L., et al. (2020). PSORTm: a bacterial and archaeal protein subcellular localization prediction tool for metagenomics data. *Bioinformatics* 36, 3043–3048. doi: 10.1093/bioinformatics/btaa136

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pfanner, N., Warscheid, B., and Wiedemann, N. (2019). Mitochondrial proteins: from biogenesis to functional networks. *Nat. Rev. Mol. Cell Biol.* 20, 267–284. doi: 10.1038/s41580-018-0092-0

Pierleoni, A., Martelli, P., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 9:392. doi: 10.1186/1471-2105-9-392

Pierleoni, A., Martelli, P. L., and Casadio, R. (2011). MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics* 27, 1224–1230. doi: 10.1093/bioinformatics/btr108

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22, e408–e416. doi: 10.1093/bioinformatics/btl222

Prieto, G., Fullaondo, A., and Rodriguez, J. A. (2014). Prediction of nuclear export signals using weighted regular expressions (Wregex). *Bioinformatics* 30, 1220–1227. doi: 10.1093/bioinformatics/btu016

Salvatore, M., Warholm, P., Shu, N., Basile, W., and Elofsson, A. (2017). SubCons: a new ensemble method for improved human subcellular localization predictions. *Bioinformatics* 33, 2464–2470. doi: 10.1093/bioinformatics/btx219

Savojardo, C., Fariselli, P., and Casadio, R. (2013). BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics* 29, 504–505. doi: 10.1093/bioinformatics/bts728

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2015). TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics* 31, 3269–3275. doi: 10.1093/bioinformatics/btv367

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2017). SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments. *Bioinformatics* 33, 347–353. doi: 10.1093/bioinformatics/btw656

Savojardo, C., Martelli, P. L., Fariselli, P., and Casadio, R. (2018a). DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics* 34, 1690–1696. doi: 10.1093/bioinformatics/btx818

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018b). BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, W459–W466. doi: 10.1093/nar/gky320

Schneider, G., Sjöling, S., Wallin, E., Wrede, P., Glaser, E., and von Heijne, G. (1998). Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins Struct. Funct. Genet.* 30, 49–60. doi: 10.1002/(SICI)1097-0134(19980101)30:1<49::AID-PROT5>3.0.CO;2-F

Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2020). Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinform.* 21, 1628–1640. doi: 10.1093/bib/bbz106

Siegel, S. D., Reardon, M. E., and Ton-That, H. (2017). Anchoring of LPXTG-like proteins to the gram-positive cell wall envelope. *Curr. Top. Microbiol. Immunol.* 404, 159–175. doi: 10.1007/82_2016_8

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4, 1581–1590. doi: 10.1002/pmic.200300776

Stekhoven, D. J., Omasits, U., Quebatte, M., Dehio, C., and Ahrens, C. H. (2014). Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. *J. Proteomics* 99, 123–137. doi: 10.1016/j.jprot.2014.01.015

Thul, P. J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017). A subcellular map of the human proteome. *Science* 356:eaal3321. doi: 10.1126/science.aal3321

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* 28, 1248–1250. doi: 10.1038/nbt1210-1248

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Vakser, I. A. (2020). Challenges in protein docking. *Curr. Opin. Struct. Biol.* 64, 160–165. doi: 10.1016/j.sbi.2020.07.001

Vögtle, F. N., Wortelkamp, S., Zahedi, R. P., Becker, D., Leidhold, C., Gevaert, K., et al. (2009). Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* 139, 428–439. doi: 10.1016/j.cell.2009.07.045

von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.* 5, 1335–1342. doi: 10.1002/j.1460-2075.1986.tb04364.x

von Heijne, G. (1990). The signal peptide. *J. Membr. Biol.* 115, 195–201. doi: 10.1007/BF01868635

Wan, S., Mak, M. W., and Kung, S. Y. (2012). mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics* 13:290. doi: 10.1186/1471-2105-13-290

Wang, X., Zhang, J., and Li, G. Z. (2015). Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. *BMC Bioinformatics* 16:S1. doi: 10.1186/1471-2105-16-S12-S1

Xu, D., Marquis, K., Pei, J., Fu, S. C., Caıatay, T., Grishin, N. V., et al. (2015). LocNES: a computational tool for locating classical NESs in CRM1 cargo proteins. *Bioinformatics* 31, 1357–1365. doi: 10.1093/bioinformatics/btu826

Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins Struct. Funct. Genet.* 64, 643–651. doi: 10.1002/prot.21018

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615. doi: 10.1093/bioinformatics/btq249

Zhang, S., Xia, X., Shen, J., Zhou, Y., and Sun, Z. (2008). DBMLoc: a database of proteins with multiple subcellular localizations. *BMC Bioinformatics* 9:127. doi: 10.1186/1471-2105-9-127

Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., et al. (2008). Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 3:e1994. doi: 10.1371/journal.pone.0001994

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 December 2020
doi: 10.3389/fgene.2020.573877

Analysis of Hub Genes Involved in Distinction Between Aged and Fetal Bone Marrow Mesenchymal Stem Cells by Robust Rank Aggregation and Multiple Functional Annotation Methods

Xiaoyao Liu[1], Mingjing Yin[1], Xinpeng Liu[1], Junlong Da[1], Kai Zhang[1], Xinjian Zhang[1], Lixue Liu[1], Jianqun Wang[1], Han Jin[1], Zhongshuang Liu[1], Bin Zhang[1,2] and Ying Li[1]**

[1] Institute of Hard Tissue Development and Regeneration, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Heilongjiang Academy of Medical Sciences, Harbin, China

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong,
Hong Kong

Reviewed by:
Bhanwar Lal Puniya,
University of Nebraska-Lincoln,
United States
Priyanka Baloni,
Institute for Systems Biology (ISB),
United States

***Correspondence:**
Bin Zhang
zhangbin@hrbmu.edu.cn
Ying Li
liying@hrbmu.edu.cn

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

Received: *18 June 2020*
Accepted: *24 November 2020*
Published: *14 December 2020*

Citation:
Liu X, Yin M, Liu X, Da J, Zhang K, Zhang X, Liu L, Wang J, Jin H, Liu Z, Zhang B and Li Y (2020) Analysis of Hub Genes Involved in Distinction Between Aged and Fetal Bone Marrow Mesenchymal Stem Cells by Robust Rank Aggregation and Multiple Functional Annotation Methods. Front. Genet. 11:573877. doi: 10.3389/fgene.2020.573877

Stem cells from fetal tissue protect against aging and possess greater proliferative capacity than their adult counterparts. These cells can more readily expand *in vitro* and senesce later in culture. However, the underlying molecular mechanisms for these differences are still not fully understood. In this study, we used a robust rank aggregation (RRA) method to discover robust differentially expressed genes (DEGs) between fetal bone marrow mesenchymal stem cells (fMSCs) and aged adult bone marrow mesenchymal stem cells (aMSCs). Multiple methods, including gene set enrichment analysis (GSEA), Gene Ontology (GO) analysis, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed for functional annotation of the robust DEGs, and the results were visualized using the R software. The hub genes and other genes with which they interacted directly were detected by protein–protein interaction (PPI) network analysis. Correlation of gene expression was measured by Pearson correlation coefficient. A total of 388 up-regulated and 289 down-regulated DEGs were identified between aMSCs and fMSCs. We found that the down-regulated genes were mainly involved in the cell cycle, telomerase activity, and stem cell proliferation. The up-regulated DEGs were associated with cell adhesion molecules, extracellular matrix (ECM)–receptor interactions, and the immune response. We screened out four hub genes, *MYC*, *KIF20A*, *HLA-DRA*, and *HLA-DPA1*, through PPI-network analysis. The *MYC* gene was negatively correlated with *TXNIP*, an age-related gene, and *KIF20A* was extensively involved in the cell cycle. The results suggested that MSCs derived from the bone marrow of an elderly donor present a pro-inflammatory phenotype compared with that of fMSCs, and the *HLA-DRA* and *HLA-DPA1* genes are related to the immune response. These findings provide new insights into the differences between aMSCs and fMSCs and may suggest novel strategies for *ex vivo* expansion and application of adult MSCs.

Keywords: bone marrow mesenchymal stem cell, fetal stem cell, adult setm cell, robust rank aggregation, hub genes

INTRODUCTION

Stem cells can be isolated at all stages of ontogeny, from the early developing embryo to the post-reproductive adult organism. Adult stem cells are less potent than embryonic stem cells, but still play a very important role in maintaining overall health (Jin, 2017). Adult bone marrow was the first source of mesenchymal stem cells (MSCs) to be identified and is still by far the best characterized (Friedenstein et al., 1987; Kolf et al., 2007; Lv et al., 2014). These cells hold great promise as seed cells in tissue engineering and regenerative medicine, based on their self-renewal, multi-differentiation, and immunoregulation abilities (Dogan et al., 2014; Wei et al., 2015; Castagnini et al., 2016; Mehrabani et al., 2016). However, there is growing evidence that demonstrates that the number of bone marrow-derived MSCs is limited and declines with the age of the donor (Dexheimer et al., 2011). Thus, long-term cell culture is needed to obtain large numbers of cells suitable for clinical applications. However, MSCs may undergo senescence, as well as impaired function during *ex vivo* expansion (Turinetto et al., 2016).

Although fetal and adult MSCs share the same morphology and surface molecules, previous studies have shown that MSCs from fetal tissues are more adaptable, with greater self-renewal capacity, both *in vivo* and *in vitro* (O'Donoghue and Fisk, 2004; Guillot et al., 2007; Ding et al., 2011). The prevalence of MSCs in fetal bone marrow is significantly higher than that in adult tissue (O'Donoghue and Fisk, 2004; Ding et al., 2011). Fetal MSCs are readily expandable *in vitro*, with a shorter doubling time, and display no obvious change in phenotype after 20 passages (Campagnoli et al., 2001). Existing research recognizes the critical role telomerase plays in self-renewal and the replicative potential of stem cells (Hayflick, 2000). Comparative studies of fetal liver hematopoietic stem cells (HSCs) and adult bone marrow HSCs have confirmed that fetal stem cells have higher telomerase activity and again suggest that proliferative potential is limited and declines with age (Verfaillie et al., 2002). In addition, telomere length is longer in fetal MSCs (Xu et al., 2016).

Another advantage of MSCs is their immunomodulatory properties (Chen et al., 2011; Andrzejewska et al., 2019). Bone marrow MSCs (BM-MSCs) from both adults and fetuses reportedly possess immune-suppressive effects. Fetal MSCs display immunological inertness and appear to have stronger immunomodulatory abilities than their adult counterparts (Götherström et al., 2005; Chang et al., 2006).

Mesenchymal stem cells isolated from fetal tissue may therefore have greater potential for clinical application, but the exact mechanisms by which they exert their effects are still not very clear. Moreover, the application of fetal tissue is not widely accepted and is still being debated. Understanding the difference between fetal and adult stem cells and their regulatory mechanisms may provide new insights for the clinical application of adult stem cells.

In this study, two existing datasets from Gene Expression Omnibus (GEO) were analyzed by the robust rank aggregation (RRA) method, which facilitates the detection of genes that are ranked consistently in multiple datasets and assigns a significance score for each gene (Reimand et al., 2010). This method was used to identify robust differentially expressed genes (DEGs) between MSCs derived from elderly adult bone marrow and fetal bone marrow. Functions of these robust DEGs were then explored by gene set enrichment analysis (GSEA), Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. Using protein–protein interaction (PPI) network analysis, we screened out four hub genes, *MYC*, *KIF20A*, *HLA-DRA*, and *HLA-DPA1*, which were closely related to function. Furthermore, GO and KEGG enrichment analyses were utilized to verify potential biological functions of hub genes and their first neighbors.

MATERIALS AND METHODS

Microarray Data Information

All available datasets were acquired from the GEO database[1]. The screening criteria of datasets were inclusion of gene expression data of BM-MSCs from fetal and aged donors. Eventually, two datasets, GSE97311 (Spitzhorn et al., 2019) and GSE68374 (Paciejewska et al., 2016), were included in the study. Series matrix files and platform information of GSE97311 and GSE68374 were downloaded from the GEO database for further study. The GSE97311 dataset contained three fetal femur-derived MSC samples and four adult MSC samples. The GSE68374 dataset included three biological replicates for both fetal and adult bone marrow-derived MSC samples.

Data Processing and Identification of Robust DEGs

The microarray data of GSE97311 and GSE68374 were initially normalized and differential expression was analyzed using the R software though the "limma" package (Ritchie et al., 2015). The results were presented on a volcano plot (Li, 2012). We then used the "RobustRankAggreg" package (Reimand et al., 2010) to integrate the differential expression results of the two datasets to identify the robust DEGs. As this RRA method screens genes ranked consistently better than expected based on null hypothesis of uncorrelated inputs, batch effect correction is not needed (Liu et al., 2018). Benjamini–Hochberg's method was used to control the false discovery rate (FDR). The *P*-value of each gene represents its ranking in the final gene list. Genes with a P-value < 0.05 and $|\log FC| > 1$ in the final list were considered significant DEGs for the next mining. The R package "pheatmap" was used to visualize expression patterns of the top 40 DEGS (top 20 up-regulated genes and top 20 down-regulated genes) from RRA analysis.

Gene Set Enrichment Analysis (GSEA)

The following sets were downloaded from the Molecular Signatures Database version 7.1[2]: H.all.v7.symbles.gmt, c2.cp.kegg.v7.1.symbols.gmt, and other interesting gene sets involved in the oxidative response, production of interleukin 6, telomerase activity, and stem cell self-renewal

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

[2] www.gsea-msigdb.org/gsea/msigdb/index.jsp

in c5.bp.v7.1.symbols.gmt. The GseaPreranked tool was then used to perform enrichment analysis for all the DEGs integrated via RRA method, which are ranked by log FC from large to small. Gene set permutations were performed 1000 times for each analysis. We then visualized the results of GSEA using "ggplot2" in the R package (Ito and Murphy, 2013).

Functional Enrichment Analysis of Robust DEGs

BinGO (Maere et al., 2005), a plug-in of Cytoscape, was used for GO enrichment. The KEGG pathway analyses were conducted by the R package "clusterprofiler" (Yu et al., 2012). The GO terms and KEGG pathways with adjusted P-value < 0.05 were considered statistically significant and visualized by the "GOplot" package (Walter et al., 2015). The Z-score was calculated, which hinted at whether the biological process (or/molecular function/cellular component) or KEGG pathway was more likely to be reduced (negative value) or increased (positive value).

$$z - score = \backslash frac\left\{(up - down)\right\}\{\sqrt{\{count\}}\}$$

Identification of Hub Genes and Their First Neighbors by PPI Network Analysis

The DEGs with $P < 0.001$ were defined as the most robust DEGs and uploaded to the STRING database to establish a PPI network (Szklarczyk et al., 2017). Interaction scores > 0.4 were set as the cut-off point. The STRING analysis results were then imported into the Cytoscape software version 3.8.0, and the network was ranked by degree and betweenness methods using the cytoHubba plug-in (Chin et al., 2014) to select hub genes. Hub genes were screened according to the degree score > 10 and ranked at the top 10 of total genes, sorted by the betweenness method. We then selected the first neighbors of hub genes that were directly related to the hub genes, to construct their sub-networks, respectively.

Correlation Analysis and Functional Enrichment Analysis

Correlations were analyzed by Pearson's correlation (Schober et al., 2018) for genes involved in the sub-network, which was built by hub gene and its first neighbors. Genes with a Pearson's correlation coefficient greater than 0.5 were considered most significant correlation with hub genes and were selected for GO annotation and KEGG pathway enrichment analysis. A total of 13 samples were included in correlation analysis and the expression of genes was obtained from GSE97311 and GSE68374 datasets.

RESULTS

Identification of Robust DEGs by the RRA Method

The expression profiles of aged adult BM-MSCs (aMSCs) were compared with those of fetal BM-MSCs (fMSCs). Based on the screening criteria of $P < 0.05$ and $|\ \text{logFC}|\ > 1$, a total of 933 DEGs were identified from GSE97311, including 553 up-regulated genes and 380 down-regulated genes (**Figure 1A**). In addition, 993 DEGs, including 496 up-regulated genes and 497 down-regulated genes, were identified from GSE68374, according to the same criteria (**Figure 1B**). We integrated the results of the two datasets using the RRA method, and obtained 14,024 up-regulated genes and 9872 down-regulated genes (**Supplementary Table S1**), which finally yielded 677 robust DEGs. A heatmap of the top 20 up-regulated robust genes and top 20 down-regulated robust genes are presented in **Figure 1C**, and a complete list of the robust DEGs is provided in **Supplementary Table S2**. The most significant up-regulated gene was *AKR1C3* (P = 4.24E-07, logFC = 4.005), followed by *FMO3* (P = 4.79E-06, logFC = 3.849), and *TMEM140* (P = 5.60E-06, logFC = 3.385). Moreover, *FBN* (P = 2.02E-06, logFC = −3.449); *SCD* (P = 4.05E-06, logFC = −3.363); *CLDN1* (P = 5.60E-06, logFC = −3.209) were the most significant down-regulated genes.

GSEA Reveals Differences Between aMSCs and fMSCs

The ranked list of DEGs is shown in **Supplementary Table S3**. Results of GSEA for KEGG pathways revealed that most up-regulated DEGs in aMSCs were related to immune-related diseases (**Figure 2A**), and down-regulated DEGs were abundant in the cell cycle, ribosome pathway, and spliceosome pathway (**Figure 2C**). Using hallmark gene sets, the interferon response and myogenesis were enriched in aMSCs (**Figure 2B**). Moreover, E2F targets, the G2M checkpoint, and MYC targets were enriched in fMSCs (**Figure 2D**). Gene sets with the highest enrichment scores were all associated with the cell cycle. All gene sets were significantly enriched at an FDR < 0.05.

Analysis of Gene Sets Related to Aging and Stem Cell Self-Renewal

To further determine the functions of robust DEGs in specific biological processes, we performed GSEA analysis related to aging and stem cell self-renewal. The GO annotation in terms of response to oxidative stress and IL-6 production enriched in the aMSC group (**Figure 3A**). The GO gene sets of telomerase activity, somatic stem cell population maintenance, stem cell division, and stem cell proliferation were all enriched in fMSCs (**Figure 3B**).

GO and KEGG Enrichment Analysis of Robust DEGs

Using the BINGO plug-in, we obtained a global perspective of the changes in gene expression patterns. The up-regulated DEGs in aMSCs were enriched in biological processes such as the immune response, response to stimulus, extracellular structure organization, cell adhesion, and biological adhesion (**Figure 4A**). In contrast, down-regulated DEGs were mainly enriched in the cell cycle and developmental processes (**Figure 4B**). Based on the results of KEGG enrichment analysis, the top five significant pathways were cell adhesion molecules, complement and coagulation cascades, *Staphylococcus*

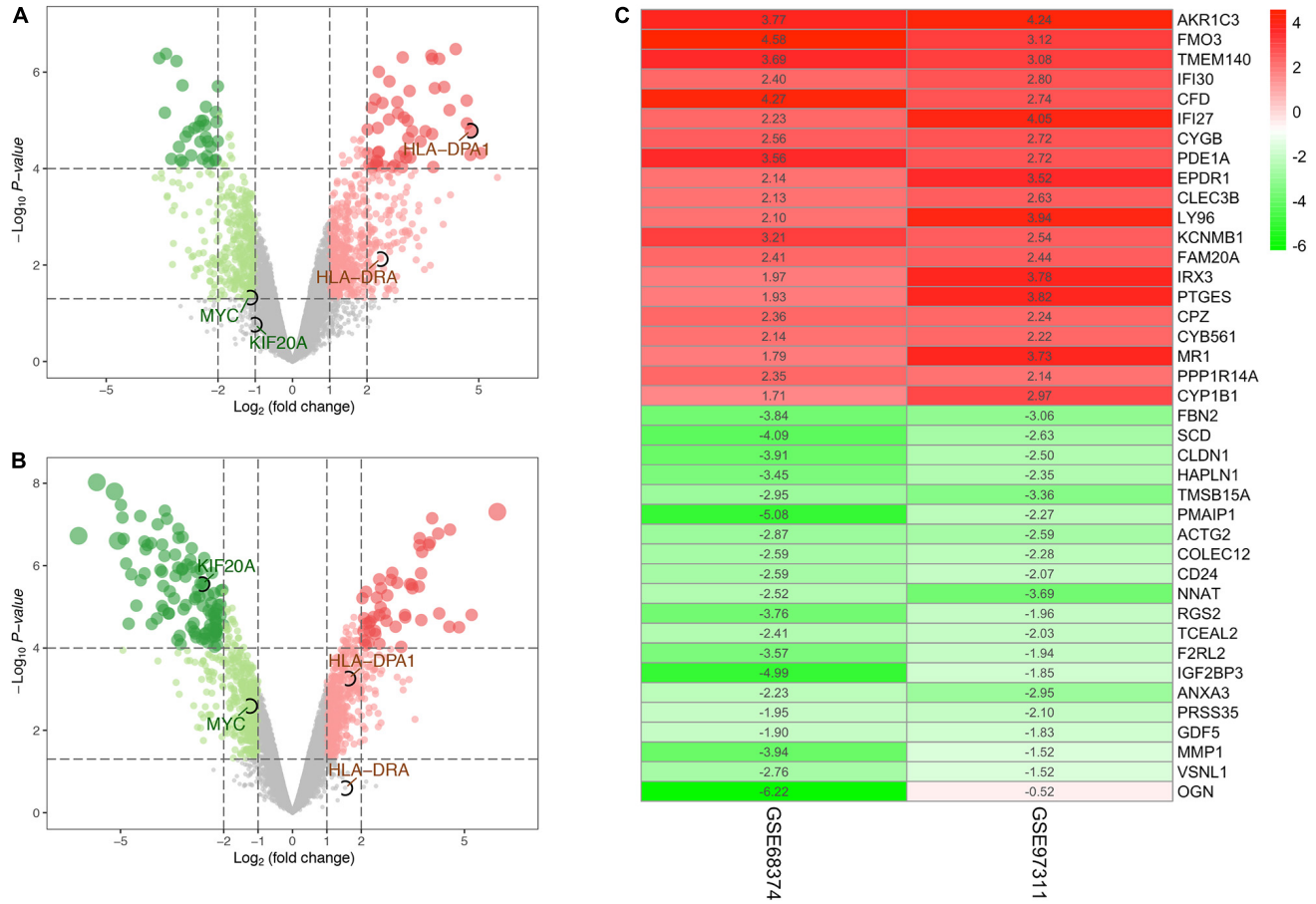


FIGURE 1 | Identification of differentially expressed genes (DEGs). **(A)** Volcano plot of DEGs for dataset GSE68374. **(B)** Volcano plot of DEGs for dataset GSE97311. **(C)** The heatmap of the top 40 (20 up-regulated and 20 down-regulated) DEGs according to *P*-value identified by the RRA method. The number in the heat map represents log FC. The gradual color ranging from red to green indicates the changing process from up- to down-regulation. Genes circle by black ellipses are the hub genes screened out by subsequent analysis.

aureus infection, hematopoietic cell lineage, and extracellular matrix (ECM)–receptor interaction. Unlike the other pathways, the peroxisome proliferator-activated receptor (PPAR) signaling pathway was more likely to be inhibited in aMSCs (**Figures 5A,B**).

Identification of Hub Genes and Their First Neighbors

The PPI network established by the most robust DEGs ($P < 0.001$) contained 144 nodes and 291 edges. We screened four hub genes, considering the degree (DC) and betweenness (BC) centrality (**Figures 6A,B**). Among them, *MYC* (DC = 26, BC = 9816.96488) and *KIF20A* (DC = 17, BC = 4690.54753) were up-regulated in the fMSC group, and *HLA-DRA* (DC = 11, BC = 2105.09486) and *HLA-DPA1* (DC = 11, BC = 2105.09486) were up-regulated in the aMSC group. We then selected their first neighbors and structured the respective sub-networks. As shown in **Figures 7A–C**, there were 26 nodes and 63 edges in the sub-network of *MYC*; and 18 nodes and 89 edges in the sub-network of *KIF20A*. Both *HLA-DRA* and *HLA-DPA1* were part of the same network, which included 12 nodes and 34 edges.

Correlation Analysis and Functional Enrichment Analysis of Sub-Networks

We analyzed gene-gene expression correlation coefficients for genes in sub-networks (**Figures 7D–F**) and filtered out genes with a Pearson correlation coefficient > 0.5. Correlation coefficient, *P*-values, and coefficient of variation for all the genes included in the correlation analysis are shown in **Supplementary Tables S4, S5**. Some interesting examples are shown in **Supplementary Figure S1**. The GO and KEGG pathway analyses were also performed for these genes (**Figures 8A–F**). For *MYC*, we observed that both *MYC* and *TXNIP*, a gene up-regulated in aMSCs, were involved in negative regulation of cell division. The KEGG analysis showed that the cell cycle, breast cancer, oocyte meiosis, and human T-cell leukemia virus 1 infection were enriched. Several biological processes and GO terms, such as miotic nuclear division, mitotic sister chromatid segregation, and nuclear division, were abundant in *KIF20A*-related genes.

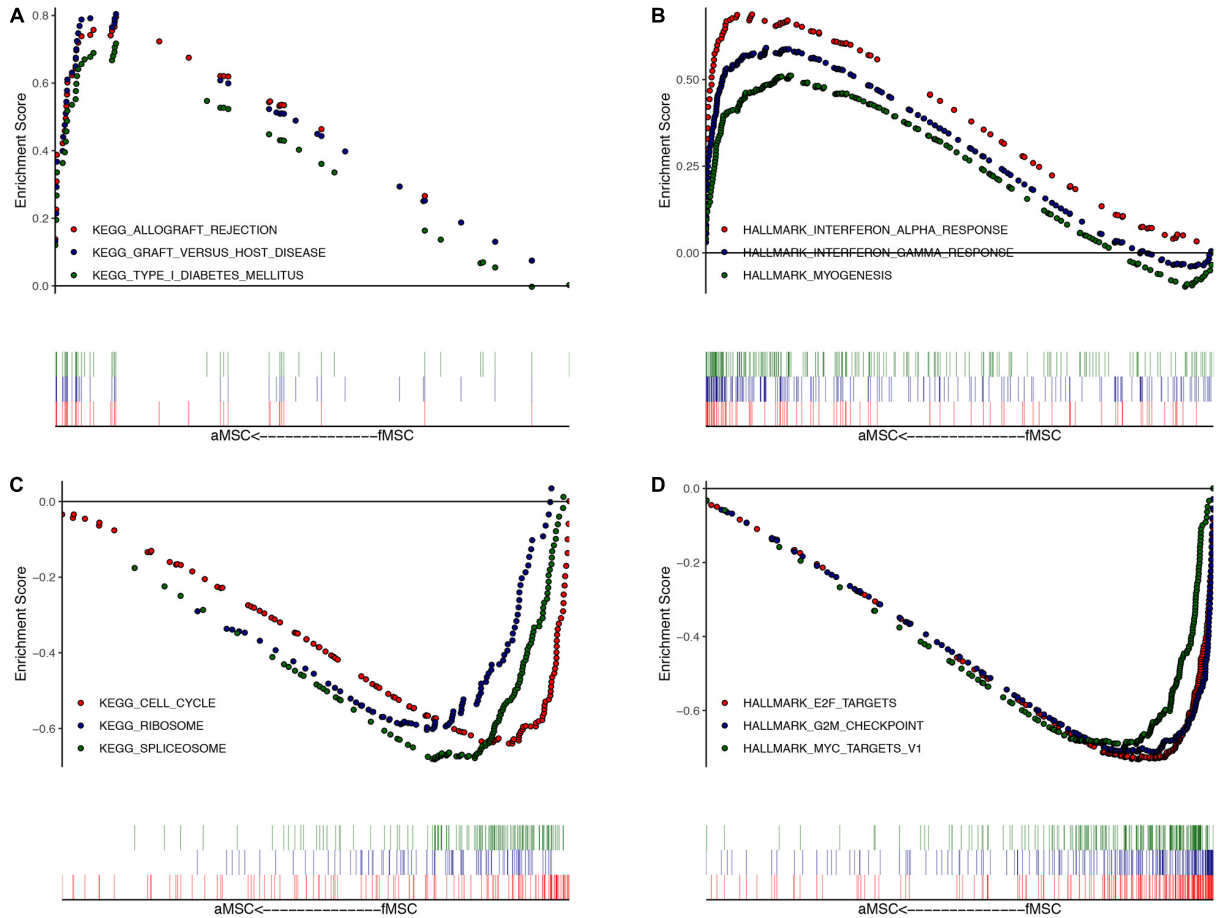


FIGURE 2 | Gene set enrichment analysis (GSEA) for DEGs. The top three enriched gene sets (according to GSEA enrichment score) in aMSC group: **(A)** for KEGG; **(B)** for HALLMARK. The top three enriched gene sets (according to GSEA enrichment score) in fMSC group: **(C)** for KEGG; **(D)** for HALLMARK. All gene sets are significantly enriched at a false discovery rate (FDR) < 0.05. The location of the vertical bar shows the occurrence of the gene in the ranked list.

Pathways implicated with these genes were similar to *MYC* and its closely related neighbors. For *HLA-DRA* and *HLA-DPA1*, GO terms such as antigen processing and presentation of peptide antigen, and the interferon gamma mediated signaling pathway were enriched; and KEGG pathways associated with antigen processing and presentation, hematopoietic cell lineage, and Th cell differentiation were enriched.

DISCUSSION

Human BM-MSCs (hBM-MSCs) are promising sources for tissue engineering and regenerative medicine. Human fetal MSCs (hf-MSCs) have more primitive expression profiles and greater proliferative capacity than their adult counterparts (Spitzhorn et al., 2019). These cells can more readily expand *in vitro* and senesce later in culture. Both aMSCs and fMSCs harbor immunomodulatory capacity and are non-immunogenic, even though some differences have been reported (Götherström et al., 2005; Chang et al., 2006; Chen et al., 2011; Andrzejewska et al., 2019). The underlying molecular mechanisms for these differences are still not fully understood.

In this study, we used bioinformatics to mine the underlying molecular mechanisms that explain the difference between aMSCs and fMSCs. To our best knowledge, this is the first study to use the RRA method to analyze the difference between aMSC and fMSC sources in human bone marrow. Götherström et al. (2005) explored the gene expression profile of MSCs derived from the fetal liver and adult bone marrow. Other studies have shown that MSCs derived from different sources possess distinct biological properties (Berebichez-Fridman and Montero-Olvera, 2018; Kozlowska et al., 2019).

Only two datasets in the GEO database, GSE97311 and GSE 68374, met our experimental requirements. We performed robust differential expression profiling analysis using the two existing GEO datasets and obtained 677 robust DEGs, including 388 up-regulated and 289 down-regulated DEGs in aMSCs compared with fMSCs. The most significantly up-regulated gene in aMSCs was *AKR1C3*. This gene may play an important role in the pathogenesis of allergic diseases such asthma and may have a role in controlling cell growth or differentiation.

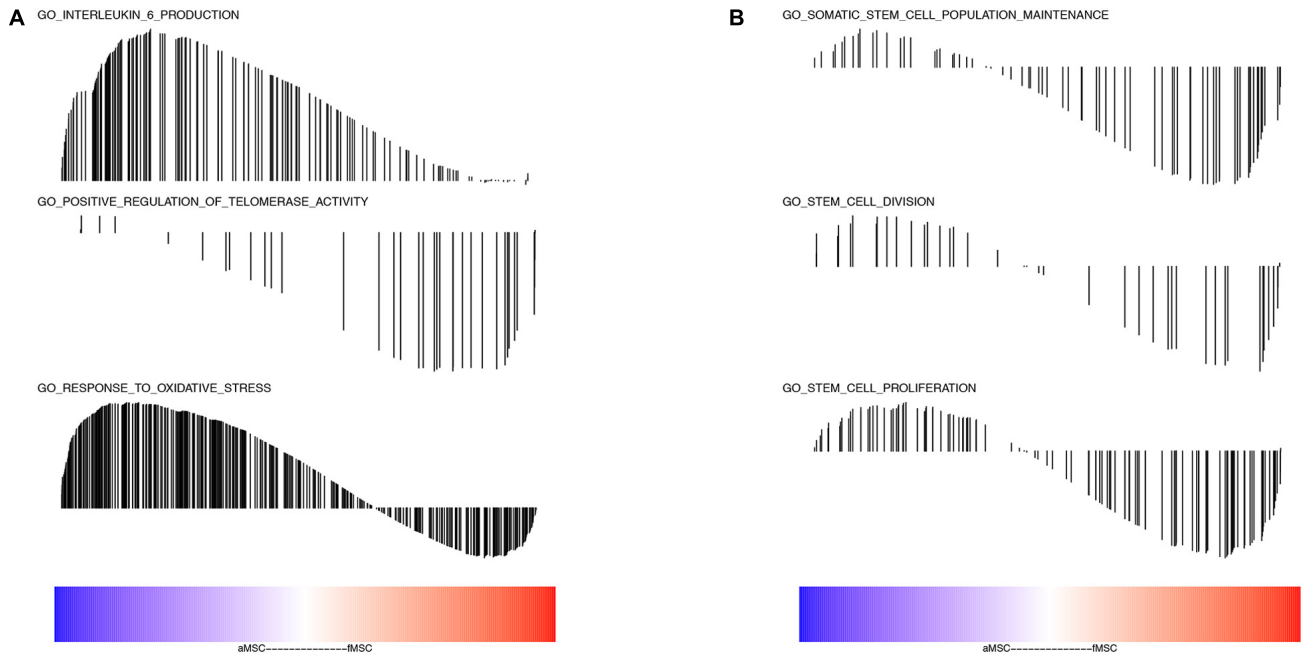


FIGURE 3 | Analyses of gene sets related to aging and stem cell self-renewal. **(A)** Interleukin 6 production (NES = 1.939, FDR q-value = 0.019); positive regulation of telomerase activity (NES = –1.616, FDR q-value = 0.021); response to oxidative stress (NES = 1.4371426, FDR q-value = 0.076). **(B)** Somatic stem cell population maintenance (NES = –1.378, FDR q-value = 0.069); stem cell division (NES = –1.207, FDR q-value = 0.158); stem cell proliferation (NES = –1.266, FDR q-value = 0.046). The color bar indicates the sort order of genes (blue represents up-regulated in aMSC group and red represents up-regulated in fMSC group). The location of vertical bar indicates the occurrence of that gene within the ranked list and the height of the bar indicated the enrichment score.

The most significantly up-regulated gene in fMSCs was *FBN2*, which regulates the early processes of elastic fiber assembly and osteoblast maturation.

The GSEA was conducted for all DEGs in the final robust rank list. Gene sets abundant in fMSCs, such as those associated with the cell cycle, E2F targets, and *MYC* targets, were all related to proliferation. These results further support the notion that fMSCs have greater self-renewal abilities than aMSCs and are consistent with earlier observations (Chen et al., 2011). One unexpected finding was the up-regulated immune response in aMSCs. We found all of the most significantly enriched gene sets in aMSCs involved in the immune response. Several reports have shown that aging leads to the pro-inflammatory phenotype, with activated innate immune responses (Danilova, 2006; Salminen et al., 2008). The chronic low-grade inflammatory state of elderly donors may be the reason for the heightened immune response of aMSCs.

As previous studies have shown, IL-6 is a reliable aging parameter and senescent MSCs release excess IL-6 (Salminen et al., 2008; Suvakov et al., 2019). Thus, we conducted GSEA on interesting biological processes, including IL-6 production, telomerase activity, oxidative stress, and stem cell self-renewal. The results showed enrichment of IL-6 production and oxidative stress in aMSCs; and enrichment of telomerase activity and stem cell proliferation related gene sets in fMSCs. These results further suggest that MSCs derived from elderly adults possess age-related characteristics. The disadvantages of aMSCs could be partially attributed to these intrinsic age-related drawbacks.

Consistent with published data, robust DEGs enriched in several KEGG pathways, such as cell adhesion and ECM-receptor, which also participate in the immune response, are reportedly down-regulated pluripotency markers that inhibit mouse embryonic stem cell self-renewal (Taleahmad et al., 2015). The PPAR pathway is down-regulated in aMSCs; and this might be explained by reduced PPAR activity related to increased inflammation levels in old age (Michalik and Wahli, 2008). Regarding GO annotation, the up-regulated DEGs in adults compared with those in fetuses were involved in the immune response, and cell–cell and cell–ECM contact; whereas down-regulated expression was observed in aMSCs compared with fMSCs in cell cycle progression and development.

The PPI network was then constructed by the most robust DEGs with P-values < 0.001 and $|\log FC| > 1$, to evaluate the relationship between these genes and identify hub genes. We detected four hub genes: *MYC*, *KIF20A*, *HLA-DRA*, and *HLA-DPA1* according to BC and DC.

Prior studies have noted the important role of *MYC* in a range of cellular processes, including proliferation, the cell cycle, and pluripotency maintenance in stem cells (Lüscher, 2001; Kumamoto et al., 2009; Chen et al., 2018). Furthermore, c-MYC can inhibit replicative senescence caused by telomere damage by promoting the expression of human telomerase reverse transcriptase (hTERT), a catalytic subunit of telomerase (Xu et al., 2001). Past research has revealed that high expression of c-MYC is associated with increased self-renewal and differentiation, which is regulated by Sox2 (Park et al., 2012). A recent study showed

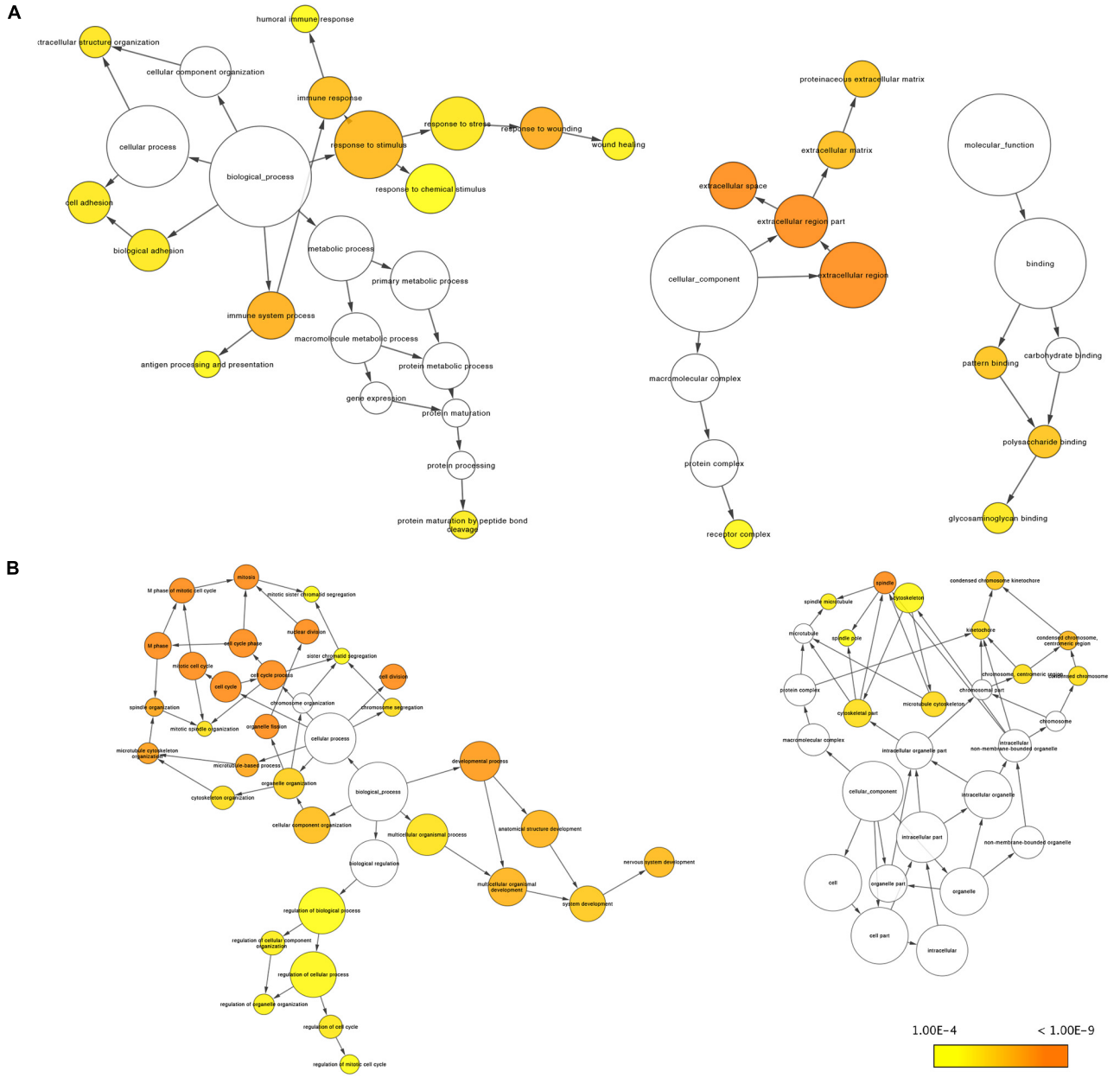
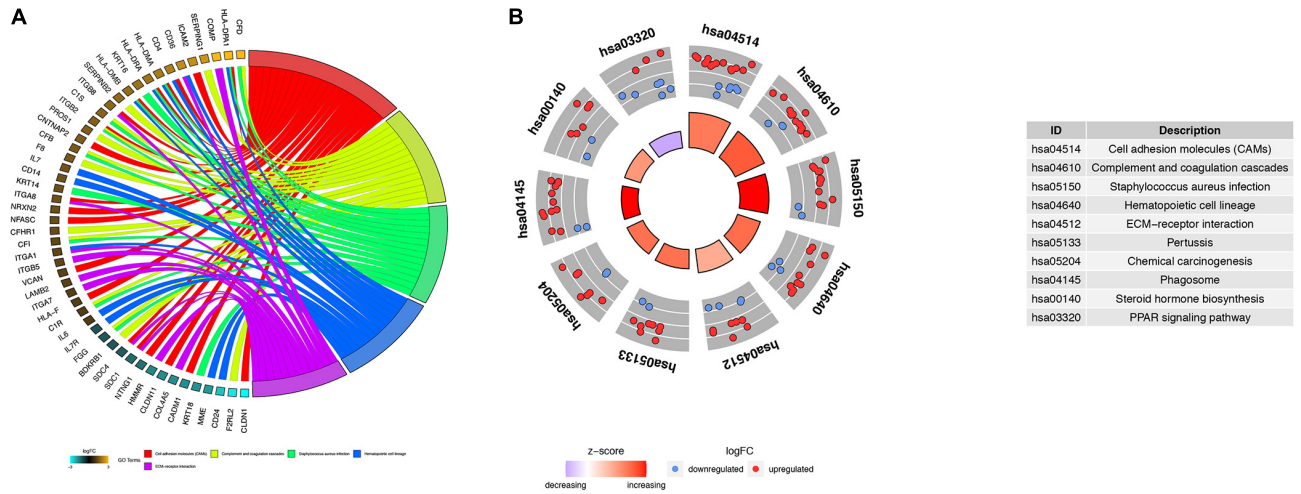


FIGURE 4 | Gene Ontology analyses of robust up-regulated **(A)** and down-regulated **(B)** DEGs by BINGO plug-in. There is no enrichment for molecular function in **B**. The size of the nodes is proportional to the number of genes in that term. Significance decreases from dark orange (p = 1.00E-9) to yellow (p = 1.00E-4).

that the proliferative capacity of human MSCs derived from bone marrow is linked to c-MYC expression (Melnik et al., 2019). Sufficient c-MYC expression may be essential to maintain high proliferative rates and an undifferentiated state of MSCs during *ex vivo* expansion.

In the present study, *MYC* as a hub gene was up-regulated in fetal BM-MSCs. We found a negative correlation between the expression of *MYC* and *TXNIP* (correlation coefficient = −0.77). Oberacker et al. (2018) showed that *TXNIP* plays a crucial role in aging processes. Age-related up-regulation of *TXNIP* results in reduced resistance to oxidative stress, and thereby accelerates aging. The present study reveals that fMSCs with high expression of *MYC* and low expression *TXNIP* may explain why fMSCs possess greater proliferative capacity and are more resistant to aging. The negative correlation between *MYC* and *TXNIP* in BM-MSCs warrants further research. Nevertheless, we must also acknowledge that high levels of *c-MYC* increase the risk of oncogenesis (Kozlowska et al., 2019).

The mitotic kinesin, *KIF20A*, is essential for central spindle organization at anaphase as well as cytokinesis regulation (Zhang et al., 2019). It is supposedly a key factor in cell proliferation and invasion in many cancers (Sheng et al., 2018; Zhang et al., 2019).



ID	Description
hsa04514	Cell adhesion molecules (CAMs)
hsa04610	Complement and coagulation cascades
hsa05150	Staphylococcus aureus infection
hsa04640	Hematopoietic cell lineage
hsa04512	ECM–receptor interaction
hsa05133	Pertussis
hsa05204	Chemical carcinogenesis
hsa04145	Phagosome
hsa00140	Steroid hormone biosynthesis
hsa03320	PPAR signaling pathway

FIGURE 5 | Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of robust DEGs. Chord plot **(A)** and circle plot **(B)** show the relationship between genes and KEGG pathways. **(A)** Chord plot depicting the relationship between genes and the top five KEGG pathways. **(B)** Circle plot depicting the distribution of genes in the top 10 KEGG pathways.

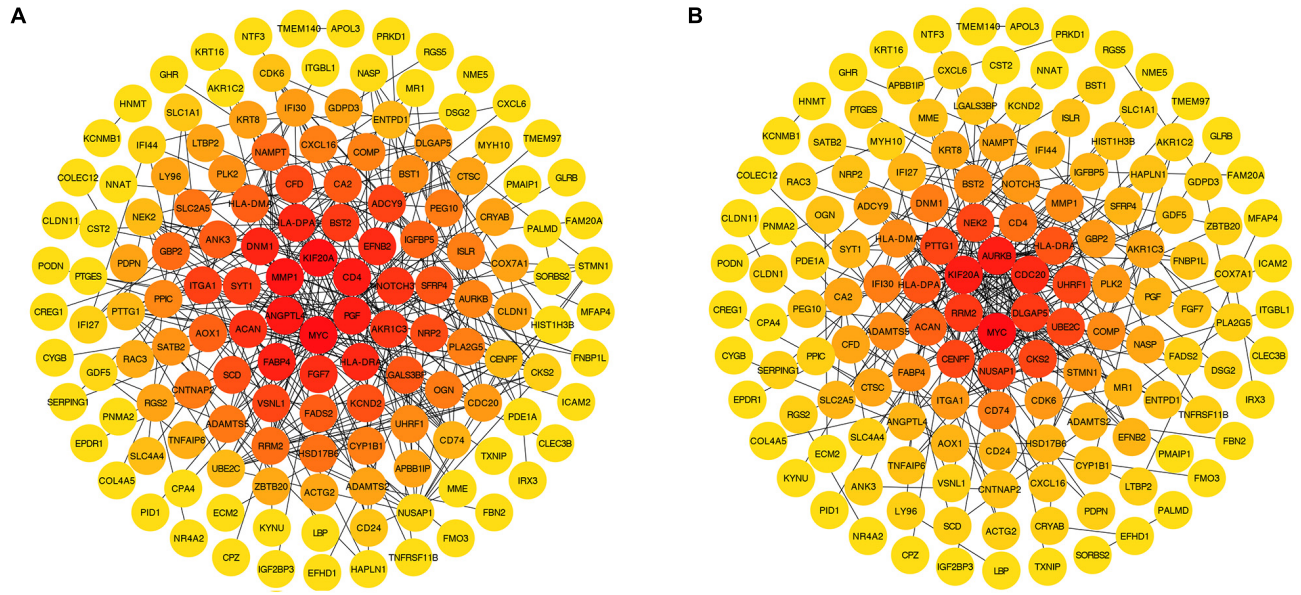


FIGURE 6 | DEGs with a *P*-value < 0.001 ranked by betweenness **(A)** and degree **(B)** methods using cytoHubba in Cytoscape. The change of color from orange red to yellow represents the change of centrality score from high to low.

A recent study showed that *KIF20A* could reportedly be regulated by *MYC* (Pan et al., 2020). There is currently no data regarding the function of *KIF20A* in adult stem cells, and the relationship between *MYC* and *KIF20A* is unclear. Our findings suggest that *KIF20A* is extensively linked with the cell cycle in BM-MSCs, and is moderately correlated (correlation coefficient = 0.41) with *MYC*.

Another major difference was observed in immunoregulatory function. Both *HLA-DRA* and *HLA-DPA1* belong to the major histocompatibility complex class II, are up-regulated in aMSCs, and mainly involved in antigen procession and presentation. The aMSCs express intermediate levels of HLA class **I** and low levels of HLA class **II**, while fMSCs express no HLA class **II** (Gotherstrom et al., 2003; O'Donoghue and Fisk, 2004). Previous reports have shown that adult MSCs contain intracellular deposits of class **II** alloantigen, and their surface expression can be induced under inflammatory conditions, such as in the presence of INFγ (Gotherstrom et al., 2003; Ryan et al., 2005). The BM-MSCs can therefore be recognized by allogeneic lymphocytes, possess immunomodulatory properties *in vitro*, and suppress the proliferation of activated lymphocytes (Ryan et al., 2005; Chen et al., 2011). However, Gallipeau and

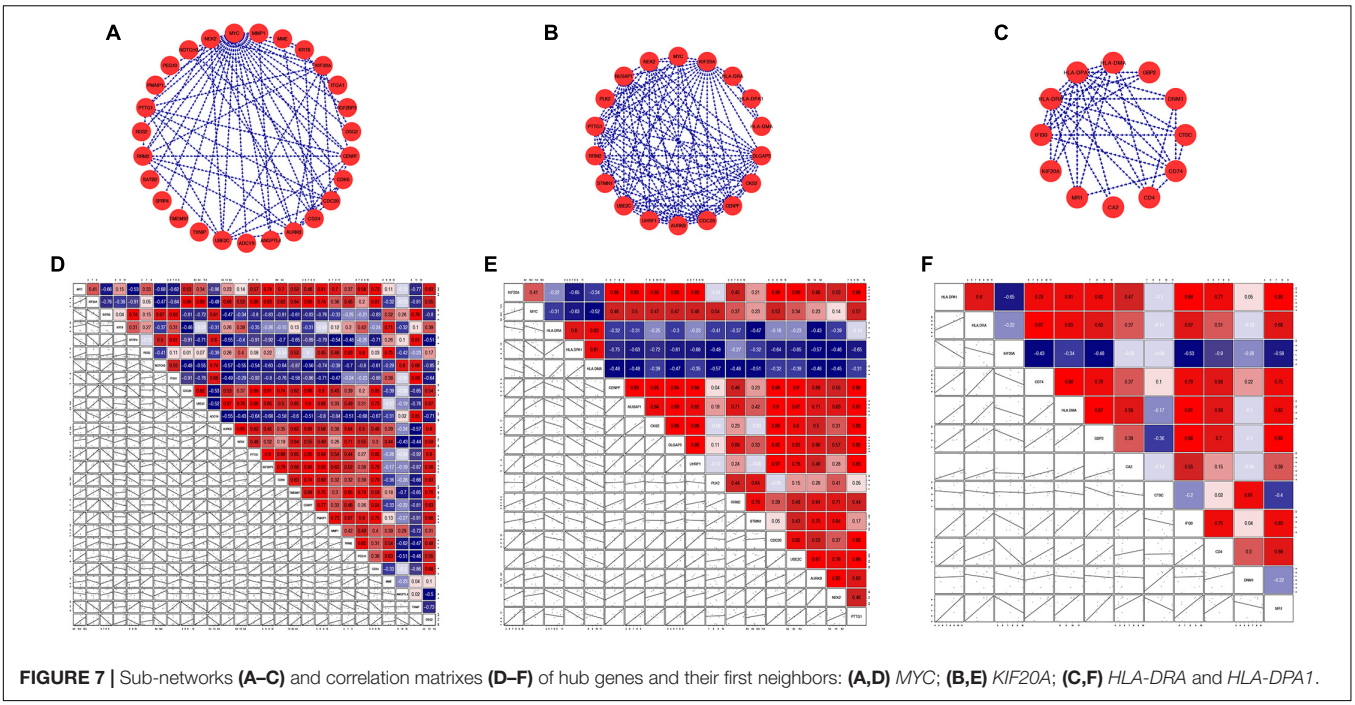


FIGURE 7 | Sub-networks **(A–C)** and correlation matrixes **(D–F)** of hub genes and their first neighbors: **(A,D)** *MYC*; **(B,E)** *KIF20A*; **(C,F)** *HLA-DRA* and *HLA-DPA1*.

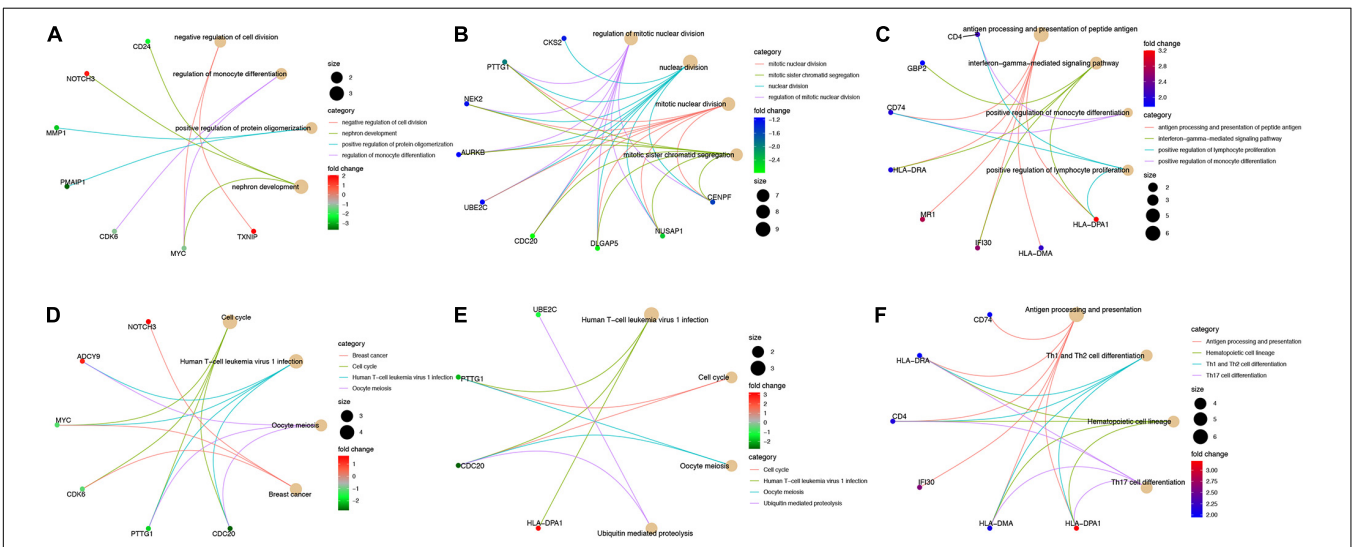


FIGURE 8 | Interactive genes of hub-genes were filtered by a Pearson correlation coefficient > 0.5 and GO **(A–C)** and KEGG pathway **(D–E)** analyses were performed for these genes. GO and KEGG analyses for **(A,D)** interactive genes of *MYC*; **(B,E)** interactive genes of *KIF20A*; **(C,F)** interactive genes of *HLA-DRA* and *HLA-DPA1*. Cut off = adjust $P < 0.05$, the top four are shown in plot.

colleagues proposed discrepancies in the immune-suppressive activities of MSCs arising from intrinsic variability of each donor source, with an average age > 65 years (Romieu-Mourez et al., 2012). Furthermore, the immunomodulation of MSCs can be regulated by inflammatory conditions; in low-level inflammatory microenvironments, BM-MSCs promote inflammation and act as antigen-presenting cells (Betancourt, 2013). The BM-MSCs from elderly donors in this study seemed to have a pro-inflammatory phenotype, which may be due to the chronic low-grade inflammatory conditions of aged donors.

As for cancerous condition, there is a growing body of evidence suggests that it plays a key role in the maintenance and progression of tumor. Fernando et al. highlight the importance of tumor microenvironment, especially for MSC, in multiple myeloma (MM) (Fernando et al., 2019). The telomeric length of MM – MSC is more lower and genes, such as CDC20, CDC6, involved in cell cycle are decrease in expression, which exhibit similar down-regulated in aMSCs. However, immune response related genes, for instance, HLA-DRA, are also down-regulated in MM – MSC, which is up-regulated in aMSCs.

CONCLUSION

Through data mining and network analysis, we detected four hub genes, *MYC*, *KIF20A*, *HLA-DRA*, and *HLA-DPA1*. Expression of the *MYC* gene was negatively correlated with that of *TXNIP*, a known senescence-associated gene. Furthermore, *KIF20A* is extensively linked with the cell cycle. The other two core genes, *HLA-DRA* and *HLA-DPA1*, are implicated in the immune response and may be induced by age-related inflammatory conditions. We infer that BM-MSCs derived from elderly donors may have age-related drawbacks. These cells show lower proliferative capacity and a pro-inflammatory phenotype. More experiments are required for further verification of these findings.

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

XYL conceived and designed the research, performed analysis, and wrote the majority of the manuscript. BZ, YL, HJ, and ZL supervised the research. MY, XPL, JD, KZ, XZ, LL, and JW contributed to data collation. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 81870736, 81801040, 81500816, and 81570951).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.573877/full#supplementary-material

Supplementary Figure 1 | The correlation profiles of **(A)** *MYC* and *TXNIP*; **(B)** *KIF20A* and *CDC20*; **(C)** *HLA-DPA1* and *MR1*; **(D)** *HLA-DRA* and *CD74*.

Supplementary Table 1 | The full list of up- and down-regulated genes integrated by RRA method.

Supplementary Table 2 | The details of all the robust DEGs.

Supplementary Table 3 | The list of DEGs ranked by log FC from large to small.

Supplementary Table 4 | Correlation coefficient and *P*-values for the correlation between genes in sub-network and hub genes.

Supplementary Table 5 | Coefficient of variation for all the genes included in the correlation analysis.

REFERENCES

Andrzejewska, A., Lukomska, B., and Janowski, M. (2019). Concise review: mesenchymal stem cells: from roots to boost. *Stem Cells* 37, 855–864. doi: 10.1002/stem.3016

Berebichez-Fridman, R., and Montero-Olvera, P. R. (2018). Sources and clinical applications of mesenchymal stem cells: state-of-the-art review. *Sultan Qaboos Univ. Med. J.* 18, e264–e277. doi: 10.18295/squmj.2018.18.03.002

Betancourt, A. M. (2013). New cell-based therapy paradigm: induction of bone marrow-derived multipotent mesenchymal stromal cells into pro-inflammatory MSC1 and Anti-inflammatory MSC2 Phenotypes. *Adv. Biochem. Eng./Biotechnol.* 130, 163–197. doi: 10.1007/10_2012_141

Campagnoli, C., Roberts, I. A., Kumar, S., Bennett, P. R., Bellantuono, I., and Fisk, N. M. (2001). Identification of mesenchymal stem/progenitor cells in human first-trimester fetal blood, liver, and bone marrow. *Blood* 98, 2396–2402. doi: 10.1182/blood.v98.8.2396

Castagnini, F., Pellegrini, C., Perazzo, L., Vannini, F., and Buda, R. (2016). Joint sparing treatments in early ankle osteoarthritis: current procedures and future perspectives. *J. Exp. Orthop.* 3:3. doi: 10.1186/s40634-016-0038-34

Chang, C. J., Yen, M. L., Chen, Y. C., Chien, C. C., Huang, H. I., Bai, C. H., et al. (2006). Placenta-derived multipotent cells exhibit immunosuppressive properties that are enhanced in the presence of interferon-gamma. *Stem Cells* 24, 2466–2477. doi: 10.1634/stemcells.2006-2071

Chen, B., Yu, J., Wang, Q., Zhao, Y., Sun, L., Xu, C., et al. (2018). Human bone marrow mesenchymal stem cells promote gastric cancer growth via regulating c-Myc. *Stem Cells Int.* 2018:9501747. doi: 10.1155/2018/9501747

Chen, P. M., Yen, M. L., Liu, K. J., Sytwu, H. K., and Yen, B. L. (2011). Immunomodulatory properties of human adult and fetal multipotent mesenchymal stem cells. *J. Biomed. Sci.* 18:49. doi: 10.1186/1423-0127-18-49

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol.* 8(Suppl. 4):S11. doi: 10.1186/1752-0509-8-S4-S11

Danilova, N. (2006). The evolution of immune mechanisms. *J. Exp. Zool. B Mol. Dev. Evol.* 306, 496–520. doi: 10.1002/jez.b.21102

Dexheimer, V., Mueller, S., Braatz, F., and Richter, W. (2011). Reduced reactivation from dormancy but maintained lineage choice of human mesenchymal stem cells with donor age. *PLoS One* 6:e22980. doi: 10.1371/journal.pone.0022980

Ding, D. C., Shyu, W. C., and Lin, S. Z. (2011). Mesenchymal stem cells. *Cell Transplant* 20, 5–14. doi: 10.3727/096368910X

Dogan, A., Demirci, S., Bayir, Y., Halici, Z., Karakus, E., Aydin, A., et al. (2014). Boron containing poly-(lactide-co-glycolide) (PLGA) scaffolds for bone tissue engineering. *Mater. Sci. Eng. C Mater. Biol. Appl.* 44, 246–253. doi: 10.1016/j.msec.2014.08.035

Fernando, R. C., Mazzotti, D. R., Azevedo, H., Sandes, A. F., Rizzatti, E. G., de Oliveira, M. B., et al. (2019). Transcriptome analysis of mesenchymal stem cells from multiple myeloma patients reveals downregulation of genes involved in cell cycle progression, immune response, and bone metabolism. *Sci. Rep.* 9:1056. doi: 10.1038/s41598-018-38314-8

Friedenstein, A. J., Chailakhyan, R. K., and Gerasimov, U. V. (1987). Bone marrow osteogenic stem cells: in vitro cultivation and transplantation in diffusion chambers. *Cell Tissue Kinet.* 20, 263–272. doi: 10.1111/j.1365-2184.1987.tb01309.x

Gotherstrom, C., Ringden, O., Westgren, M., Tammik, C., and Le Blanc, K. (2003). Immunomodulatory effects of human foetal liver-derived mesenchymal stem cells. *Bone Marrow Transplant* 32, 265–272. doi: 10.1038/sj.bmt.1704111

Götherström, C., West, A., Liden, J., Uzunel, M., Lahesmaa, R., and Le, Blanc K (2005). Difference in gene expression between human fetal liver and adult bone marrow mesenchymal stem cells. *Haematologica* 90, 1017–1026.

Guillot, P. V., Gotherstrom, C., Chan, J., Kurata, H., and Fisk, N. M. (2007). Human first-trimester fetal MSC express pluripotency markers and grow faster and have longer telomeres than adult MSC. *Stem Cells* 25, 646–654. doi: 10.1634/stemcells.2006-2208

Hayflick, L. (2000). The illusion of cell immortality. *Br. J. Cancer* 83, 841–846. doi: 10.1054/bjoc.2000.1296

Ito, K., and Murphy, D. (2013). Application of ggplot2 to pharmacometric graphics. *CPT Pharmacometrics Syst Pharmacol.* 2:e79. doi: 10.1038/psp.2013.56

Jin, J. (2017). Stem cell treatments. *JAMA* 317:330. doi: 10.1001/jama.2016

Kolf, C. M., Cho, E., and Tuan, R. S. (2007). Mesenchymal stromal cells. Biology of adult mesenchymal stem cells: regulation of niche, self-renewal and differentiation. *Arthritis. Res. Ther.* 9:204. doi: 10.1186/ar2116

Kozlowska, U., Krawczenko, A., Futoma, K., Jurek, T., Rorat, M., Patrzalek, D., et al. (2019). Similarities and differences between mesenchymal stem/progenitor cells derived from various human tissues. *World J. Stem Cells* 11, 347–374. doi: 10.4252/wjsc.v11.i6.347

Kumamoto, M., Nishiwaki, T., Matsuo, N., Kimura, H., and Matsushima, K. (2009). Minimally cultured bone marrow mesenchymal stem cells ameliorate fibrotic lung injury. *Eur. Respir. J.* 34, 740–748. doi: 10.1183/09031936.00128508

Li, W. (2012). Volcano plots in analyzing differential expressions with mRNA microarrays. *J. Bioinform. Comput. Biol.* 10:1231003. doi: 10.1142/S0219720012310038

Liu, X., Wu, J., Zhang, D., Bing, Z., Tian, J., Ni, M., et al. (2018). Identification of potential key genes associated with the pathogenesis and prognosis of gastric cancer based on integrated bioinformatics analysis. *Front. Genet.* 9:265. doi: 10.3389/fgene.2018.00265

Lüscher, B. (2001). Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene* 277, 1–14. doi: 10.1016/s0378-1119(01)00697-697

Lv, F. J., Tuan, R. S., Cheung, K. M., and Leung, V. Y. (2014). Concise review: the surface markers and identity of human mesenchymal stem cells. *Stem Cells* 32, 1408–1419. doi: 10.1002/stem.1681

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551

Mehrabani, D., Mojtahed, J. F., Zakerinia, M., Hadianfard, M. J. J., Tanideh, R. N., et al. (2016). The healing effect of bone marrow-derived stem cells in knee osteoarthritis: a case report. *World J. plastic Surg.* 5, 168–174.

Melnik, S., Werth, N., Boeuf, S., Hahn, E. M., Gotterbarm, T., Anton, M., et al. (2019). Impact of c-MYC expression on proliferation, differentiation, and risk of neoplastic transformation of human mesenchymal stromal cells. *Stem Cell Res. Ther.* 10:73. doi: 10.1186/s13287-019-1187-z

Michalik, L., and Wahli, W. (2008). PPARs mediate lipid signaling in inflammation and cancer. *PPAR Res.* 2008:134059. doi: 10.1155/2008/134059

Oberacker, T., Bajorat, J., Ziola, S., Schroeder, A., Roth, D., Kastl, L., et al. (2018). Enhanced expression of thioredoxin-interacting-protein regulates oxidative DNA damage and aging. *FEBS Lett.* 592, 2297–2307. doi: 10.1002/1873-3468.13156

O'Donoghue, K., and Fisk, N. M. (2004). Fetal stem cells. *Best Pract. Res. Clin. Obstet Gynaecol.* 18, 853–875. doi: 10.1016/j.bpobgyn.2004.06.010

Paciejewska, M. M., Maijenburg, M. W., Gilissen, C., Kleijer, M., Vermeul, K., Weijer, K., et al. (2016). Different balance of wnt signaling in adult and fetal bone marrow-derived mesenchymal stromal cells. *Stem Cells Dev.* 25, 934–947. doi: 10.1089/scd.2015.0263

Pan, X., Liu, W., Chai, Y., Hu, L., Wang, J., and Zhang, Y. (2020). Identification of hub genes in atypical teratoid/rhabdoid tumor by bioinformatics analyses. *J. Mol. Neurosci.* 70, 1906–1913. doi: 10.1007/s12031-020-01587-1588

Park, S. B., Seo, K. W., So, A. Y., Seo, M. S., Yu, K. R., Kang, S. K., et al. (2012). SOX2 has a crucial role in the lineage determination and proliferation of mesenchymal stem cells through Dickkopf-1 and c-MYC. *Cell Death Differ.* 19, 534–545. doi: 10.1038/cdd.2011.137

Reimand, J., Vaquerizas, J. M., Todd, A. E., Vilo, J., and Luscombe, N. M. (2010). Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Res.* 38, 4768–4777. doi: 10.1093/nar/gkq232

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Romieu-Mourez, R., Coutu, D. L., and Galipeau, J. (2012). The immune plasticity of mesenchymal stromal cells from mice and men: concordances and discrepancies. *Front. Biosci. (Elite edition)* 4, 824–837. doi: 10.2741/e422

Ryan, J. M., Barry, F. P., Murphy, J. M., and Mahon, B. P. (2005). Mesenchymal stem cells avoid allogeneic rejection. *J. Inflamm. (Lond)* 2:8. doi: 10.1186/1476-9255-2-8

Salminen, A., Huuskonen, J., Ojala, J., Kauppinen, A., Kaarniranta, K., and Suuronen, T. (2008). Activation of innate immunity system during aging: NF-kB signaling is the molecular culprit of inflamm-aging. *Ageing Res. Rev.* 7, 83–105. doi: 10.1016/j.arr.2007.09.002

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesth Analg.* 126, 1763–1768. doi: 10.1213/ANE.0000000000002864

Sheng, Y., Wang, W., Hong, B., Jiang, X., Sun, R., Yan, Q., et al. (2018). Upregulation of KIF20A correlates with poor prognosis in gastric cancer. *Cancer Manag. Res.* 10, 6205–6216. doi: 10.2147/CMAR.S176147

Spitzhorn, L. S., Megges, M., Wruck, W., Rahman, M. S., Otte, J., Degistirici, O., et al. (2019). Human iPSC-derived MSCs (iMSCs) from aged individuals acquire a rejuvenation signature. *Stem Cell Res. Ther.* 10:100. doi: 10.1186/s13287-019-1209-x

Suvakov, S., Cubro, H., White, W. M., Butler Tobah, Y. S., Weissgerber, T. L., Jordan, K. L., et al. (2019). Targeting senescence improves angiogenic potential of adipose-derived mesenchymal stem cells in patients with preeclampsia. *Biol. Sex Differ.* 10:49. doi: 10.1186/s13293-019-0263-265

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Taleahmad, S., Mirzaei, M., Parker, L. M., Hassani, S. N., Mollamohammadi, S., Sharifi-Zarchi, A., et al. (2015). Proteome analysis of ground state pluripotency. *Sci. Rep.* 5:17985. doi: 10.1038/srep17985

Turinetto, V., Vitale, E., and Giachino, C. (2016). Senescence in human mesenchymal stem cells: functional changes and implications in stem cell-based therapy. *Int. J. Mol. Sci.* 17:1164. doi: 10.3390/ijms17071164

Verfaillie, C. M., Pera, M. F., and Lansdorp, P. M. (2002). Stem cells: hype and reality. *Hematol. Am. Soc. Hematol. Educ. Prog.* 2002, 369–391. doi: 10.1182/asheducation-2002.1.369

Walter, W., Sanchez-Cabo, F., and Ricote, M. (2015). GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* 31, 2912–2914. doi: 10.1093/bioinformatics/btv300

Wei, B., Yao, Q., Guo, Y., Mao, F., Liu, S., Xu, Y., et al. (2015). Three-dimensional polycaprolactone-hydroxyapatite scaffolds combined with bone marrow cells for cartilage tissue engineering. *J. Biomater. Appl.* 30, 160–170. doi: 10.1177/0885328215575762

Xu, D., Popov, N., Hou, M., Wang, Q., Björkholm, M., Gruber, A., et al. (2001). Switch from Myc/Max to Mad1/Max binding and decrease in histone acetylation at the telomerase reverse transcriptase promoter during differentiation of HL60 cells. *Proc. Natl. Acad. Sci. U S A.* 98, 3826–3831. doi: 10.1073/071043198

Xu, J., Wang, B., Sun, Y., Wu, T., Liu, Y., Zhang, J., et al. (2016). Human fetal mesenchymal stem cell secretome enhances bone consolidation in distraction osteogenesis. *Stem Cell Res. Ther.* 7:134. doi: 10.1186/s13287-016-0392-392

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

Zhang, Z., Chai, C., Shen, T., Li, X., Ji, J., Li, C., et al. (2019). Aberrant KIF20A expression is associated with adverse clinical outcome and promotes tumor progression in prostate cancer. *Dis. Markers* 2019:4782730. doi: 10.1155/2019/4782730

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 January 2021
doi: 10.3389/fgene.2020.596826

Identification of Key MicroRNAs and Mechanisms in Prostate Cancer Evolution Based on Biomarker Prioritization Model and Carcinogenic Survey

Yuxin Lin[1†], *Zhijun Miao*[2†], *Xuefeng Zhang*[1], *Xuedong Wei*[1], *Jianquan Hou*[1], *Yuhua Huang*[1*] and *Bairong Shen*[3*]

[1] Department of Urology, The First Affiliated Hospital of Soochow University, Suzhou, China, [2] Department of Urology, Suzhou Dushuhu Public Hospital, Suzhou, China, [3] Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, China

OPEN ACCESS

Edited by:
Xiaogang Wu,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Leda Torres,
National Institute of Pediatrics, Mexico
Haiyun Wang,
Tongji University, China

***Correspondence:**
Bairong Shen
bairong.shen@scu.edu.cn
Yuhua Huang
sdfyyhyh@163.com

[†] These authors have contributed equally to this work

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

Received: 20 August 2020
Accepted: 21 December 2020
Published: 15 January 2021

Citation:
Lin Y, Miao Z, Zhang X, Wei X, Hou J, Huang Y and Shen B (2021) Identification of Key MicroRNAs and Mechanisms in Prostate Cancer Evolution Based on Biomarker Prioritization Model and Carcinogenic Survey. Front. Genet. 11:596826. doi: 10.3389/fgene.2020.596826

Background: Prostate cancer (PCa) is occurred with increasing incidence and heterogeneous pathogenesis. Although clinical strategies are accumulated for PCa prevention, there is still a lack of sensitive biomarkers for the holistic management in PCa occurrence and progression. Based on systems biology and artificial intelligence, translational informatics provides new perspectives for PCa biomarker prioritization and carcinogenic survey.

Methods: In this study, gene expression and miRNA-mRNA association data were integrated to construct conditional networks specific to PCa occurrence and progression, respectively. Based on network modeling, hub miRNAs with significantly strong single-line regulatory power were topologically identified and those shared by the condition-specific network systems were chosen as candidate biomarkers for computational validation and functional enrichment analysis.

Results: Nine miRNAs, i.e., *hsa-miR-1-3p*, *hsa-miR-125b-5p*, *hsa-miR-145-5p*, *hsa-miR-182-5p*, *hsa-miR-198*, *hsa-miR-22-3p*, *hsa-miR-24-3p*, *hsa-miR-34a-5p*, and *hsa-miR-499a-5p*, were prioritized as key players for PCa management. Most of these miRNAs achieved high AUC values (AUC $>$ 0.70) in differentiating different prostate samples. Among them, seven of the miRNAs have been previously reported as PCa biomarkers, which indicated the performance of the proposed model. The remaining *hsa-miR-22-3p* and *hsa-miR-499a-5p* could serve as novel candidates for PCa predicting and monitoring. In particular, key miRNA-mRNA regulations were extracted for pathogenetic understanding. Here *hsa-miR-145-5p* was selected as the case and *hsa-miR-145-5p/NDRG2/AR* and *hsa-miR-145-5p/KLF5/AR* axis were found to be putative mechanisms during PCa evolution. In addition, *Wnt* signaling, prostate cancer, microRNAs in cancer etc. were significantly enriched by the identified miRNAs-mRNAs, demonstrating the functional role of the identified miRNAs in PCa genesis.

Conclusion: Biomarker miRNAs together with the associated miRNA-mRNA relations were computationally identified and analyzed for PCa management and carcinogenic deciphering. Further experimental and clinical validations using low-throughput techniques and human samples are expected for future translational studies.

Keywords: miRNA biomarker, prostate cancer management, miRNA-mRNA network modeling, miRNA regulatory pattern, systems biology

INTRODUCTION

Prostate cancer (PCa) is a kind of malignant tumors which ranks first in the incidence of male leading cancer types according to the reports from Cancer Statistics 2020 (Siegel et al., 2020). It has been acknowledged that the occurrence and progression of PCa are highly heterogeneous, resulting in the difficulty in PCa precision medicine and personalized healthcare. In clinical practice, although the level of serum prostate-specific antigen (*PSA*) and multi parameter Magnetic Resonance Imaging (MRI) techniques are widely tested for PCa screening, the sensitivity and specificity are still need to be measured for improving positive detection rate and avoiding unnecessary biopsy.

As a class of post-transcriptional regulators, microRNAs (miRNAs) are found to be active in carcinogenesis, including PCa (Khanmi et al., 2015). Extensive efforts showed that miRNAs could regulate down-stream messenger RNAs (mRNAs) though complementary base pairing and eventually affect the signal transmission of pathways and the function of cellular activities (Esteller, 2011). Currently, the identification and prioritization of miRNAs as biomarkers for PCa theranostics is of clinical interest, which would help the early diagnosis, prognosis tracking and targeted therapy of PCa patients (Bhagirath et al., 2018; Wei et al., 2020).

In the era of artificial intelligence and biomedical informatics, data-driven translational PCa research brings a new frontier for systems modeling of complex genetic interactions (Lin et al., 2020). The structural characteristics within biological networks offer great opportunities for understanding cancer heterogeneity at systems biology level (Liu Y. Y. et al., 2011). Accumulating evidences have demonstrated the functional importance of hub nodes in gene network for prioritizing key players during PCa development. For example, Zhu et al. identified five miRNAs and seven genes for predicting the biochemical recurrence-free survival of PCa by evaluating the significance of differential expression and the location of genes in the network (Zhu et al., 2020). Analogously, Tu et al. proposed an integrated framework that considers both dynamical changes of gene expression and static features in protein-protein network for extracting key miRNA-mRNA pairs associated with docetaxel resistance in PCa (Tu et al., 2019).

It is reasonable that hub genes are located in the center of the network by directly connect more partner genes to control the information flow. In addition to such regulatory pattern, special structures hidden in the network are still worth being explored for investigating the strength of genes in network stability. Biomarkers hold the power to indicate the dynamical alternations in biological systems, searching for vulnerable hallmarks from miRNA-gene regulations would therefore provide crucial clues for cancer biomarker discovery (Lin et al., 2019). In our previous studies, we found that a certain number of genes were independently regulated by single miRNAs and a new parameter, i.e., NSR (number of single-line regulation) was defined based on network vulnerability theory to quantify the single-line regulation of miRNAs in miRNA-mRNA network. According to statistical evidences, miRNAs with higher NSR values were structurally important to serve as candidate biomarkers for cancer management (Lin et al., 2018b), and five miRNAs were computationally screened and validated for PCa metastasis (Lin et al., 2018a).

On the basis of our previous findings, in this study we expand our research interest and update the bioinformatics framework to identify key miRNAs functionally important in the whole process during PCa evolution as both of the diagnosis and metastasis monitoring are hot topics for precision medicine. In methodology, two PCa condition-specific miRNA-mRNA networks, i.e., PCa occurrence and progression-specific network, are respectively constructed and characterized based on the integration of novel gene expression and network topological signatures. Meanwhile potential miRNA-mRNA pairs in PCa evolution are deciphered for functional survey and multi-level carcinogenesis understanding. In particular, the traditional hub property is improved by combing and measuring the single-line regulatory power of miRNAs in the computational simulation process, which would enhance the overall predictive performance and biological significance of the bioinformatics model. The schematic pipeline is shown in **Figure 1**.

MATERIALS AND METHODS

Dataset Collection and Processing

The miRNA and mRNA datasets were both collected from gene expression omnibus (GEO) (Edgar et al., 2002), where the super-series GSE21032 provides the integrative genomic profiling of human PCa (Taylor et al., 2010), including the clinically localized primary PCa (pPCa), metastatic PCa (mPCa) and the normal adjacent benign prostate samples. Here the normalized datasets of sub-series GSE21036 and GSE21034 were downloaded for further analysis. As illustrated in **Table 1**, GSE21036 contains a total of 99 pPCa, 14 mPCa and 28 normal miRNA samples screened by Agilent-019118 Human microRNA Microarray 2.0 G4470B platform, whereas GSE21034 consists of the whole-transcript expression data for 131 pPCa, 19 mPCa, and 29 normal control prostate samples profiled on

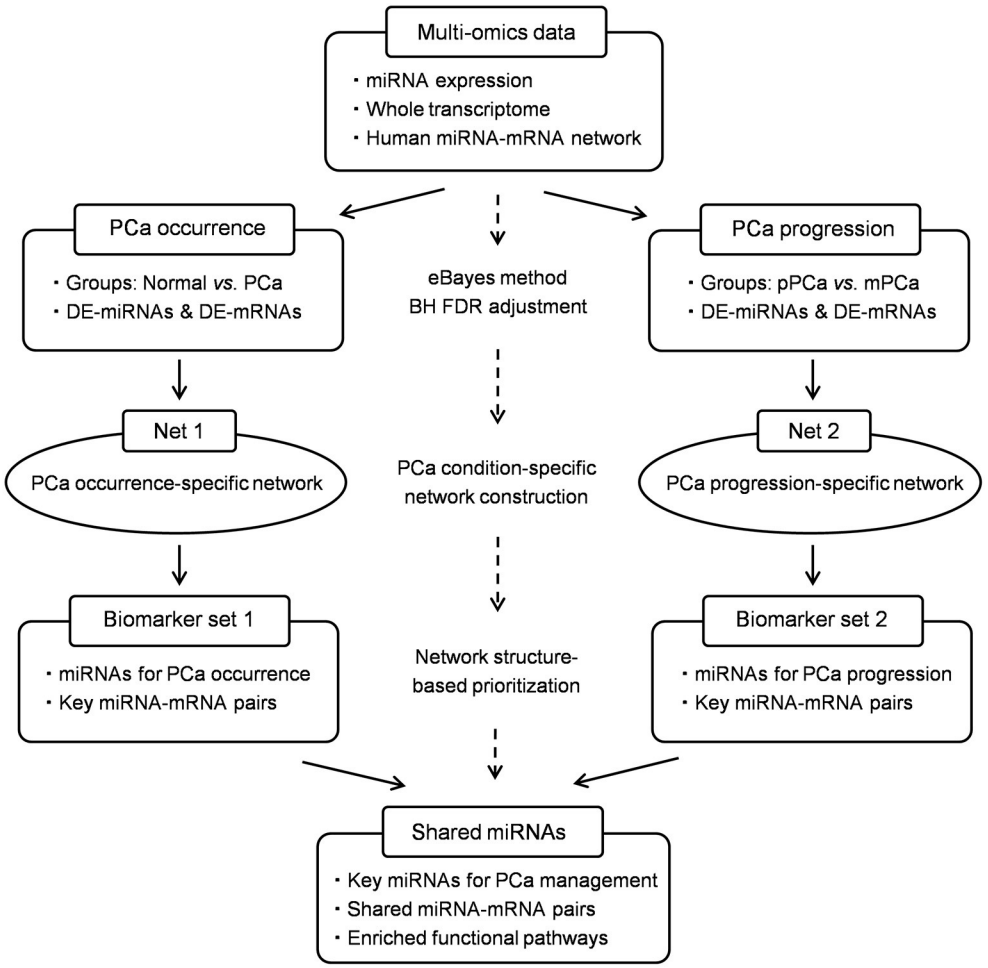


FIGURE 1 | The schematic pipeline of this study. PCa, prostate cancer; pPCa, clinically localized primary prostate cancer; mPCa, metastatic prostate cancer; DE, differentially expressed; eBayes, empirical bayes; BH FDR, Benjamini-Hochberg false discovery rate.

Affymetrix Human Exon 1.0 ST Array. In addition, GSE54516 with miRNA expression data measured in prostate benign and tumor tissues using miRNA Taqman plates was chosen as an independent dataset for result validation (Gu et al., 2015).

To ensure the specificity of the RNAs in PCa genesis, differential expression analysis was performed and compared among different sample groups, i.e., normal vs. PCa and pPCa vs. mPCa. Based on the evaluation of statistics approaches for generating differentially expressed (DE) genes from Microarray data (Jeffery et al., 2006), the empirical bayes (eBayes) method was chosen for raw *p*-value calculation (Smyth, 2004) and the Benjamini-Hochberg false discovery rate (FDR) was then applied to adjust raw *p*-values. For the gene associated with multiple probes, the probe with the most significant variation was selected and assigned. The criterion for DE-miRNA and DE-mRNA identification was defined as the adjusted *p*-value (adj. *p*-value) < 0.05.

Model Development Based on Network Construction and Characterization

The bioinformatics model was developed based on the characterization of miRNA regulation in PCa condition-specific miRNA-mRNA network. As shown in **Figure 1**, a human global miRNA-mRNA network was first constructed as the reference by integrating both experimentally validated and computationally predicted miRNA-mRNA pairs from public databases and software tools (Lin et al., 2018b). Then DE-miRNAs and DE-mRNAs were mapped onto the given network to extract PCa condition-specific networks. In this study a total of two networks, i.e., PCa occurrence-specific and progression-specific networks, were measured, respectively, where the former described the role of miRNAs in PCa occurrence process and the latter simulated miRNA regulation during PCa progression and metastasis.

To quantity the regulatory pattern of miRNAs in the network, the feature parameters NTG (number of targeted genes) and NSR were defined and used for biomarker prioritization. Among

TABLE 1 | Datasets used in this study.

RNA type	GEO accession	Platform	Sample source	Normal sample	PCa sample	
					pPCa	mPCa
miRNA	GSE21036	GPL8227	Human prostate	28	99	14
mRNA	GSE21034	GPL10264	Human prostate	29	131	19
miRNA	GSE54516	GPL18234	Human prostate	48	51	

PCa, prostate cancer; pPCa, clinically localized primary prostate cancer; mPCa, metastatic prostate cancer; GEO, Gene Expression Omnibus.

TABLE 2 | Statistics and topological features of the identified miRNAs.

miRNA	Expression	adj. *p*-value		NTG		NSR		NSR/NTG	
		I	II	I	II	I	II	I	II
hsa-miR-1-3p	Down	4.10e-3	1.10e-30	69	75	15	17	0.2174	0.2267
hsa-miR-125b-5p	Down	3.03e-3	8.98e-16	45	48	12	8	0.2667	0.1667
hsa-miR-145-5p	Down	1.96e-7	8.00e-25	56	57	11	17	0.1964	0.2982
hsa-miR-182-5p	Up	1.25e-9	2.46e-2	47	44	10	7	0.2128	0.1591
hsa-miR-198	Up	9.19e-3	7.67e-5	50	47	14	9	0.28	0.1915
hsa-miR-22-3p	Down	4.11e-3	2.45e-2	83	82	19	14	0.2289	0.1707
hsa-miR-24-3p	Down	5.82e-6	1.12e-9	39	43	7	10	0.1795	0.2326
hsa-miR-34a-5p	Down	1.20e-2	2.24e-2	58	59	9	14	0.1552	0.2373
hsa-miR-499a-5p	Down	4.22e-4	3.13e-2	73	76	10	10	0.1370	0.1316

adj.p-value, adjusted p-value; NTG, number of targeted genes; NSR, number of single-line regulation; I, Normal vs. PCa; II, pPCa vs. mPCa.

them, NTG represents the number of genes targeted by certain miRNAs. According to the theory of network sciences, hub nodes with more links in the network are functionally important in biological systems. Meanwhile our previous findings have demonstrated that biomarker miRNAs held strong single-line regulatory power in the network since the single-line points are vulnerable and the dysregulation in such sites are likely to cause the disorder at the systems level (Lin et al., 2018b, 2019). Thus, NSR parameter is set to indicate the number of genes independently regulated by a given miRNA. In addition, the ratio NSR/NTG was calculated to further evaluate the significance of miRNAs on gene regulation.

Based on above network systems and feature parameters, miRNAs with significantly high NTG, NSR, and NSR/NTG values (p-value < 0.05, Wilcoxon signed-rank test) were prioritized in each PCa-specific network and those shared by the two networks were selected as key players for predicting the occurrence and progression of PCa. Moreover, the shared miRNA-mRNA pairs were also identified for functional and carcinogenic survey.

Performance Evaluation and Comparison

The receiver-operating characteristic (ROC) and clustering analysis were performed based on the expression data of the identified miRNAs using ROCR and Pheatmap package in R program, respectively. Here the area under ROC curve (AUC) was calculated for each miRNA to evaluate and compare the biomarker potential on differentiating prostate samples, i.e., normal vs. PCa and pPCa vs. mPCa. Moreover, an additional index called prediction precision was defined as the percentage of literature-reported PCa miRNA biomarkers in the whole predicted set to validate and compare the performance of the proposed model.

Functional Exploration and Carcinogenic Analysis

The functional carcinogenesis of the identified miRNAs in PCa evolution was investigated based on the research paradigm of miRNA-gene-pathway axis. First the targets of miRNAs were retrieved from each PCa condition-specific network and miRNAs-mRNAs shared by the two networks were then collected as key regulations for gene ontology (GO) annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using the online tool Database for Annotation, Visualization and Integrated Discovery (DAVID, version 6.8) (Kanehisa and Goto, 2000; Huang da et al., 2009). The top ten significant terms with p-value < 0.05 were chosen for pathogenic understanding of their associations with cellular proliferation, invasion, metastasis and the responses to PCa treatment through literature exploration.

RESULTS

Biomarker miRNAs Identified for PCa Management

In this study, two PCa condition-specific networks, i.e., occurrence-specific and progression-specific network, were respectively, extracted based on human miRNA-mRNA reference

network and the selected sample datasets. Among them, the occurrence-specific network comprised 6,063 regulatory pairs associated with 138 DE-miRNAs and 2,035 DE-mRNAs between normal and PCa samples. In the progression-specific network, a total of 7,510 regulations among 169 DE-miRNAs and 2,238 DE-mRNAs with the expression change in PCa progression and metastasis were statistically identified.

After network structure-based filtration, 17 and 19 miRNAs with significantly high NTG, NSR and NSR/NTG values were computationally screened in PCa occurrence-specific and progression-specific network, respectively (see **Supplementary Table 1**), and the shared nine miRNAs, i.e., *hsa-miR-1-3p*, *hsa-miR-125b-5p*, *hsa-miR-145-5p*, *hsa-miR-182-5p*, *hsa-miR-198*, *hsa-miR-22-3p*, *hsa-miR-24-3p*, *hsa-miR-34a-5p*, and *hsa-miR-499a-5p*, were finally collected as key players during PCa evolution. As illustrated in **Table 2**, *hsa-miR-182-5p* and *hsa-miR-198* were over-expressed in the initiation and metastasis processes of PCa, whereas the remaining seven miRNAs were down-regulated during PCa development.

As shown in **Figure 2A**, the ROC analysis strengthened the power of the identified miRNAs for classifying different prostate samples, i.e., normal vs. PCa and pPCa vs. mPCa. For example, the average AUC between the groups of normal and PCa was 0.7862 (ranged from 0.6802 to 0.9447), and it reached 0.8057 (ranged from 0.6291 to 0.9986) for discriminating pPCa and mPCa samples. In the validation set, the average AUC was 0.8010 (ranged from 0.5960 to 0.9608), which was comparable with that in the prediction set. In particular, *hsa-miR-145-5p* achieved the overall best performance on PCa predicting and subtyping (both AUC $>$ 0.9), demonstrating its prospects for carcinogenic study and future clinical translation. However, as shown in **Figure 2B** the three groups were not well distinguished by combing the expression signature of these miRNAs, which indicated that the identified miRNAs would not be suitable to serve as potential biomarker combinations.

Literature-Based Functional Annotation and Validation

According to the review of citations in PubMed, seven of the identified miRNAs (77.8%, 7/9), i.e., *hsa-miR-1-3p*, *hsa-miR-125b-5p*, *hsa-miR-145-5p*, *hsa-miR-182-5p*, *hsa-miR-198*, *hsa-miR-24-3p*, and *hsa-miR-34a-5p* have been previously reported as biomarkers or molecular tools for PCa prediction. For example, Xie et al., investigated the diagnostic value and carcinogenic mechanisms of *hsa-miR-1* (namely *hsa-miR-1-3p*) in PCa. The result of meta-analyses and bioinformatics studies showed that this miRNA was significantly down-expressed in PCa samples and it could regulate pathways associated with androgen receptor (*AR*) activities in PCa development (Xie et al., 2018). Hudson et al. performed *in vitro* analysis and proved the tumor suppressor function of *hsa-miR-1* in PCa cell proliferation and motility. Besides the down-regulation in pPCa samples, the expression of this miRNAs was found to be reduced in distant metastasis, which indicated its power for prediction PCa progression and recurrence (Hudson et al., 2012). Zhu et al. identified *hsa-miR-125b* (namely *hsa-miR-125b-5p*) as an independent factor indicating castration resistant in PCa (Zhu et al., 2015), and this miRNA could improve the prediction of PCa status on the basis of serum *PSA* screening (Roberts et al., 2015). Xu et al. evaluated the level of *hsa-miR-145* expression (namely *hsa-miR-145-5p*) in urinary extracellular vesicles between PCa and healthy or benign prostate hyperplasia subjects. They found that the expression of this miRNAs was significantly altered in the urine of PCa patients, which highlighted its potential for PCa non-invasive diagnosis (Xu et al., 2017). Moreover, this miRNA was both detected in our previous studies using different network systems and computational models, demonstrating its significance on tumor regulation in PCa invasion, metastasis, and castration resistance (Zhu et al., 2015; Lin et al., 2018a). Based on RT-qPCR testing and validation, Bidarra et al. reported that the level of *hsa-miR-182-5p* was related to the advanced stage of PCa pathogenesis, and it was over-expressed in the plasma samples of patients with metastasis (Bidarra et al., 2019). Similarly, the expression of *hsa-miR-198* was found to be increased especially in the cohorts of high-grade (Gleason score $\geq$ 8) PCa (Walter et al., 2013), which was consistent with the result in this study. In addition, *hsa-miR-24-3p* and *hsa-miR-34a-5p* were also functional regulators in progression and therapeutic intervention by targeting PCa-related genes. Lynch et al. used the PCR to investigate the role of *hsa-miR-24* (namely *hsa-miR-24-3p*) in PCa cell lines. Compared with the normal prostate epithelial cell line, *hsa-miR-24* was down-regulated in PCa. This pattern was closely correlated with higher level of serum *PSA* and other clinical indices for PCa monitoring. Moreover, *p27* (*CDKN1B*) and *p16* (*CDK2NA*) were confirmed as targets of this miRNA in PCa cells (Lynch et al., 2016). Liu et al. announced that *hsa-miR-34a* (namely *hsa-miR-34a-5p*) was a tumor suppressor gene and it could inhibit the stem cells and metastasis by directly targeting *CD44* (Liu C. et al., 2011). Meanwhile this miRNA was a predictive biomarker for docetaxel responses associated with PCa therapy (Corcoran et al., 2014).

Although the remaining miRNAs, i.e., *hsa-miR-22-3p* and *hsa-miR-499a-5p*, have not been reported as PCa biomarkers yet, they were also powerful for PCa classifying and subtyping based on ROC analysis of this study. Hence these two miRNAs could serve as novel candidates for PCa diagnosis and prognosis. In summary, in this study the proposed bioinformatics model outperformed our previous method by increasing the prediction precision from 40 to 77.8% (Lin et al., 2018a), and more experimental validations using wet-lab approaches are needed in the future work.

Key miRNA-mRNA Regulations in PCa Carcinogenesis: A Focused Study on *hsa-miR-145-5p*

A total of 194 miRNA-mRNA regulations among the nine miRNA biomarkers and 172 dysfunctional mRNAs were extracted from PCa occurrence and progression networks to investigate their carcinogenetic role in PCa evolution at the gene level. As shown in **Figure 3A**, approximately half of the mRNAs, single-line or co-regulated by the identified miRNA candidates, were involved in PCa genesis according to literature reports. In

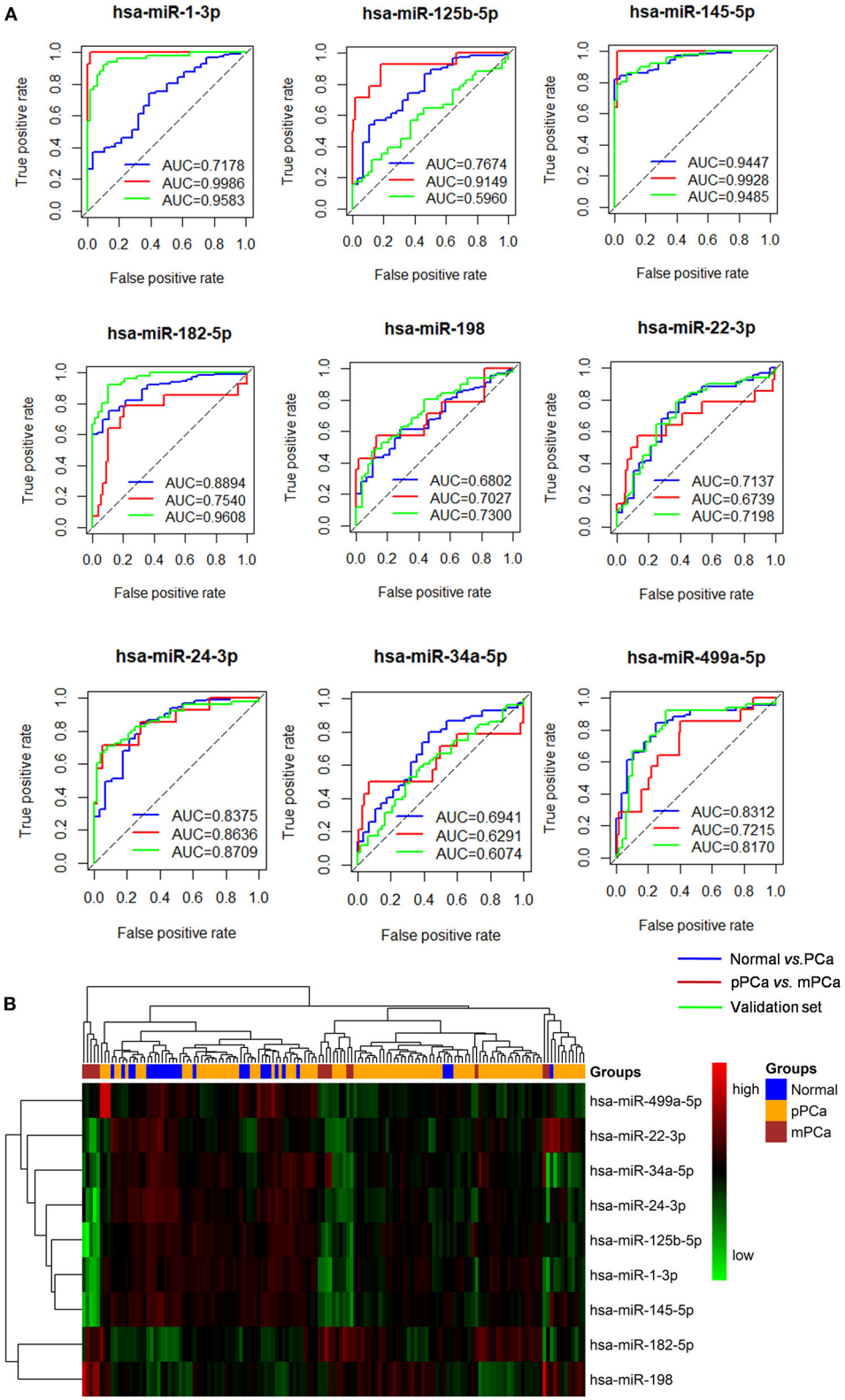


FIGURE 2 | The ROC and clustering analysis for the identified miRNAs. **(A)** ROC analysis. Blue curve: Normal vs. PCa for diagnostic performance evaluation; Red curve: pPCa vs. mPCa for prognostic and subtyping performance evaluation. Green curve: validation based on an independent dataset. **(B)** Clustering analysis. The

(Continued)

FIGURE 2 | blue, orange and red blocks represent normal, pPCa and PCa sample, respectively, and the degree of green and red in the map indicates the relative expression level of miRNAs from low to high, respectively. ROC, the receiver-operating characteristic curve; AUC, area under the ROC curve; PCa, prostate cancer; pPCa, clinically localized primary prostate cancer; mPCa, metastatic prostate cancer.

TABLE 3 | Functional mechanisms of *hsa-miR-145-5p* in PCa carcinogenesis.

miRNA function	Potential target	Gene function
• Suppressing *AR* signaling in PCa cells. • The level is inversely correlated with *PSA* alternation, the occurrence of metastasis phenotype and the response to androgen deprivation therapy. • PMID: 25969144	*NDRG2*	• A prognostic biomarker and regulator downstream of *AR*. • Associated with PCa malignant and metastatic progression. • Affecting the growth of androgen-dependent and castration-resistant PCa. • PMID: 24222185, 25756511
	KLF5	• A functional factor for *Androgen-AR* signaling. • Promoting cell proliferation in PCa cells. • PMID: 32245249
	IRS1	• G972R variant is associated with the risk of PCa occurrence. • PMID: 15678496
	ZFP36	• The expression change is associated with the overall survival and indicates the PCa biochemical recurrence. • PMID: 26563146
	GOLM1	• Promoting the progression of PCa via activating *PI3K-AKT-mTOR* signaling. • PMID: 29181846
	MYO6	• The knockdown inhibits the growth and results in the apoptosis of PCa cells. • PMID: 27431378
	ILK	• The inhibition suppresses the activation of *B/Akt* and induces apoptosis of *PTEN*-mutant PCa cells. • PMID: 10716737

PCa, prostate cancer; PMID, PubMed ID.

this study the regulations between *hsa-miR-145-5p* and known PCa-related genes were further analyzed since *hsa-miR-145-5p* was highly prioritized with the overall best performance on PCa prediction and subtyping in our ROC validation. As summarized in **Table 3**, *hsa-miR-145-5p* is a tumor suppressor in PCa development. It inhibited the *AR* signaling in PCa cells and the expression was inversely correlated with the change of *AR* and serum *PSA* level. In a well-characterized PCa cohort, this miRNA was found to be associated with the metastasis phenotype and could indicate the survival and the response to androgen deprivation therapy (Larne et al., 2015). Based on network extraction, seven PCa-related genes, i.e., *NDRG2*, *KLF5*, *IRS1*, *ZFP36*, *GOLM1*, *MYO6*, and *ILK*, were identified as potential targets in PCa carcinogenesis. Among them, *NDRG2* is a prognostic biomarker and negative regulator downstream of *AR* (Ren et al., 2014; Yu et al., 2015). It predicts PCa clinicopathologic features such as malignant and metastatic progression and affects the growth of androgen-dependent and castration-resistant PCa (Yu et al., 2015). Li et al. showed that *KLF5* was a key factor in androgen-*AR* signaling. It promoted the proliferation of PCa cells and could serve as a therapeutic target for PCa treatment (Li et al., 2020). As a famous star driving the carcinogenesis of PCa from normal prostate tissue to cancer biology, the *AR* signaling has already been explored across different researches, and the results in this study computationally demonstrated new insights and improved the understandings in *AR*-mediated PCa genesis. As described in **Figure 3B**, *hsa-miR-145-5p/NDRG2/AR* and *hsa-miR-145-5p/KLF5/AR* were inferred to be correlated with PCa evolution, which would be helpful for PCa diagnosis and therapy. In addition, *hsa-miR-145-5p* may regulate PCa cell proliferation, invasion, metastasis and apoptosis through other functional genes and associated pathways, which highlighted the underlying diversity and complexity in miRNA-PCa interaction (Persad et al., 2000; Neuhausen et al., 2005; Wang et al., 2016; Zhu et al., 2016; Yan et al., 2018).

GO and Pathway Enrichment Analysis

Functional enrichment analyses were performed on biomarker miRNA targets shared by PCa condition-specific networks to help decipher the miRNA-PCa carcinogenetic relationships at the GO and pathway level. In terms of the GO analysis, three domains, i.e., biological process (BP), cellular component (CC) and molecular function (MF), were annotated, respectively. As listed in **Figure 4A** and **Supplementary Table 2**, some of the significant BP terms were associated with the positive or negative regulation of cell proliferation and migration, indicating the potential role of the identified miRNAs in PCa invasion and metastasis. In the domain of CC, cytoplasm, nucleus, nucleoplasm, cytosol and extracellular exosome were the top-five ranked items

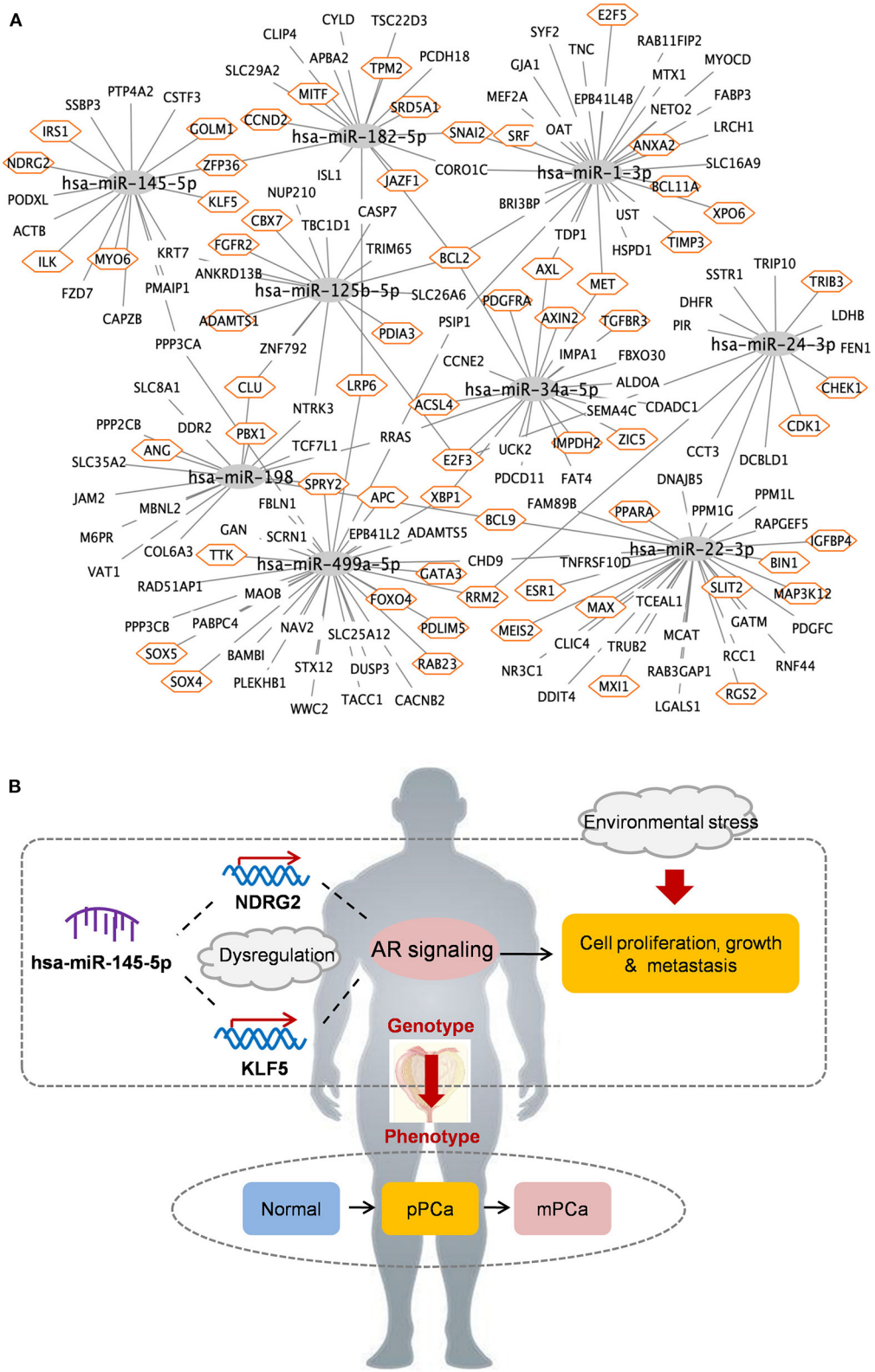


FIGURE 3 | The identified miRNAs and key miRNA-mRNA regulations in PCa occurrence and progression. **(A)** Overview of regulatory associations. Orange hexagon: PCa-related genes. **(B)** Potential mechanisms of *hsa-miR-145-5p* in PCa evolution. pPCa, clinically localized primary prostate cancer; mPCa, metastatic prostate cancer.

and at the MF level, transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding, protein binding, transmembrane receptor protein tyrosine kinase activity, and transcription factor binding etc. uncovered important clues for molecular function understanding.

To better investigate the mechanisms of the identified miRNAs in PCa, pathway enrichment was conducted and analyzed. As illustrated in **Figure 4B** and **Supplementary Table 2**, most of the significantly enriched terms were closely involved in PCa development, such as *Wnt* signaling, prostate cancer, microRNAs in cancer, *p53* signaling, *PI3K-AKT* signaling, and *MAPK* signaling etc. For example, *Wnt* signaling is implicated in PCa-related osteoblast differentiation as a key driver. It could activate *AR*-mediated transcription and promote cell proliferation of androgen-independent PCa (Seo et al., 2017; Wang et al., 2020). The genetic variants or abnormal regulations in *Wnt* signaling provide new approaches for predicting the aggressive behavior of PCa (Shu et al., 2016), and bring candidate targets for PCa personalized therapy (Nandana et al., 2017). Currently, extensive efforts have demonstrated that miRNAs could influence PCa cell activities by regulating genes to inactivate *Wnt* signaling pathway (Du et al., 2019; Ghafouri-Fard et al., 2020). Compared with our previous research, the prostate cancer signaling was enriched higher in this study (Lin et al., 2018a). As shown in **Supplementary Figure 1**, *GF* and *GFR* controlling the signal transduction from outside to inside of cell membrane were regulated by the identified miRNAs, and the remaining targets were potentially associated with PCa cell proliferation and survival. *p53* is a well-known tumor suppressor protein responding to various cellular stresses such as the growth, invasion and metastasis in PCa development (Takayama et al., 2018; Zhang et al., 2020). As another two cancer-related pathways, *PI3K-AKT* and *MAPK* signaling have been widely reported as targets of miRNAs and genes during PCa activation (Wu et al., 2019; Zheng et al., 2019). In particular, the crosstalk and signaling cascades among *I3K-AKT-mTOR*, *MAPK*, *AR,* and *Wnt* improve the mechanistic insights into PCa tumorigenesis and accelerate the understanding in androgen-deprivation therapeutics for precision medicine and personalized healthcare of PCa patients (Shorning et al., 2020).

DISCUSSION

PCa is developed with increasing incidence and high heterogeneity. In clinical practice, strategies used for PCa screening and monitoring have improved over the years, however, it still lacks sensitive factors to indicate the dynamical changes within prostate signals at the early stage. As a member of non-coding RNAs, miRNAs are reported to regulate gene expression in various biological processes including PCa carcinogenesis, which provide an attractive direction for PCa precision medicine and personalized healthcare.

In the era of translational informatics and intelligent medicine, systems biology creates unprecedented opportunities to integrate multi-dimensional data for computer-aided knowledge discovery. In this study, we collected gene expression and miRNA-mRNA association datasets to identify and explore key miRNAs as candidate biomarkers for PCa holistic management at the network level. Compared with our previous studies solely considering the metastatic or castration-resistant status of PCa (Zhu et al., 2015; Lin et al., 2018a), this study updated the datasets and bioinformatics parameters for model refining, and focused on the functional role of miRNAs associated with the whole development process in PCa, therefore two condition-specific miRNA-mRNA networks were respectively constructed to describe the regulatory pattern and dynamical change between PCa occurrence and progression. In particular, the hub theory and single-line regulation pattern of miRNAs were integrated for the first time to measure the regulatory power, and miRNAs locating at hub sites to independently regulate genes were extracted from each network system based on the definition and characterization of network topologies. Finally, nine miRNAs shared by two networks, i.e., *hsa-miR-1-3p*, *hsa-miR-125b-5p*, *hsa-miR-145-5p*, *hsa-miR-182-5p*, *hsa-miR-198*, *hsa-miR-22-3p*, *hsa-miR-24-3p*, *hsa-miR-34a-5p*, and *hsa-miR-499a-5p*, were chosen as candidate biomarkers during PCa evolution for performance evaluation and carcinogenic survey.

To validate the potential of the identified miRNAs, ROC and clustering analysis were sequentially performed to test the ability in PCa diagnosis and prognosis. Fortunately, most of these miRNAs achieved higher AUC both in differentiating normal vs. PCa and pPCa vs. mPCa samples based on the prediction and an independent validation dataset, which indicated the predictive power of the miRNAs. Among them, *hsa-miR-145-5p* was top-ranked as the key factor and the result was highly consistent with that in our previous findings using different training datasets and network analysis strategies (Lin et al., 2018a). According to PubMed literature searching, seven of the miRNAs have been reported to be associated with PCa genesis and could serve as biomarkers or therapeutic targets for PCa prevention. Combing with computational prediction, key miRNA-mRNA were screened to decode the relationships between miRNA genotypes and PCa phenotypes at the gene and pathway level, respectively. In particular, *hsa-miR-145-5p/NDRG2/AR* and *hsa-miR-145-5p/KLF5/AR* axis were inferred to be latent mechanisms during PCa occurrence and progression according to bioinformatics identification and literature validation. Moreover, pathways including *Wnt* signaling, prostate cancer, microRNAs in cancer, *p53* signaling, *PI3K-AKT* signaling, and *MAPK* signaling etc. were significantly enriched for pathogenesis understanding.

It should be admitted that several limitations still need to be considered. First, the structural robustness is comprehensively weighted in this study, however, the proposed model lacks sufficient information related to the biological function of miRNAs and mRNAs. Hence the computational framework can be updated by reasonably adding PCa-associated genes as prior knowledge to improve the specificity of miRNAs in PCa carcinogenesis. Second, the complexity and diversity of the background network is not powerful enough. The development of PCa is a dynamical process, so that changeable

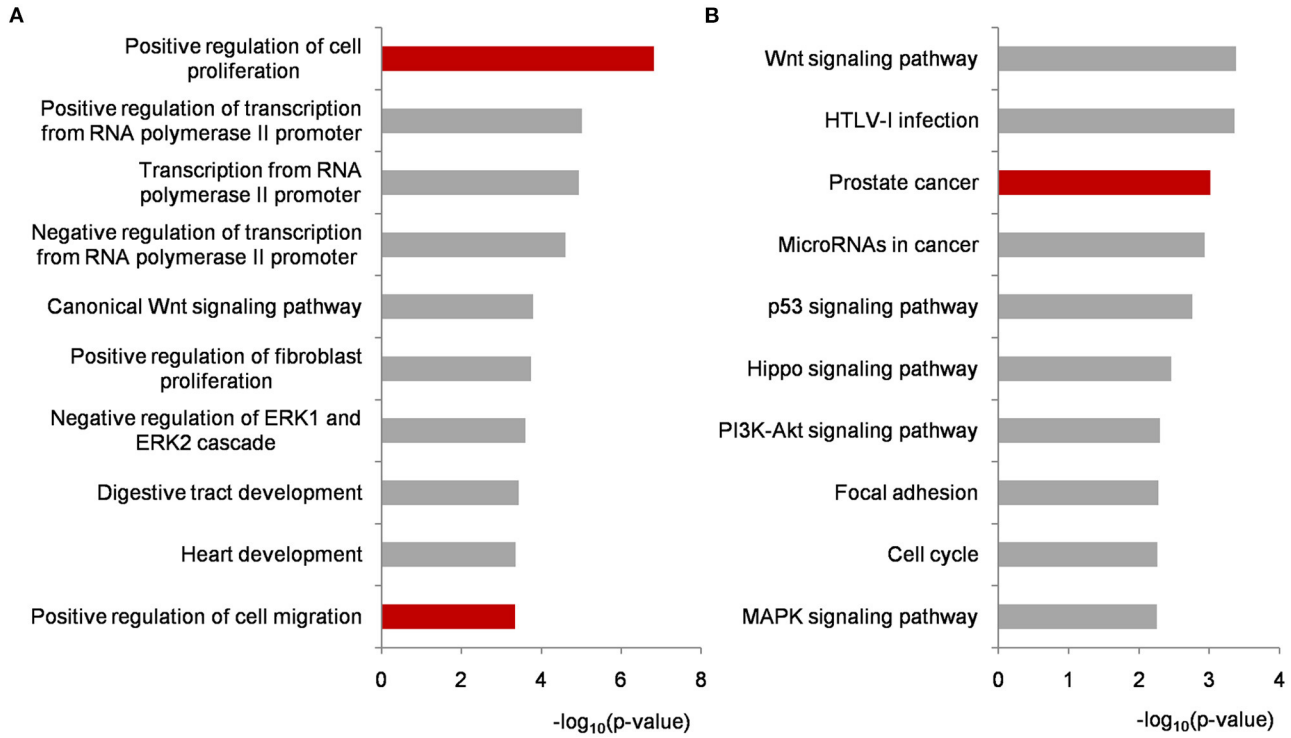


FIGURE 4 | Functional enrichment analyses on biomarker miRNA targets shared by PCa occurrence and progression-specific networks. **(A)** Top ten significantly enriched BP terms. **(B)** Top ten significantly enriched KEGG pathways. All of the *p*-values were negative 10-based log transformed. PCa, prostate cancer; BP, biological process; KEGG, Kyoto Encyclopedia of Genes and Genomes.

signals in the network system from normal to different PCa states are meaningful for capturing and comparison. To better understand the heterogeneity across PCa stages, miRNAs specific to PCa conditions, e.g., the occurrence, invasion, metastasis and therapeutic intervention, should be respectively analyzed. Last but the most important, it is difficult to collect enough samples of advanced or mPCa in a short period of time due to the inoperability of these patients. In our future work, low-throughput experiments using wet-lab approaches such as qPCR and western blot will be conducted to validate the identified miRNAs and miRNA-mRNA regulations for biomarker measurement and long-range clinical translation.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

BS, YL, and YH designed this study. YL and ZM collected the data. YL and BS proposed the bioinformatics model. YL, ZM, XZ, and XW performed the analysis. YL, JH, YH, and BS wrote and revised the manuscript. BS and YH conceived and supervised the study jointly. All the authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (31670851), the Key Research and Development Program of Jiangsu Province (BE2020655), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (20KJB180010).

ACKNOWLEDGMENTS

The authors gratefully thank the editors and reviewers for their constructive suggestions to improve this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.596826/full#supplementary-material

REFERENCES

Bhagirath, D., Yang, T. L., Bucay, N., Sekhon, K., Majid, S., Shahryari, V., et al. (2018). microRNA-1246 is an exosomal biomarker for aggressive prostate cancer. *Cancer Res* 78, 1833–1844. doi: 10.1158/0008-5472.CAN-17-2069

Bidarra, D., Constancio, V., Barros-Silva, D., Ramalho-Carvalho, J., Moreira-Barbosa, C., Antunes, L., et al. (2019). Circulating MicroRNAs as biomarkers for prostate cancer detection and metastasis development prediction. *Front. Oncol.* 9:900. doi: 10.3389/fonc.2019.00900

Corcoran, C., Rani, S., and O'Driscoll, L. (2014). miR-34a is an intracellular and exosomal predictive biomarker for response to docetaxel with clinical relevance to prostate cancer progression. *Prostate* 74, 1320–1334. doi: 10.1002/pros.22848

Du, H., Wang, X., Dong, R., Hu, D., and Xiong, Y. (2019). miR-601 inhibits proliferation, migration and invasion of prostate cancer stem cells by targeting KRT5 to inactivate the Wnt signaling pathway. *Int. J. Clin. Exp. Pathol.* 12, 4361–4379.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874. doi: 10.1038/nrg3074

Ghafouri-Fard, S., Shoorei, H., and Taheri, M. (2020). Role of microRNAs in the development, prognosis and therapeutic response of patients with prostate cancer. *Gene* 759, 144995. doi: 10.1016/j.gene.2020.144995

Gu, L., Frommel, S. C., Oakes, C. C., Simon, R., Grupp, K., Gerig, C. Y., et al. (2015). BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence. *Nat. Genet.* 47:22. doi: 10.1038/ng.3165

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

Hudson, R. S., Yi, M., Esposito, D., Watkins, S. K., Hurwitz, A. A., Yfantis, H. G., et al. (2012). MicroRNA-1 is a candidate tumor suppressor and prognostic marker in human prostate cancer. *Nucleic Acids Res.* 40, 3689–3703. doi: 10.1093/nar/gkr1222

Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 7:359. doi: 10.1186/1471-2105-7-359

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

Khanmi, K., Ignacimuthu, S., and Paulraj, M. G. (2015). MicroRNA in prostate cancer. *Clin Chim Acta.* 451(Pt B), 154–160. doi: 10.1016/j.cca.2015.09.022

Larne, O., Hagman, Z., Lilja, H., Bjartell, A., Edsjo, A., and Ceder, Y. (2015). miR-145 suppress the androgen receptor in prostate cancer cells and correlates to prostate cancer prognosis. *Carcinogenesis* 36, 858–866. doi: 10.1093/carcin/bgv063

Li, J., Zhang, B., Liu, M., Fu, X., Ci, X., A, J., et al. (2020). KLF5 is crucial for androgen-AR signaling to transactivate genes and promote cell proliferation in prostate cancer cells. *Cancers* 12:3. doi: 10.3390/cancers12030748

Lin, Y., Chen, F., Shen, L., Tang, X., Du, C., Sun, Z., et al. (2018a). Biomarker microRNAs for prostate cancer metastasis: screened with a network vulnerability analysis model. *J. Transl. Med.* 16:134. doi: 10.1186/s12967-018-1506-7

Lin, Y., Qian, F., Shen, L., Chen, F., Chen, J., and Shen, B. (2019). Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Brief. Bioinformatics* 20, 952–975. doi: 10.1093/bib/bbx158

Lin, Y., Wu, W., Sun, Z., Shen, L., and Shen, B. (2018b). MiRNA-BD: an evidence-based bioinformatics model and software tool for microRNA biomarker discovery. *RNA Biol.* 15, 1093–1105. doi: 10.1080/15476286.2018.1502590

Lin, Y., Zhao, X., Miao, Z., Ling, Z., Wei, X., Pu, J., et al. (2020). Data-driven translational prostate cancer research: from biomarker discovery to clinical decision. *J. Transl. Med.* 18:119. doi: 10.1186/s12967-020-02281-4

Liu, C., Kelnar, K., Liu, B., Chen, X., Calhoun-Davis, T., Li, H., et al. (2011). The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. *Nat. Med.* 17, 211–215. doi: 10.1038/nm.2284

Liu, Y. Y., Slotine, J. J., and Barabasi, A. L. (2011). Controllability of complex networks. *Nature* 473, 167–173. doi: 10.1038/nature10011

Lynch, S. M., McKenna, M. M., Walsh, C. P., and McKenna, D. J. (2016). miR-24 regulates CDKN1B/p27 expression in prostate cancer. *Prostate* 76, 637–648. doi: 10.1002/pros.23156

Nandana, S., Tripathi, M., Duan, P., Chu, C. Y., Mishra, R., Liu, C., et al. (2017). Bone metastasis of prostate cancer can be therapeutically targeted at the TBX2-WNT signaling axis. *Cancer Res.* 77, 1331–1344. doi: 10.1158/0008-5472.CAN-16-0497

Neuhausen, S. L., Slattery, M. L., Garner, C. P., Ding, Y. C., Hoffman, M., and Brothman, A. R. (2005). Prostate cancer risk and IRS1, IRS2, IGF1, and INS polymorphisms: strong association of IRS1 G972R variant and cancer risk. *Prostate* 64, 168–174. doi: 10.1002/pros.20216

Persad, S., Attwell, S., Gray, V., Delcommenne, M., Troussard, A., Sanghera, J., et al. (2000). Inhibition of integrin-linked kinase (ILK) suppresses activation of protein kinase B/Akt and induces cell cycle arrest and apoptosis of PTEN-mutant prostate cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 97, 3207–3212. doi: 10.1073/pnas.060579697

Ren, G. F., Tang, L., Yang, A. Q., Jiang, W. W., and Huang, Y. M. (2014). Prognostic impact of NDRG2 and NDRG3 in prostate cancer patients undergoing radical prostatectomy. *Histol. Histopathol.* 29, 535–542. doi: 10.14670/HH-29.10.535

Roberts, M. J., Chow, C. W., Schirra, H. J., Richards, R., Buck, M., Selth, L. A., et al. (2015). Diagnostic performance of expression of PCA3, Hepsin and miR biomarkers inejaculate in combination with serum PSA for the detection of prostate cancer. *Prostate* 75, 539–549. doi: 10.1002/pros.22942

Seo, W. I., Park, S., Gwak, J., Ju, B. G., Chung, J. I., Kang, P. M., et al. (2017). Wnt signaling promotes androgen-independent prostate cancer cell proliferation through up-regulation of the hippo pathway effector YAP. *Biochem. Biophys. Res. Commun.* 486, 1034–1039. doi: 10.1016/j.bbrc.2017.03.158

Shorning, B. Y., Dass, M. S., Smalley, M. J., and Pearson, H. B. (2020). The PI3K-AKT-mTOR pathway and prostate cancer: at the crossroads of AR, MAPK, and WNT signaling. *Int. J. Mol. Sci.* 21:12. doi: 10.3390/ijms21124507

Shu, X., Ye, Y., Gu, J., He, Y., Davis, J. W., Thompson, T. C., et al. (2016). Genetic variants of the Wnt signaling pathway as predictors of aggressive disease and reclassification in men with early stage prostate cancer on active surveillance. *Carcinogenesis* 37, 965–971. doi: 10.1093/carcin/bgw082

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

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3:3. doi: 10.2202/1544-6115.1027

Takayama, K. I., Suzuki, T., Tanaka, T., Fujimura, T., Takahashi, S., Urano, T., et al. (2018). TRIM25 enhances cell growth and cell survival by modulating p53 signals via interaction with G3BP2 in prostate cancer. *Oncogene* 37, 2165–2180. doi: 10.1038/s41388-017-0095-x

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., et al. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell* 18, 11–22. doi: 10.1016/j.ccr.2010.05.026

Tu, J., Peng, Q., Shen, Y., Hong, Y., Zhu, J., Feng, Z., et al. (2019). Identification of biomarker microRNA-mRNA regulatory pairs for predicting the docetaxel resistance in prostate cancer. *J. Cancer* 10, 5469–5482. doi: 10.7150/jca.29032

Walter, B. A., Valera, V. A., Pinto, P. A., and Merino, M. J. (2013). Comprehensive microRNA Profiling of Prostate Cancer. *J. Cancer* 4, 350–357. doi: 10.7150/jca.6394

Wang, D., Zhu, L., Liao, M., Zeng, T., Zhuo, W., Yang, S., et al. (2016). MYO6 knockdown inhibits the growth and induces the apoptosis of prostate cancer cells by decreasing the phosphorylation of ERK1/2 and PRAS40. *Oncol. Rep.* 36, 1285–1292. doi: 10.3892/or.2016.4910

Wang, Y., Singhal, U., Qiao, Y., Kasputis, T., Chung, J. S., Zhao, H., et al. (2020). Wnt signaling drives prostate cancer bone metastatic tropism and invasion. *Transl. Oncol.* 13:100747. doi: 10.1016/j.tranon.2020.100747

Wei, J., Yin, Y., Deng, Q., Zhou, J., Wang, Y., Yin, G., et al. (2020). Integrative analysis of MicroRNA and gene interactions for revealing candidate signatures in prostate cancer. *Front. Genet.* 11:176. doi: 10.3389/fgene.2020.00176

Wu, X., Xiao, Y., Yan, W., Ji, Z., and Zheng, G. (2019). The human oncogene SCL/TAL1 interrupting locus (STIL) promotes tumor growth through MAPK/ERK, PI3K/Akt and AMPK pathways in prostate cancer. *Gene* 686, 220–227. doi: 10.1016/j.gene.2018.11.048

Xie, Z. C., Huang, J. C., Zhang, L. J., Gan, B. L., Wen, D. Y., Chen, G., et al. (2018). Exploration of the diagnostic value and molecular mechanism of miR1

in prostate cancer: A study based on metaanalyses and bioinformatics. *Mol. Med. Rep.* 18, 5630–5646. doi: 10.3892/mmr.2018.9598

Xu, Y., Qin, S., An, T., Tang, Y., Huang, Y., and Zheng, L. (2017). MiR-145 detection in urinary extracellular vesicles increase diagnostic efficiency of prostate cancer based on hydrostatic filtration dialysis method. *Prostate* 77, 1167–1175. doi: 10.1002/pros.23376

Yan, G., Ru, Y., Wu, K., Yan, F., Wang, Q., Wang, J., et al. (2018). GOLM1 promotes prostate cancer progression through activating PI3K-AKT-mTOR signaling. *Prostate* 78, 166–177. doi: 10.1002/pros.23461

Yu, C., Wu, G., Li, R., Gao, L., Yang, F., Zhao, Y., et al. (2015). NDRG2 acts as a negative regulator downstream of androgen receptor and inhibits the growth of androgen-dependent and castration-resistant prostate cancer. *Cancer Biol. Ther.* 16, 287–296. doi: 10.1080/15384047.2014.1002348

Zhang, S., Yu, J., Sun, B. F., Hou, G. Z., Yu, Z. J., and Luo, H. (2020). MicroRNA-92a targets SERTAD3 and regulates the growth, invasion, and migration of prostate cancer cells via the P53 pathway. *Onco. Targets. Ther.* 13, 5495–5514. doi: 10.2147/OTT.S249168

Zheng, X. M., Zhang, P., Liu, M. H., Chen, P., and Zhang, W. B. (2019). MicroRNA-30e inhibits adhesion, migration, invasion and cell cycle progression of prostate cancer cells via inhibition of the activation of the MAPK signaling pathway by downregulating CHRM3. *Int. J. Oncol.* 54, 443–454. doi: 10.3892/ijo.2018.4647

Zhu, J., Wang, S., Zhang, W., Qiu, J., Shan, Y., Yang, D., et al. (2015). Screening key microRNAs for castration-resistant prostate cancer based on miRNA/mRNA functional synergistic network. *Oncotarget* 6, 43819–43830. doi: 10.18632/oncotarget.6102

Zhu, J. G., Yuan, D. B., Chen, W. H., Han, Z. D., Liang, Y. X., Chen, G., et al. (2016). Prognostic value of ZFP36 and SOCS3 expressions in human prostate cancer. *Clin. Transl. Oncol.* 18, 782–791. doi: 10.1007/s12094-015-1432-6

Zhu, Z., Wen, Y., Xuan, C., Chen, Q., Xiang, Q., Wang, J., et al. (2020). Identifying the key genes and microRNAs in prostate cancer bone metastasis by bioinformatics analysis. *FEBS Open Bio* 10, 674–688. doi: 10.1002/2211-5463.12805

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 February 2021
doi: 10.3389/fgene.2021.617915

The Shared Use of Extended Phenotypes Increases the Fitness of Simulated Populations

*Guilherme F. de Araújo[1], Renan C. Moioli[1] and Sandro J. de Souza[1,2,3]**

[1]*Bioinformatics Multidisciplinary Environment, Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal, Brazil,* [2]*Brain Institute, Universidade Federal do Rio Grande do Norte, Natal, Brazil,* [3]*Institutes for Systems Genetics, West China Hospital, University of Sichuan, Chengdu, China*

Extended phenotypes are manifestations of genes that occur outside of the organism that possess those genes. In spite of their widespread occurrence, the role of extended phenotypes in evolutionary biology is still a matter of debate. Here, we explore the indirect effects of extended phenotypes, especially their shared use, in the fitness of simulated individuals and populations. A computer simulation platform was developed in which different populations were compared regarding their ability to produce, use, and share extended phenotypes. Our results show that populations that produce and share extended phenotypes outrun populations that only produce them. A specific parameter in the simulations, a bonus for sharing extended phenotypes among conspecifics, has a more significant impact in defining which population will prevail. All these findings strongly support the view, postulated by the extended fitness hypothesis (EFH) that extended phenotypes play a significant role at the population level and their shared use increases population fitness. Our simulation platform is available at https://github.com/guilherme-araujo/gsop-dist.

Keywords: extended phenotypes, Moran process, simulated platform, system biology, network

OPEN ACCESS

Edited by:
Josselin Noirel, Conservatoire National des Arts et Métiers (CNAM), France

Reviewed by:
Sean Blamires, University of New South Wales, Australia
Alexey Goltsov, Moscow State Institute of Radio Engineering, Electronics, and Automation, Russia

****Correspondence:***
Sandro J. de Souza
sandro@neuro.ufrn.br

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

Received: *15 October 2020*
Accepted: *15 January 2021*
Published: *03 February 2021*

Citation:
de Araújo GF, Moioli RC and de Souza SJ (2021) The Shared Use of Extended Phenotypes Increases the Fitness of Simulated Populations. Front. Genet. 12:617915. doi: 10.3389/fgene.2021.617915

INTRODUCTION

The main idea behind the extended phenotype (Dawkins, 1982) lies in how far a gene effect can reach. According to Dawkins (1982), a gene can have its effect outside of the physical body of the bearer with several types of consequences, including environmental ones. In that way, a gene extends its effect in, for example, a beaver's dam, a spider's web, or a bird's nest. Although the examples above represent physical structures, extended phenotypes are also seen as signals (Schaedelin and Taborsky, 2009), social interactions (Wang et al., 2008), or manipulations of behaviors (Hoover et al., 2011). Extended phenotypes are described in all taxonomic kingdoms, from viruses (Hoover et al., 2011) to humans (Dixson, 2019). Although the widespread existence of extended phenotypes is clearly established in contemporaneous evolutionary biology (reviewed in Bailey, 2012), the degree and intensity of its effects are still controversial (Hunter, 2009; Bailey, 2012).

Besides the obvious effect of the extended phenotypes in the fitness of the organism who generated it, many authors have discussed their indirect genetic effects (Laland, 2004; Wang et al., 2008; de Souza, 2013; Fisher et al., 2019). One type of indirect genetic effect is through social interactions mediated by extended phenotypes (Wang et al., 2008). Extended phenotypes could also affect other parties' fitness through niche construction, as discussed by Laland (2004). A few years ago, the extended fitness hypothesis (EFH) had been proposed, which states that extended phenotypes serve as a link between individual and group selection (de Souza, 2013). Suppose the following scenario: a spider web is abandoned by the individual who built it. A different spider from the same species can then use that web, which in turn contributes to the fitness of the new owner. Remarkably, the spider web silk can vary within the same species depending on environmental factors, and protein-deprived spiders produce silk that is more efficient at capturing preys than that produced by protein-fed members of the same species (conspecifics; Blamires et al., 2017, 2018), resulting in "silk performance landscapes across nutrient space" (Blamires et al., 2016). In another example, a bird's nest shape impact on its thermal profile, which in turn, has been shown to influence offspring fitness (Olson et al., 2006; Martin et al., 2017). Thus, using an extended phenotype built by others may have greater advantage than simply reducing the costs associated to building the phenotype. However, the fitness effects of such biological plasticity mechanisms and their impact on individual and group selection are not fully understood.

The basis of the EFH is the fact that individuals can use extended phenotypes built by conspecifics. Thus, extended phenotypes possess indirect genetic effects in individuals who can use them. Group selection emerges naturally as a consequence of such shared use of extended phenotypes by members of the same species/group. As discussed by de Souza (2013), there are several examples of the use of available extended phenotypes by conspecifics, including cases with spiders (Schuck-Paim and Alonso, 2001), cichlids (Schaedelin and Taborsky, 2009), and wasps (Brockmann et al., 1979). For instance, a beaver's dam may cause a significant environmental change that goes beyond the immediate ecosystem (Gurnell, 1998; Rosell et al., 2005). More recently, Fisher et al. (2019) showed that food hoards, identified as extended phenotypes, built by red squirrels outlived the individuals who built them and were subsequently used by conspecifics. More interestingly, different features of the food hoards, like size, affected the fitness of the subsequent owner. While the data from Fisher et al. (2019) fit predictions made by EFH, such empirical models are hard to find and study. One alternative is the use of computer simulations to either compare distinct evolutionary scenarios or to study the role of a given parameter, in this case extended phenotypes, in the evolutionary process.

This led us to develop a computer simulation framework to test some premises of the EFH. Here, we show that extended phenotypes can, *per se*, increase the fitness of individuals who produce them. More importantly, however, populations that produce and share extended phenotypes outrun populations that only produce them. A mathematical treatment allowed us to derive variables that can be evaluated regarding their role in the fitness of the tested populations. A bonus linked to the shared use of extended phenotypes is strongly associated with winning populations in our simulations. All these findings support the view that the shared use of extended phenotypes is an important contributor to selection at population level. We made our simulation platform available at https://github.com/guilherme-araujo/gsop-dist.

MATERIALS AND METHODS

Simulation Steps

The simulation protocol consists of three steps: graph generation, main simulation, and plot/analysis. At the graph generation step, the type of graph, the number of nodes, and the density of the graph are defined (Steger and Wormwald, 1999). The generated graph is then read by the main simulation program, which accepts parameters related to the bonuses, maximum number of cycles, and number of samples and states in which each node type can transition into. Finally, the output is processed and the plots generated using the scripts available at the corresponding folder of the public repository of this simulation platform. All simulations were run in the High-Performance Computing Unit of the Federal University of Rio Grande do Norte, consisting of 64 blade computational nodes, each with two 16-core Intel Xeon E5-2698v3 processors and 128GB DDR4 RAM.

Graph Generation

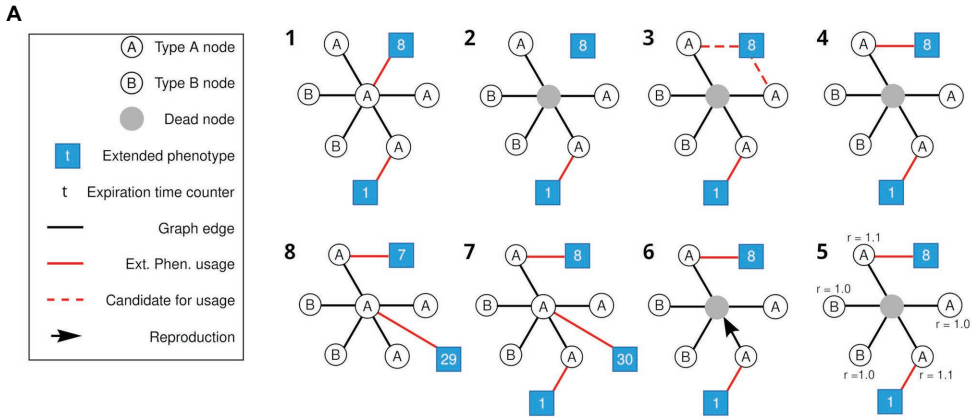
Graphs were generated using the *newtworkx* (Hagberg et al., 2008) package for the Python programming language. All graphs in the simulation described in this paper were generated with the *barabasi_albert_graph* function of this package, with parameters $n = 500$ and $m = 4$.

Main Simulation

The algorithm for the first simulation is described as a pseudocode in **Figure 1B**. The first simulation implements the framework described in **Figure 1A**, and generates data for plotting **Figure 2**. The algorithm for the second simulation is described as a pseudocode in **Figure 3B**. The second simulation implements the framework described in **Figure 3A**, and generates data for plotting **Figure 4**.

For the first algorithm, two sets of data were generated, the first for the simulation where B individuals do not generate any extended phenotype, and the second where both types generate extended phenotypes, but only type A individuals can reuse them. These two sets of data resulted in the plots seen in **Figure 2**. The second algorithm was used to generate another two sets of data, which resulted in the data seen in **Figure 4**.

In the first pseudocode (**Figure 3B**), the nodes are first (line 1) load from the graph generated in the first simulation



B

```
1  nodes = Read graph
2  for i = 0...nodes.size() do
3      if i < nodes.size()/2 then
4          nodes[i].type = A
5      else
6          nodes[i].type = B
7  Add extended phenotypes to 50% of type A individuals in nodes
8  if Type B is Extended Phenotype Generator then
9      Add extended phenotypes to 50% of type B individuals in nodes
10 for i = 0...Number of cycles do
11     for d = 0... Number of death/birth events by cycle do
12         neighbors = Neighbors of nodes[d]
13         chosenNeighbor = Random choice among neighbors weighted
            by relative fitness
14         newNode = Create new node
15         Assign chosenNeighbor type to newNode
16         if newNode generates an extended phenotype then
17             Assign new extended phenotype to newNode
18         abandoned = Extended phenotype of nodes[d]
19         Replace nodes[d] with newNode
20         try:
21             Assign abandoned to random individual in neighbors
22         catch No neighbor available for assignment:
23             try:
24                 Assign abandoned to random individual in nodes
25         Decrease extended phenotype timers
26         Remove expired extended phenotypes
27 Print result
```

FIGURE 1 | Schematic view of the simulation framework and respective pseudocode. **(A)** All steps (1–8) of a cycle of the framework are depicted. After the initial setup of the network (1), random individuals are selected to die (like the gray node in 2). Its associated extended phenotype becomes available (3) and one of the neighbors of the same type (in this case, type A) and without an associated extended phenotype is selected to gain the available extended phenotype (4). Selection of a node to duplicate and occupy the position of the dead node is based on a weight matrix (5, 6), as described in the text. A new node has a chance to generate an extended phenotype attached to itself (7). Each extended phenotype has an expiration time (*t*) represented by the number in the respective squares (7, 8). Step 8 represents the step 1 of the new cycle. For clarity, only the central node is represented with all its connections. **(B)** Pseudocode for the simulation framework described above **(A)**.

step, described in "Graph generation." In lines 2–6, nodes are initialized with type A or B. In lines 7–9, the extended phenotypes are initialized, and 50% of all individuals start with an extended phenotype. The verification at line 8 is related to the difference between the simulations that generated **Figures 2A,B**, since in the first set of data type B individuals do not generate extended phenotypes. Next, for each cycle (line 10) at each death/birth event (line 11), a random

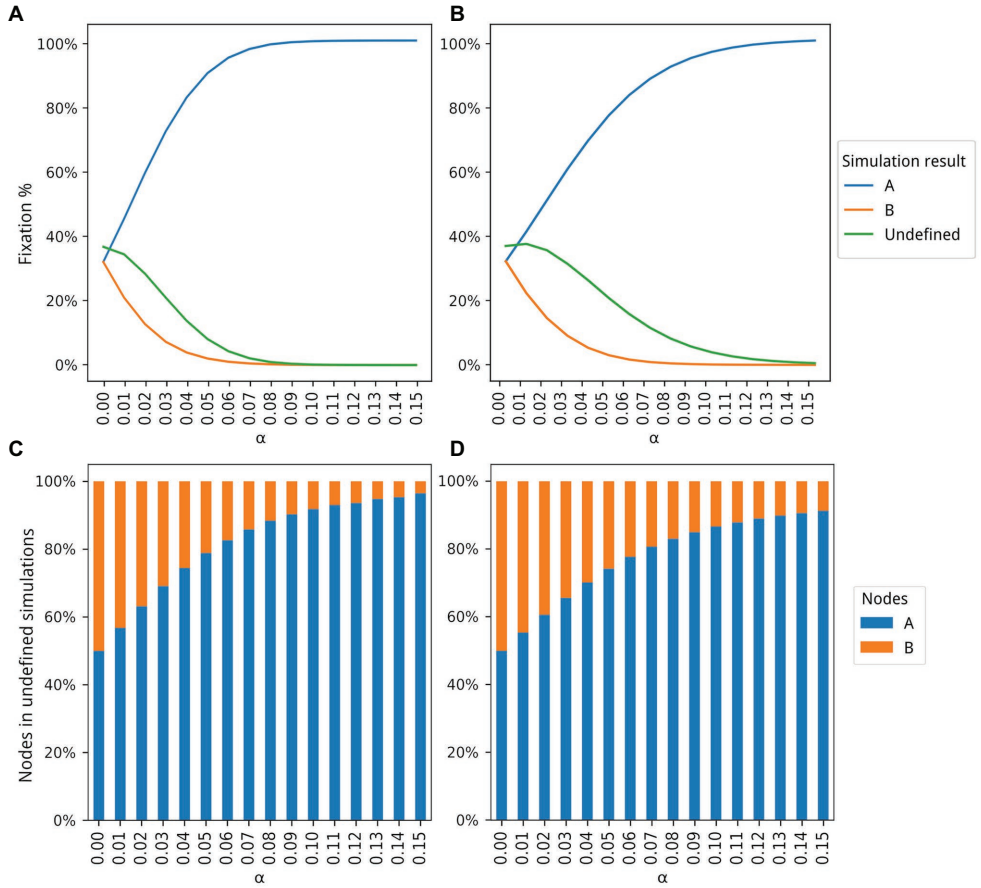


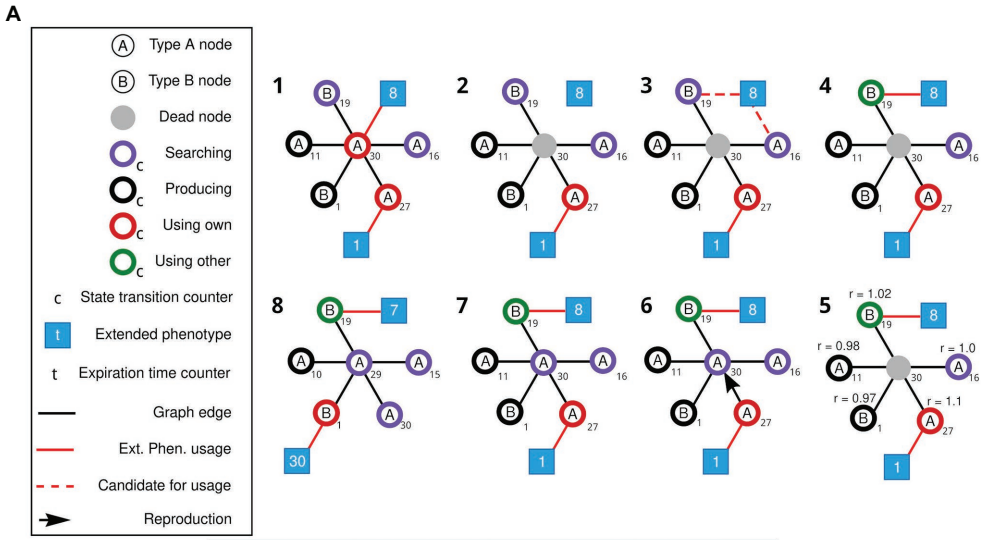
FIGURE 2 | Dynamics of populations A and B according to 5 million simulations for different values of α. The blue and orange lines in **(A,B)** show how many simulations ended with the fixation of types A and B, respectively. The green line in **(A,B)** shows how many simulations ended without the fixation of either type, that is, undefined simulations. Proportions of type A and B individuals in the undefined simulations are shown in **(C,D)**. **(A)** Only population A is able to produce and share extended phenotypes. **(B)** Both populations can produce extended phenotypes but only population A is able to share extended phenotypes. **(C)** Proportions of type A and B individuals for the simulations represented by the green curve shown in **(A)**. **(D)** Proportion of type A and B individuals for the simulations represented by the green curve shown in **(B)**.

neighbor of the dying individual is chosen weighted according to its relative fitness (lines 12–13). The new individual is created having the same type of the chosen neighbor (lines 14–15), and if it belongs to a type which generates extended phenotypes – only A in the first, and both types in the second simulation – it occupies the extended phenotype (lines 16–17). If the dying individual had an extended phenotype, an attempt is made to assign it to one of the other neighbors or a random individual in the population (lines 18 and 20–24). The dying individual is then replaced by the new one (line 19) and finally the extended phenotype timers are decreased and those who expired are removed (lines 25–26).

The second algorithm starts like the previous one (lines 1–6), then initializes the states according to parameters defined by the user (lines 7–8). Then, for each death/birth event at each cycle (lines 9–11), a random neighbor of the dying individual is chosen weighted according to its relative fitness (lines 12–13) and a new node is created, always at the searching state, with no extended phenotype attached (lines 14–18). If the dying individual had an extended phenotype, an attempt is made to assign it to one of the other neighbors or a random individual in the population (lines 19 and 21–27). This individual will be transitioned to the "using other" state (lines 23 and 27). The dying individual is then replaced by the new one (line 19). Finally, all states and extended phenotype time counters are updated, its states transitioned and expired extended phenotypes removed (lines 28–37).

All simulations are provided only with the random graphs generated in the previous step and the parameters described in section "Simulation parameters." For each of the four sets of data described previously, 1,000 graphs were generated, and 5,000 samples were generated with each of the 1,000 graphs, resulting in 5,000,000 total samples for each x-axis data point of each set of data.

The sets of data for the first and second simulation, which generated data for **Figure 2**, had 15 subsets of data each, varying the α value for A from 0.0 to 0.15 in both simulations,



B

```
1  nodes = Read graph
2  for i = 0...nodes.size() do
3      if i < nodes.size()/2 then
4          nodes[i].type = A
5      else
6          nodes[i].type = B
7  Initialize states searching, producing, using and using other in all nodes
8  Initialize extended phenotypes for all nodes in states [using, using other]
9  for i = 0...Number of cycles do
10     count = Number of death/birth events by cycle
11     for d in [count random values in nodes.size()] do
12         neighbors = Neighbors of nodes[d]
13         chosenNeighbor = Random choice among neighbors weighted
            by relative fitness
14         newNode = Create new node
15         Assign chosenNeighbor type to newNode
16         Assign newNode state to searching
17         Assing newNode timer to 30
18         Assign newNode extended phenotype to NULL
19         abandoned = Extended phenotype of nodes[d]
20         Replace nodes[d] with newNode
21         try:
22             Assign abandoned to random individual in neighbors
23             Change state of chosen among neighbors to using other
24         catch No neighbor available for assignment:
25             try:
26                 Assign abandoned to random individual in nodes
27                 Change state of chosen among nodes to using other
28     Decrease timers of nodes in states [searching, producing]
29     for node in nodes with timers == 0 do
30         if State of node == searching then
31             Change state of node to producing
32         if State of node == producing then
33             Assign new extended phenotype to node
34             Change state of node to using
35     Decrease extended phenotype timers
36     Remove expired extended phenotypes
37     Set state of nodes whose extended phenotypes expired to searching
38 Print result
```

FIGURE 3 | Schematic view of the modified simulation framework and respective pseudocode. **(A)** Nodes of types A and B can search, produce, and use its own or use other extended phenotypes. After the initial setup of the network (1), random individuals are selected to die (2). The associated extended phenotype becomes available, and one of the neighboring nodes in the Searching state is selected to gain the available extended phenotype (3, 4). Selection of a node to duplicate and occupy the position of the dead node is based on a weight matrix (5, 6), according to the state of each node. Node state transition and expiration time counters (*t*) are updated, and states and extended phenotypes are adjusted accordingly (7, 8). Step 8 represents step 1 of the new cycle. **(B)** Pseudocode for the simulation framework described above **(A)**.

and for B in the second simulation. In the first simulation, the α value of B is fixed at 0.0. Every subset consists of 5,000,000 samples, generated in the previously described way, with the intention of removing any influence a particular characteristic of a randomly generated graph could have had on the final result.

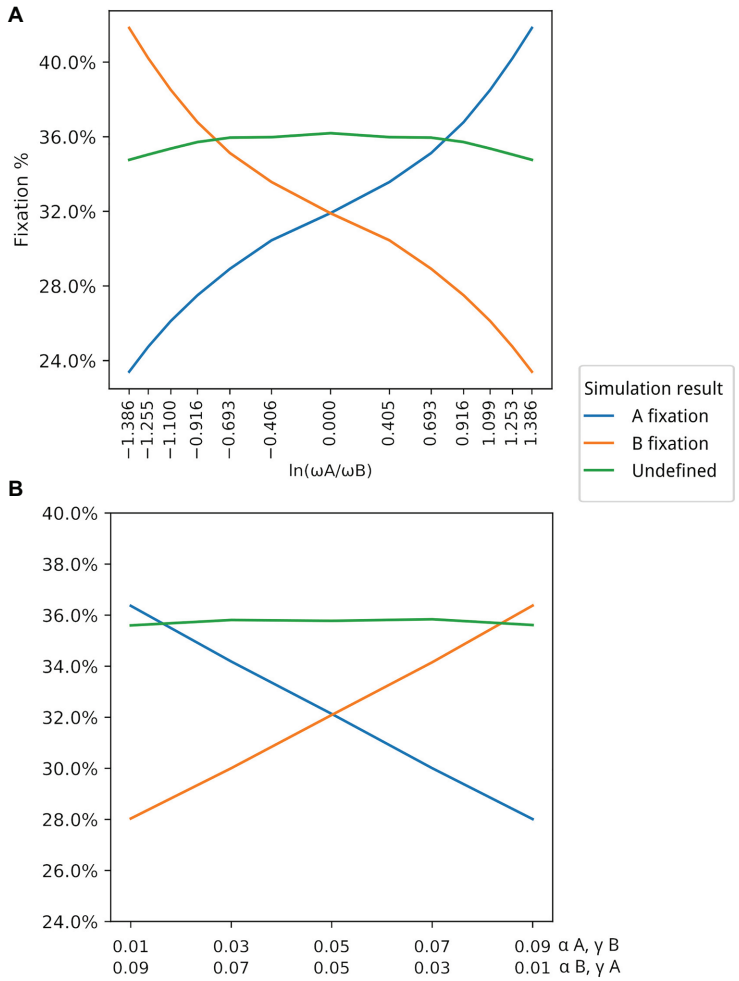


FIGURE 4 | Association between ω and winning populations. *Y* axis in both graphs represents the average fixation % in the corresponding simulations. **(A)** Association between ω and winning populations (those with higher fixation rate). For this simulation, $\beta = 0.03$ and $\gamma = 0.02$ for both populations. **(B)** In this simulation, both α and γ are changed under the restriction that $\omega_A = \omega_B$ Populations with $\gamma > \alpha$ are winners in situations where $\omega_A = \omega_B$. Values in the first line in the *X* axis correspond to α_A and γ_B . Values in the second line of the X axis correspond to α_B and γ_A .

The third set of data, which resulted in **Figure 4A**, varied the values of α and γ for A, in order to change the values of ω for the population of type A individuals, such that ω_A divided by ω_B varies from 0.25 to 4, resulting in the plotted log values seen on the *x*-axis of **Figure 4A**. The 13 subsets of data generated each point in the *x*-axis scale of this figure.

The fourth set of data resulted in **Figure 4B**. The α and γ values for A and B were set between 0.01 and 0.09, as seen on the label of the *x*-axis of **Figure 4B** in all five subsets of data.

Plot and Analysis

The plots were generated using the data sets previously described. **Figures 2A–D**, were generated from the data produced by the first and second simulations, respectively, both based on the first simulation framework. **Figures 4A,B** depict the data generated from the latter data sets, produced by the simulation configured according to the second simulation framework. All plots were generated using the *matplotlib* (Hunter, 2007) package for the Python programming language.

Simulation Parameters

Parameters for each simulation and the scripts used to generate them – those who resulted in **Figures 2, 4** – are available at the public repository of this simulation platform. Below there is a brief description of each parameter from the main simulation program.

1. Samples – number of full simulations to be run with the currently loaded graph. The used value was 5,000 for all simulations.
2. Cycles – Simulation cycle limit. The simulation is considered undefined if it ends without fixation of either type A or B.
3. α values for types A and B.

4. β values for types A and B. If set to −1, this node type will not transition into "producing" state. The values are set to −1 in the simulations based on the first framework.
5. γ values for types A and B. If set to −1, this node type will not reuse abandoned extended phenotypes. This is the case for type B individuals on the sets of simulations based on the first framework.
6. Percentage of nodes at each state at the beginning of the simulation. For the simulations based on the framework described by **Figure 1**, only "with" or "without" extended phenotype states are available. This is achieved by setting percentages for "producing" and "using other" to zero.
7. Extended phenotype time. After generation, an extended phenotype will last a number of cycles before it expires. This time counter continues even after the extended phenotype is reused. If it has, for example, 10 cycles remaining when its original occupier dies, it will still have 10 cycles left whether it is reused or not. All simulations in this work had this parameter set to 30 cycles.
8. State time. For the simulations based on the framework described by **Figure 3**, states "searching" and "producing" last for a certain number of cycles before transitioning. The other two states, "using" and "using shared," depend on the extended phenotype time. All simulations based on the second framework had it set to 30 cycles.
9. Extended phenotype birth generation chance. This parameter is relevant for simulations based on the framework described by **Figure 1**. It defines the chance of a new node having an extended phenotype attached to it, and was set to 50% on those simulations. On simulations based on the framework described by **Figure 3**, it is set to zero, since in these simulations, the extended phenotypes are generated by nodes transitioning from the "producing" state, instead of at birth.

See **Supplementary Material** for a more detailed description of the values passed to each parameter at each simulation set.

RESULTS

Simulation Framework

An established approach for modeling the evolution of populations is the Moran Process (Moran, 1958). It is a simple stochastic model used to describe finite populations and can be used to simulate events, such as mutation and genetic drift by describing the probabilistic dynamics in a population containing two alleles, one of which can ultimately dominate the population. More recently, random scale-free graphs have been used to adapt the Moran Process to a more friendly simulation framework (Lieberman et al., 2005). These graphs share many characteristics of naturally occurring populations, such as in natural and artificial networks of relationships (Barabási and Albert, 1999). Therefore, it is suitable for modeling a generic population providing the conditions to test the EFH.

Thus, a population of individuals was modeled under the Barabási-Albert network model. This model generates a random graph that follows a power-law distribution of node degree, favoring the formation of clusters of highly connected nodes. The network grows according to preferential attachment, where new edges are more likely to be linked to nodes with higher degrees. In the original Moran Process adapted by Lieberman et al. (2005), all nodes begin with the same status. An individual of a different status is introduced into this population and by neutral drift or selection all other individuals can become bearers of the second status. This is achieved by a death-birth process, where an individual is randomly chosen to die, and in its place, a new individual is born. This individual is chosen based on a probability matrix calculated according to the neighbors of the dead node, weighted by their relative fitness, which translates into a numerical value representing its ability to reproduce. Regular nodes have relative fitness $r = 1$, and the "mutant" individuals have a relative fitness $r = 1 + \alpha$, where α is the bonus/penalty provided by the mutation.

Here, a similar model was used to evaluate the effect of the shared use of extended phenotypes in the fitness of a population (see **Figure 1A** for a schematic view of our simulation framework). We started by generating a network with 500 nodes (individuals) with a parameter $m = 4$, which is the minimum number of edges for any given node. One modification of the Moran process implemented in the present model is that nodes are classified either as a type A (250 nodes) or B (250 nodes) since the start of the simulation. Our framework was designed to compare two populations composed of either type A or type B individuals. Here, the death-birth process was adapted for the extended fitness context by taking into consideration the production and use of extended phenotypes. For each set of parameters, we run 5 million simulations (first, 1,000 random Barabási-Albert networks were designed and then for each one of them 5,000 simulations were run). A pseudocode for this first algorithm is presented as **Figure 1B** and detailed in Methods (section Main simulation).

Extended Phenotypes Increase the Fitness of Populations

In our first experiment, type A individuals were modeled as individuals who can produce and use their own extended phenotypes, and reuse extended phenotypes left behind by dead conspecifics. Type B individuals do not produce or use extended phenotypes. The initial setup for all executed simulations comprised a start ratio of 1:1 for type A and B individuals and a renewal rate of 4%, where at each generation 4% of all nodes are selected to die, and new nodes are placed in their locations in the graph according to the probability matrix explained in the previous section. The α value represents the bonus, the adaptive advantage of the extended phenotype, and was set between 0.0 and 0.15 (0.01 step) for each batch of simulations.

As illustrated in **Figure 1**, individuals of type A start with a 50% probability of having an extended phenotype already attached. Only individuals with attached extended phenotypes are given the bonus value in their relative fitness. Newborn individuals of type A have also a 50% chance of generating new extended phenotypes attached to themselves. When type A individuals leave behind an extended phenotype after death,

this can be occupied by one of their type A neighbors chosen at random with equal chance, as long as it is not already occupying an extended phenotype. If there is no neighbor of type A or all of them already have their own extended phenotype, a random individual of type A with an unattached extended phenotype is chosen anywhere in the graph, in case such an individual exists. Otherwise, the extended phenotype vanishes.

Figure 2A shows the results for all 5 million simulations for each bonus value (see **Supplementary Material** for details). With $\alpha = 0$, both populations reach fixation at the same rate, as expected, with a higher number of simulations undefined. A simulation is classified as undefined when no fixation of either node type is achieved. As α increases, a higher number of fixations of type A occurs until almost the totality of experiments ends with the fixation of type A individuals. A plateau, close to the upper limit of 5 million simulations, is reached around $\alpha = 0.08$. The number of undefined simulations also decreases, and it is also possible to observe that even in those simulations, there is a larger number of type A individuals as the bonus increases (**Figure 2C**). For example, at $\alpha = 0.04$, 75% of all undefined simulations had a higher proportion of type A individuals.

The Reuse of Extended Phenotypes Increases the Fitness of Populations

While the data in **Figure 2A** show that production and use of extended phenotypes increase the fitness of populations (See also **Supplementary Figure 2**), predictions of the EFH remained untested, namely that selection would favor groups where extended phenotypes are shared between conspecifics. Some of the simulation parameters were thus modified to perform such tests. Now, both types produce extended phenotypes but only type A individuals are able to reuse a given extended phenotype when it becomes available.

As before, individuals of types A and B start with a 50% probability of having an extended phenotype already attached, and newborn individuals of both types also have a chance of 50% of generating new extended phenotypes attached to themselves. The major difference between individuals of type A and B happens at death: type A individuals can leave behind an existing extended phenotype, which can be preferentially occupied by one of their type A neighbors as described in the previous simulation. On the other hand, the death of type B individuals causes the vanishing of the corresponding associated extended phenotypes and no reuse ever happens in this case. To eliminate the saturation effect, the extended phenotype half-life is the same for both individual types.

Figure 2B shows the results with this second proposed simulation. The number of simulations ending with the fixation of A still grows with rising α values, but at a slower pace, given that now type B individuals also produce and benefit from extended phenotypes. However, the observed advantage of type A individuals is still dramatic, even when both types generate extended phenotypes with the same bonus values. The major difference between the two simulation sets seems to be the number of undefined simulations, which is slightly higher in the second set of simulations (**Figure 2B**), where both populations produce extended phenotypes but only type A individuals are able to share them. For example, with $\alpha = 0.05$ only 10% of all simulations in **Figure 2A** are classified as undefined while the same number is 25% in **Figure 2B**. Like in **Figure 2C**, the proportion of type A individuals in the undefined simulations shows a positive association with α (**Figure 2D**). The occupancy rate of type A individuals with extended phenotypes also increases, since now the ones abandoned by dead individuals can be occupied by them. This effect can be seen in more detail in **Supplementary Figure 2**.

The Bonus Gained for Sharing Extended Phenotypes Has a Higher Impact in the Fitness of the Population

The previous simulations only considered a fitness bonus (α) for individuals that occupy an extended phenotype. This first, simple simulation can be enhanced to include parameters that reflect broader effects of extended phenotypes in both individual and group selection: (i) the benefit of using an extended phenotype built by yourself (α); (ii) the cost of building an extended phenotype (β), and (iii) the benefit of using an extended phenotype built by another individual (γ). One could think of a cost for searching for an extended phenotype previously built by someone else and now available, but there is no evidence that such behavior exists, and these encounters seem fortuitous. Based on the above, it is reasonable to think that the shared use of extended phenotypes will be favored when:

$$\frac{\gamma}{\alpha-\beta} > 1 \qquad [1]$$

However, when comparing two populations (A and B in our simulations), a more appropriate equation is:

$$\omega_i = (\alpha_i - \beta_i) + \gamma_i \qquad [2]$$

where ω_i is the absolute fitness of population A or B. Although selection parameters in equation [2] are described from the perspective of the individual, they are considered here at the population level. They represent average effects across the whole population. In that sense, after defining ω_i, one could estimate the abundance of a given phenotype using an equation like 3:

$$n(g+1) = \omega_i n(g) \qquad [3]$$

where $n(g)$ is the abundance of the phenotype in generation g. For the sake of simplicity, we will focus on equation 2 for the remaining simulations.

A new simulation (schematically viewed in **Figure 3A**) was modeled to test equation [2]. The four behaviors previously described were translated into four states: "searching" (searching for an extended phenotype), "producing" (producing an extended phenotype), "using own" (using your own extended phenotype), and "using other" (using an existing extended phenotype built by someone else). Each of these states has a different associated

relative fitness value at each simulation step. The "searching" state describes the default behavior. The individual is neither using nor producing an extended phenotype and if by chance, it encounters an unused one, it will occupy it. This state is the baseline behavior, with relative fitness $r = 1$. If the individual stays in the "searching" state for a specific amount of time (a specific number of cycles in the simulation – see Methods), it will transition into a "producing" state, meaning it searched for some time for an unoccupied extended phenotype and now started producing its own. Its relative fitness will be negatively affected by a β modifier since that individual will be spending time and resources building its own extended phenotype. After a certain amount of time, a producing node will transition into a "using own" state. It receives an α bonus for occupying an extended phenotype, as in previous simulations. After its extended phenotype time expires, the individual returns to a "searching" state. The "using other" state is set to individuals who, similarly to the previous simulations, are using extended phenotypes abandoned by dead individuals. The individuals in the "using other" state will receive a γ bonus. A pseudocode for this second algorithm is presented as **Figure 3B** and detailed in Methods (section Main simulation).

We first had to evaluate whether simulation data fit equation (2). By varying both ω_A and ω_B in different simulations (by changing the corresponding α bonus for each population while keeping β and γ fixed), we observed that there is indeed a strong positive association between ω and the winning population (the population with higher fixation rate), as can be seen in **Figure 4A**. Although there is a strong association between the value of ω and the winning population, there are also winning populations when ω for both populations have the same value ($\omega_A / \omega_B = 1$). This suggests that other parameters may have an impact on the simulation output. Thus, we decided to test the effect of each variable in the outcome of the simulations by exploring different values for each variable but always keeping $\omega_A = \omega_B$. This allowed us to evaluate the impact of each individual parameter, especially α and γ. **Table 1** shows all parameter values for each set of simulations. Our data show that β does not seem to affect the outcome of the simulations in terms of proportions of A and B (**Supplementary Figure 1**). This is probably due to the fact that the value of β is the same for both populations. On the other hand, the values of α and γ are critical in defining which population dominates the simulation. In all simulated scenarios, the population with a higher γ wins, as shown in **Figure 4B**, suggesting that the fitness gained for using an existing extended phenotype has a more significant impact than the fitness for using your own extended phenotype.

TABLE 1 | Values of ω for different values of α, β, and γ.

α_A, γB; α_B, γA	β = 0.01	β = 0.03	β = 0.05	β = 0.07	β = 0.09
0.01; 0.08	0.08	0.06	0.04	0.02	0.00
0.02; 0.07	0.08	0.06	0.04	0.02	0.00
0.03; 0.06	0.08	0.06	0.04	0.02	0.00
0.04; 0.05	0.08	0.06	0.04	0.02	0.00
0.05; 0.04	0.08	0.06	0.04	0.02	0.00
0.06; 0.03	0.08	0.06	0.04	0.02	0.00
0.07; 0.02	0.08	0.06	0.04	0.02	0.00
0.08; 0.01	0.08	0.06	0.04	0.02	0.00

DISCUSSION

Extended phenotypes have received significant interest since the original concept emerged in the early 80's (Dawkins, 1982), especially their indirect effects in other individuals or environments (Dawkins, 2004; Bailey, 2012; de Souza, 2013; Blamires et al., 2018; Fisher et al., 2019). Research in the field has been limited by the paucity of empirical models in which extended phenotypes can be manipulated and different evolutionary models be compared. We have, therefore, generated a computer simulation platform to evaluate the effects of the production and shared use of extended phenotypes on the fitness of simulated populations. We were particularly interested in testing the EFH as proposed by de Souza (2013).

The platform is flexible and can be easily adapted to study different real biosystems. For example, population interaction is structured with graphs, whose topology can be reconfigured to accommodate different ecological networks. Also, evolutionary dynamics can be manipulated by changing the probabilities of encounter, interaction, production and reuse of extended phenotypes, and the bonus/penalty associated with each behavior. This flexible architecture can thus be used to study, formulate, and test hypotheses in diverse areas, from plant-soil-microbial communities (Terhorst and Zee, 2016) to cancer evolution (Ewald and Swain Ewald, 2013). In fact, the extended phenotype hypothesis has been linked to a myriad of phenomena and has recently sparked interest (Hunter, 2018), partly due to novel computer simulations and data processing techniques. In this way, we believe that our work, more than testing aspects of the EFH, expands the toolbox to unveil evolutionary dynamics.

Nevertheless, there are several issues regarding extended phenotypes that could be explored using our simulation platform. Extended phenotype plasticity and its effect on the fitness of individuals and populations (Blamires, 2010; Bailey, 2012; Katz et al., 2017; Blamires et al., 2018) is an example that could be explored in our computational framework. Furthermore, the interplay between extended phenotype plasticity and other features, like for example dietary conditions, as observed by Blamires et al. (2018) and Katz et al. (2017) could likewise be studied in the computational setup presented here.

We show that the shared use of extended phenotypes has a significant contribution to the absolute fitness of a given population. This gives support to the EFH. One interesting aspect of the EFH is the fact that it does not advocate mutually exclusive fundamental evolutionary processes. As discussed by de Souza (2013), the effect of EFH at the group level is a natural consequence of the shared use of extended phenotypes by conspecifics. Furthermore, this shared use of extended phenotypes does not involve cooperation since the two parties likely never met, as discussed by Fisher et al. (2019).

The mathematical treatment provided here, although simple, allowed us to evaluate quantitatively the influence of different parameters in the fitness of the respective populations. In all scenarios tested, the shared use of extended phenotypes (quantified by the parameter γ) had a stronger influence on the fitness of the respective populations.

It is important to emphasize the assumptions and limitations of the strategy used in this report. There are, of course, intrinsic limitations derived from the simulated nature of the data. The different types of extended phenotypes (ranging from different physical structures to behaviors) bring also some challenges for an approach based on computational simulations. For example, the type of network used here (the Barabasi-Albert graph) may be more appropriate for some types of extended phenotype (like a web or a nest), while a regular network (where all nodes have the same degree) may be more appropriate for the study of the effect of a biofilm on the fitness of a bacterial population. Furthermore, few assumptions made in our simulations have the potential to affect our conclusions. First, no cost for searching for an existing extended phenotype was set in our simulations. This is a reasonable assumption since, to our knowledge, no such behavior has been described so far, and it is likely that the encounters are fortuitous. Furthermore, we have not taken into consideration the emergence of cheaters in our system (i.e., genetic variants that stop producing their own extended phenotypes and only use extended phenotypes of other individuals), which could also affect the evolutionary dynamics of the corresponding population. de Souza (2013) has discussed this issue but a formal evaluation through computer simulations needs to be done. Another interesting possibility, not explored here, is the modification of an existing extended phenotype by the individual who occupied it. These issues should be explored in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/guilherme-araujo/gsop-dist.

AUTHOR CONTRIBUTIONS

SS conceived the presented idea and wrote the first version of the manuscript. GA wrote all the scripts of the simulation platform and carried out the simulations. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grant 305233/2015-7 from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and by Center of Excellence – International Cooperation Initiative Grant, West China Hospital, Sichuan University (serial number 139200032), both to SS.

ACKNOWLEDGMENTS

We are indebted to Beatriz Stransky and César Renno Costa for a critical review of the manuscript. We thank NPAD/UFRN for computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617915/full#supplementary-material

REFERENCES

Bailey, N. W. (2012). Evolutionary models of extended phenotypes. *Trends Ecol. Evol.* 27, 561–569. doi: 10.1016/j.tree.2012.05.011

Barabási, A. -L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512. doi: 10.1126/science.286.5439.509

Blamires, S. J. (2010). Plasticity in extended phenotypes: orb web architectural responses to variations in prey parameters. *J. Exp. Biol.* 213, 3207–3212. doi: 10.1242/jeb.045583

Blamires, S. J., Blackledge, T. A., and Tso, I. M. (2017). Physicochemical property variation in spider silk: ecology, evolution, and synthetic production. *Annu. Rev. Entomol.* 62, 443–460. doi: 10.1146/annurev-ento-031616-035615

Blamires, S. J., Martens, P. J., and Kasumovic, M. M. (2018). Fitness consequences of plasticity in an extended phenotype. *J. Exp. Biol.* 221:jeb167288. doi: 10.1242/jeb.167288

Blamires, S. J., Tseng, Y. H., Wu, C. L., Toft, S., Raubenheimer, D., and Tso, I. M. (2016). Spider web and silk performance landscapes across nutrient space. *Sci. Rep.* 6:26383. doi: 10.1038/srep26383

Brockmann, H. J., Grafen, A., and Dawkins, R. (1979). Evolutionary stable nesting strategy in a digger wasp. *J. Theor. Biol.* 77, 473–496. doi: 10.1016/0022-5193(79)90021-3

Dawkins, R. (1982). *The extended phenotype: The long reach of the gene*. Oxford: Oxford University Press.

Dawkins, R. (2004). Extended phenotypes - but not too extended. A reply to Laland, Turner and Jablonka. *Biol. Philos.* 19, 377–396. doi: 10.1023/B:BIPH.0000036180.14904.96

de Souza, S. J. (2013). "Extended fitness" hypothesis: a link between individual and group selection. *Genet. Mol. Res.* 12, 4625–4629. doi: 10.4238/2013.October.17.5

Dixson, B. J. W. (2019). Sexual selection and extended phenotypes in humans. *Adapt. Hum. Behav. Physiol.* 5, 103–107. doi: 10.1007/s40750-018-0106-3

Ewald, P. W., and Swain Ewald, H. A. (2013). Toward a general evolutionary theory of oncogenesis. *Evol. Appl.* 6, 70–81. doi: 10.1111/eva.12023

Fisher, D. N., Haines, J. A., Boutin, S., Dantzer, B., Lane, J. E., Coltman, D. W., et al. (2019). Indirect effects on fitness between individuals that have never met via an extended phenotype. *Ecol. Lett.* 22, 697–706. doi: 10.1111/ele.13230

Gurnell, A. M. (1998). The hydrogeomorphological e•ects of beaver dam-building activity. *Prog. Phys. Geo. Earth Environ.* 22, 167–189. doi: 10.1177/030913339802200202

Hagberg, A., Schult, A., and Swart, P. (2008). "Exploring network structure, dynamics, and function using NetworkX". in *Proceedings of the 7th Python in Science Conference*; August 19-24, 2008; 11–15.

Hoover, K., Grove, M., Gardner, M., Hughes, D. P., McNeil, J., and Slavicek, J. (2011). A gene for an extended phenotype. *Science* 333:1401. doi: 10.1126/science.1209199

Hunter, P. (2009). Extended phenotype redux. How far can the reach of genes extend in manipulating the environment of an organism? *EMBO Rep.* 10, 212–215. doi: 10.1038/embor.2009.18

Hunter, P. (2018). The revival of the extended phenotype. *EMBO Rep.* 19:e46477. doi: 10.15252/embr.201846477

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi: 10.1109/MCSE.2007.55

Katz, N., Shavit, R., Pruitt, J. N., and Scharf, I. (2017). Group dynamics and relocation decisions of a trap-building predator are differentially affected by biotic and abiotic factors. *Curr. Zool.* 63, 647–655. doi: 10.1093/cz/zow120

Laland, K. (2004). Extending the extended phenotypes. *Biol. Philos.* 19, 313–325. doi: 10.1023/B:BIPH.0000036113.38737.d8

Lieberman, E., Hauert, C., and Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature* 433, 312–316. doi: 10.1038/nature03204

Martin, T. E., Boyce, A. J., Fierro-Calderón, K., Mitchell, A. E., Armstad, C. E., Mouton, J. C., et al. (2017). Enclosed nests may provide greater thermal than nest predation benefits compared with open nests across latitudes. *Funct. Ecol.* 31, 1231–1240. doi: 10.1111/1365-2435.12819

Moran, P. A. P. (1958). Random processes in genetics. *Proc. Camb. Phil. Soc.* 54, 60–71.

Olson, R. C., Vleck, C. M., and Vleck, D. (2006). Periodic cooling of bird eggs reduces embryonic growth efficiency. *Physiol. Biochem. Zool.* 79, 927–936. doi: 10.1086/506003

Rosell, F., Bozsér, O., Collen, P., and Parker, H. (2005). Ecological impact of beavers Castor fiber and Castor canadensis and their ability to modify ecosystems. *Mammal Rev.* 35, 248–276. doi: 10.1111/j.1365-2907.2005.00067.x

Schaedelin, F. C., and Taborsky, M. (2009). Extended phenotypes as signals. *Biol. Rev.* 84, 293–313. doi: 10.1111/j.1469-185X.2008.00075.x

Schuck-Paim, C., and Alonso, W. J. (2001). Deciding where to settle: conspecific attraction and web-site selection in the orb-web spider *Nephilengys cruentata*. *Anim. Behav.* 62, 1007–1012. doi: 10.1006/anbe.2001.1841

Steger, A., and Wormwald, N. (1999). Generating random regular graphs quickly. *Probab. Comput.* 8, 377–396. doi: 10.1017/S0963548399003867

Terhorst, C. P., and Zee, P. C. (2016). Eco-evolutionary dynamics in plant-soil feedbacks. *Funct. Ecol.* 30, 1062–1072. doi: 10.1111/1365-2435.12671

Wang, J., Ross, K. G., and Keller, L. (2008). Genome-wide expression patterns and the genetic architecture of a fundamental social trait. *PLoS Genet.* 4:e1000127. doi: 10.1371/journal.pgen.1000127

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 February 2021
doi: 10.3389/fgene.2020.607798

DriverSubNet: A Novel Algorithm for Identifying Cancer Driver Genes by Subnetwork Enrichment Analysis

Di Zhang[1] *and Yannan Bin*[2]*

[1] College of Information Engineering, Shaoguan University, Shaoguan, China, [2] Institutes of Physical Science and Information Technology, Anhui University, Hefei, China

Identification of driver genes from mass non-functional passenger genes in cancers is still a critical challenge. Here, an effective and no parameter algorithm, named DriverSubNet, is presented for detecting driver genes by effectively mining the mutation and gene expression information based on subnetwork enrichment analysis. Compared with the existing classic methods, DriverSubNet can rank driver genes and filter out passenger genes more efficiently in terms of precision, recall, and F1 score, as indicated by the analysis of four cancer datasets. The method recovered about 50% more known cancer driver genes in the top 100 detected genes than those found in other algorithms. Intriguingly, DriverSubNet was able to find these unknown cancer driver genes which could act as potential therapeutic targets and useful prognostic biomarkers for cancer patients. Therefore, DriverSubNet may act as a useful tool for the identification of driver genes by subnetwork enrichment analysis.

Keywords: cancer, driver gene, multi-omics data, neighbor network, TCGA

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong, Hong Kong

Reviewed by:
Hugo Tovar,
Instituto Nacional de Medicina Genómica (INMEGEN), Mexico
Minghui Li,
Soochow University, China

***Correspondence:**
Yannan Bin
ynbin@ahu.edu.cn

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

Received: *18 September 2020*
Accepted: *30 December 2020*
Published: *19 February 2021*

Citation:
Zhang D and Bin Y (2021) DriverSubNet: A Novel Algorithm for Identifying Cancer Driver Genes by Subnetwork Enrichment Analysis. Front. Genet. 11:607798. doi: 10.3389/fgene.2020.607798

INTRODUCTION

Cancer is a globally prevalent threat to the overall survival of patients, and is driven by a few important cancer genes, viz., driver genes (Dinstag and Shamir, 2019). Oncogenic mutations on driver genes contribute to abnormal cell proliferation and tumor development. Most other genes undergoing mutations due to genomic instability caused by driver genes, termed passenger genes, are neutral, and do not lead to any cancerous growth (Di Zhang et al., 2016; Yue et al., 2018). Thus, increasing efforts are being made to recognize these driver genes for developing a better elucidation regarding cancer initiation and progression. There are some systemic cancer genomics research projects, such as The Cancer Genome Atlas (TCGA), which is a public free platform and provides data on 33 cancer types for cancer research.

Computational patterns have been developed to screen driver genes by distinguishing them from passenger genes through mutation frequency. For instance, MuSiC adopts a statistical approach to detect driver genes with significantly high mutative rates (Dees et al., 2012). DeepDriver employs deep learning to identify driver genes by estimating the functional impact of mutations (Luo et al., 2019). However, these methods are based on mutation frequency, and do not uncover driver genes which carry few variants. Recently, researchers realize that genes cooperate with each other in cancer progression through biological pathways, and detection of driver genes by pathway- or network-based pipelines is emerging with a high speed (Hou et al., 2018). These studies revealed that functional networks could be available for identifying driver genes without consideration of mutation frequency. They concentrate on uncovering cancer associated core modules consisting

of gene-sets rather than a single gene critical to tumor progression. The lack of prioritization in this approach is a shortcoming from the considerations of clinical treatment, particularly when the predicted driver gene set contains more than one gene.

To solve this situation, many algorithms have been developed to rank the candidate genes (Hou and Ma, 2014; Dinstag and Shamir, 2019; Hristov et al., 2020). For instance, HotNet2 identifies rare mutations across pathways and protein-protein interaction (PPI) networks using the heat-diffusion theory (Leiserson et al., 2015). DriverNet also consolidates PPI and gene expression data to uncover driver genes (Bashashati et al., 2012). DawnRank method adopts Google's PageRank algorithm and ranks an individual's mutated gene profile by means of measuring the effect of each mutated gene on the differentially expressed genes (DEGs) (Hou and Ma, 2014). MUFFINN algorithm evaluates the significance of mutations on neighboring genes in the specific network, demonstrating excellent predictive performance in a large number of patients (Cho et al., 2016). MaxMIF tries to find driver genes by evaluating the impact of single nucleotide variants on transcriptional networks (Hou et al., 2018). Nevertheless, the false positive rates of the current existing computational algorithms need to be further reduced.

Here, we have designed an effective algorithm, called DriverSubNet, which has the ability of prioritizing driver genes. In this approach, the driver genes were scored by combining their influence on DEGs in each neighbor subnetwork and their mutation frequency. These pipelines are based on enrichment of subnetworks, where each subnetwork may reflect the situation of dysregulated biological process in tumor. Thus, the extent to which a given driver gene explains multiple functional biological process deregulations serves as a proxy for the likelihood that this gene is indeed the driver. Our algorithm views that driver genes affect the deregulations of other genes in the functional biological processes. Besides, mutation recurrence makes a vital contribution on detecting high frequency mutated drivers. In fact, the true cancer drivers have good connectivity to these functional biological processes, and our algorithm aims to measure such connections directly via subnetwork enrichment and the impact of mutations.

MATERIALS AND METHODS

Data Collection

For four cancer types, including thyroid carcinoma (THCA), kidney renal clear cell carcinoma (KIRC), and breast cancer (BRCA) and Head-Neck Squamous Carcinoma (HNSC), somatic mutations, somatic copy number alterations (SCNAs), and RNA-seq expression data belong to the TCGA (Weinstein et al., 2013) platform, downloaded from the UCSC data portal (http://xena.ucsc.edu/) (Rosenbloom et al., 2015). Undirected interaction network information was collected from the Human Protein Reference Database (HPRD) release 9 (Keshava Prasad et al., 2009). HPRD is a comprehensive resource for studying the human proteome, and the proteins have been manually extracted from the literature by expert biologists. In the mutation matrix, where a row denotes a gene, and a column denotes a patient, if a gene exists the mutations (e.g., SCNAs, small insertions, and small deletions), which was marked as one, otherwise marked as zero. Gene expression profiles from control samples were also used for differential expression analysis. The details of the data can be seen in **Supplementary Table 1**. To evaluate the performance of our results, we obtained the set of all 723 entries from the Cancer Gene Census (CGC, Accessed on: 01/30/2020) (Tate et al., 2019). Functional gene sets were collected from literature (Ge et al., 2018; Malta et al., 2018; Sanchezvega et al., 2018) and the Atlas of Cancer Signaling Network website (https://acsn.curie.fr/ACSN2/ACSN2.html), which includes data for various pathways including ubiquitin pathway, DNA repair pathway, TGF-beta signaling, and oncogenic signaling pathway. Finally, we used the Functional Set (FG) with 3,681 functional genes to represent the functional biological processes.

Evaluation Criteria

The performance of algorithms for prioritizing candidate genes was widely adopted the following criteria: precision, recall, and the F1 score (Bashashati et al., 2012; Hou and Ma, 2014). MUFFINN, Dawnrank, and DriverNet were the state of art methods to be compared with other algorithms. We use MUFFINN algorithm based on NDmax and HumanNet. One hundred top-ranked candidate genes were selected to compare the state-of-art methods (Hui et al., 2019). The following evaluation criteria were used to assess the ability of a method to identify real driver genes from the top-ranked candidates.

$$Precision = \frac{(\#\ Genes\ in\ CGC) \cap (\#Genes\ found\ in\ our\ method)}{(\#Genes\ found\ in\ our\ method)}$$

$$Recall = \frac{(\#Genes\ in\ CGC) \cap (\#Genes\ found\ in\ our\ method)}{(\#Genes\ in\ CGC)}$$

$$F1\ sore = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Scoring Scheme of DriverSubNet

A schematic diagram of our DriverSubNet pipeline consists of four steps (**Figure 1**). Firstly, differential expression analysis was carried out statistical analysis by using the DEseq2 package in R (version 3.6). All genes with adjusted $p < 0.05$ were considered as DEGs. Secondly, DEGs and mutated genes were mapped to HPRD interaction network. For each mutated gene in HPRD network, mutated gene and its directly connected neighbor genes consist of the adjacent neighbor subnetwork, and the central gene is mutated gene in subnetwork. Thirdly, for each subnetwork, we want to evaluate whether the subnetwork have an impact on vital biological process. For DEGs in subnetwork, we measure whether these DEGs were enriched the FG. If these DEGs were significantly enriched FG, it represents that the subnetwork tends to play a crucial role in cancer progression. In our result,

Abbreviations: TCGA, The Cancer Genome Atlas; SCNAs, Somatic Copy Number Alterations; THCA, Thyroid Carcinoma; KIRC, Clear Cell Kidney Carcinoma; BRCA, Breast Cancer; HNSC, Head-Neck Squamous Carcinoma; CGC, Cancer Gene Census; FG, Functional Set; DEGs, Differentially Expressed Genes; PPI, Protein-Protein Interaction.

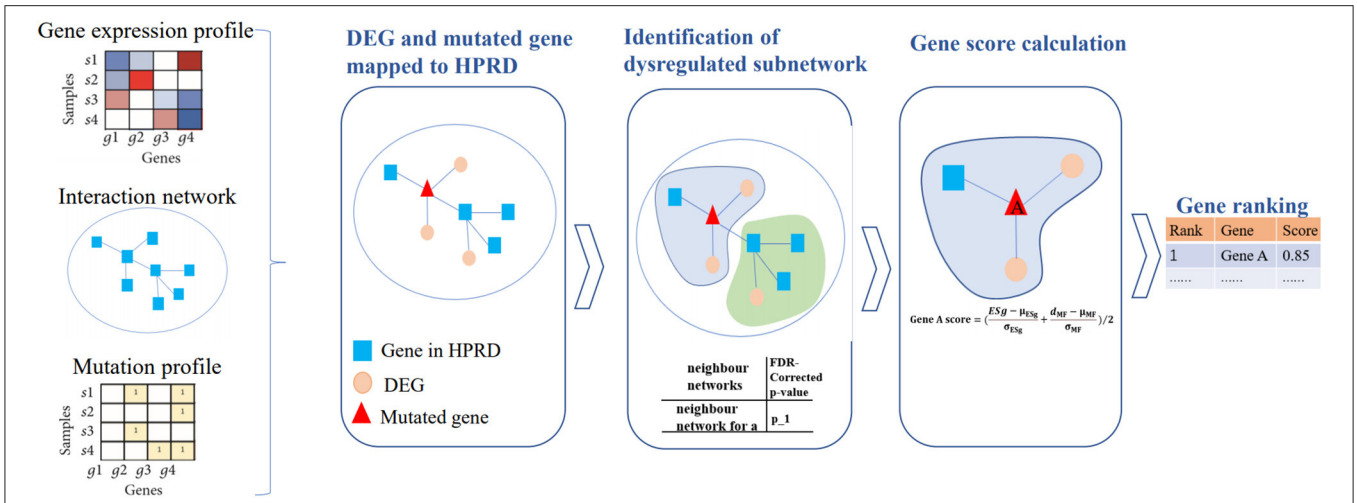


FIGURE 1 | The pipeline of DriverSubNet. Differential expression analysis was performed using the DESeq2 package. All genes with adjusted $p < 0.05$ were considered as DEGs. DEGs and mutated genes were mapped to the network. We extract a neighbor subnetwork from the protein interaction network for each mutated gene, where a gene and its neighbor genes together consisted of an adjacent neighbor subnetwork. We then evaluate whether the subnetwork is deregulated subnetwork by enrichment analysis. An enrichment score ESg was obtained by evaluating the functional significance of the central gene in each neighbor subnetwork. Each center gene with an enrichment score *ESg* was calculated. Finally, candidate genes were ranked according to their overall mutational influence scores and *ESg*.

enrichment *p*-value of DEGs was set as 5E-6 across four datasets and the recall value of recognizing known cancer genes in the top 100 genes achieved high. If the enrichment *p*-value of DEGs <5E-6 and the subnetwork consist of more than two genes, the subnetwork was regarded as a deregulated subnetwork. To assess the impact of mutated gene in the deregulated subnetwork, we calculated the mutated impact score *ESg*. We performed the enrichment analysis using the fisher.test function in R (version 3.6), and then transformed it using -log function. It was computed as follows:

$$ESg = -log\left(1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}\right)$$

where N represents the total genes in each subnetwork, n represents the number of DEGs in the subnetwork, M represents the overlap with DEGs and functional gene set in each subnetwork, and i represents the overlap with DEGs and functional gene set.

Then, in view of combing the effect of gene expression and gene mutations can improve the performance of algorithms (Hou and Ma, 2014), and mutation recurrence makes a vital contribution on detecting high frequency mutated drivers, we also considered mutation frequency in our approach to uncover the most functional drivers in a large number of patients. We evaluated the significance of mutated genes based on the mutation frequency. We calculate the number of mutations according to the mutation matrix, then we normalized the number of mutations, then the value is range 0–1. Finally, we computed a score for every candidate gene by averaging the normalized ESg gene score in the deregulated subnetwork and the normalized gene mutational scores. Candidate genes were ranked according to their overall scores. The score of candidate driver gene score was calculated as follows:

$$Score = \left(\frac{ESg - \mu_{ESg}}{\sigma_{ESg}} + \frac{d_{MF} - \mu_{MF}}{\sigma_{MF}}\right)/2$$

where μ_{ESg} is the expected mean of *ESg*, and σ_{ESg} is the standard deviation of *ESg*, d_{MF} is the number of patients with mutated genes, μ_{MF} is the expected mean of d_{MF}, and σ_{MF} is the standard deviation of d_{MF}.

Functional Enrichment Analysis

To understand the features detected in our results, we used the R package and found significant enrichment of these uncovered top 100 genes in terms of biological process. Briefly, biological process terms were annotated according to statistical significance. Enrichment was calculated through the hypergeometric test with $p < 0.05$, and following which top 100 most significant categories were selected.

Survival and Drug Analysis

We used the online tool for analyzing patient survival via its standard processing pipeline GEPIA (Zefang et al., 2017). The drug information for genes was obtained from the Drug Gene Interaction database (DGIdb) (Cotto et al., 2018). DGIdb is comprehensive catalog of druggable genes (i.e., genes with directed pharmacotherapy) and drug-gene interactions database, which integrates existing 30 sources (DrugBank, PharmGKB, Chembl, Drug Target Commons, TTD, and others) and collects 56,309 drug-gene interactions. Drug-gene interactions represents that genes or gene products are known or predicted to interact with drugs, and the gene might be targeted therapeutically. In our study, we use DGIdb to analyze whether these identified genes are clinically relevant genes.

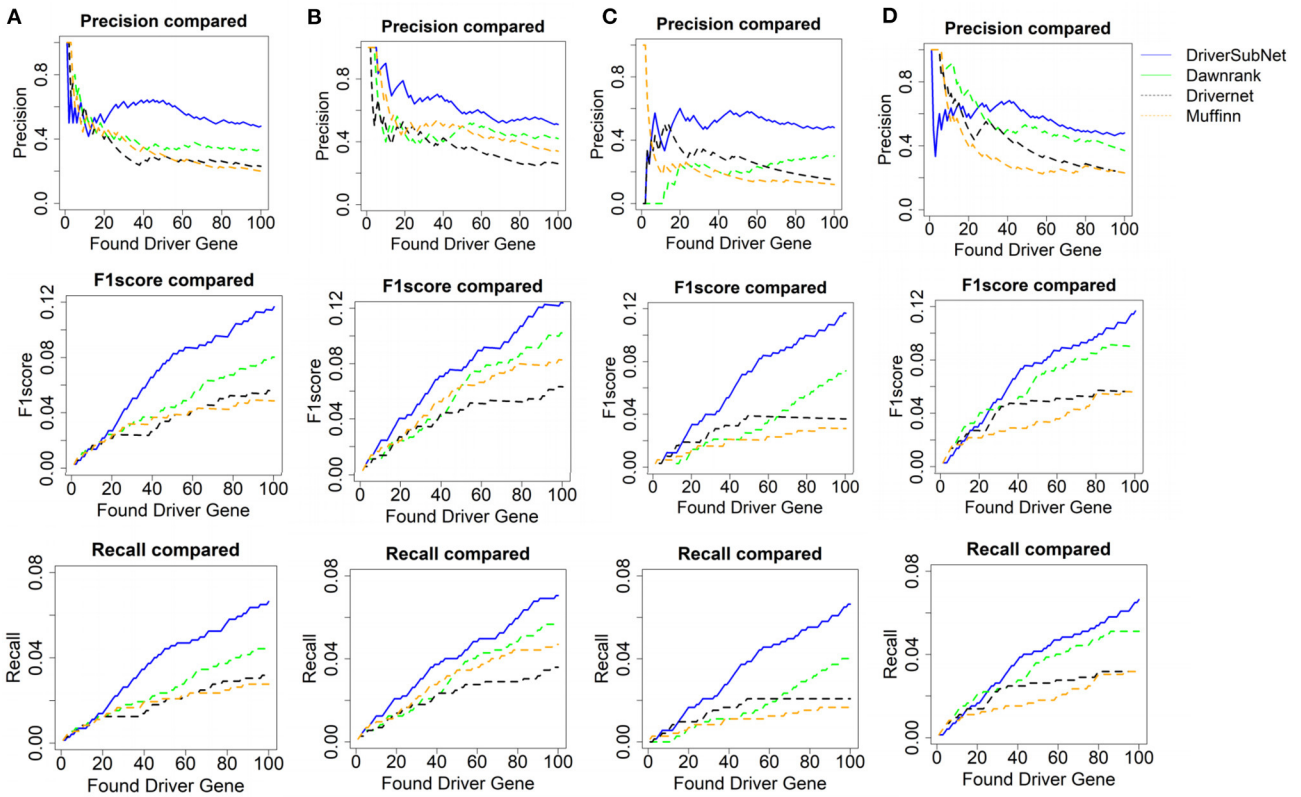


FIGURE 2 | Performance comparison with CGC in terms of precision, *F*1 score, and recall of Dawnrank and DriverNet methods on **(A)** BRCA, **(B)** HNSC, **(C)** KIRC, and **(D)** THCA datasets.

RESULTS

Performance Evaluation for Known Cancer-Related Genes

Here, we adopt a subnetwork analysis with PPI information. The core of algorithm is a local subnetwork model, which views that a driver gene can be detected by aggregating its involvement in functional biological process from a central gene and its direct neighbor DEGs. We applied DriverSubNet to four datasets from BRCA, THCA, KIRC, and HNSC, respectively, which the cancer type is randomly chose. Then, we evaluate the effectiveness of our method, MUFFINN, Dawnrank, and DriverNet algorithms.

The performances of DriverSubNet, MUFFINN, Dawnrank, and DriverNet methods were compared on the basis of precision, recall, and F1 scores for the top 100 genes. In general, DriverSubNet outperformed MUFFINN, Dawnrank, and DriverNet methods in all four cancer datasets with gold standard CGC dataset (**Figure 2**). Especially the most of candidate genes were overlapped with CGC in the top 100 driver genes using the DriverSubNet method across four datasets. It suggests that DriverSubNet is robust and has an excellent ability of identifying driver genes. Although the Dawnrank method performed better ability than other algorithms in ranking the top 12 genes in THCA, it had a poorer ability in KIRC. The reason for this phenomenon may be the different number of gene mutations and the variety of gene expression levels across the four cancer types. DriverSubNet is easier to evade the number of mutation noise and expression than other methods. For example, DriverSubNet was able to recover most of known cancer driver genes in the top 100 detected genes across four datasets, while the percentage of known cancer driver genes in the top 100 detected genes using Dawnrank and DriverNet is sensitive to cancer type. This may lead to Dawnrank have a good performance in THCA, while bad performance in KIRC. In KIRC, although some known drivers were found by these three methods, DriverSubNet uncovered significant famous driver genes, such as *EGFR*, which was ranked the 16th, and it were not detected by either Dawnrank or DriverNet or MUFFINN method as the top ranking drivers.

Novel Candidate Genes Predicted by DriverSubNet

To evaluate the performance of algorithm, precision, recall, and F1 score are widely used to analyze the top 100 genes. In our result, we identified some genes that were not known cancer driver genes. It is essential to explore whether these genes have a potential relationship with cancer. Previous study has suggested that high-ranking unknown cancer driver genes have a potential to be novel driver genes (Hou and Ma, 2014). In our study, we used the top 10 genes to detect some unknown cancer driver genes which have a potential to be novel driver genes.

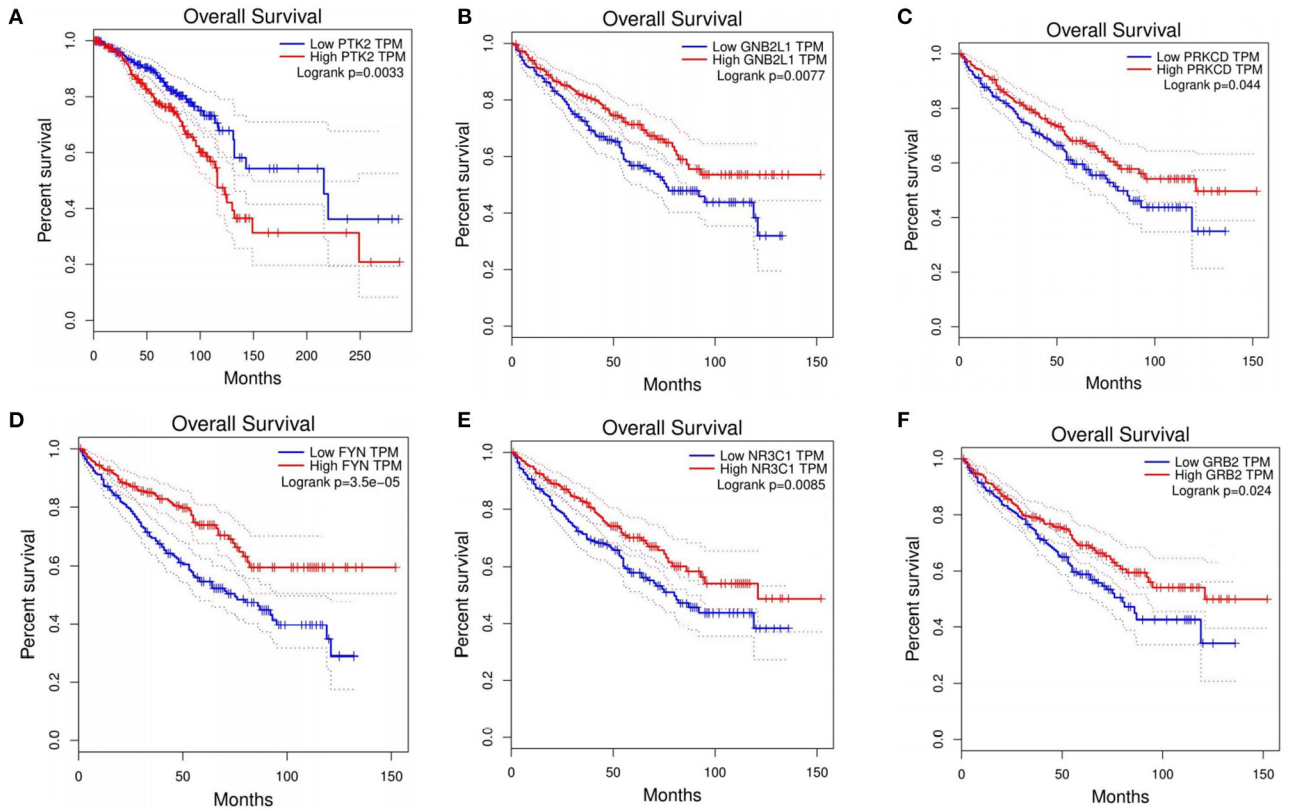


FIGURE 3 | Prognostic value of six genes in cancer patients. **(A)** *PTK2* in BRCA. **(B)** *GNB2L1* in KIRC. **(C)** *PRKCD* in KIRC. **(D)** *FYN* in KIRC. **(E)** *NR3C1* in KIRC. **(F)** *GRB2* in KIRC.

For the BRCA dataset, 48 genes overlapped with CGC for the top 100 candidate driver genes (**Supplementary Table 2**). Among the top 10 ranking genes in BRCA, *CREBBP*, *EP300*, *MYC*, *SRC*, and *TP53* overlapped with the cancer genes in CGC, whereas the other five genes, (*CDK1*, *GRB2*, *YWHAZ*, SHC1, and *PTK2*) did not include in CGC. These five genes were differentially expressed in BRCA. To investigate whether these five genes were involved in BRCA, we explored the correction between these five genes and overall survival in BRCA. Through Kaplan-Meier analysis using an online GEPIA, *PTK2* showed high expression was corrected with a shorter overall survival in BRCA patients (**Figure 3A**). *CDK1, GRB2,* and *PTK2* were the druggable genes in DGIdb. We concluded that *CDK1*, *GRB2,* and *PTK2* were more likely to be involved in pathogenesis of BRCA, simultaneously, which have a great potential to be therapeutic targets. Through analysis, *PTK2* can be applied to predict survival of BRCA patients.

For the HNSC dataset, 51 genes overlapped with the genes in CGC for the top 100 candidate driver genes (**Supplementary Table 2**). Among the top 10 ranking genes in HNSC, *CREBBP*, *CTNNB1*, *EGFR*, *EP300*, MAPK1, *SMAD2*, *SMAD3*, *SRC*, and *TP53* overlapped with the genes in CGC, whereas the other one *GRB2* did not. To investigate whether *GRB2* was involved in HNSC, we explored the correction between *GRB2* and overall survival in HNSC. Through Kaplan-Meier analysis, *GRB2* was not corrected with shorter overall survival in HNSC patients. *GRB2* was the druggable gene in DGIdb and more likely to be involved in the pathogenesis of HNSC.

For the KIRC dataset, 48 genes overlapped with CGC for the top 100 candidate driver genes (**Supplementary Table 2**). Among the top 10 ranking genes in KIRC, *CTNNB1*, *EP300*, *SRC*, and *TP53* were found in CGC. Other six genes (*PRKCA, PRKCD, GNB2L1, FYN, NR3C1*, and *GRB2*) did not present in CGC. To investigate whether these genes were involved in KIRC, we explored the correction between these six genes and overall survival in KIRC. Through Kaplan-Meier analysis, five out of the six genes (*PRKCD, GNB2L1, FYN, NR3C1,* and *GRB2*) showed high expression were corrected with shorter overall survival in KIRC patients (**Figures 3B–F**). It was concluded that these five genes had a great ability to participate in pathogenesis of KIRC, and were possible therapeutic targets. Besides, through the analysis, these five genes can be applied to predict the overall survival of KIRC patients.

For the THCA dataset, 48 genes overlapped with the genes in CGC for the top 100 candidate driver genes. The top 10 ranking genes in THCA were accessed in the **Supplementary Table 2**. Among these genes, *BRAF*, *CREBBP*, *EGFR*, *EP300*, *MAPK1*, *SMAD3*, *SRC*, and *TP53* overlapped with the genes in CGC. These eight genes were known to participate in cancer progression. The other two genes (*FYN* and *GRB2*) did not match with the CGC database. *GRB2* belongs to druggable genes according to DGIdb.

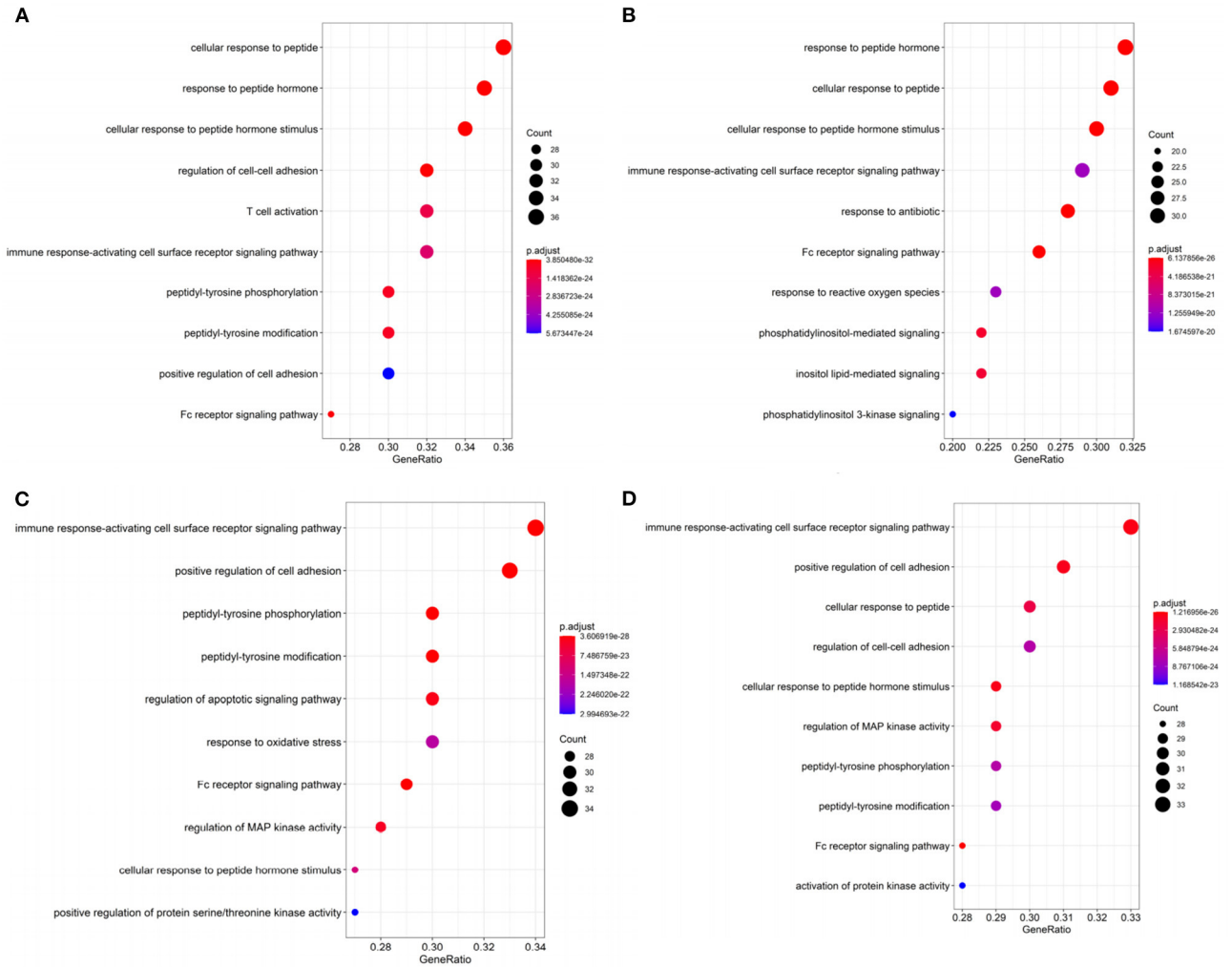


FIGURE 4 | The top 10 gene ontology (GO) terms enrichment of **(A)** BRCA, **(B)** HNSC, **(C)** KIRC, and **(D)** THCA by significant genes with $p < 0.05$ in DriverSubNet.

We concluded that *GRB2* had a great ability to participate in the pathogenesis of THCA, and was a possible therapeutic target.

Enrichment Analysis

KEGG and GO enrichment analysis displayed that the top 100 uncovered genes of cancers were significantly enriched in vital KEGG and GO terms, as shown in **Supplementary Figure 1**, **Figure 4,** respectively.

In BRCA, the most significantly enriched KEGG term was "Proteoglycans in cancer" (**Supplementary Figure 1**). Proteoglycans are implicated in regulating cellular growth and differentiation (Filmus et al., 2008). Other enriched terms (e.g., Viral carcinogenesis, ErbB signaling pathway, chronic myeloid leukemia, and prostate cancer) are also related to cancer. The top ranked significantly enriched GO term was peptide associated (**Figure 4A**). Peptide hormone can negatively regulate iron efflux and is crucial for modulating the growth of breast tumors (Blanchette-Farra et al., 2018). Other enriched terms (e.g., Fc receptor signaling pathway, adhesion) are also related to cancer. Fc receptor can be acted as an indicator for prognosis in many cancers, such as colorectal and lung cancer (Cadena Castaneda et al., 2020). The roles of Fc receptor signaling pathway in BRCA brings forward the need for further studies.

In HNSC, the significantly enriched KEGG term was cancer related, such as proteoglycans in cancer, viral carcinogenesis, and pancreatic cancer. In **Figure 4B**, "Response to reactive oxygen species" was the enrichment GO term, which can induce oxidative stress (Ma, 2013). Increased reactive oxygen species production involved in multiple cancers through various mechanisms, for example, they can express pro-tumorigenic signaling, and lead to tumor abnormal survival and proliferation, and avail to DNA damage and genetic instability (Moloney and Cotter, 2017). Oxidative stress can contribute to the maintenance of genomic instability during the progression phase of cancer (Hassani et al., 2019) remove. This suggests that oxidative stress has a clinical significance in cancer remove. Moreover, the cellular response to oxidative stress plays crucial roles in cellular adaptation to hypoxic stress

remove. Other terms including immune response-activating cell surface receptor signaling pathway, phosphatidylinositol-mediated signaling, Fc receptor signaling pathway, and so on. Moreover, Fc receptor plays a crucial role in NK cell maturation and tumor immunosurveillance (Cadena Castaneda et al., 2020). Immune system play a vital role in HNSC (Mirza et al., 2019). Thus, the top 100 genes in HNSC that we identified were significantly related to cancer.

In KIRC, KEGG pathway annotation indicated that the pathways most enriched in chemokine signaling pathway, neurotrophin signaling pathway, ErbB signaling pathway (**Supplementary Figure 1**). The top ranked GO term in KIRC was "immune response-activating cell surface (**Figure 4C)**. The top 100 genes identified in KIRC were significantly related to cancer. Other terms including regulation of apoptotic signaling pathway, and Fc receptor signaling pathway, regulation of MAP kinase activity, positive regulation of protein serine/threonine kinase activity were also recorded. Deregulation in apoptotic is a hallmark of cancer (Pistritto et al., 2016). Apoptosis alteration is responsible for tumor development and progression (Pistritto et al., 2016). Other terms, such as response to oxidative stress, cell-cell adhesion, and Fc-gamma receptor signaling pathway, were involved in cancer progression. Through above analysis, these top 100 genes identified in KIRC were related to cancer.

In THCA, KEGG pathway analysis revealed that the top 100 genes were linked with proteoglycans in cancer, chemokine signaling pathway, ErbB signaling pathway, and so on (**Supplementary Figure 1**). The most significantly enriched GO term was "immune response-activating cell surface receptor signaling pathway" (**Figure 4D**). This means that the top 100 genes in THCA make a contribution to modulate immune system in cancer. Other enriched terms, such as regulation of cell-cell adhesion and Fc receptor signaling pathway, regulation of MAP kinase activity are associated with cancer progression. Thus, the top 100 genes that we identified were significantly related to cancer.

Actionable Druggable Genes

DriverSubNet's rankings can guide scientists to decide on drug development and clinical treatment. The top 100 driver genes for BRCA, HNSC, KIRC, and THCA, respectively, were looked-up in DGIdb. Genes with target drug information were considered as druggable driver genes, and the others as undruggable driver genes. The results (**Figure 5**) indicated that most of the identified driver genes were druggable driver genes. In **Figure 5**, it was obvious that the proportions of druggable genes increased substantially when the number of genes were increased. Hence, DriverSubNet has the ability of uncovering potential therapeutic targets, tailored to the clinical treatment.

DISCUSSION

Many methods have been designed to screen driver genes by distinguishing them from passenger genes, but almost all of them have limited sensitivity and specificity. To solve this shortcoming, we constructed the DriverSubNet, which effectively mined the mutation and expression information in PPI network. The algorithm takes into effect of central gene on neighboring DEGs, and mutated frequency. Comparing DriverSubNet with Dawnrank and DriverNet on the four cancer datasets, our results reveal that DriverSubNet achieves better performance than Dawnrank and DriverNet methods in the top 100 gene set. DriverSubNet was able to find well-known genes, such as *EGFR*. In addition, DriverSubNet could also found functional driver genes which have a low mutation rate.

Indeed, to explore the non-CGC candidate genes in the top 100 candidate driver genes by DriverSubNet, we performed literature search, and found that most of non-CGC candidate genes with experimental evidence revealing their relation with cancer. Among the top 10 driver genes identified in BRCA, HNSC, KIRC, and THCA (**Supplementary Table 2**), overall, seven unique genes (*CDK1, GRB2, YWHAG, SHC1 and PTK2, FYN,* and *TRAF2*) were detected as non-CGC genes. *YWHAG* is critical for maintaining several canonical pathways. miRNAs can directly target *YWHAG*, which has been reported as a tumor suppressor, and participates in the progression in breast cancer, glioblastoma, and lung cancer (Yoo et al., 2016; Wang et al., 2017a,b). *GRB2* encodes protein can activate cell surface receptors in signaling transduction (Giubellino et al., 2008). *GRB2* signaling is associated with cell motility, angiogenesis, and vasculogenesis (Giubellino et al., 2008). These functions make *GRB2* a potential target biomarker to hinder tumor metastasis and local invasion (Giubellino et al., 2008). *SHC1* encoding protein is recruited to tyrosine kinases, which is essential for breast cancer initiation, progression, and metastasis (Ahn et al., 2017). It has implicated that *SHC1* mediate several key signaling pathways in breast cancer (Wright et al., 2019). *PTK2* is a highly phosphorylated kinases in breast cancer (Mertins et al., 2016). Substantial evidence has shown that activated PTK2 expression level links to tumor progression (Fan et al., 2019). In our result, *PTK2* is highly expressed (Fold Change = 1.39) in BRCA samples, which suggests that high PTK2 expression leads to BRCA growth and metastasis. *FYN* is differentially expressed in multiple cancers, and has a correction with cancer progression by controlling cellular motility, cell growth, and death (Elias and Ditzel, 2015). *FYN* is a promising candidate therapeutic marker and may be applied to Fyn-targeted therapy (Elias and Ditzel, 2015). *TRAF2* is reported as an NF-κB-activating oncogene (Shen et al., 2015). *CDK1* can regulate cell cycle progression by executing the G2/M phase transition (Asghar et al., 2015). *CDK1* is the central regulator of cell proliferation and a promising therapeutic target for BRCA (Galindomoreno et al., 2017). Knockout of *CDK1* in mouse experiments revealed that *CDK1* contributed to cellular proliferation (Santamaría et al., 2007). *DLG1* expression associates with the progress of cervical disease (Cavatorta et al., 2017). Through the above analysis, we may find that cancer is heterogeneity that the same driver gene has differential function across cancers, for example, *GRB2* is identified driver gene in four dataset, and *GRB2* expression has a significant survival rate in KIRC, while not in other three cancer types. The findings from this analysis indicate that six genes (**Figure 3**) which are not in CGC or the independent predictor of poor survival or therapeutic target genes, may contribute to cancer through other mechanisms. Namely, DriverSubNet was

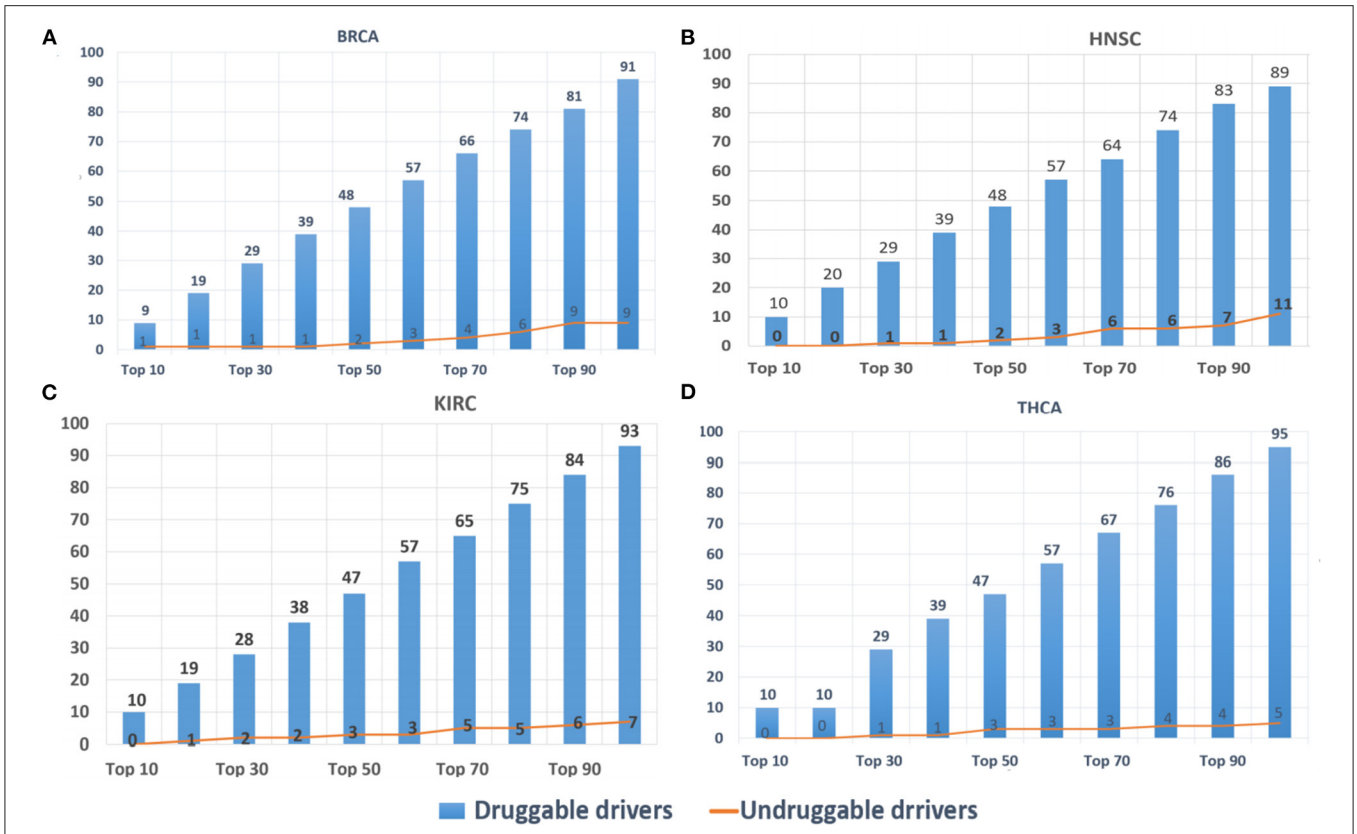


FIGURE 5 | Distribution of top 100 candidate-driver genes from the four cancer gene databases in druggable genes databases. **(A)** BRCA, **(B)** HNSC, **(C)** KIRC, and **(D)** THCA.

able to find these unknown cancer driver genes which could act as potential therapeutic targets and useful prognostic biomarkers for overall survival of patients.

Through performing the KEGG and GO enrichment of these top 100 ranked genes in BRCA, HNSC, KIRC, and THCA, respectively, these drivers were involved in oxidative stress, immune response-regulating cell surface receptor signaling pathway, apoptotic signaling pathway, and immune response-activating cell surface receptor signaling pathway. All of the KEGG and GO terms play important roles in the response to cancer.

Although the present study shows various positive results, it has certain limitations as well. Future validation using multiple cancer types is warranted. In addition, the present study did not attempt to use the synonymous mutations (Wen et al., 2016) and indels (insertions and deletions) (Yue et al., 2019), which have been found to regulate tumorigenesis via various mechanisms (Yue et al., 2019; Zhang and Xia, 2020). We will attempt to integrate these somatic mutation data in our future work.

In conclusion, we have designed an effective and no parameter algorithm, termed DriverSubNet, for prioritizing cancer driver genes by integrating somatic mutational, expression, and PPI network. As indicated by the evaluation of four cancer datasets, DriverSubNet consistently outperformed Dawnrank and DriverNet methods in terms of precision, recall, and F1 score. Further, it was able to identify potential driver genes that have not been documented, but might be important driver genes. Thus, DriverSubNet acted as a useful tool for the identification of driver genes by subnetwork enrichment analysis. However, studies with larger multiple cancer types and by including synonymous mutations and indels will be helpful in further development of this method.

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

DZ conceived the algorithm, designed the method, carried out the experiments, analyzed the data, and drafted the manuscript. DZ and YB refined the idea, polished the English expression and revised the paper, and participated in the design and revision of the research. All authors read and approved the final manuscript.

FUNDING

This work was supported by the grants from the National Natural Science Foundation of China (21601001) and Shaoguan City Science and Technology Project (2019sn082).

ACKNOWLEDGMENTS

We acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.607798/full#supplementary-material

Supplementary Figure 1 | The top 10 gene ontology (KEGG) terms enrichment of **(A)** BRCA, **(B)** HNSC, **(C)** KIRC, and **(D)** THCA by significant genes with $p < 0.05$ in DriverSubNet.

Supplementary Table 1 | The details of the dataset.

Supplementary Table 2 | A list of top 100 candidate-driver genes of four datasets.

REFERENCES

Ahn, R., Sabourin, V., Bolt, A. M., Hebert, S., Totten, S., De Jay, N., et al. (2017). The Shc1 adaptor simultaneously balances Stat1 and Stat3 activity to promote breast cancer immune suppression. *Nat. Commun.* 8:14638. doi: 10.1038/ncomms14638

Asghar, U., Witkiewicz, A. K., Turner, N. C., and Knudsen, E. S. (2015). The history and future of targeting cyclin-dependent kinases in cancer therapy. *Nat. Rev. Drug Disc.* 14, 130–146. doi: 10.1038/nrd4504

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., et al. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* 13, 1–14. doi: 10.1186/gb-2012-13-12-r124

Blanchette-Farra, N., Kita, D., Konstorum, A., Tesfay, L., Lemler, D., Hegde, P., et al. (2018). Contribution of three-dimensional architecture and tumor-associated fibroblasts to hepcidin regulation in breast cancer. *Oncogene* 37, 4013–4032. doi: 10.1038/s41388-018-0243-y

Cadena Castaneda, D., Brachet, G., Goupille, C., Ouldamer, L., and Gouilleux-Gruart, V. (2020). The neonatal Fc receptor in cancer FcRn in cancer. *Cancer Med.* 9, 4736–4742. doi: 10.1002/cam4.3067

Cavatorta, A. L., Gregorio, A. D., Valdano, M. P. B., Marziali, F. E., Cabral, M., Bottai, H., et al. (2017). DLG1 polarity protein expression associates with the disease progress of low-grade cervical intraepithelial lesions. *Exp. Mol. Pathol.* 102, 65–69. doi: 10.1016/j.yexmp.2016.12.008

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

Cotto, K. C., Wagner, A. H., Feng, Y., Kiwala, S., Coffman, A. C., Spies, G., et al. (2018). DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res.* 46, D1068–D1073. doi: 10.1093/nar/gkx1143

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Di Zhang, P. C., Zheng, C.-H., and Xia, J. (2016). Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. *Oncotarget* 7:4298. doi: 10.18632/oncotarget.6774

Dinstag, G., and Shamir, R. (2019). PRODIGY: personalized prioritization of driver genes. *Bioinformatics* 36, 1831–1839. doi: 10.1093/bioinformatics/btz815

Elias, D., and Ditzel, H. J. (2015). Fyn is an important molecule in cancer pathogenesis and drug resistance. *Pharmacol. Res.* 100, 250–254. doi: 10.1016/j.phrs.2015.08.010

Fan, Z., Duan, J., Wang, L., Xiao, S., Li, L., Yan, X., et al. (2019). PTK2 promotes cancer stem cell traits in hepatocellular carcinoma by activating Wnt/β-catenin signaling. *Cancer Lett.* 450, 132–143. doi: 10.1016/j.canlet.2019.02.040

Filmus, J., Capurro, M., and Rast, J. (2008). Glypicans. *Genome Biol.* 9:224. doi: 10.1186/gb-2008-9-5-224

Galindomoreno, M., Giraldez, S., Saez, C., Japon, M. A., Tortolero, M., and Romero, F. (2017). Both p62/SQSTM1-HDAC6-dependent autophagy and the aggresome pathway mediate CDK1 degradation in human breast cancer. *Scient. Rep.* 7, 10078–10078. doi: 10.1038/s41598-017-10506-8

Ge, Z., Leighton, J., Wang, Y., Peng, X., Chen, Z., Chen, H., et al. (2018). Integrated genomic analysis of the ubiquitin pathway across cancer types. *Cell Rep.* 23:213. doi: 10.1016/j.celrep.2018.03.047

Giubellino, A., Burke, T. R., and Bottaro, D. P. (2008). Grb2 signaling in cell motility and cancer. *Expert Opin. Therap. Targets* 12, 1021–1033. doi: 10.1517/14728222.12.8.1021

Hassani, R. A. E., Buffet, C., Leboulleux, S., and Dupuy, C. (2019). Oxidative stress in thyroid carcinomas: biological and clinical significance. *Endocrine-Related Cancer* 26, R131–R143. doi: 10.1530/ERC-18-0476

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

Hou, Y., Gao, B., Li, G., and Su, Z. (2018). MaxMIF: a new method for identifying cancer driver genes through effective data integration. *Adv. Sci.* 5:1800640. doi: 10.1002/advs.201800640

Hristov, B. H., Chazelle, B., and Singh, M. (2020). A guided network propagation approach to identify disease genes that combines prior and new information. *Lect. Notes Comput. Sci.* 12074, 251–252. doi: 10.1007/978-3-030-45257-5_25

Hui, Y., Wei, P., Xia, J., Wang, Y., and Zheng, C. (2019). MECoRank: cancer driver genes discovery simultaneously evaluating the impact of SNVs and differential expression on transcriptional networks. *BMC Med. Genomics* 12, 1–10. doi: 10.1186/s12920-019-0582-8

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database—2009 update. *Nucleic Acids Res.* 37(Database issue), D767–72. doi: 10.1093/nar/gkn892

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

Luo, P., Ding, Y., Lei, X., and Wu, F.-X. (2019). deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.* 10:13. doi: 10.3389/fgene.2019.00013

Ma, Q. (2013). Role of nrf2 in oxidative stress and toxicity. *Annu. Rev. Pharmacol. Toxicol.* 53, 401–426. doi: 10.1146/annurev-pharmtox-011112-140320

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L. M., Weinstein, J. N., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 173, 338–354. doi: 10.1016/j.cell.2018.03.034

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55–62. doi: 10.1038/nature18003

Mirza, A. H., Thomas, G., Ottensmeier, C. H., and King, E. V. (2019). Importance of the immune system in head and neck cancer. *Head Neck.* 41, 2789–2800. doi: 10.1002/hed.25716

Moloney, J. N., and Cotter, T. G. (2017). ROS signalling in the biology of cancer. *Sem. Cell Dev. Biol.* 80, 50–64. doi: 10.1016/j.semcdb.2017.05.023

Pistritto, G., Trisciuoglio, D., Ceci, C., Garufi, A., and D'Orazi, G. (2016). Apoptosis as anticancer mechanism: function and dysfunction of its modulators and targeted therapeutic strategies. *Aging* 8, 603–619. doi: 10.18632/aging.100934

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., et al. (2015). The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* 43(Database issue), D670–81. doi: 10.1093/nar/gku1177

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

Santamaría, D., Barrière, C., Cerqueira, A., Hunt, S., Tardy, C., Newton, K., et al. (2007). Cdk1 is sufficient to drive the mammalian cell cycle. *Nature* 448, 811–815. doi: 10.1038/nature06046

Shen, R. R., Zhou, A. Y., Kim, E., Oconnell, J. T., Hagerstrand, D., Beroukhim, R., et al. (2015). TRAF2 is an NF-κB-activating oncogene in epithelial cancers. *Oncogene* 34, 209–216. doi: 10.1038/onc.2013.543

Tate, J. G., Bamford, S., Jubb, H., Sondka, Z., Beare, D., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015

Wang, H., Zhi, H., Ma, D., and Li, T. (2017a). MiR-217 promoted the proliferation and invasion of glioblastoma by repressing YWHAG. *Cytokine* 92, 93–102. doi: 10.1016/j.cyto.2016.12.013

Wang, P., Deng, Y., and Fu, X. (2017b). MiR-509-5p suppresses the proliferation, migration, and invasion of non-small cell lung cancer by targeting YWHAG. *Bioche. Biophys. Res. Commun.* 482, 935–941. doi: 10.1016/j.bbrc.2016.11.136

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

Wen, P., Xiao, P., and Xia, J. (2016). dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics* 32, 1914–1916. doi: 10.1093/bioinformatics/btw086

Wright, K. D., Miller, B. S., Elmeanawy, S., Tsaih, S., Banerjee, A., Geurts, A. M., et al. (2019). The p52 isoform of SHC1 is a key driver of breast cancer initiation. *Breast Cancer Res.* 21, 74. doi: 10.1186/s13058-019-1155-7

Yoo, J.-O., Kwak, S.-Y., An, H.-J., Bae, I.-H., Park, M.-J., and Han, Y.-H. (2016). miR-181b-3p promotes epithelial–mesenchymal transition in breast cancer cells through Snail stabilization by directly targeting YWHAG. *Biochim. Biophys. Acta (BBA)-Mol. Cell Res.* 1863, 1601–1611. doi: 10.1016/j.bbamcr.2016.04.016

Yue, Z., Zhao, L., Cheng, N., Yan, H., and Xia, J. (2019). dbCID: a manually curated resource for exploring the driver indels in human cancer. *Briefings Bioinform.* 20, 1925–1933. doi: 10.1093/bib/bby059

Yue, Z., Zhao, L., and Xia, J. (2018). dbCPM: a manually curated database for exploring the cancer passenger mutations. *Brief. Bioinform.* 21, 309–317. doi: 10.1093/bib/bby105

Zefang, T., Chenwei, L., Boxi, K., Ge, G., Cheng, L., and Zemin, 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

Zhang, D., and Xia, J. (2020). Somatic synonymous mutations in regulatory elements contribute to the genetic aetiology of melanoma. *BMC Med. Genomics* 13 (Suppl. 5), 43. doi: 10.1186/s12920-020-0685-2

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.

frontiers in Genetics

REVIEW
published: 12 April 2021
doi: 10.3389/fgene.2021.607471

Intelligent Health Care: Applications of Deep Learning in Computational Medicine

Sijie Yang[1], Fei Zhu[1], Xinghong Ling[1,2], Quan Liu[1] and Peiyao Zhao[1]*

[1] School of Computer Science and Technology, Soochow University, Suzhou, China, [2] WenZheng College of Soochow University, Suzhou, China

With the progress of medical technology, biomedical field ushered in the era of big data, based on which and driven by artificial intelligence technology, computational medicine has emerged. People need to extract the effective information contained in these big biomedical data to promote the development of precision medicine. Traditionally, the machine learning methods are used to dig out biomedical data to find the features from data, which generally rely on feature engineering and domain knowledge of experts, requiring tremendous time and human resources. Different from traditional approaches, deep learning, as a cutting-edge machine learning branch, can automatically learn complex and robust feature from raw data without the need for feature engineering. The applications of deep learning in medical image, electronic health record, genomics, and drug development are studied, where the suggestion is that deep learning has obvious advantage in making full use of biomedical data and improving medical health level. Deep learning plays an increasingly important role in the field of medical health and has a broad prospect of application. However, the problems and challenges of deep learning in computational medical health still exist, including insufficient data, interpretability, data privacy, and heterogeneity. Analysis and discussion on these problems provide a reference to improve the application of deep learning in medical health.

Keywords: deep learning, computational medicine, health care, medical imaging, genomics, electronic health records, drug development

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong,
Hong Kong

Reviewed by:
Padhmanand Sudhakar,
KU Leuven, Belgium
Jiajia Chen,
Suzhou University of Science and Technology, China
Jianhong Zhou,
Public Library of Science,
United States

***Correspondence:**
Xinghong Ling
lingxinghong@suda.edu.cn

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

Received: *17 September 2020*
Accepted: *05 March 2021*
Published: *12 April 2021*

Citation:
Yang S, Zhu F, Ling X, Liu Q and Zhao P (2021) Intelligent Health Care: Applications of Deep Learning in Computational Medicine. Front. Genet. 12:607471. doi: 10.3389/fgene.2021.607471

INTRODUCTION

In recent years, with the explosive growth of biomedical data and the rapid development of medical technology and computer technology, the field of medical health ushered in the era of big data (Miotto et al., 2018). In this context, computational medicine began to appear as a new subject. Based on big biomedical data and computer technology, computational medicine is an interdisciplinary subject combining medicine, computer science, biology, mathematics, etc. It uses the method of artificial intelligence to intelligently understand the principle and physiological mechanism of human diseases by analyzing big data and provides useful information and guidance for disease prediction, clinical diagnosis, and medical services. Taking the pharmaceutical industry as an example, traditionally, new drug research suffers from long periods, considerable investment, and high failure rate. In contrast, the research based on computational medicine can complete the preclinical drug research and development in an average of 1–2 years, with high success rate and

low resource consumption indicating that the field of medical health is gradually entering the era of intelligence and digitization.

However, biomedical datasets are high-dimensional, jumbled, noisy, and sparse, making it difficult to mine the rich information behind these datasets effectively. Therefore, an appropriate approach is needed to process large amounts of biomedical data to obtain efficient information. At present, in the community of machine learning, deep learning is a bright pearl in the field of artificial intelligence. As a branch of machine learning, it has been proved that deep learning is an effective method and surpasses the traditional machine learning in areas such as computer vision (He et al., 2016), natural language processing (Lan et al., 2020), and speech recognition (Abdel-Hamid et al., 2014). The key step of the machine learning method, called feature engineering, is to artificially use expert and domain knowledge to distill features from data and further analyze the features by machine learning models (such as support vector machine, random forest, etc.). In the era of big data, the manual extraction is insufficient and biased such that it cannot establish a high-performance model for specific tasks.

Unlike traditional machine learning approaches, deep learning spares the need to extract features manually, which improves time and resource efficiency. Deep learning is implemented by neural networks consisting of neurons. Each layer of neural networks is composed of a large number of neurons, and the output of the upper layer is regarded as the input of the next layer. Through the connection between layers and the nonlinear processing method, the neural network can convert the original input to the output. More importantly, the high-level network can automatically learn more abstract and generalized features from the data, which overcomes the shortcoming that machine learning needs to extract features manually.

As the most advanced artificial intelligence method, deep learning provides a method for computational medicine, so it is a trend to apply deep learning method to biomedical data analysis. **Figure 1** is a schematic diagram of deep learning in computational medicine. However, biomedical data are not as clean, easy to process, and easy to obtain as data in other fields, so it is a challenge to give a full play to the role of deep learning in computational medicine.

In this article, we first introduce several popular deep learning frameworks. Second, we survey the application of deep learning in clinical imaging, electronic health records, genomics, and drug development. Finally, we point out that the application of deep learning in the field of medical and health faces challenges such as insufficient data, model interpretability, data privacy, and heterogeneity.

INTRODUCTION TO DEEP LEARNING

Deep learning is a part of machine learning, which is inspired by neurons in the human brain: there are tens of millions of neurons in the human brain, and there are more than 100,000 connections between them. The deep learning method is called artificial neural network. As shown in **Figure 2**, the neural network is composed of the input layer, hidden layer, and output layer. Each layer is composed of several neurons, and the hidden layer may consist of many layers. According to different task types, there are different numbers of neurons in the output layer of the neural networks. For example, there are three neurons in the output layer in the three-classification problem, and each neuron represents the probability of belonging to a certain category.

We take a neuron in the hidden layer as an example to illustrate the calculation method of the neuron. As shown in **Figure 3**, the calculation of the neuron is as follows:

The equation of the calculation of the neuron is as follows:

$$y = f(\sum_{j} w_{ij}{}^{*} x_{j}) \tag{1}$$

where x_j is the jth input of the neuron, y is the output of the neuron, w_{ij} is the weight of the ith neuron and the jth input, b_i is the bias of the ith neuron, and f is the activation function to perform a nonlinear transformation on the output of the neuron. The common activation functions are sigmoid, ReLU, tanh, softmax, etc.

The training of the neural network depends on forward propagation algorithm and back-propagation algorithm. Forward propagation refers to the whole process of data propagation from the input layer to the output layer, where the neural network calculates intermediate variables of each neuron

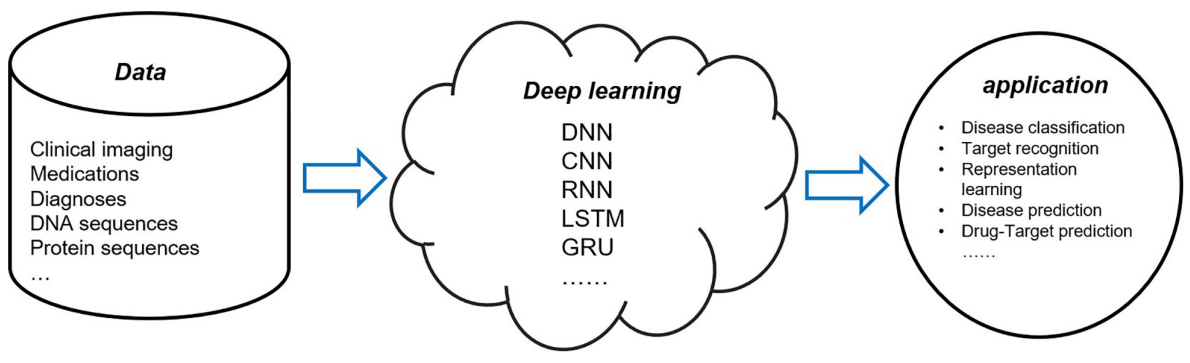


FIGURE 1 | Application of deep learning models in computational medicine.

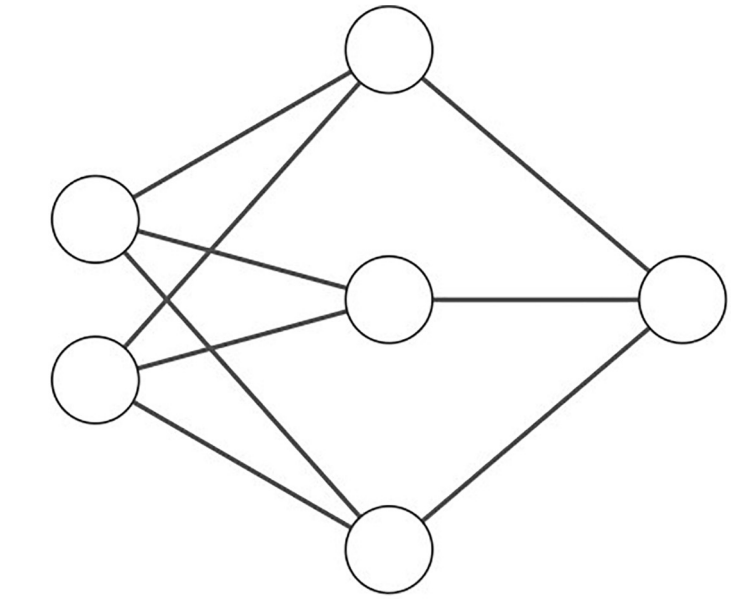


FIGURE 2 | Illustration of neural network architecture.

in turn. Back propagation refers to the parameter optimization process of the neural network. According to the intermediate variables calculated by forward propagation, the parameters of the neural network are updated by gradient descent. The common gradient descent algorithms include random gradient descent, Adam (Kingma and Ba, 2015), RMSprop, Adadelta (Zeiler, 2012), Adagrad (Duchi et al., 2011), etc.

According to the connection and calculation methods, there are different types of the neural networks. Here, we discuss only the most common and basic neural networks, which constitute the basic deep learning methods. **Table 1** lists a summary of the neural networks.

Fully Connected Neural Network

As the name implies, a fully connected neural network means that the neurons in the layers of the neural networks are completely connected, as shown in **Figure 4**. The fully connected neural network consists of the input layer, the hidden layer, and the output layer. The input layer is responsible for receiving input data. The hidden layer is composed of many neural network layers for feature extraction. The output layer outputs the final prediction result.

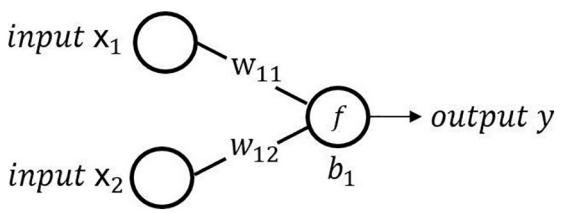


FIGURE 3 | Schematic diagram of neural network calculation.

Assume that the input is $_{X\in R^{n*d}}$, where the number of samples is n, and each sample consists of d features. Assume that the first hidden layer *Hidden layer 1* contains h neurons, that is, the *Hidden layer 1* contains h outputs, and then the weight matrix of the first hidden layer is denoted as $_{W_1\in R^{d*h}}$, the bias is denoted as $_{b_1\in R^{1*h}}$, and the calculation of output $Output_1$ is $_{Output_1=f(XW_1+b_1)}$.

After that, the output $Output_1$ of *Hidden layer 1* is used as the input of *Hidden layer 2*, and so on, until the output layer of the fully connected network outputs the final calculation result. The fully connected neural network is the most basic neural network. Combined with other neural networks, it is widely used to integrate high-level features and output prediction results. Depending on the task type, the final output can be either a probability distribution or a task-related value.

Convolutional Neural Network

In recent years, the convolutional neural network has made remarkable achievements in image recognition. The convolutional neural networks adopt the method of local connection and weight sharing, which reduces the complexity of the network and enables the network to directly use the image as input. The convolution neural network has two important characteristics: first, the features learned from the

TABLE 1 | A summary of the neural networks.

Neural networks	Advantages	Disadvantages	Biomedical tasks
Fully connected neural network	It is widely used at the end of the other neural network models to integrate features and make predictions	It is not easy to process high-dimensional data	Combined with other neural networks, it is widely used in many fields
Convolutional neural network	It can extract highly abstract and complex features from images	It has too many parameters, and the training speed is slow	It is suitable for processing imaging-related tasks, such as clinical imaging
Recurrent neural network	It has a memory function and can effectively process data about sequence and time	Training procedure is difficult and computationally intensive	It is suitable for processing sequence related biomedical data, such as DNA sequence, protein sequence, electronic health records
Autoencoder	It can perform unsupervised learning without using labeled data	It needs a pretraining phase	It is suitable for feature dimensionality reduction or learning effective features from data, such as clinical imaging and genomics
Deep belief network	It can be used for both supervised learning and unsupervised learning	The training process is computationally intensive	It is suitable for automatic feature extraction tasks, such as genomics and drug development

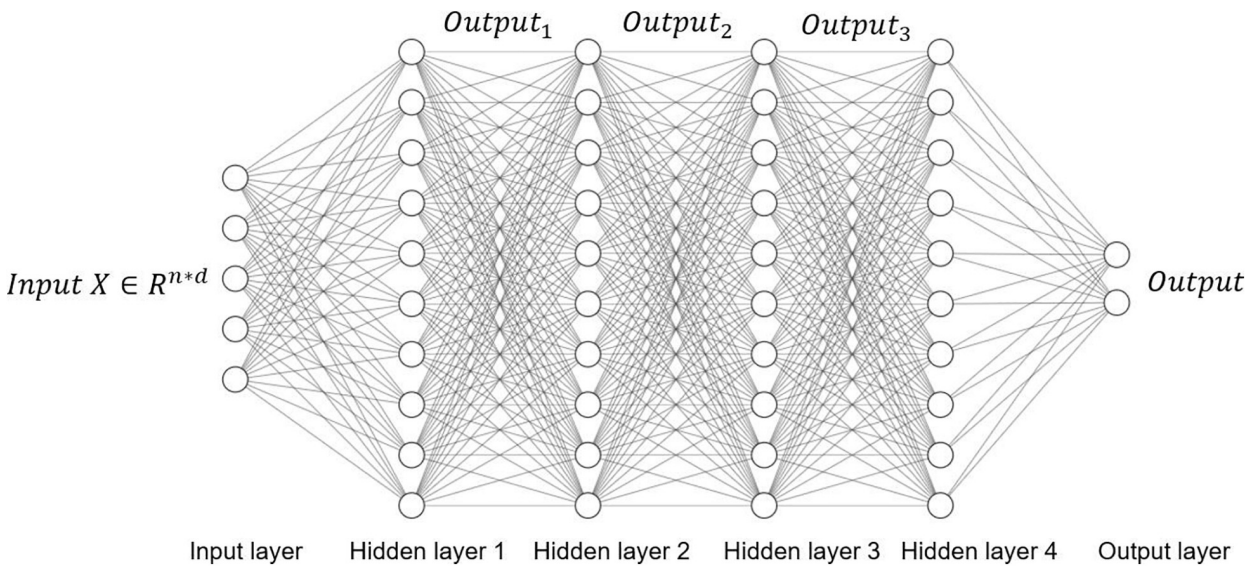


FIGURE 4 | Fully connected neural network.

image are translational and nondeformable; second, the higher the convolution layer, the more abstract and complex the features extracted. The convolution neural network is composed of the convolution layer, the pooling layer, and the fully connected layer. The convolution layer is composed of filters. Each filter is equivalent to a small window. These small windows move on the image to learn features from the image. Then, the learned features are subsampled by pooling operation to extract more representative features and improve the robustness and accuracy of the model. Finally, the fully connected layer outputs the prediction result. A convolutional neural network framework for lung pattern recognition (Anthimopoulos et al., 2016) is shown in **Figure 5**.

Recurrent Neural Network

Another common neural network is called the recurrent neural network, which is very suitable for processing sequential data, such as time-dependent data.

In the fully connected neural networks and the convolutional neural networks, their inputs are independent. In contrast, in the recurrent neural network, the former input and the latter input are dependent and have sequence relation. Just like analyzing a sentence, because the current word depends on the front and back words, it means that analyzing each independent word will not produce good results. The structure of the recurrent neural network is shown in **Figure 6**. At time t, the input of the neural network is x_t. The output of the neural network is y_t, which is calculated from the hidden layer state s_t that depends not only on the input x_t at the current time t, but also on the state s_{t-1} at the time t-1, which makes the recurrent neural network have memory, and the state of the last moment can affect the effectiveness of the current time.

The variants of recurrent neural networks include Gated Recurrent Unit (GRU) (Chung et al., 2014) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), and so on.

Autoencoder

The fourth deep learning framework is called the autoencoder, which is often used in unsupervised learning.

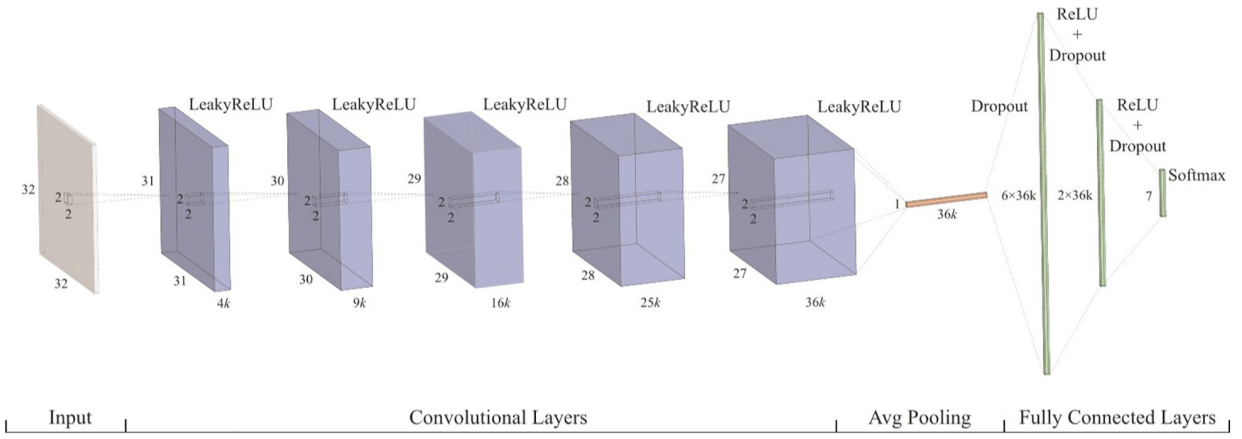


FIGURE 5 | Convolutional neural network.

The autoencoder can be used to reduce dimension and learn feature. The structure of the autoencoder is shown in **Figure 7**. The autoencoder is composed of an encoder and a decoder. Encoders and decoders can be any neural networks models. In general, the number of neurons in the middle-hidden layer is less than that in the input layer and the output layer, which is useful for compressing data and learning effective features from data. The number of neurons in the input layer and the output layer in the autoencoder is the same. Specifically, the encoder reduces the dimension of the original data to get a new representation. Then, the decoder restores the input data through this new representation.

The deformation of autoencoder includes stacked autoencoder (Bengio et al., 2006), denoising autoencoder (Vincent et al., 2008), variational autoencoder, etc. The stacked autoencoder is a hierarchical deep neural network structure composed of multilayer autoencoders. It has deeper depth and stronger learning ability. The denoising autoencoder adds random noise to the input data and then uses the data with noise to train the autoencoder. The autoencoder trained in this way is stronger and has better antinoise ability. Variational autoencoder adds some restrictions in the encoding process, which makes the generated vectors follow the standard normal distribution. The encoding method makes the automatic encoder more effective.

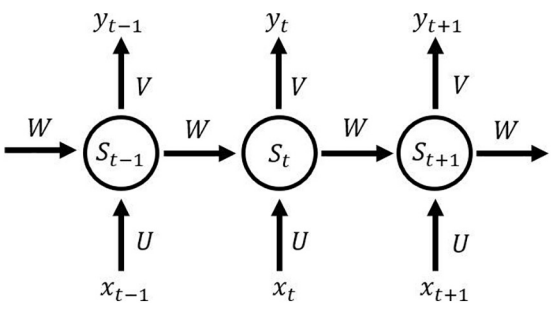


FIGURE 6 | Illustration of recurrent neural network.

Deep Belief Network

Deep belief network (Hinton et al., 2006) is a probability generation model based on the restricted Boltzmann machine, which establishes a joint probability distribution between data and label.

As shown in **Figure 8**, the restricted Boltzmann machine has only two layers: the visible layer composed of visible units and the hidden layer composed of hidden units. The visible layer is used for the input of training data, whereas the hidden layer is used as a feature detector. Each layer can be represented as a vector, each dimension by each neuron. Neurons are independent of each other. The advantage of

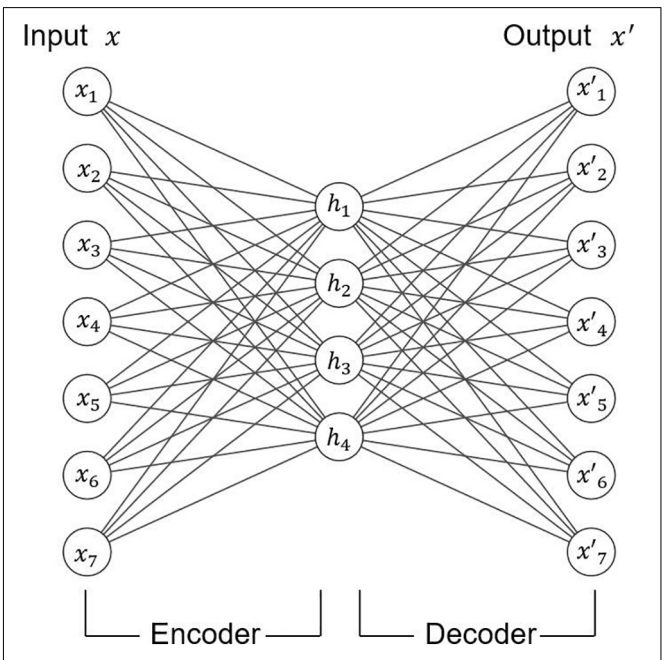


FIGURE 7 | Illustration of autoencoder.

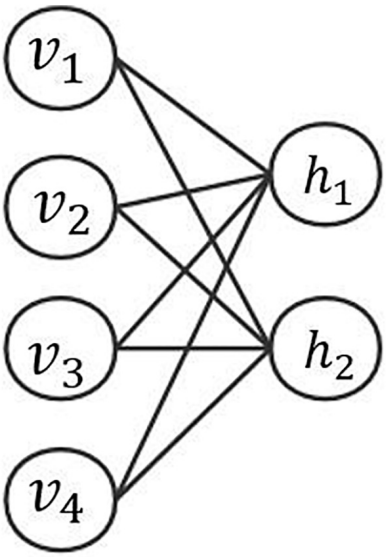


FIGURE 8 | Illustration of restricted Boltzmann machine.

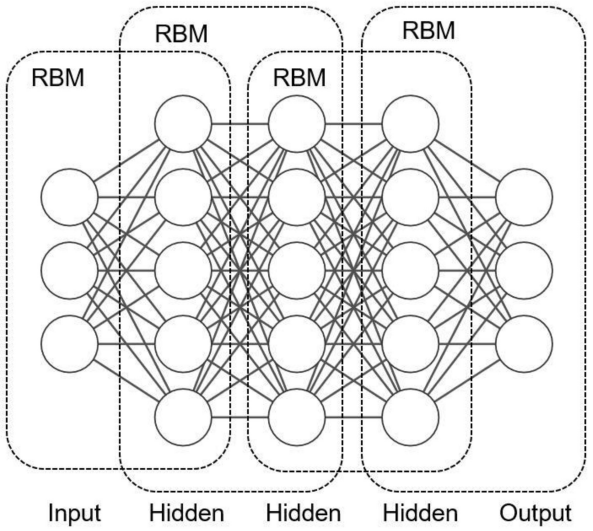


FIGURE 9 | Illustration of deep belief network.

this is that given the values of all the explicit elements, the values of each implicit element are independent of each other. The trained restricted Boltzmann machine can extract the features of the explicit layer more accurately or restore the explicit layer according to the features represented by the implicit layer.

As shown in **Figure 9**, several restricted Boltzmann machines are connected to form a deep belief network in which the hidden layer of the previously restricted Boltzmann machine is the visible layer of the next restricted Boltzmann machine. It means that the output of the previously restricted Boltzmann machine is the input of the next restricted Boltzmann machine. In the training process, it is necessary to fully train the restricted Boltzmann machine in the upper layer before training the restricted Boltzmann machine in the current layer. The procedure continues until the last layer.

Software/Hardware Support

Owing to the explosion of data and the support of GPU hardware acceleration technology, deep learning has developed rapidly in recent years. At present, there are many deep learning libraries such as PyTorch, Keras, TensorFlow, Theano, Caffe, and so on. **Table 2** lists the representative libraries and packages that provide a convenient and efficient tool for researchers who need to develop deep learning programs and greatly promote the application and development of deep learning in various fields (including computational medicine).

TABLE 2 | Some frequently used deep learning packages.

Name	Interface	URL
Keras	Python	https://keras.io/
PyTorch	Python	https://pytorch.org/
TensorFlow	Python	https://www.tensorflow.org/
Caffe	C++/Python/MATLAB	https://caffe.berkeleyvision.org/
Theano	Python	http://deeplearning.net/software/theano/
Torch	LuaJIT/C	http://torch.ch/

APPLICATION OF DEEP LEARNING IN COMPUTATIONAL MEDICINE

Search Strategy and Selection Criteria

Because Google Scholar provides researchers with a convenient and quick way to search the literature, we searched Google Scholar for published researches from 2015 to 2020 that contained the keyword "deep learning," which combines the terms of corresponding fields for deep learning in each different application field until September 2020. All searched researches are published in English. Specifically, for the application of deep learning in the clinical imaging, the combination of "deep learning" and "medical image" was used to search. For the application of deep learning in the field of the electronic medical records, the combination of "deep learning" and "electronic health record" or "electronic medical record" was used to search. For the research of deep learning in genomics, the combination of "deep learning" and "genomics" or "gene" was used to search. For the research of deep learning in drug development, we used the combination of "deep learning" and "drug development" or "drug repositioning" or "drug repurposing." For the literature found, we also conducted manual screening to check whether the content of the article is about the application of

deep learning in computational medicine. Finally, this review includes 107 articles.

Medical Image

The medical image plays a key role in medical diagnosis and treatment providing an important basis for understanding a patient's disease and helping physicians make decisions. As medical devices become more advanced, and the career of medical health is rapidly growing, more and more medical image data are generated, such as magnetic resonance imaging, computed tomography (CT), and so on. Huge amounts of medical imaging data require much more time if experts analyzed the data alone. And the analysis of medical image data may produce erroneous or biased results due to varying degrees of experience, knowledge, and other factors that the experts themselves have. Machine learning algorithms are, to some extent, able to assist specialists in automated analysis, but may not have the ability to process and achieve high accuracy when faced with such vast data and complex problems.

Deep learning has been successful in the field of image processing: it carries out tasks such as image classification, target recognition, and target segmentation by analyzing images. Therefore, the application of deep learning in the task of medical image analysis has become a trend in medical research in recent years. Researchers used the artificial intelligence methods to help physicians make accurate diagnoses and decisions. Many aspects are involved in these tasks, such as detecting retinopathy, bone age, skin cancer identification, etc. Deep learning achieves expert level in these tasks. The convolutional neural network is a powerful deep learning method. The convolutional neural networks follow the principle of translational invariance and parameter sharing, which is very suitable for automatically extracting image features from the original image.

Figure 10 shows a convolutional neural network structure for detecting pneumonia with chest X-ray images. The convolution neural network is composed of the convolution layer, the pooling layer, and the fully connected layer. First, the filtering window in a convolution layer will move step by step in the image to learn local features from the image. Second, the pooling layer will sample the learned features to reduce the parameters and overfitting to improve the performance of the network. Finally, these features will output the final results through the fully connected layer.

In a convolution neural network, there are usually multiple convolution layers such that the convolution layer at the bottom can learn the local features in the image, and the high-level convolution layer can integrate these local features and learn the overall features. Usually, for the same symptom, the location and shape of lesions or tumors are different in different pictures, making it very difficult to analyze. The advantage of the convolutional neural network is that it can automatically learn local features from images and integrate them into global features. Therefore, the convolutional neural network is very suitable for clinical image processing.

Researchers have analyzed the effects of different structural convolutional neural network applications on clinical images. Shin et al. (2016) pointed out that even if the available training dataset is limited, a convolutional neural network architecture with a depth of 8 or even 22 layers may be useful. They also proved that it was beneficial to migrate models trained from large-scale annotated natural image datasets (ImageNet) (Russakovsky et al., 2015) to computer-aided diagnosis problems in experiments. It can be seen that the convolutional neural network with few layers can obtain satisfactory results in limited datasets and has great potential in clinical imaging.

Many researchers have used convolutional neural networks on the task of the fundus image and achieved good results (Grinsven et al., 2016; Gulshan et al., 2016; Dai et al., 2018). Poplin et al. (2018) predicted cardiovascular risk factors from retinal fundus images. In order to better understand how neural network models can predict, a deep learning technology called "soft attention" is used, which can identify the parts that affect the prediction of the model. Kermany et al. (2018) developed a deep learning system to effectively classify the images of macular degeneration and diabetic retinopathy. Fauw et al. (2018) used a three-dimensional U-Net (Ronneberger et al., 2015; Çiçek et al., 2016) architecture to segment the original optical coherence tomography images into 15 types of tissue maps. Experimental results showed that the proposed method achieved and even surpassed the performance of experts in referral decision-making and disease prediction.

Other researchers used convolutional neural networks to detect or classify diseases on chest X-ray datasets. For example, Cao et al. (2016) used the convolutional neural network to X-ray

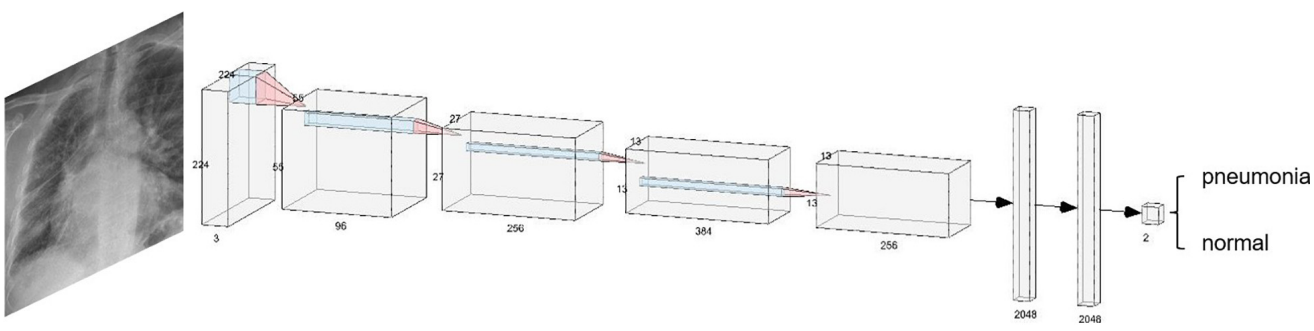


FIGURE 10 | Convolutional neural network structure for detecting pneumonia with chest X-ray images.

images collected from mobile devices to diagnose tuberculosis. The accuracy of the binary classification task reached 89.6%, which proved the learning ability of the convolutional neural network in the medical field. Cicero et al. (2017) used convolutional neural networks to detect and classify chest abnormalities. Yuan et al. (2019) conducted an experiment with collaborative deep learning on chest X-ray images. Through collaborative deep learning, the accuracy was improved by approximately 19%. Liu et al. (2019) proposed a deep fusion network based on segmentation to obtain the features of the whole chest X-ray image and local lung region image from the image. Kleesiek et al. (2016) used convolutional neural networks to analyze the brain from magnetic resonance imaging. The method could deal with any number of patterns, which proved the feasibility of the convolution neural network in large-scale research.

In other clinical image tasks, the convolution neural network also achieved the performance of doctors including skin cancer diagnosis, knee osteoarthritis diagnosis, bone age assessment, etc. (Esteva et al., 2017; Haenssle et al., 2018; Iglovikov et al., 2018; Rakhlin et al., 2018; Tiulpin et al., 2018).

In addition to tasks of routine disease detection, the convolution neural network can also be used to evaluate the operation of doctors to help doctors improve the operation effect. Jin et al. (2018) used a convolutional neural network to automatically evaluate the performance of surgeons by tracking and analyzing tool movements in surgical videos. Shvets et al. (2018) introduced a method of robot instrument segmentation from surgical images based on deep learning. They used four different deep learning frameworks: a modified U-Net, two modified TernausNet (Iglovikov and Shvets, 2018), and a modified LinkNet (Chaurasia and Culurciello, 2017). The modified TernausNet performed best in binary segmentation experiments and partial segmentation experiments.

Some researchers also use data augmentation technology to solve the problem of data sparsity to prevent model overfitting (Anthimopoulos et al., 2016; Avendi et al., 2016; Kooi et al., 2017). Data augmentation refers to using some methods such as flipping, rotation, translation, clipping, changing contrast to transform the original image to generate more training data from the existing training samples, solving the problem of insufficient data, and improving the ability of the model.

In addition to the convolution neural networks, other neural networks methods such as the autoencoder, the deformation model of autoencoder, and the recurrent neural network are also used in medical imaging research. For example, Hu et al. (2016) constructed an autoencoder architecture to classify patients and predict Alzheimer disease. Compared with the support vector machine, the prediction accuracy is improved by approximately 25%. Cheng J. Z. et al., 2016) used the stacked denoising autoencoder structure for the differential diagnosis of breast lesions in ultrasound images and pulmonary nodules. Mansoor et al. (2016) used the stacked autoencoder to locate anterior visual pathway segmentation and to create a model to capture local appearance features of anterior visual pathway segmentation. Ortiz et al. (2016) used the set of deep belief networks for early detection of Alzheimer disease. Andermatt et al. (2016) proposed multidimensional GRU for brain segmentation, which could accurately segment three-dimensional data.

In conclusion, in the field of clinical imaging, the most used deep learning network is convolutional neural networks, other neural networks such as recurrent neural networks and autoencoder have also been used. To make the models capture the patterns and features, some measures have also been taken by researchers, such as regularization, dropout, and expansion of the dataset methods. Among them, the most common approach for expanding dataset is data augmentation techniques, which play an important role in improving the performance of models. These experiments demonstrate that deep learning performs better than traditional machine learning on clinical imaging tasks. Deep learning provides doctors with automated technology to analyze pictures, videos, etc., to help biomedical careers develop rapidly.

Although the effect of deep learning on medical imaging is better than that of machine learning, and deep learning achieves human-level performance, there are still some limitations:

(1) It is difficult to collect sufficient labeled data. Training a convolutional neural network with good performance requires a large number of parameters and samples. Sometimes it is difficult to collect sufficient and labeled training data. In addition to the differences in features, patterns, colors, values, and shapes in real medical image data, it is difficult to train a suitable network.
There are two ways to deal with the problem. The first way is the strategy of data augmentation, which is a very powerful technology to reduce overfitting. It generates a new image through a series of transformations (such as translation, flipping, changing contrast) of the original images to expand the dataset, so that the model has better generalization ability. The other is to use transfer learning, which uses a model trained in other training data in advance and then transfers the model to the medical image data to fine-tune the model, so as to get a model with strong generalization ability.

(2) Convolution neural networks cannot explain the hierarchical and positional relationship between features extracted from images. For example, neurons can capture the dataset feature, but neurons cannot well capture the spatial relationship between these features. For this reason, Sabour et al. (2017) proposed the Capsules Network. In this structure, the input and output of the capsules are not a scalar, but a vector instead of a traditional neuron. The length of the vector means the probability of the existence of instances, while the value of the vector can represent the relationship between features.
At present, there are few types of research on the Capsules Network in medical imaging. Jiménez-Sánchez et al. (2018) have studied the application of Capsules Network in clinical imaging. The experimental results showed that Capsules Network could be trained with fewer data to obtain the same or better performance, and it was more robust for unbalanced class distribution. This

result undoubtedly brought new ideas and directions to the application of deep learning in medical imaging.

(3) Convolutional neural networks are suitable for processing two-dimensional image data, but the images produced by magnetic resonance imaging or CT image have the inherent three-dimensional structure. If the convolutional neural network is used to process these medical images, key information will be lost.

Electronic Health Record

One-dimensional convolutional neural network, recurrent neural network, LSTM, GRU, and other neural networks in deep learning have been widely used in the natural language processing community and have achieved great success. These networks are very suitable for processing sequence-related data, such as sentence, voice, time series, and so on. Similarly, natural language processing technology is also used in the field of computational medicine, which uses these neural networks to process electronic medical records.

In recent years, electronic health record has received more and more attention. Electronic health record stores the treatment information of patients in electronic form. The information includes structured demographic information, diagnosis information, drug information, operation process information, experimental test results, and unstructured clinical text (Jensen et al., 2012). Mining electronic health records can improve the efficiency and quality of diagnosis and promote medical development. For example, it provides timely treatment for patients by mining the data in the electronic health record to predict the disease, or it analyzes the hidden relationship between diseases and diseases, diseases and drugs, and drugs and drugs in the electronic health records to provide help for doctors in decision-making.

In short, for patients, the use of electronic health records can better help patients understand their physical condition and disease status; for the medical staff, the use of electronic health records can help them better analyze problems and provide effective solutions.

Traditionally, machine learning is used to analyze electronic health record data. Usually, we need to extract the features manually and then input them into the model. This feature extraction method often depends on the professional domain knowledge of the extractor, and it may be difficult to find the hidden relationship in the data. Therefore, the quality of the model prediction results is affected by the quality of the manually extracted features. Moreover, this method causes huge human and time loss and affects the research efficiency.

Deep learning overcomes the disadvantage of traditional machine learning, which needs manual feature extraction. However, because of the particularity and complexity of electronic health record data, there are some problems when using deep learning method to deal with them. There are many clinical concepts in the electronic health records, which contain rich information. These concepts are recorded in the form of coding, such as diagnostic coding, disease coding, drug coding, etc. Different medical ontologies formulate the rules of coding and the meanings they represent. At the same time, doctors record these clinical concepts in chronological order. However, it is difficult to find and explore the relationship between these concepts simply by coding the patient's condition.

The traditional method is one hot coding whose dimension represents the number of medical concepts. The coding method does not reflect the relationship between concepts. Thus, researchers put forward representation learning, which studies how to map the clinical concepts represented by coding in electronic health records into low dimensional space and transform them into low dimensional features (embedding) for representation. At the same time, these low-dimensional features reflect the relationship between different concepts. After getting the representation of clinical concepts, researchers can find out the relationship between them through analysis, such as the relationship between diseases, or diseases and clinical events, or use these concepts in further tasks to provide useful information for doctors and help doctors make decisions. These studies included analysis of mortality, prediction of rehospitalization, prediction of disease, etc. As shown in **Figure 11**, the electronic health record can be transformed into patient representation by using the deep learning method (Miotto et al., 2016).

Most of the researches using deep learning to represent the clinical concept of electronic health records use the skip-gram structure of the word2vec model (Mikolov et al., 2013). The skip-gram approach assumes that the meaning of a concept depends on its context (or neighbors). Therefore, given the sequence of a concept, the skip-gram method predicts the context of a target concept when it selects it. After getting the representation of clinical concepts, researchers can find out the relationship between them through analysis, such as the relationship between diseases, or diseases and clinical events. Researchers can also use these concepts in further tasks to provide useful information for doctors and help doctors make decisions (Choi Y. et al., 2016; Choi et al., 2016a,b,c).

Attention mechanisms are also used in the analysis of electronic medical records. The mechanism of attention enables deep learning models to focus from the multitude of information to more critical and important information that is more closely connected to the task. Using the attention mechanisms, it is possible to know what information in the data contributes to the model's predictions. Zhang J. et al., 2018 established a framework by using recurrent neural networks and attention mechanisms to learn the representation of patients from the temporal electronic health record data. Then, they applied the model to the risk prediction task of future hospitalization. The experimental results showed that deep learning model can achieve a more accurate prediction effect.

Several works have shown that using the combination of different neural networks or different methods can improve the accuracy and efficiency of models. For example, Miotto et al. (2016) used an unsupervised deep learning method to learn the patient's representation from the electronic health records. They used a three-layer stacked denoising autoencoder to capture the hierarchical relationship and dependence between the data. Ma et al. (2018) proposed a deep learning framework, which is composed of recurrent neural networks and convolutional neural

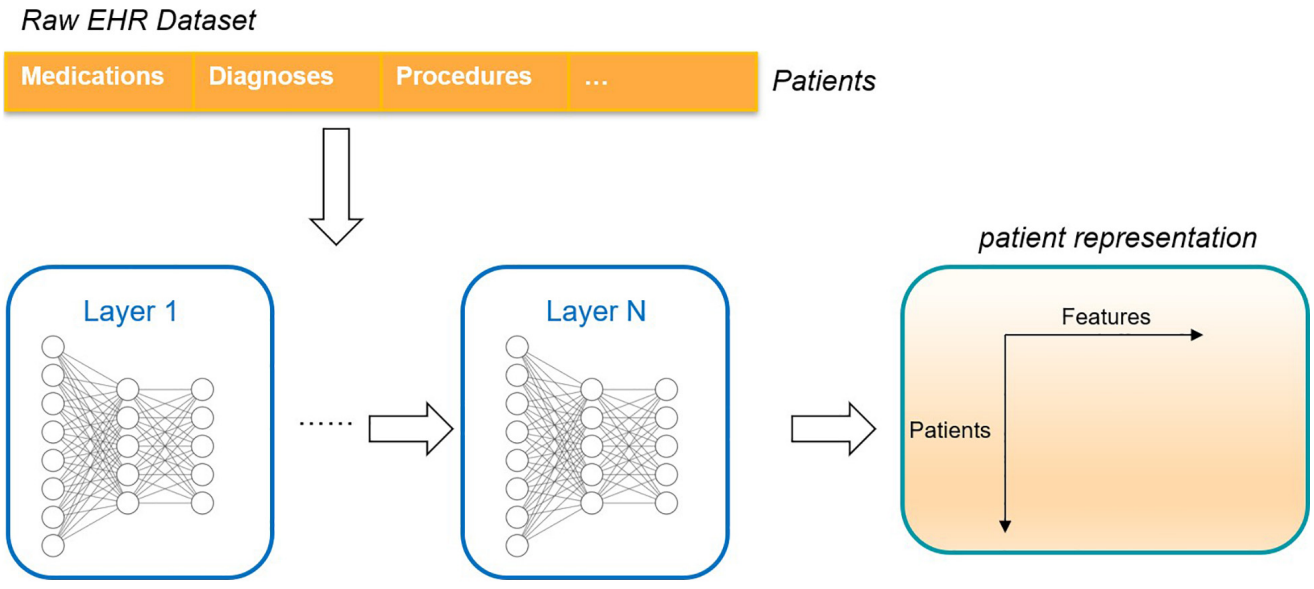


FIGURE 11 | An unsupervised deep learning framework that converts the raw electronic health record into a deep representation of the patient.

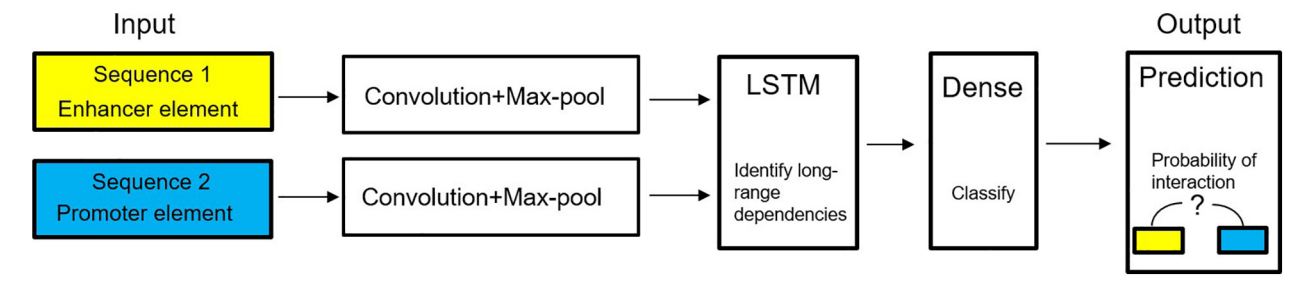


FIGURE 12 | Diagram of deep learning model to predict enhancer–promoter interactions (Chen Y. et al., 2016).

networks to extract patient information patterns. Rajkomar et al. (2018) used LSTM, the attention mechanism, and the single-layer decision tree to learn information from the dataset. In all tasks, the performance of the deep learning model is better than that of the traditional clinical prediction model.

In addition to using deep learning to represent clinical concepts in electronic health records, researchers also use deep learning to carry out disease prediction tasks. Disease prediction refers to predicting whether a patient will suffer from a certain disease or find out the factors related to the disease according to the patient's electronic health record information. Because the electronic health record contains a wealth of patient information, some indicators and characteristics in the information can be used as a reference to predict whether the patient has a disease. Nickerson et al. (2016) explored two neural networks architectures to deal with issues related to postoperative pain management. Pham et al. (2016) introduced a deep dynamic neural network for tasks such as predicting the next stage of illness. The results are competitive compared with other excellent methods at present. Nguyen et al. (2017) introduced a deep learning system to learn how to extract features from the electronic health record and automatically predict future disease risk. Compared with the traditional technology, the system can detect meaningful clinical patterns and reveal the potential structure of the disease and intervention space.

To deal with the missing values in medical records, Che et al. (2017) developed a model that was based on the GRU. The model captures the observations and their dependence by applying masking and time interval methods to the input and network state. Cheng Y. et al., 2016 used convolutional neural networks to extract phenotypes. They verified the validity of the proposed model on real and virtual electronic health records.

It can be seen that as the electronic health record is sequential, recurrent neural networks such as LSTM and GRU have a very wide range of applications in the field of electronic medical records. Recurrent neural networks are well suited for processing electronic medical records and achieve better results than traditional methods. When using deep learning method to mine the information of the electronic health records, most of the methods are the supervised learning methods; that is, the data are labeled. Some researchers use unsupervised learning to study electronic health records. With the deep learning methods, the

patterns in the electronic health record are analyzed. Then the learned patterns are used in disease prediction, event prediction, incidence prediction, and other tasks, which is undoubtedly the trend of the application of deep learning in the field of the electronic medical record.

Many studies have proven the effectiveness of deep learning in the electronic health record. However, there are still some challenges that hinder the further application of deep learning in the electronic health record:

(1) There are many types of data in the electronic health record, which are heterogeneous such that it is difficult to use the data in electronic health records in medical applications. There are five types of electronic health record data types: numerical type, such as body mass index; date–time object type, such as patient admission date; category type, such as race and international disease code; text type of natural languages, such as patient discharge summary; and time-series type, such as patient history (Shickel et al., 2018). In addition, the electronic health record also has the characteristics of high dimension, noise, complexity, and sparsity. How to use a suitable model for different types of the electronic health record is a big challenge when using deep learning methods to process electronic health record data.

(2) The coding in the electronic health record is different because of the difference in medical ontology in reality, which also brings challenges to applying of deep learning in the electronic health record. For example, medical ontology has the Unified Medical Language System (Unified Medical Language System, 2031), *International Classification of Diseases, Ninth Revision* (*ICD-9*) (ICD9, 2032), *ICD-10* (ICD10, 2033), National Drug Code, and other codes. In the specific implementation, different regions or different hospitals do not strictly abide by the coding rules of the medical oncology, so there are nonstandard records. And sometimes, the same disease phenotype can be represented by different medical ontology. For example, in the electronic health record, patients diagnosed with "type 2 diabetes mellitus" can be identified by the laboratory value of hemoglobin $A_{1c} > 7.0$, the code of 250.00 in *ICD-9*, and the writing method of "type 2 diabetes mellitus" in the clinical text (Miotto et al., 2018). The above problems increase the difficulty of data processing. In addition, the mapping between these codes also brings difficulties to researchers.

(3) The information in electronic health records may have a long-time range, which makes the application of deep learning more difficult. It is very difficult to find a wide range of patients in electronic medical records. For such a long time series of information, it is very difficult to confirm the mapping relationship between symptoms and seizures.

Genomics

Genomics studies the function, structure, editing, and performance of genes. Because of its powerful ability to process data and automatic feature extraction, many researchers have applied it to the field of genomics to discover deeper patterns.

Compared with traditional machine learning methods, the deep learning methods can extract the higher dimensional features, richer information, and more complex structure from biological data. In recent years, deep learning has been widely used in genomics, such as gene expression, gene slicing, RNA measurement, and other tasks. Deep learning brings new methods to bioinformatics and helps to understand the principles of human diseases further.

In genomics, deep learning can effectively understand the cause and development process of diseases from the molecular level, and the interaction between genes and environment, and understand the factors leading to disease. The deep learning method can capture the relationship between disease and gene from the high-throughput biological dataset. Doctors can understand the disease more comprehensively, make accurate decisions, and provide patients with more appropriate treatment and diagnosis. The application of deep learning in genomics has greatly promoted the development of personalized therapy and precision medicine. A model predicting enhancer-promoter interactions is shown in **Figure 12**.

Several studies have demonstrated the effectiveness of deep learning applications on genomics tasks better than traditional machine learning methods. Many researchers have researched the task of gene expression using the deep learning method. For example, Singh et al. (2016) proposed a deep learning model to automatically extract complex histone modifications between essential functions to classify gene expression. Chen Y. et al., 2016 proposed a multitask multilayer neural network to infer the expression of the target gene from the expression of the marker gene. In terms of the average absolute error of all genes, deep learning is better than linear regression, and the performance is improved by 15.33%. Tan et al. (2016) used the gene expression denoising autoencoder to integrate various gene expression data and capture patterns corresponding to biological states. Singh et al. (2017) proposed a deep learning method to discover the interaction between each chromatin marker signal. Badsha et al. (2020) developed the method to understand the dependent structures between genes stored in parameter estimates. Kong and Yu (2018) integrated the external relationship information of gene expression features into the deep neural network. The application of real data proved the practicability of the new model in classification and biological feature selection. Mostavi et al. (2020) proposed three unique convolutional neural network architectures for different data formats. These architectures used high-dimensional gene expression inputs and predicted cancer types while considering their origin tissues.

In addition, other researchers also used deep learning for other genomic tasks, such as predicting the binding sites of RNA-binding proteins (Zhang et al., 2016), predicting DNA methylation status (Angermueller et al., 2017), predicting enhancer–promoter interactions (Singh et al., 2019), and predicting mRNA abundance (Washburn et al., 2019; Agarwal and Shendure, 2020). Besides, Kelley et al. (2016) used the convolutional neural networks to understand the functional activity of DNA sequences from genomic data. Qin and Feng

(2017) developed the model to predict cell-specific transcription factor binding based on the available ChIP-seq data. Compared with the existing methods, the model can achieve considerable accuracy in the combination of transcription factors and cell lines with ChIP-seq data. Jha et al. (2017) used an autoencoder to integrate other types of experimental data into the spliced code model for selective splicing. Chen Y. et al. (2016) used an autoencoder to more than 1,000 yeast microarrays to understand the coding system of the yeast transcriptome mechanism and the hierarchical organization of the yeast transcriptome mechanism. Zhou et al. (2018) proposed a deep convolution neural network model that predicts tissue-specific transcriptional effects of mutations, including rare or undetected mutations.

Researchers also used natural language processing technology to represent gene sequence information. Zou et al. (2019) used word embedding technology to represent mRNA subsequences. Then, they used the convolutional neural networks to predict N6-methyladenosine from mRNA sequence.

Some researchers also use the convolutional neural networks for other tasks. Gao et al. (2020) used convolutional neural networks to analyze the global activity of splicing modified compounds and identify new therapeutic targets. Experiments have shown that the deep learning methods could recognize the sequence features and predict the response to drug regulation. Yuan and Bar-Joseph (2019) proposed the convolutional neural network model to infer the relationship between the different expression levels encoded in the genes. In many aspects of performance, convolutional neural network was better than the previous methods.

Researchers have also used several deep learning networks such as convolutional neural networks, recurrent neural networks, autoencoders, and deep belief networks in genomics. For example, in tasks involving DNA sequences and protein sequences, one-dimensional convolutional neural networks and recurrent neural networks are the two most commonly used networks. The one-dimensional convolutional neural network extracts high-level features from the genome sequence data through the movement of the convolution window. The recurrent neural networks can effectively process the information about the sequence and discover specific patterns from the genome sequence data through its memory characteristics. Also, the methods used in natural language processing are also applied to the field of genomics, which provides innovative ideas and methods for the study of genomic data. To make full use of the information in the data, multimodal learning using different data sources is also one of the methods used by researchers. In conclusion, deep learning offers the possibility of mining features from genomic data to help develop a deeper understanding of the disease.

Although deep learning has made impressive achievements in genomics, some problems still exist. The problems of deep learning in genomics are as follows:

(1) Deep learning training usually requires a large number of datasets, and the quality of these datasets is required to be high so that the deep learning model can learn distinguishing features and patterns from the data. However, data insufficiency still exists in genomics, so the model cannot learn from sufficient data and cannot provide key information for researchers or doctors.
There are many parameters in the deep learning model, and there are a lot of deep learning frameworks. If the amount of data is not enough, it is difficult for researchers to find a deep learning model suitable for the current task. It is easy to see that the model has poor performance in the dataset. Because the deep learning model only remembers the characteristics and patterns of the data in the training set but does not learn the deep relationship between them from the data.
Generally speaking, there are three measures to solve the problem of insufficient data: The first measure is to expand the dataset from the data layer, such as using data enhancement technology to increase the sample size; the second measure is to use regularization and dropout methods to improve the performance of the deep learning model and increase the generalization ability of the model; the third measure is to use the transfer learning technology to train the model on an unrelated but large number of datasets, and then the trained model is used in the task of their concern, and the parameters of the model are adjusted to get the suitable model. In the above studies, we can see that some researchers have used these techniques to solve the problem of insufficient data in the field of genomics, thus improving the model effect.
(2) Biomedical data are complex and need professional domain knowledge to analyze. Unlike other fields, in genomics, the structure and function of the genome are a very complex model, which puts forward higher requirements for the interpretability of the deep learning model. In recent years, the multimodal learning has become an attempt to improve the interpretability and accuracy of the model. The so-called multimodal learning refers to the combination of data from different input sources and the establishment of different types of deep learning models for different types of data to make full use of the relationship and characteristics of different types of data, so the model can make a more comprehensive and accurate prediction. Many researchers have begun to combine different types of data, such as gene sequences with other types of data such as electronic health records and medical imaging, to expand the field of knowledge and provide more insights for doctors.

Drug Development

In recent years, with the rapid growth of biomedical data, deep learning technology has become a new method in drug development. The application of deep learning in the field of drug development can help researchers effectively carry out drug development and disease treatment research and greatly promote the development of precision medicine.

Drug development is a complex process. Traditionally, it takes at least 10 years from the development of a new drug to its marketing, which is a very long and resource-consuming process. Traditional drug development is divided into two methods: one

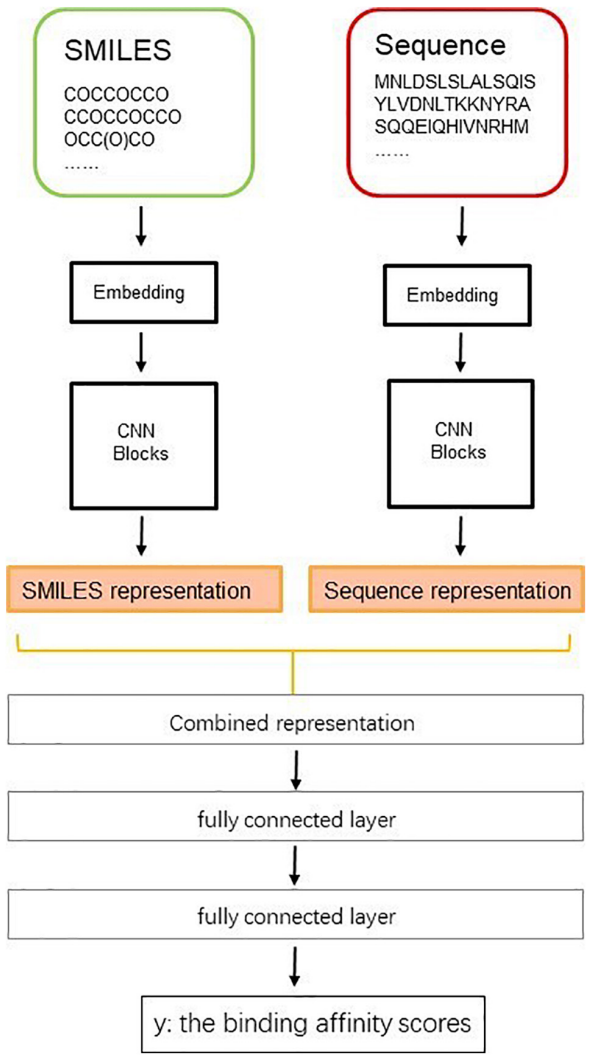


FIGURE 13 | Diagram of deep learning model to predict the binding affinity scores of drugs–targets.

is the experimental method, which not only consumes time and efficiency but also causes huge cost; the other is the computational method, which can save time and reduce loss.

In drug development and design, identifying the interaction between drug and target is an important step. This step can save resources and accelerate the time from drug development to market. In recent years, the biomedical data have increased significantly, which provides a data basis for applying deep learning in drug development. Many researchers have begun to apply deep learning to explore the relationship between drugs and targets. **Figure 13** shows a diagram of a deep learning model predicting the binding affinity scores of drug–targets.

Many researchers have begun to apply deep learning to explore the relationship between drugs and targets. They have achieved good results, demonstrating the great ability of deep learning in the field of drug development. For example, Wen et al. (2017) used deep belief networks to predict the interaction between drugs and targets. The model automatically extracted the characteristics of drugs and targets from the simple chemical substructure and sequence information in an unsupervised way. Wang et al. (2017) predicted potential unknown drug–target interactions from drug molecular structures and protein sequences. They used a stacked autoencoder to learn useful information of protein sequences automatically. Öztürk et al. (2018) used two convolutional neural network frameworks to learn features from protein sequences and compounds' SMILES strings and then combined the learned features into a fully connected layer to predict the interaction between drugs and targets binding affinity. Zong et al. (2019) used the deep learning methods to calculate the vertex similarity and then input the similarity into two rule-based inference methods. Hu S. et al., 2019 represented each drug target pair by connecting the coding vector of the target descriptor and the drug descriptor. Then, they inputted the representations into the convolutional neural network. They evaluated the ability of the model on the DrugBank dataset, with an accuracy of 0.88 and an area under the curve value of 0.95. The results show that the model can be used to distinguish drug and target interactions.

Other investigators have also made efforts to understand drug–target interaction. Hu P. et al., 2019 used the autoencoder with a cascade structure to obtain the deep expression of the fusion network and identify the drug–target interaction. Wang et al. (2020) combined evolutionary characteristics of proteins with drug molecular structure fingerprints to form the feature vector of drug–target pairs. Then, they used LSTM on the feature space to predict the interaction between drug and target. Zeng et al. (2020) developed a deep learning model for target recognition and drug reuse by learning low dimensional vector representations of drugs and targets. Zhao et al. (2020) used two generative adversarial networks to calculate binding affinity between drug and target in an unsupervised way. Then, a convolution neural network is used for prediction. The experimental results show that the proposed method can make full use of the unlabeled data and obtain competitive performance.

In addition to using the deep learning method to predict the interaction between drugs and targets, some researchers also use deep learning to design molecules from scratch, predict the pharmacological properties and synergistic effects of drugs, etc. For example, Aliper et al. (2016) proposed the deep learning model to map transcriptome data to therapeutic categories. In the experiments, the model achieved high classification accuracy and is better than the support vector machine model. Gupta et al. (2018) used the deep learning method to design molecular weight, which captured the semantics of molecular representation and carried out a virtual compound design. Preuer et al. (2018) proposed DeepSynergy to predict the synergistic effect of anticancer drugs. Experiments showed that the model can explore the synergistic effect of drugs and novel combinations in cell line space with high precision. Zhang X. et al., 2018 learned the representation of each molecule in an unsupervised way. Zeng et al. (2019) developed a model that learns advanced features of drugs from different networks through a multimode autoencoder.

Representation learning techniques also have applications in drug development. By combining unsupervised representation learning with deep learning techniques, Wan et al. (2019) used the word2vec to learn the low dimensional representation of compounds and proteins to predict the interaction between unstructured compounds and proteins. Zhang et al. (2020) used deep learning methods to represent molecules as vectors to identify potential drugs, peptides, or small ligands targeting protein targets in the 2019-nCoV virus. The model had high speed and high accuracy, which was suitable for screening thousands of drugs in a short time in some emergencies. Karimi et al. (2020) developed the deep generation model for drug combination design. They used hierarchical variational graph autoencoders to jointly embed gene–gene, gene–disease, and disease–disease networks.

As one of the state-of-the-art methods in artificial intelligence, deep learning provides an important opportunity for drug development. It not only saves drug development time and human resources but also enables the efficient mining of information that is difficult for people. Although modern medical careers progress rapidly nowadays, there are still diseases for which it is still difficult to find drugs to treat them. The deep learning method for drug development has provided a possible breakthrough in finding drugs to treat these diseases. It can be seen that in drug development, researchers used convolutional neural networks, recurrent neural networks, autoencoders, and fully connected neural networks. To deal with the small labeled datasets, there are also studies using unsupervised learning methods for application in the field of drug development. For example, several researches use unsupervised methods to predict the interaction of drug targets and predict the binding affinity of drug targets. In unsupervised learning, autoencoders and their variants, generative adversarial networks, and deep belief networks are commonly used neural networks. They can recover data well or find advanced features from the data for the deeper application.

Although deep learning has broad prospects in drug development, limitations still exist:

(1) It is known that deep learning requires a lot of labeled data. But in the field of drug development, the labeled data are limited. At present, many researchers use semisupervised learning or unsupervised learning to find key information from unlabeled biomedical data. However, even if semisupervised learning or unsupervised learning is used, it is difficult to find useful information in these unlabeled data.
(2) For deep learning, it is difficult to scientifically explain the reasons for making predictions because the occurrence and process of disease are a very complex biomedical field. Deep learning is considered as a "black box" method. If deep learning cannot provide a good explanation, it will be difficult for doctors to believe the prediction results given by deep learning, so they cannot make decisions. It is possible to integrate other types of data and information into drug development to solve the problem of model interpretability.

Deep Learning Research on Longitudinal Datasets

In this section, we introduce the application of deep learning in longitudinal datasets. Longitudinal data track and record the patient's long-term condition. Using longitudinal datasets, we can perform tasks such as predicting disease-related risks, predicting the trajectory of relevant biomarkers at different disease development stages, or conducting survival analysis. Some longitudinal data studies, such as Framingham Heart Study (Araki et al., 2016) or the UK Cystic Fibrosis Registry (Taylor-Robinson et al., 2018), provide useful datasets for longitudinal data research.

The patient's medical history may contain some information about the future disease, so it is very useful to study the history and infer the future disease development from the past information. It requires doctors to make a prediction as soon as possible so that doctors can take measures to prevent the trend of disease onset or deterioration. However, the use of longitudinal data for research will also have difficulties. For example, for predicting disease trajectory, for a patient, his disease status may develop slowly, which increases the difficulty of related research. For example, a patient with a chronic disease such as diabetes may have different conditions over time. How to apply appropriate deep learning to longitudinal datasets has become a research direction.

Because of its memory function and ability to remember the information of data, recurrent neural networks such as LSTM and GRU have become the main methods to process longitudinal datasets. In general, some researchers use joint models to predict disease trajectories over time using longitudinal and time–event datasets. However, it can also reduce the accuracy of the model by applying it to large datasets. To solve this problem, Lim and van der Schaar (2018) developed a joint framework using recurrent neural networks to capture the relationship between disease trajectories using shared representations. Lee et al. (2020) combined recurrent neural networks with the attention mechanism to learn the complex relationship between the trajectory and survival probability and learn the distribution of time–event without making any assumptions about the stochastic models of longitudinal and time–event processes. Beaulieu-Jones et al. (2018) used autoencoder to represent patient care events in low dimensional vector space. They then used LSTM to predict survival rates from the sequence of nursing events learned. Lee et al. (2019) used various types of data related to Alzheimer disease to predict mild cognitive impairment. They proposed a model that used the deep learning method to learn task-related feature representation from data.

Table 3 shows some deep learning models used in computational medicine.

PROBLEMS AND CHALLENGES

Although deep learning has achieved better results than machine learning in the medical and health field, there are still some challenges and problems. Here, we highlight the following problems and challenges and explore some solutions to them.

TABLE 3 | Some deep learning models used in computational medicine.

Domain	Model	Brief introduction	URL
	Breast cancer type classification (Rakhlin et al., 2018)	A simple and effective method for the classification of hematoxylin and eosin–stained histological breast cancer images	https://github.com/alexander-rakhlin/ICIAR2018
Clinical image	DeepKnee (Tiulpin et al., 2018)	An automatic pipeline for osteoarthritis severity assessment from plain radiographs	https://github.com/MIPT-Oulu/DeepKnee
	Robotic instrument segmentation (Shvets et al., 2018)	A robotic instrument segmentation approach based on the deep learning network architecture	https://github.com/ternaus/robot-surgery-segmentation
	Segmentation of the left ventricle (Avendi et al., 2016)	An automatic segmentation approach of the left ventricle using deep learning and deformable model	https://github.com/alexattia/Medical-Image-Analysis
Electronic health record	Embeddings (Choi Y. et al., 2016)	A deep learning method that learns low-dimensional representations of concepts in medicine	https://github.com/clinicalml/embeddings
	Med2Vec (Choi Y. et al., 2016c)	A representation learning model for learning code representations and visit representations	https://github.com/mp2893/med2vec
	Doctor AI (Choi Y. et al., 2016b)	An automatic diagnosis machine that predicts medical codes	https://github.com/pckuo/doctorai
	Patient2Vec (Zhang X. et al., 2018a)	A deep learning method that learns an interpretable deep representation of longitudinal electronic health records data	https://github.com/BarnesLab/Patient2Vec
	DeepCare (Pham et al., 2016)	A deep learning model that reads electronic health record data and infers disease progression and predicts future outcome	https://github.com/trangptm/DeepCare
	GRU-D (Che et al., 2017)	Captures the informative missingness	https://github.com/fteufel/PyTorch-GRU-D
Genomics	DeepChrome (Singh et al., 2016)	A deep learning framework that learns combinatorial interactions among histone modification marks to predict the gene expression	https://github.com/QData/DeepChrome
	D-GEX (Chen Y. et al., 2016b)	A deep learning method that infers the expression of the target gene from the expression of the marker gene	https://github.com/uci-cbcl/D-GEX
	ADAGE (Tan et al., 2016)	Analysis using Denoising Autoencoders for Gene Expression	https://github.com/greenelab/adage
	AttentiveChrome (Singh et al., 2017)	A unified architecture that models and interprets dependencies among chromatin factors for controlling gene regulation	https://github.com/QData/AttentiveChrome
	GEDFN (Kong and Yu, 2018)	A deep learning classifier embedding feature graph information	https://github.com/yunchuankong/GEDFN
	CancerTypePrediction (Mostavi et al., 2020)	A model that uses gene expression inputs and predicts cancer types	https://github.com/chenlabgccri/CancerTypePrediction
	Deepnet-RBQ (Zhang et al., 2016)	A multimodal deep belief network that predicts the target sites of RNA-binding proteins	https://github.com/thucombio/deepnet-rbp
	DeepCpG (Angermueller et al., 2017)	A model for predicting the methylation state of CpG dinucleotides in multiple cells	https://github.com/PMBio/deepcpg
	SPEID (Singh et al., 2019)	A deep neural network for predicting enhancer–promoter interactions from sequence data	https://github.com/ma-compbio/SPEID
	Xpresso (Agarwal and Shendure, 2020)	Deep learning models for predicting gene expression levels from genomic sequence	https://github.com/vagarwal87/Xpresso
	Basset (Kelley et al., 2016)	A tool for learning highly accurate models of DNA sequence activity	https://github.com/davek44/Basset
	Integrative deep models for alternative splicing (Jha et al., 2017)	Deep learning models for alternative splicing	https://majiq.biociphers.org/jha_et_al_2017/
	ExPecto (Zhou et al., 2018)	A deep learning framework for predicting expression effects of human genome variants *ab initio* from sequence	https://github.com/FunctionLab/ExPecto
	Gene2vec (Zou et al., 2019)	A deep learning neural embedding for prediction of mammalian N6-methyladenosine sites	http://server.malab.cn/Gene2vec/

(Continued)

TABLE 3 | Continued

Domain	Model	Brief introduction	URL
	CNNC (Yuan and Bar-Joseph, 2019)	A deep learning method for inferring gene relationships from single-cell expression data	https://github.com/xiaoyeye/CNNC
Drug development	DeepDTIs (Wen et al., 2017)	A deep belief network for predicting the interaction between drugs and targets	https://github.com/Bjoux2/DeepDTIs
	DeepDTA (Öztürk et al., 2018)	The convolutional neural networks for predicting the binding affinity value of drug–target pairs	https://github.com/hkmztrk/DeepDTA
	deepDTnet (Zeng et al., 2020)	A deep learning method for predicting drug–target interactions.	https://github.com/ChengF-Lab/deepDTnet
	MLP (Aliper et al., 2016)	A deep learning model that predicts pharmacological properties of drugs and drug repurposing	https://github.com/alvarouc/mlp
	DeepSynergy (Preuer et al., 2018)	A deep learning approach for predicting the synergy of drug combinations	http://www.bioinf.jku.at/software/DeepSynergy/
	deepDR (Zeng et al., 2019)	A deep learning approach for inferring new drug–disease relationships for in silicon drug repurposing	https://github.com/ChengF-Lab/deepDR
	DeepCPI (Wan et al., 2019)	A deep learning framework for large-scale *in silico* drug screening	https://github.com/FangpingWan/DeepCPI
	Drug-Combo-Generator (Karimi et al., 2020)	Deep generative models for drug combination generation	https://github.com/Shen-Lab/Drug-Combo-Generator

Data Insufficiency

Deep learning is a data-driven approach. Generally, there are many parameters in the neural networks, which need to be learned, updated, and optimized from the data. With the advent of the era of big data, sufficient data provide a data basis for the development of deep learning. Therefore, deep learning has achieved great success in many data fields, such as image recognition, natural language processing, and computer vision. However, in the field of health care, medical datasets are usually limited and biased. Because the number of health samples is far more than the number of disease cases, or the number of images in each category is uneven, the application of deep learning in this field is difficult.

Insufficient data will limit the parameter optimization of deep learning and lead to the problem of overfitting. The performance of the learned model on the training set is good, but the performance on the data that have never been seen is very poor. The generalization ability of the model is poor. Usually, dropout and regularization are two common methods to solve overfitting. Besides, an increasing dataset is also a common means to suppress overfitting. In clinical imaging, data enhancement is a method to expand datasets. The data enhancement is to use translation, rotation, clipping, scaling, changing contrast, and other methods to generate new images. For example, Rakhlin et al. (2018) used data enhancement on a small breast cancer histological image dataset to improve the robustness of the model.

Transfer learning is also an effective way to solve the problem of insufficient data. Transfer learning means that the model first learns on a task with sufficient data and then applies the learned model to another related task and then fine-tunes the model parameters. For example, it can transfer the neural network model trained on a large number of natural images to the task of small sample medical images. Many experiments have proven that transfer learning is an effective method. For example, Shin et al. (2016) have proven that the transfer learning using datasets from large-scale annotated natural image datasets (ImageNet) to computer-aided diagnosis datasets is beneficial in experiments.

In recent years, multimodal learning has become a trend to solve this problem. Multimodal learning can learn different types of data simultaneously, which uses different types of data as input, such as electronic health records, medical images, and genomic data. According to the characteristics of different types of data, different models are developed. Finally, the information is integrated to provide the model ability. For example, Dai et al. (2018) integrated clinical reports with fundus images to detect retinal microaneurysms. Combining expert domain knowledge and image information solves the problem of training neural networks under extremely unbalanced data distribution.

Model Interpretability

Deep learning is often considered as a "black box" because of its lack of explaining ability. In some areas, such as image recognition, the lack of interpretability may not be a big problem. Still, in the field of health care, the interpretability of models is very important. Because if a model can provide sufficient and reliable information, doctors will trust the results of the model and make correct and appropriate decisions; at the same time, an interpretable model can also provide a comprehensive understanding for patients.

Some researchers have been exploring the interpretive problems of neural networks. Lanchantin et al. (2016) proposed an optimization strategy to extract features or patterns to visualize gene sequence classification. In the form of statistical physics, Finnegan and Song (2017) extracted and interpreted the features of sequences learned by the network. Choi E. et al., 2016 used the attention model to detect the part of electronic health records that affected the predictive ability of the model.

Privacy and Ethical Issues

Data privacy is an important aspect in the medical and health field. The improper use, misuse, and even abuse of patient data

will bring disastrous consequences. As we know, deep learning training requires a large number of representative datasets. These datasets are very valuable, but they can be very sensitive.

At present, in computational medicine, many researchers have developed and publicly shared their deep learning models for others to use. There are many parameters in the deep learning model, which may contain sensitive information of data. Some people with ulterior motives may design some ways carefully to attack deep learning models. They can infer these parameters from the deep learning model and even infer the sensitive information in the dataset, thus violating the privacy of the model and patients. Phong et al. (2018) pointed out that even a small part of the gradient stored on cloud services can cause information leakage of local data. So, they use a homomorphic encryption mechanism to solve the problem of information leakage. Hitaj et al. (2017) pointed out that the generation of countermeasures network can recover the information in the data. Shokri and Shmatikov (2015) pointed out that the parameters of neural networks might leak the information of the training set when training neural networks.

In recent years, some researchers have explored the safety of deep learning. For example, Abadi et al. (2016) proposed a differential privacy random gradient descent algorithm by combining deep learning with differential privacy. By developing new technologies, they improved the computational efficiency of differential privacy in deep learning. These technologies include an efficient gradient algorithm, a differential privacy projection in the input layer, and so on. Their method is universal and can be applied to many classical optimization algorithms to solve the privacy problem in deep learning. Lecuyer et al. (2019) proposed PixelDP, which is the first certified defense and can be extended to large networks and datasets and widely applied to any type of deep learning model. They did a quantitative analysis for the robustness of the deep neural network to counter samples and achieved a good defense effect.

Due to the explosive growth of data, some users will put their data on the cloud, which brings challenges for deep learning to cloud computing the data provided by different data owners. In order to solve the privacy problem of collaborative deep learning in cloud computing, Li et al. (2017) provided two schemes to protect the privacy of deep learning. One is the basic scheme, which is based on multi-key fully homomorphic encryption; the other is the advanced scheme, which combines double decryption mechanism with fully homomorphic encryption. They have proven that the two schemes are secure in maintaining multi-key privacy on encrypted data. Yuan et al. (2019) used the differential privacy method to add Gaussian noise to the shared parameters to solve the problem of privacy leakage caused by shared parameters.

Patient data contains very sensitive information, which brings challenges to the application and development of deep learning in the field of computational medicine, resulting in a vicious circle. A hospital or researcher has a huge amount of patient privacy information, once the information is leaked, it will cause incalculable loss and bad influence. On the contrary, hospitals and researchers are unwilling to disclose their patient information and data because they are worried about the risk of data leakage, which will lead to the problem that deep learning cannot take advantage of large-scale data.

Because of the privacy of patients' data, the sharing of medical data has become a very complex and difficult problem. It involves not only moral and legal issues but also technical issues. Some countries have adopted laws and regulations to regulate the use of user sensitive information by organizations. For example, on May 25, 2018, the European Union(EU) issued a very strict privacy protection regulation, the General Data Protection Regulation (GDPR). The application scope of the regulation is very wide. Any organization dealing with the data of EU users must comply with the regulation. Anyone who violates the regulation will face a great degree of punishment. GDPR defines user data as personal identifier information. As long as the data can locate users, it is considered as personal identifier information, and this data must be strictly protected. The promulgation of GDPR is a start. With the increasing importance of personal privacy today, other countries or regions will also issue similar policies to protect people's privacy. These policies will bring a profound impact on the field of artificial intelligence driven by big data.

How to make full use of the advantages of big data to promote the development of deep learning under the premise of protecting patients' privacy data is a problem that must be considered in the application of deep learning, that is, artificial intelligence in the field of computational medicine.

Heterogeneity

The data in the field of health care are full of heterogeneity. There are both unstructured data and structured data. In addition, the data in the field of health care are noisy, high-dimensional, and of low quality.

Because of the existence of these heterogeneous data, it is difficult to find a suitable deep learning model. We know that the input data of neural networks must be processed and converted into a numerical value. How to properly preprocess the structured and unstructured biomedical data is a problem that researchers should first consider when training the neural networks. Therefore, processing these data is also one of the challenges faced by applying deep learning in the medical field. Some researchers have explored the processing of medical imaging datasets. For example, Li et al. (2020) intensively annotated medical images based on anatomical and pathological features. Their research expanded the label of the dataset and effectively solved the problem of small data sample size. The results show that the algorithms trained on the densely annotated medical imaging datasets they used have significantly higher diagnostic accuracy.

CONCLUSION

We surveyed the application of deep learning in computational medicine such as clinical imaging, electronic health record, genomics, and drug development. Using deep learning to process big biomedical data can mine the information in the data to better provide guidance for doctors and improve the level of medical health. At the same time, this article also points out

the problems and challenges in the application of deep learning in computational medicine. It provides a reference and way to improve the application of deep learning in the medical and health field in the future.

AUTHOR CONTRIBUTIONS

XL conceived and designed the survey. SY analyzed the data and contributed reagents, materials, and analysis tools. All authors wrote the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61303108), the Natural Science Foundation of Jiangsu Higher Education Institutions of China (17KJA520004), Program of the Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) (KJS1524), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., et al. (2016). *Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. New York City, NY: ACM Inc., 308–318. doi: 10.1145/2976749.2978318

Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Trans. Audio Speech Lang. Proc.* 22, 1533–1545. doi: 10.1109/TASLP.2014.2339736

Agarwal, V., and Shendure, J. (2020). Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell. Rep.* 31:107663. doi: 10.1016/j.celrep.2020.107663

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharmaceutics* 13, 2524–2530. doi: 10.1021/acs.molpharmaceut.6b00248

Andermatt, S., Pezold, S., and Cattin, P. C. (2016). *Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-Data. International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. New York City, NY: Springer, 142–151. doi: 10.1007/978-3-319-46976-8_15

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

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

Araki, T., Putman, R. K., Hatabu, H., Gao, W., Dupuis, J., Latourelle, J. C., et al. (2016). Development and progression of interstitial lung abnormalities in the framingham heart study. *Am. J. Resp. Critical Care Med.* 194, 1514–1522. doi: 10.1164/rccm.201512-2523oc

Avendi, M. R., Kheradvar, A., and Jafarkhani, H. (2016). A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med. Image Anal.* 30, 108–119. doi: 10.1016/j.media.2016.01.005

Badsha, M. B., Li, R., Liu, B., Li, Y. I., Xian, M., Banovich, N. E., et al. (2020). Imputation of single-cell gene expression with an autoencoder neural network. *Quantit. Biol.* 8, 78–94. doi: 10.1007/s40484-019-0192-7

Beaulieu-Jones, B. K., Orzechowski, P., and Moore, J. H. (2018). Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database. *Pac. Symp. Biocomput.* 23, 123–132. doi: 10.1142/9789813235533_0012

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2006). Greedy layer-wise training of deep networks. *Adv. Neural Inform. Proc. Syst.* 19, 153–160.

Cao, Y., Liu, C., Liu, B., Brunette, M. J., Zhang, N., Sun, T., et al. (2016). *Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. In 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. New York, NY: IEEE, 274–281. doi: 10.1109/CHASE.2016.18

Chaurasia, A., and Culurciello, E. (2017). *LinkNet: Exploiting encoder representations for efficient semantic segmentation. In 2017 IEEE Visual Communications and Image Processing (VCIP)*. New York, NY: IEEE, 1–4. doi: 10.1109/VCIP.2017.8305148

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2017). Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* 8, 6085–6085. doi: 10.1038/s41598-018-24271-9

Chen, L., Cai, C., Chen, V., and Lu, X. (2016). Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinform.* 17:9. doi: 10.1186/s12859-015-0852-1

Chen, Y., Li, Y., Narayan, R., Subramanian, A., and Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics* 32, 1832–1839. doi: 10.1093/bioinformatics/btw074

Cheng, J. Z., Ni, D., Chou, Y. H., Qin, J., Tiu, C. M., Chang, Y. C., et al. (2016). Computer-Aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* 6, 24454–24454. doi: 10.1038/srep24454

Cheng, Y., Wang, F., Zhang, P., and Hu, J. (2016). *Risk prediction with electronic health records: a deep learning approach. In Proceedings of the 2016 SIAM International Conference on Data Mining*. Philadelphia, USA: SIAM, 432–440. doi: 10.1137/1.9781611974348.49

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., and Sun, J. (2016a). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Adv. Neural Inform. Proc. Syst.* 2016, 3504–3512.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016b). *Doctor AI: Predicting clinical events via recurrent neural networks. Proceedings of the 1st Machine Learning for Healthcare Conference.* 301–318.

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., et al. (2016c). *Multi-layer representation learning for medical concepts. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: KDD '16, 1495–1504.

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. (2016). *ArXiv*. [Preprint].

Choi, Y., Chiu, C. Y.-I., and Sontag, D. A. (2016). Learning low-dimensional representations of medical concepts. *AMIA Joint Summits Transl. Sci. Proc.* 2016, 41–50.

Chung, J., Gülçehre, Ç, Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*,

Çiçek, Ö, Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science*, eds S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells (Cham: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Cicero, M., Bilbily, A., Colak, E., Dowdell, T., Gray, B., Perampaladas, K., et al. (2017). Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investigat. Radiol.* 52, 281–287. doi: 10.1097/rli.0000000000000341

Dai, L., Fang, R., Li, H., Hou, X., Sheng, B., Wu, Q., et al. (2018). Clinical report guided retinal microaneurysm detection with multi-sieving deep learning. *IEEE Trans. Med. Imag.* 37, 1149–1161. doi: 10.1109/tmi.2018.2794988

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159.

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

Fauw, J. D., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* 24, 1342–1350. doi: 10.1038/s41591-018-0107-6

Finnegan, A., and Song, J. S. (2017). Maximum entropy methods for extracting the learned features of deep neural networks. *PLoS Comput. Biol.* 13:e1005836. doi: 10.1371/journal.pcbi.1005836

Gao, D., Morini, E., Salani, M., Krauson, A. J., Ragavendran, A., Erdin, S., et al. (2020). A deep learning approach to identify new gene targets of a novel therapeutic for human splicing disorders. *BioRxiv*,

Grinsven, M. J. J. P., van Ginneken, B., van, Hoyng, C. B., Theelen, T., and Sanchez, C. I. (2016). Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images. *IEEE Trans. Med. Imag.* 35, 1273–1284. doi: 10.1109/tmi.2016.2526689

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., and Schneider, G. (2018). Generative recurrent networks for de novo drug design. *Mol. Inform.* 37:1700111. doi: 10.1002/minf.201700111

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 29, 1836–1842.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). *Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* . New York, NY: IEEE, 770–778.

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

Hitaj, B., Ateniese, G., and Perez-Cruz, F. (2017). *Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. New York City, NY: ACM Inc., 603–618.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.

Hu, C., Ju, R., Shen, Y., Zhou, P., and Li, Q. (2016). *Clinical decision support for Alzheimer's disease based on deep learning and brain network. In 2016 IEEE International Conference on Communications (ICC).* New York, NY: IEEE, 1–6.

Hu, P., Huang, Y., You, Z., Li, S., Chan, K. C. C., Leung, H., et al. (2019). "Learning from deep representations of multiple networks for predicting drug–target interactions," in *Intelligent Computing Theories and Application. ICIC 2019. Lecture Notes in Computer Science*, Vol. 11644, eds D. S. Huang, K. H. Jo, and Z. K. Huang (Cham: Springer), 151–161. doi: 10.1007/978-3-030-26969-2_14

Hu, S., Xia, D., Su, B., Chen, P., Wang, B., and Li, J. (2019). *A convolutional neural network system to discriminate drug-target interactions. IEEE/ACM Transactions on Computational Biology and Bioinformatics*. New York, NY: IEEE.

ICD10 (2033). Available online at: https://www.cdc.gov/nchs/icd/icd10.htm (accessed September 10, 2020).

ICD9 (2032). Available online at: https://www.cdc.gov/nchs/icd/icd9.htm (accessed September 10, 2020).

Iglovikov, V., and Shvets, A. (2018). Ternausnet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation. *ArXiv*, [Preprint].

Iglovikov, V. I., Rakhlin, A., Kalinin, A. A., and Shvets, A. A. (2018). "Paediatric bone age assessment using deep convolutional neural networks: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, granada, spain, september 20, 2018, proceedings," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision*, (New York City, NY: Springer International Publishing), 300–308. doi: 10.1007/978-3-030-00889-5_34

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* 13, 395–405. doi: 10.1038/nrg3208

Jha, A., Gazzara, M. R., and Barash, Y. (2017). Integrative deep models for alternative splicing. *Bioinformatics* 33, 274–282.

Jiménez-Sánchez, A., Albarqouni, S., and Mateus, D. (2018). *Capsule networks against medical imaging data challenges*. Cham: Springer, 150–160.

Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., et al. (2018). *Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. New York, NY: IEEE, 691–699.

Karimi, M., Hasanzadeh, A., and Shen, Y. (2020). Network-principled deep generative models for designing drug combinations as graph sets. *Bioinformatics* 36, 445–454.

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genom. Res.* 26, 990–999. doi: 10.1101/gr.200535.115

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131. doi: 10.1016/j.cell.2018.02.010

Kingma, D. P., and Ba, J. L. (2015). *Adam: A method for stochastic optimization. In ICLR 2015?: International Conference on Learning Representations*. La Jolla: ICLR.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K. H., Bendszus, M., et al. (2016). Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. *NeuroImage* 129, 460–469. doi: 10.1016/j.neuroimage.2016.01.024

Kong, Y., and Yu, T. (2018). A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics* 34, 3727–3737. doi: 10.1093/bioinformatics/bty429

Kooi, T., Litjens, G. J. S., Ginneken, B., van Gubern-Mérida, A., Sánchez, C. I., Mann, R., et al. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* 35, 303–312. doi: 10.1016/j.media.2016.07.007

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). *ALBERT: A lite bert for self-supervised learning of language representations. In ICLR 2020?: Eighth International Conference on Learning Representations*. La Jolla: ICLR.

Lanchantin, J., Singh, R., Lin, Z., and Qi, Y. (2016). *Deep Motif: Visualizing genomic sequence classifications*. New Orleans, LA: ICLR.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2019). *Certified robustness to adversarial examples with differential privacy. In 2019 IEEE Symposium on Security and Privacy (SP)*. New York: IEEE, 656–672.

Lee, C., Yoon, J., and van der Schaar, M. (2020). Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.* 67, 122–133. doi: 10.1109/tbme.2019.2909027

Lee, G., Kang, B., Nho, K., Sohn, K.-A., and Kim, D. (2019). MildInt: deep learning-based multimodal longitudinal data integration framework. *Front. Genet.* 10:617. doi: 10.3389/fgene.2019.00617

Li, P., Li, J., Huang, Z., Li, T., Gao, C.-Z., Yiu, S.-M., et al. (2017). Multi-key privacy-preserving deep learning in cloud computing. *Future Gen. Comp. Syst.* 74, 76–85. doi: 10.1016/j.future.2017.02.006

Li, W., Yang, Y., Zhang, K., Long, E., He, L., Zhang, L., et al. (2020). Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. *Nat. Biomed. Eng.* 4, 767–777. doi: 10.1038/s41551-020-0577-y

Lim, B., and van der Schaar, M. (2018). Disease-atlas: navigating disease trajectories using deep learning. *Mach. Learn. Healthcare Confer.* 2018, 137–160.

Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., and Pu, J. (2019). SDFN: segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Comp. Med. Imag. Graphics* 75, 66–73. doi: 10.1016/j.compmedimag.2019.05.005

Ma, T., Xiao, C., and Wang, F. (2018). *Health-ATM: A deep architecture for multifaceted patient health record representation and risk prediction. In Proceedings of the 2018 SIAM International Conference on Data Mining (SDM)*. Philadelphia, USA: SIAM, 261–269.

Mansoor, A., Cerrolaza, J. J., Idrees, R., Biggs, E., Alsharid, M. A., Avery, R. A., et al. (2016). Deep learning guided partitioned shape model for anterior visual

pathway segmentation. *IEEE Trans. Med.* 35, 1856–1865. doi: 10.1109/tmi.2016.2535222

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). *Efficient estimation of word representations in vector space*. La Jolla,CA: ICLR.

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

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings Bioinform.* 19, 1236–1246. doi: 10.1093/bib/bbx044

Mostavi, M., Chiu, Y.-C., Huang, Y., and Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med. Genom.* 13:44. doi: 10.1186/s12920-020-0677-2

Nguyen, P., Tran, T., Wickramasinghe, N., and Venkatesh, S. (2017). Deepr: a convolutional net for medical records. *IEEE J. Biomed. Health Inform.* 21, 22–30. doi: 10.1109/jbhi.2016.2633963

Nickerson, P., Tighe, P., Shickel, B., and Rashidi, P. (2016). *Deep neural network architectures for forecasting analgesic response. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. New York, NY: IEEE, 2966–2969.

Ortiz, A., Munilla, J., Górriz, J. M., and Ramírez, J. (2016). Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *Int. J. Neural. Syst.* 26:1650025. doi: 10.1142/s0129065716500258

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34, 821–829.

Pham, T., Tran, T., Phung, D., and Venkatesh, S. (2016). "DeepCare: A deep dynamic memory model for predictive medicine," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2016. Lecture Notes in Computer Science*, eds J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Huang, and R. Wang (Cham: Springer).

Phong, L. T., Aono, Y., Hayashi, T., Wang, L., and Moriai, S. (2018). Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans. Inform. Forensics Security* 13, 1333–1345. doi: 10.1109/tifs.2017.2787987

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* 2, 158–164. doi: 10.1038/s41551-018-0195-0

Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. (2018). Deep synergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 34, 1538–1546. doi: 10.1093/bioinformatics/btx806

Qin, Q., and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PLoS Comput. Biol.* 13:e1005403. doi: 10.1371/journal.pcbi.1005403

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Med.* 1:18.

Rakhlin, A., Shvets, A., Iglovikov, V., and Kalinin, A. A. (2018). "Deep convolutional neural networks for breast cancer histology image analysis," in *International Conference on Image Analysis and Recognition*, (Montreal: ICIAR), 737–744. doi: 10.1007/978-3-319-93000-8_83

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, eds N. Navab, J. Hornegger, W. Wells, and A. Frangi (Cham: Springer), 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comp. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Adv. Neural. Inform. Proc. Syst.* 2017, 3856–3866.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* 22, 1589–1604. doi: 10.1109/jbhi.2017.2767063

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imag.* 35, 1285–1298. doi: 10.1109/tmi.2016.2528162

Shokri, R., and Shmatikov, V. (2015). *Privacy-preserving deep learning. In 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. New York: IEEE, 909–910.

Shvets, A. A., Rakhlin, A., Kalinin, A. A., and Iglovikov, V. I. (2018). *Automatic instrument segmentation in robot-assisted surgery using deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. New York, NY: IEEE, 624–628.

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). Deep chrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32, 639–648.

Singh, R., Lanchantin, J., Sekhon, A., and Qi, Y. (2017). Attend and predict: understanding gene regulation by selective attention on chromatin. *Adv. Neural. Inform. Proc. Syst.* 30, 6785–6795.

Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantit. Biol.* 7, 122–137. doi: 10.1007/s40484-019-0154-0

Tan, J., Hammond, J. H., Hogan, D. A., and Greene, C. S. (2016). ADAGE-Based integration of publicly available *pseudomonas aeruginosa* gene expression data with denoising autoencoders illuminates microbe-host interactions. *mSystems* 1, 1–15.

Taylor-Robinson, D., Archangelidi, O., Carr, S. B., Cosgriff, R., Gunn, E., Keogh, R. H., et al. (2018). Data resource profile: the UK cystic fibrosis registry. *Int. J. Epidemiol.* 47, 9–10.

Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., and Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci. Rep.* 8:1727.

Unified Medical Language System(UMLS). (2031). Availabl e Online at: https://www.nlm.nih.gov/research/umls/index.html

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). *Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning*. San Diego, CA: ICML, 1096–1103.

Wan, F., Zhu, Y., Hu, H., Dai, A., Cai, X., Chen, L., et al. (2019). DeepCPI: A deep learning-based framework for large-scale in silico drug screening. *Genom. Proteomics Bioinform.* 17, 478–495. doi: 10.1016/j.gpb.2019.04.003

Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., et al. (2017). A computational-based method for predicting drug-target interactions by using stacked autoencoder deep neural network. *J. Comput. Biol.* 25, 361–373. doi: 10.1089/cmb.2017.0135

Wang, Y.-B., You, Z.-H., Yang, S., Yi, H.-C., Chen, Z.-H., and Zheng, K. (2020). A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Med. Inform. Decis. Mak.* 20:49. doi: 10.1186/s12911-020-1052-0

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., et al. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci. U S A.* 116, 5542–5549. doi: 10.1073/pnas.1814551116

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., et al. (2017). Deep learning-based drug-target interaction prediction. *J. Proteome Res.* 16, 1401–1409.

Yuan, D., Zhu, X., Wei, M., and Ma, J. (2019). *Collaborative deep learning for medical image analysis with differential privacy. In 2019 IEEE Global Communications Conference (GLOBECOM)*. New York, NY: IEEE, 1–6.

Yuan, Y., and Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. U S A.* 116, 27151–27158. doi: 10.1073/pnas.1911536116

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *ArXiv*, [Preprint].

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi: 10.1039/c9sc04336e

Zhang, H., Saravanan, K. M., Yang, Y., Hossain, T., Li, J., Ren, X., et al. (2020). Deep learning based drug screening for novel coronavirus 2019-nCov. *Interdiscip. Sci.* 12, 368–376. doi: 10.1007/s12539-020-00376-6

Zhang, J., Kowsari, K., Harrison, J. H., Lobo, J. M., and Barnes, L. E. (2018). Patient2Vec: a personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6, 65333–65346. doi: 10.1109/access.2018.2875677

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

Zhang, X., Wang, S., Zhu, F., Xu, Z., Wang, Y., and Huang, J. (2018). *Seq3seq Fingerprint: Towards end-to-end semi-supervised deep drug discovery. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York City, NY: ACM, 404–413.

Zhao, L., Wang, J., Pang, L., Liu, Y., and Zhang, J. (2020). GANsDTA: Predicting drug-target binding affinity using GANs. *Front. Genet.* 10:1243. doi: 10.3389/fgene.2019.01243

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* 50, 1171–1179. doi: 10.1038/s41588-018-0160-6

Zong, N., Wong, R. S. N., and Ngo, V. (2019). Tripartite network-based repurposing method using deep learning to compute similarities for drug-target prediction. *Methods Mol. Biol.* 1903, 317–328. doi: 10.1007/978-1-4939-8955-3_19

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 April 2021
doi: 10.3389/fgene.2021.608042

NEM-Tar: A Probabilistic Graphical Model for Cancer Regulatory Network Inference and Prioritization of Potential Therapeutic Targets From Multi-Omics Data

*Yuchen Zhang[1], Lina Zhu[1] and Xin Wang[1,2]**

[1] Department of Biomedical Sciences, City University of Hong Kong, Hong Kong, China, [2] Key Laboratory of Biochip Technology, Biotech and Health Centre, Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China

OPEN ACCESS

Edited by:
Bairong Shen, Sichuan University, China

Reviewed by:
Yuedong Yang, Sun Yat-sen University, China
Rodrigo Juliani Siqueira Dalmolin, Federal University of Rio Grande do Norte, Brazil
Yuxin Lin, Soochow University, China
Wenying Yan, Soochow University, China

***Correspondence:**
Xin Wang
xin.wang@cityu.edu.hk

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

Received: *18 September 2020*
Accepted: *22 March 2021*
Published: *22 April 2021*

Citation:
Zhang Y, Zhu L and Wang X (2021) NEM-Tar: A Probabilistic Graphical Model for Cancer Regulatory Network Inference and Prioritization of Potential Therapeutic Targets From Multi-Omics Data. Front. Genet. 12:608042. doi: 10.3389/fgene.2021.608042

Targeted therapy has been widely adopted as an effective treatment strategy to battle against cancer. However, cancers are not single disease entities, but comprising multiple molecularly distinct subtypes, and the heterogeneity nature prevents precise selection of patients for optimized therapy. Dissecting cancer subtype-specific signaling pathways is crucial to pinpointing dysregulated genes for the prioritization of novel therapeutic targets. Nested effects models (NEMs) are a group of graphical models that encode subset relations between observed downstream effects under perturbations to upstream signaling genes, providing a prototype for mapping the inner workings of the cell. In this study, we developed NEM-Tar, which extends the original NEMs to predict drug targets by incorporating causal information of (epi)genetic aberrations for signaling pathway inference. An information theory-based score, weighted information gain (WIG), was proposed to assess the impact of signaling genes on a specific downstream biological process of interest. Subsequently, we conducted simulation studies to compare three inference methods and found that the greedy hill-climbing algorithm demonstrated the highest accuracy and robustness to noise. Furthermore, two case studies were conducted using multi-omics data for colorectal cancer (CRC) and gastric cancer (GC) in the TCGA database. Using NEM-Tar, we inferred signaling networks driving the poor-prognosis subtypes of CRC and GC, respectively. Our model prioritized not only potential individual drug targets such as HER2, for which FDA-approved inhibitors are available but also the combinations of multiple targets potentially useful for the design of combination therapies.

Keywords: nested effects model, molecular subtype, regulatory network, drug targets, combination therapy, cancer

INTRODUCTION

Cancers are always discovered with diverse molecular properties and heterogeneous clinical outcomes, even when occurring in the same tissues or organs. The last decade has witnessed tremendous progress in the emerging field of precision medicine for more accurate patient stratification for more optimized therapeutic treatment. However, it remains challenging to

dissect the mechanism underlying cancer heterogeneity to identify novel drug targets for further development of targeted therapies. Targeted cancer therapy has been accepted as an effective weapon to conquer cancer (Green, 2004; Polyak and Garber, 2011), aiming to inhibit or reverse the activation patterns of particular cancer signaling pathways. Unfortunately, pathway redundancies, complex feedback, and crosstalk present in cancer cells often result in drug resistance, leading to treatment failure (Bernards, 2012; Yamaguchi et al., 2014). Therefore, a key task of precision medicine is excavating the causally wired relationship among the regulatory elements contributing to specific cancer molecular subtypes.

The identification of cancer therapeutic targets has long been based on biological knowledge and experience, which lacks a global functional overview and efficiency. Mathematical modeling could be established to predict potential drug targets in a more systematic and efficient way (**Supplementary Table 1**). Studies like iODA (Yu et al., 2020) integrated basic bioinformatic analysis and statistical methods to prioritize consistent molecular signatures at the pathway level for further investigation of cancer pathogenesis. Methods such as MiRNA-BD (Lin et al., 2018) focused on the discovery of novel miRNA biomarkers in diseases such as cancers without training or prior knowledge. Graphical models (e.g., Mezlini and Goldenberg, 2017; Manatakis et al., 2018; Kotiang and Eslami, 2020) were also proposed to infer the regulatory relationship and key driver genes, but the networks mainly encode gene expression associations, without support of multi-omics input. Other methods such as the miRNA-TF-mRNA network (Pham et al., 2019) and bipartite graphs (Bashashati et al., 2012) employed complex structures and multi-omics data to identify cancer driver genes as potential therapeutic targets. Furthermore, computational models were also proposed for the personalized prediction of potential target genes (Hou and Ma, 2014; Guo et al., 2018). All the previous methods have demonstrated their usefulness in various applications, very few of them infer causal regulatory relationships. To study the dysregulation of pathways and discover causal regulation relationships, typical approaches are Bayesian Networks, which encode conditional independence between genes on edges [e.g., (Sachs et al., 2005)]. However, the major limitation of Bayesian networks lies in their requirement of direct observations (e.g., protein activities) of perturbation effects on other pathway components, which are often not available. Besides, these methods require a large sample size to distinguish signal from noise and only capture parts of biologically relevant networks (Markowetz and Spang, 2007). Nested effects models (NEMs) (Markowetz et al., 2005, 2007) are specifically tailored to reconstruct signaling networks from indirect observations of experimental interventions. In each experiment, one component (e.g., kinase, transcription factor) in the pathway is perturbed, and multi-dimensional downstream effects are observed (e.g., gene expression or cell imaging data) (Siebourg-Polster et al., 2015). Different from other graphical models, NEMs encode subset relations between the observed downstream effects reporter genes under perturbations to signaling genes.

Nested effects models have been successfully applied to various biological scenarios to infer the causal network of signaling components (Markowetz et al., 2005; Fröhlich et al., 2009; MacNeil et al., 2015). Several extensions of NEMs have been proposed to adapt to different experimental designs or data types. For instance, Boolean NEMs (Pirkl et al., 2016) creatively model the data observed from arbitrary experimental combinations (excitation or inhibition) to infer a full Boolean network and further integrate the information from the literature. Epistatic NEMs (Pirkl et al., 2017) infer epistasis from phenotyping screens of double knock-downs systematically to test the hypothesis that complex relationships between a gene pair can be explained by the action of a third gene that modulates the interaction. Dynamic NEMs (Anchang et al., 2009; Fröhlich et al., 2011) infer the rate of the signal flow within the network from time-series data, while Hidden Markov NEMs (Wang et al., 2014) model the evolution of the network itself over time. Motivated by a recent experiment investigating epithelial-mesenchymal transition (EMT) in murine mammary gland cells, a method for mapping a non-interventional time series onto a static NEM has been proposed (Cardner et al., 2019). Furthermore, with the rapid development of single-cell sequencing technologies, a mixture of NEMs (M&NEM) tailored explicitly for single-cell data has been proposed (Pirkl and Beerenwinkel, 2018), which is capable of identifying different cellular subpopulations and inferring their corresponding causal networks simultaneously.

To prioritize potential therapeutic targets based on tissue-derived multi-omics profiles from cancer patients, we extended the classic NEMs to model the causal effects of genetic and epigenetic aberrations of various regulatory components (kinases, transcriptional factors, and miRNAs) on downstream genes. Importantly, the computational evaluation was conducted on the regulatory components (mainly on kinases) to prioritize potential therapeutic targets. **Figure 1** illustrated the framework and major steps of NEM-Tar, which is featured with the following highlights: (1) Different from pre-existing NEMs developed for phenotyping screens derived from experimental perturbations, NEM-Tar integrates natural perturbations (e.g., somatic mutations, DNA hyper- or hypo-methylation, copy number alterations) at multiple levels of gene regulations for cancer-related signaling network inference; (2) We proposed a scoring method based on information theory, named weighted information gain (WIG), which could prioritize not only individual therapeutic targets but also evaluate potential combination therapies; (3) NEM-Tar is a versatile framework for dissecting the cancer molecular heterogeneity by inferring cancer subtype-specific signaling network. In our case studies, we specifically focused on the 'EMT' subtype in gastric cancer and the CMS4-mesenchymal subtype in colorectal cancer (Cristescu et al., 2015; Guinney et al., 2015), which are associated with a higher risk of recurrence and poor prognosis. Potential drug targets are evaluated specifically on the epithelial-mesenchymal transition (EMT) pathway, which is directly associated with cancer metastasis.

In the 'Methods and Materials' section, we introduce the design of NEM-Tar and the inference strategies in detail. Subsequently, we test the effectiveness of NEM-Tar in a simulation study ('Results on Simulated Data') and demonstrate

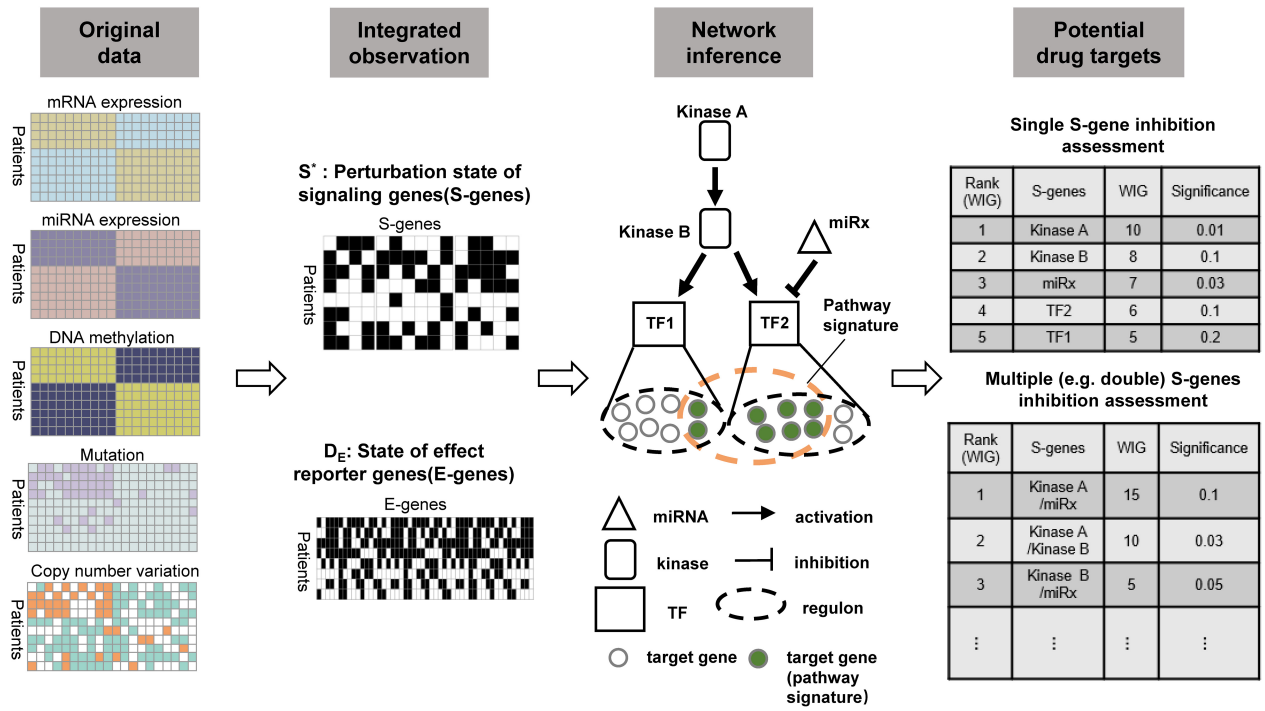


FIGURE 1 | The workflow of NEM-Tar for cancer regulatory network inference and potential drug targets prioritization. Observations of the states of S-genes and E-genes could be obtained after the preprocessing of multi-omics data. The signaling network regulating a specific cancer subtype will subsequently be inferred. Finally, based on quantification of the causal impact and specificity to downstream genes using WIG, potential drug targets could be prioritized for single and double perturbations.

its potential by real case studies on colorectal cancer and gastric cancer ('Results on Case Studies').

MATERIALS AND METHODS

The Original Nested Effects Model (NEM)

We first review the original nested effect model (NEM), before we explain in detail how we extend the original model design to fit multi-omics high-throughput profiles of cancer samples.

The structure of a NEM is illustrated in **Figure 2A**. The goal is to infer a signaling network G, represented as a directed acyclic graph involving the regulators, also referred to as signaling genes (S-genes), denoted S_j for j∈{1,2,....,m}. In the initial phenotypic screening experiments, the S-genes are individually perturbed during RNAi experiments, but their effects are indirectly measured by the expression level of effect reporter genes (E-genes) denoted E_i for i∈{1,2,....,n}. The attachment of E-genes to S-genes is denoted by Θ, within which $\theta_{ij} = 1$, if E-gene i is attached to S-gene j. The initial NEMs assumed that each E-gene can be attached to at most one S-gene, but this constraint has been relaxed thereafter. Tresch and Markowetz have proposed to add a null S-gene, which predicts no effects to account for uninformative features (Tresch and Markowetz, 2008). Due to the nested effects, it is assumed that the signaling network G is transitively closed; for instance, in **Figure 2A**, the signaling information flow is S2→S3→S4, then S2→S4 also exists.

We calculate the expected E-gene profiles for a given model $(G;\Theta)$ as the matrix product with E_{ij} the predicted state of E-gene i under knock-down of S-gene j, namely $E_{\exp} = G\Theta$. In practice, we cannot neglect potential noise in the data, which requires probabilistic modeling to infer an optimal G to interpret the observation of E-genes. Suppose that we have a candidate network structure G, which is a directed acyclic graph (DAG) of S-genes. What matters ultimately is the posterior probability of the model:

$$P(G|E) = \frac{P(E|G)P(G)}{P(E)} \tag{1}$$

where the denominator does not depend on G and cannot be taken into consideration for model comparison. Since almost nothing is known about the signaling network without reliable knowledge, we use a uniform prior P(G). Thus, we focus entirely on the likelihood P(E|G). It can be computed by marginalizing over E-gene attachments Θ, or by employing the maximum *a posteriori* (MAP) estimate of Θ (Tresch and Markowetz, 2008). The former choice is more intuitive, and the marginal likelihood can be deduced as:

$$P(G|E) = \int P(E|G,\Theta)P(\Theta,G)d\Theta$$

$$= \frac{1}{m^n}\prod_{i=1}^{n}\sum_{j=1}^{m}\prod_{k=1}^{l}P(e_{ik}|G,\theta_i = j) \tag{2}$$

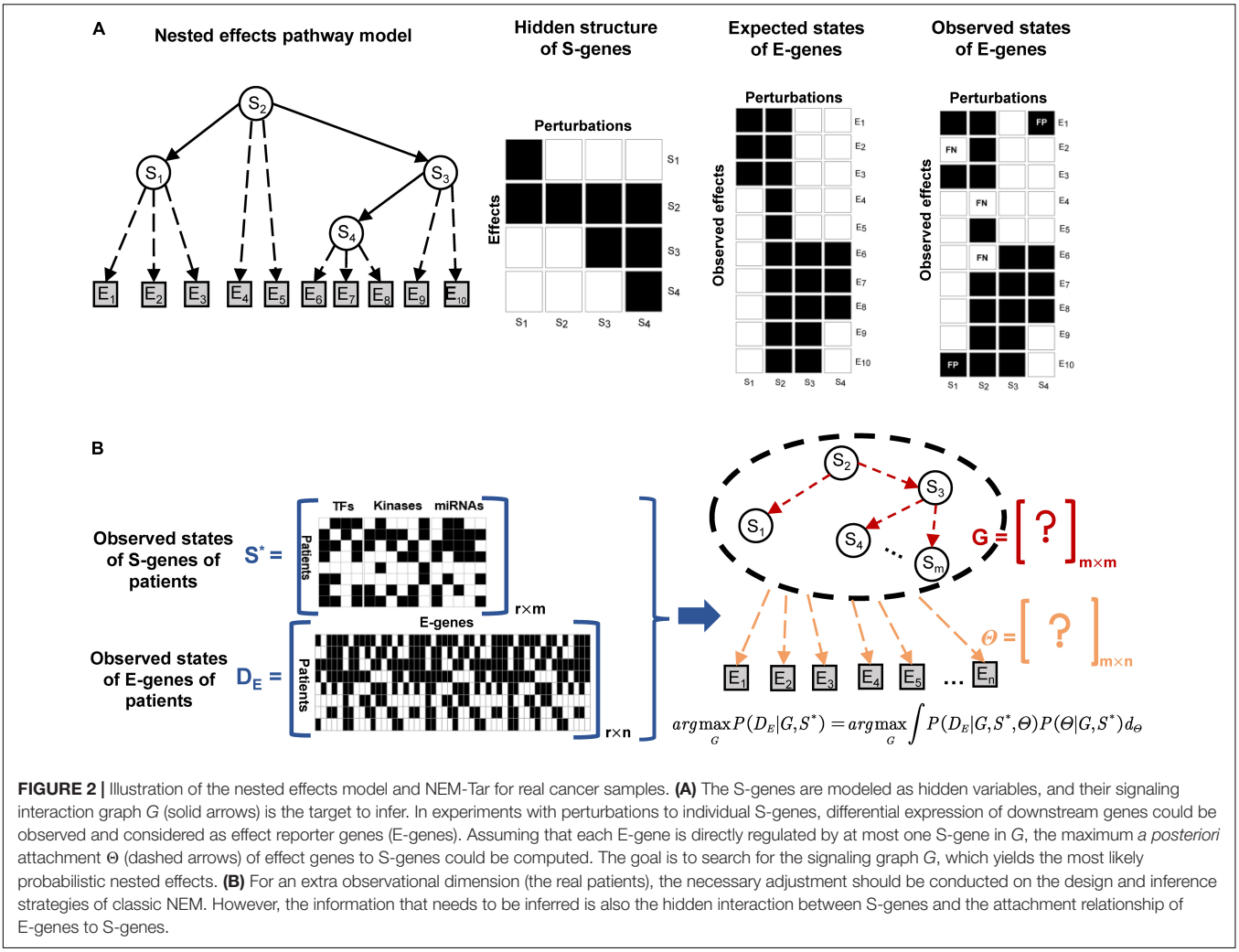


FIGURE 2 | Illustration of the nested effects model and NEM-Tar for real cancer samples. **(A)** The S-genes are modeled as hidden variables, and their signaling interaction graph G (solid arrows) is the target to infer. In experiments with perturbations to individual S-genes, differential expression of downstream genes could be observed and considered as effect reporter genes (E-genes). Assuming that each E-gene is directly regulated by at most one S-gene in G, the maximum *a posteriori* attachment Θ (dashed arrows) of effect genes to S-genes could be computed. The goal is to search for the signaling graph G, which yields the most likely probabilistic nested effects. **(B)** For an extra observational dimension (the real patients), the necessary adjustment should be conducted on the design and inference strategies of classic NEM. However, the information that needs to be inferred is also the hidden interaction between S-genes and the attachment relationship of E-genes to S-genes.

The term $P(e_{ik}|G, \theta_i = j)$ in Eq. 2 reflects the noise rate of the real binary observation e_{ik}. The distribution of e_{ik} is determined by the network structure G and the error probabilities α and β. For all E-genes and targets of perturbation, the conditional probability of the E-gene state e_{ik} given the network structure G can then be written as:

$$P(e_{ik}|G, \theta_i = j) = \begin{matrix} e_{ik}=1 & e_{ik=0} \\ \begin{cases} \alpha & 1-\alpha \\ 1-\beta & \beta \end{cases} \end{matrix} \quad (3)$$

Equation 3 means that if E_i is not an influenced target of the S-gene perturbed in experiment k, the probability of observing $e_{ik} = 1$ is α (probability of false alarm, type-I error); the probability to miss an effect and observe $e_{ik} = 0$ even though E_i is an influenced target is β (type-II error).

NEM-Tar for Multi-Omics Data

Figure 1 illustrated the NEM-Tar framework and the major steps involved to infer a signaling network using a toy example, and a comparison was made with a classic NEM in observation (**Figure 2B**). We model copy number variations or mutations (e.g., copy number gain/mutation in kinase A/B, mutation in transcription factor TF1), hyper/hypo methylation (e.g., hypermethylation of miRx) as 'natural' perturbations in tumors, which are different from experimental perturbations such as RNA interference and CRISPR-Cas9 knockout modeled in the classic NEMs. Regulators considered in the network are master regulators (TFs and miRNAs) and modulators (kinases) resulting from the reported literature (Fessler et al., 2016; Kiyozumi et al., 2018; Xie et al., 2020) and our prioritized candidates. Let $S^* = [s_{kj}]$ denote the state matrix of regulators, where s_{kj} represents whether regulator j is aberrant in sample k or not. Let G represent the signaling network of interactions between kinases, TFs, and miRNAs, and Θ be the set of interactions between regulators and their target genes. Let $D = [e_{ki}]$ be the observed data, where e_{ki} denotes whether the E-gene i is differentially expressed in patient k ($e_{ki} = 1$) or not ($e_{ki} = 0$). Our goal is to infer the optimal G that maximizes the following marginal likelihood:

$$\arg\max_G P(D|G, S^*) = \arg\max_G \int P(D_E|G, S^*, \Theta)P(\Theta|G, S^*)d\Theta \quad (4)$$

It should be noted that Eq 4 is similar to the original likelihood function of NEMs (Eq. 2), except the state matrix of regulators (S-genes) in our model.

When the optimal S-genes structure G^* is determined, we could compute the posterior probability for the edge between S_j and E_i.

$$P(\theta_i = j|G^*, S^*, D) = \frac{1}{Z}\prod_{k=1}^{l} P(e_{ki}|G^*, S^*, \theta_i = j) \quad (5)$$

where Z is a constant and does not rely on G^*.

When using NEM-Tar in real-world applications, we recommend the following criteria to select E-genes and S-genes. E-genes can be prioritized based on genes that are significantly upregulated ($\log_2$FC $>$ 1, FDR $<$ 0.01) in a specific cancer subtype of interest. If the selected E-genes are too few (e.g., only 238 E-genes for the EMT subtype of GC based on the above criteria), the cutoff on $\log_2$FC may be relaxed to 0.5. The prioritization of S-genes can be based on the following criteria. First, subtype-specific miRNAs and TFs can be prioritized based on differentially expressed genes. By default, we recommend selecting TFs that are significantly upregulated ($\log_2$FC $>$ 1 and FDR $<$ 0.01) and miRNAs that are significantly downregulated ($\log_2$FC $<$ -1 and FDR $<$ 0.01). However, due to the heterogeneity between different cancer subtypes, the number of candidate miRNAs or TFs may be limited. In the situation, the cutoff on $\log_2$FC may also be relaxed to 0.5. Second, the selection of S-genes should also satisfy the following perturbation criteria: (1) Mutation: the cutoff on mutation frequency in kinases/TFs should be $>$5%. When the overall mutation frequency of candidate S-genes is lower than 5%, the cutoff might also be relaxed appropriately. (2) copy number variations (CNVs): kinases and membrane proteins with $>$5% frequency of copy number gains. (3) DNA methylation: miRNAs with significant hypermethylation (delta-beta $>$0.1, BH-adjusted $P < 0.001$).

Inference Methods of NEM-Tar

The original NEM performs an exhaustive search over all transitively closed graphs to identify the optimal graph by the maximum likelihood estimation (Markowetz et al., 2005). Since the number of candidate network structure G grows exponentially with the number of nodes, an exhaustive enumeration is not feasible for signaling networks with more than five S-genes. In real applications, it is always necessary to search for a larger network, where heuristics are more appropriate to explore the network space. Many heuristic inference methods have been proposed, with respective advantages as well as limitations. To determine the optimal inference strategy for NEM-Tar, we investigated the triple relations, greedy hill-climbing, and MCMC sampling methods.

Instead of scoring the whole network, the model could be learned using a pairwise method (Markowetz et al., 2007). For a pair of genes A and B, their relationship could be determined by maximum *a posteriori* (MAP) from four possible models: $A \cdot B$ (unconnected), $A \rightarrow B$ (effects of A are a superset of effects of B), $A \leftarrow B$ (subset), and $A \leftrightarrow B$ (undistinguishable effects). However, the pairwise learning assumes independence of edges, which is not true in transitively closed graphs. Hence, the natural extension of pairwise learning is the inference from the triples of nodes (Markowetz et al., 2007), which comprises two steps. First, for each triple (x,y,z) in the graph with n nodes, all 29 possible quasi-orders are scored, and the MAP model is selected. Edgewise model averaging was subsequently employed to combine all models into the final graph.

Greedy hill-climbing is a more straightforward optimization strategy known from the literature (Russell and Norvig, 2016). Given an initial network hypothesis (usually an empty graph), a local maximum of the likelihood function could be reached by successively adding an edge. This procedure is continued until no improving edge can be found anymore. We also evaluated the performance of greedy hill-climbing for the benchmark in our simulation study.

Furthermore, Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution. Niederberger et al. (2012) proposed an inference method by combining MCMC sampling with an Expectation-Maximization (EM) algorithm. For reconstructing evolving signaling networks, MCMC sampling was also an important procedure in HM-NEM (Wang et al., 2014). In our simulation study, we also examined MCMC sampling, and the detailed pseudocode is in **Supplementary Figure 1**.

Weighted Information Gain (WIG) for Evaluation of the Causal Impact of S-Genes on Downstream Reporter Genes

Given the inferred optimal network G^* and interactions between regulators and target genes Θ^*, we sought to quantify the causal impact that a regulator has on downstream reporter genes, especially signature genes for a particular biological process of interest such as epithelial–mesenchymal transition (or EMT) (Nieto et al., 2016; Lambert et al., 2017). The fundamental assumptions of the assessment criteria for the impact should satisfy: (1) The more E-genes related to a particular pathway are affected by an S-gene, the more significant the influence is; (2) The more likely a particular E-gene is attached to an S-gene, the higher the global influence of the S-gene is. On the basis of the above assumptions, we defined a score called **Weighted Information Gain (WIG)** on every E-gene within the regulons of S-genes based on KL divergence (Kullback and Leibler, 1951) in information theory, which measures the information gain after network inference.

$$\begin{aligned} WIG(S_j) &= \textstyle\sum_{i=1}^{r} WIG(S_j \rightarrow E_i) \\ &= \textstyle\sum_{i=1}^{r} P(S_j \rightarrow E_i)\log[(m+1)P(S_j \rightarrow E_i)] \end{aligned} \quad (6)$$

As shown in **Figure 3**, before the network inference, for every E-gene, we assume that the probability of an E-gene attached to an S-gene is uniformly distributed, which could be denoted as$P(\theta_i = j|G) = 1/m + 1$, if we set a 'null' S-gene and no particular prior knowledge is involved. While after the inference, the posterior distribution of nested effect positions of E-genes changes into$P(S_j \rightarrow E_i) = p(\theta_i = j|G^*, S^*, D)$. According to the

original definition of KL divergence, the increase of the information of the attachment of an E-gene could be computed, like the highlighted $WIG(S_3 \rightarrow E_3)$ and $WIG(S_3 \rightarrow E_{14})$. As for an S-gene, the global causal impact over all the E-genes or some signature genes of key pathways could be obtained by summing up the WIG of related E-genes, as shown in Eq. 6. The statistical significance for the specificity of WIG on key pathways could be estimated by the bootstrap of the same number as the pathway signature genes of arbitrary E-genes within the regulon of a S-gene.

Ultimately, kinases/TFs/miRNAs with top causal WIG and/or enough significance will be prioritized as potential drug targets. For more convenient drug design, kinases, or membrane proteins are preferred.

Data Source

In our case studies, we analyzed multi-omics data for colorectal cancer (CRC) and gastric cancer (GC) patients from TCGA, including the following data types: (1) whole-genome gene expression data for 382 CRC and 415 GC patients based on RNA sequencing platform; (2) copy number variation data (scores on gene level) for 374 CRC patients and 268 GC patients; (3) somatic mutations profiles for 423 CRC patients and 433 GC patients; (4) miRNA expression data for 297 CRC and 446 gastric tumors based on Illumina sequencing platform; (5) DNA methylation data for 396 CRC and 395 GC tumor samples based on Infinium Methylation 450K platform.

RESULTS

Results on Simulated Data

Generation of *in silico* Data

The simulations evaluating the inference strategies of NEM-Tar were performed on datasets generated with varying network sizes and noise levels. The generation of simulated data is described in detail as follows.

(1) **S-gene graph generation**: We first randomly generated a graph of m S-genes, m$\in$ {6,8,10,12,15,20,30}. These graphs of S-genes were transformed to transitively closed graphs.

(2) **S-gene state generation**: For each S-gene graph generated, we simulated patient samples with a random fraction of S-genes perturbed according to the real proportions of S-genes with genetic and epigenetic alterations in the gastric cancer case study. An S-gene state matrix was subsequently generated according to the S-gene graph and simulated perturbations.

(3) **Attachment of E-genes to S-genes**: In each S-gene graph simulated, we attached effect reporter genes (or E-genes) to each S-gene, and the number of E-genes per S-gene was roughly equivalent to the average number of E-genes in the gastric cancer case study.

(4) **Generation of E-gene observations**: For each simulated graph, with the corresponding S-gene state matrix and E-gene attachment, we next generated the corresponding E-gene observation matrix. For E-genes without downstream effects expected, observations were sampled from a null distribution, or otherwise from an alternative distribution. In the simplest case, we only sampled binary data, where 1 indicated an effect and 0 no effect, according to the Type-I error α_{sim} (FP) and Type-II error β_{sim} (FN).

Using the simulation strategy, we generated data to test the performance of NEM-Tar:

(1) Scalability. Fix $\alpha_{sim} = \beta_{sim} = 0.05$ and vary the number of S-genes from 6 to 30, representing the size of a typical signaling pathway. For each number of S-genes, 200 different random S-gene networks were generated, and the simulated S-gene network structures were inferred using MCMC sampling, triple relations, and greedy hill-climbing, respectively;

(2) Robustness to noise. Fix $\beta_{sim} = 0.05$ and the number of S-genes $m = 12$ (medium size) and vary α_{sim} from 0.05 to 0.5. For the inference of S-gene network, we set $\alpha = 0.2$, $\beta = 0.1$, which were arbitrarily chosen and different from the α_{sim} and β_{sim} used for the generation of E-gene data. The evaluation criteria of their performance were TPR = TP/(TP + FN), TNR = TN/(TN + FP), Accuracy = (TP + TN)/(TP + FN + TN + FP) and Precision = TP/(TP + FP).

Benchmark the Performance of Inference Methods

The simulation results are shown in **Figure 4**. Using MCMC sampling (**Figure 4A**), although the performance showed a decreasing trend due to the increase of the size of the network, the magnitude of decrease was quite significant (e.g., the averaged TPR of 200 networks decreased from 0.867 to 0.136). Especially, the most concerned measure 'Precision' was unacceptable in real applications, no matter for smaller networks (S-genes $\leq$ 10) or larger networks (S-genes $>$ 10). Even for smaller networks with only six S-genes the instability of MCMC was evident, as the median of Precision (0.845) was much larger than the mean (0.770). For relatively large networks (e.g., 20 S-genes), the averaged Precision was too low (0.328) to accept. Using the triple relations inference, the result was slightly better than MCMC sampling (**Figure 4B**), but a dramatic decrease of Precision was also observed for networks with $\geq$ 10 S-genes. A special observation on the triple relations is that the performance on the networks with a medium size (10-15) showed fluctuating TPR, TNR, and Accuracy rather than a steady decrease, suggesting that the inference based on triple relations was also unstable.

Compared to MCMC sampling and triple relations methods, greedy hill-climbing showed much higher performance (**Figure 4C-D**). For small and medium networks (6-15 S-genes), the median of all the evaluation metrics were close to 1. Even for relatively large networks, the TPR and Precision were still reliable. Though, in essence, the greedy hill-climbing is likely to be trapped in a local optimum, at least for the graphs with less than 30 S-genes, the performance is reasonably good. The robustness for the inference with varying α_{sim} based on greedy hill-climbing is also stable and acceptable. Even for the very noisy condition ($\alpha_{sim} = 0.5$), the averaged TPR and Precision could still reach 0.947 and 0.766, respectively. Furthermore, compared to the other two methods, the greedy hill-climbing algorithm was not only superior in the performance, but also less time consuming (**Supplementary Table 2**).

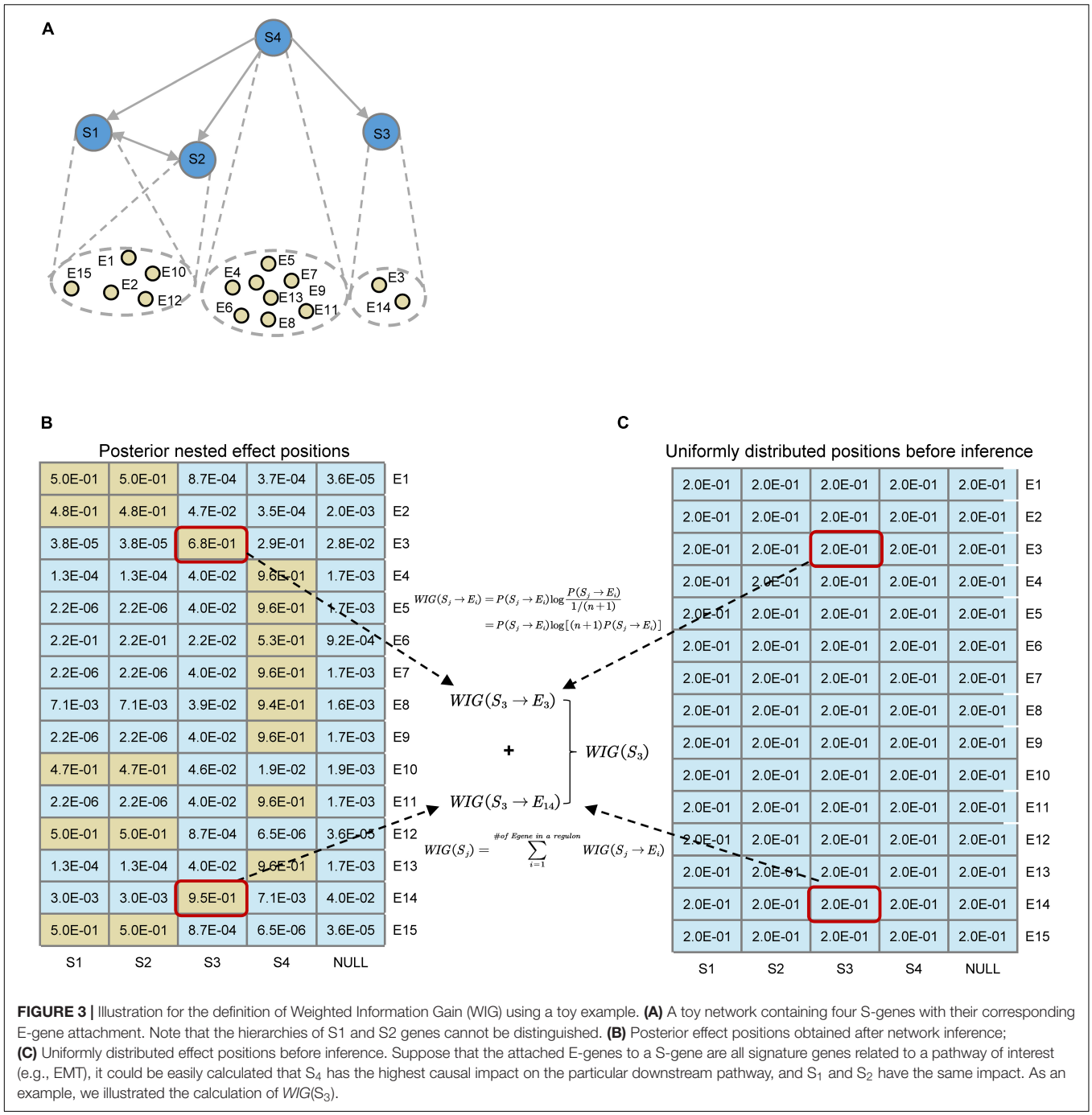


FIGURE 3 | Illustration for the definition of Weighted Information Gain (WIG) using a toy example. **(A)** A toy network containing four S-genes with their corresponding E-gene attachment. Note that the hierarchies of S1 and S2 genes cannot be distinguished. **(B)** Posterior effect positions obtained after network inference; **(C)** Uniformly distributed effect positions before inference. Suppose that the attached E-genes to a S-gene are all signature genes related to a pathway of interest (e.g., EMT), it could be easily calculated that S_4 has the highest causal impact on the particular downstream pathway, and S_1 and S_2 have the same impact. As an example, we illustrated the calculation of *WIG*(S_3).

Therefore, based on the simulation study, we employed greedy hill-climbing as the inference method for the following case studies.

Results on Case Studies

To exemplify NEM-Tar for inference of cancer subtype-specific signaling network and prioritization of potential therapeutic targets, we did two case studies using multi-omics data in gastric cancer and colorectal cancer, respectively.

Inferring the Signaling Network Driving the EMT Subtype of Gastric Cancer and Prioritization of Potential Drug Targets

Gastric cancer (GC), a leading cause of cancer-related deaths, is known to be a heterogeneous disease. The presence of molecular heterogeneity in GC has been shown through the existence of subtypes with distinct genetic/epigenetic aberrations associated with clinical outcomes. Based on 300 primary gastric cancer tumor specimens, the Asian Cancer Research Group (ACRG) identified four molecular

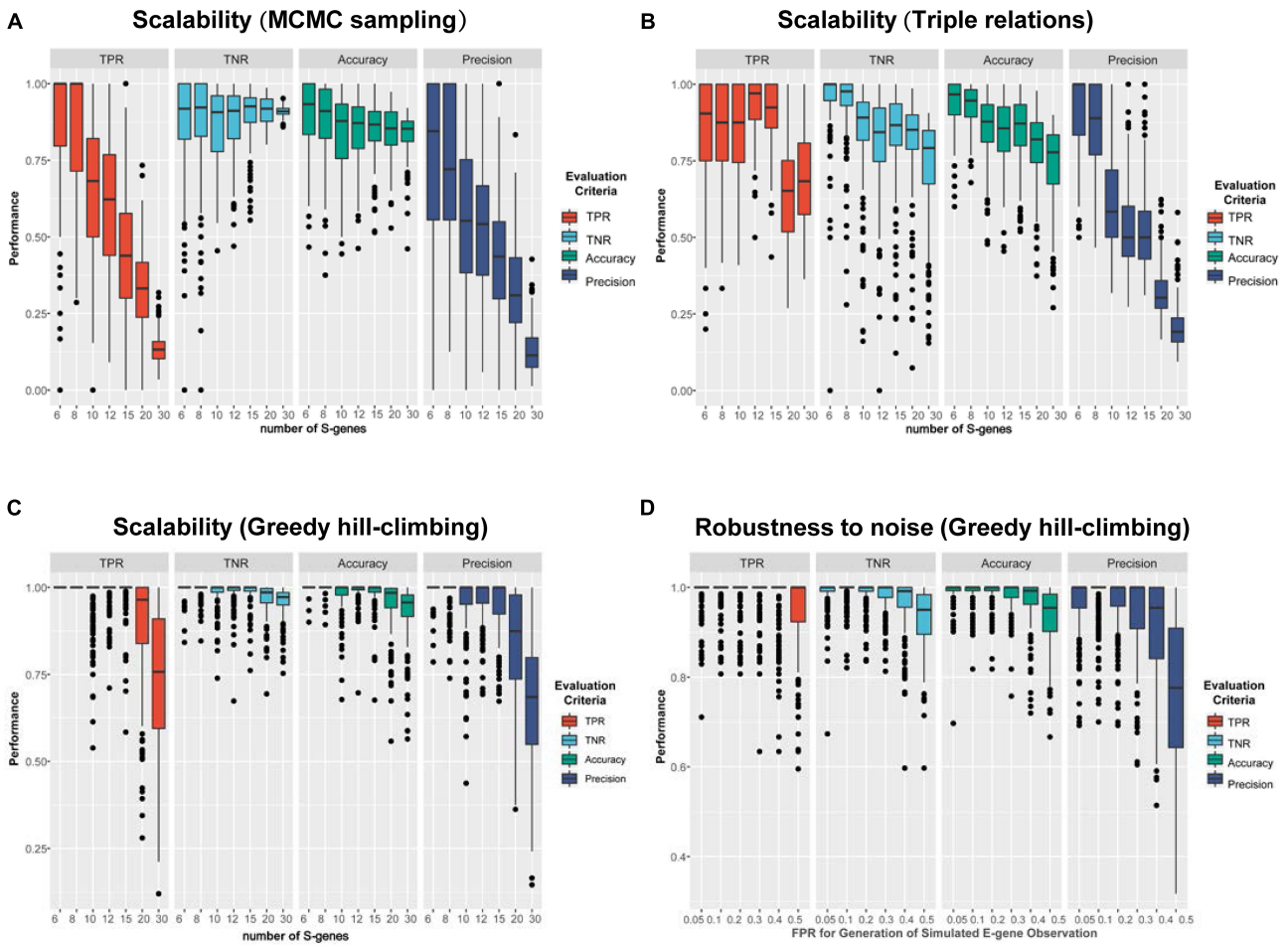


FIGURE 4 | A comparison of the performance of three representative network inference strategies. **(A–C)** The performance of NEM-Tar based on **(A)** MCMC sampling, **(B)** triple relations, and **(C)** greedy hill-climbing, respectively, on simulated data for varying numbers of S-genes. For each method, we generated 200 random signaling networks and inferred their structures using NEM-Tar from the simulated E-gene data. **(D)** The performance of NEM-Tar based on greedy hill-climbing testing its robustness to simulated data with different levels of noise.

subtypes with distinct patterns of molecular alterations and clinical outcomes (Cristescu et al., 2015). Among these four subtypes, patients classified to the MSS/EMT (in short, EMT) subtype showed the worst prognosis. Despite the extensive subtyping studies published, the regulatory mechanism underlying specific molecular subtypes has not been fully explored explicitly. Here, we employed NEM-Tar to infer the signaling network driving the EMT gastric cancer and quantitatively evaluate single and double perturbations to prioritize potential drug targets.

For the choice of the regulatory elements, we focused on the signature genes of the MAP-kinase pathway (KRAS, BRAF), frequently mutated kinases/TFs (TP53, ARID1A, CDH1, and ERBB2) (Gastric Adenocarcinoma - My Cancer Genome) and significantly upregulated TFs ($\log_2$FC $>$ 0.5, BH-adjusted $P < 0.01$) as well as downregulated miRNAs ($\log_2$FC < -1, BH-adjusted $P < 0.01$) in the EMT subtype. The regulatory elements were filtered through the integration with the somatic mutation profiles. More specifically, we kept the kinases and TFs with mutation frequency $>$ 5% and ZEB2 (Dai et al., 2012) and KRAS (Yoon et al., 2019), which were well characterized before for their roles in EMT regulation. As a result, we included nine kinases/TFs in 177 patient samples for the following analysis. The perturbations to miRNAs were measured by DNA methylation in the promoters, and six miRNAs were selected with highly significant hypermethylation (delta-beta $>$ 0.1, BH-adjusted $P < 0.001$) in the samples of the EMT subtype. Since copy number variations (CNVs) were frequently found in kinases and membrane proteins in many cancer types, we also incorporated copy number gains as a type of perturbation in the case study. Furthermore, 1194 genes significantly upregulated in the EMT subtype ($\log_2$FC $>$ 0.5, BH-adjusted $P < 0.01$) were selected as E-genes for the following analysis.

In the classic NEMs, E-genes' states are the production of individually perturbed S-genes, while for NEM-Tar E-genes' states can be the production of multiple S-genes with genetic and/or epigenetic perturbations in a tumor sample. Therefore, the concepts of positive and negative controls

for the discretization of E-genes' states should be revised accordingly. A positive control referred to the patients belonging to a particular subtype (e.g., EMT subtype) but without any (epi)genetic aberrations in the S-genes. In contrast, a negative control referred to patients not assigned to a particular subtype (e.g., Non-EMT subtypes) and had no aberrations in any S-genes. Using the strategy, we transformed the continuous gene expression data into binary observations. Denote C_{ik} as the continuous expression level of E_i of patient k. Let μ_i^+ be the mean of positive controls for E_i, and μ_i^- the mean of negative controls. To derive binary data E_{ik}, we defined individual cutoffs for every gene E_i by:

$$E_{ik} = \begin{cases} 1 \;\; if \;\; C_{ik} < \sigma \cdot \mu_i^+ + (1-\sigma) \cdot \mu_i^- \\ 0 \;\; else \end{cases} \quad (7)$$

Based on the method introduced in Markowetz et al. (2005), to make a balance between Type-I and Type-II errors, we set $\sigma = 0.5$ for the discretization. As a result, we obtained the estimated error rates $\alpha = 0.07$, $\beta = 0.08$.

Using the discretized E-gene data, we inferred the S-gene network regulating the EMT subtype of GC using NEM-Tar (**Figure 5A**). Interestingly, CDH1, ERBB2 (HER2), and KRAS were predicted to be sitting at the top hierarchies in the signaling network. Indeed, Trastuzumab, a monoclonal antibody for human epidermal growth factor receptor 2 (HER2), has already been established with chemotherapy as a first-line treatment for HER2-positive metastatic advanced GC patients (Bang et al., 2010). Besides, CDH1, coding for the E-cadherin protein, was reported to be linked to GC susceptibility and tumor invasion, and preliminary studies indicated the potential clinical value to employ CDH1 haplotypes in metastatic GC to stratify patients that will benefit from Trastuzumab-based treatments (Caggiari et al., 2017). NEM-Tar further supports the important discovery by computationally predicting and statistically evaluating the potential drug targets. Summarizing the single and double S-gene perturbations (to kinases only) with top WIGs, we found that both CDH1 and HER2 had a strong causal impact on the signature genes of epithelial-mesenchymal transition (EMT) (Zhao et al., 2019). More importantly, the causal effect was statistically significant and specific to the EMT pathway only (**Table 1**), as quantified by permutation tests, i.e., random sampling of E-genes with the same number of EMT signature genes in the regulon of a S-gene, and calculating the frequency of observing a same or higher WIG from the sampled E-gene sequences. Moreover, the combinatorial perturbations (**Table 2**) to CDH1 and ERBB2, CDH1 and KRAS or CDH1 and BRAF had the strongest and specific causal effect on the EMT pathway among all possible combinations.

Inferring the Signaling Network Driving the CMS4-Mesenchymal Subtype of Colorectal Cancer and Prioritization of Potential Drug Targets

Similar to gastric cancer, colorectal cancer (CRC) is also a heterogeneous disease posing a challenge for accurate classification and treatment of this malignancy. Recently, CRC patients have been categorized using unsupervised classification of gene expression profiling, which resulted in distinct CRC subtypes. In order to generate unified subtyping of CRC, based on a large panel of CRC patients (n = 4151), the CRC Subtyping Consortium identified four consensus molecular subtypes (CMSs) (Guinney et al., 2015). Linking the subtypes to disease outcomes revealed that the mesenchymal subtype CMS4 displayed a worse prognosis, highlighting the clinical relevance of the CMS taxonomy. As another case study, we employed NEM-Tar to infer the signaling network driving the CMS4 CRC and

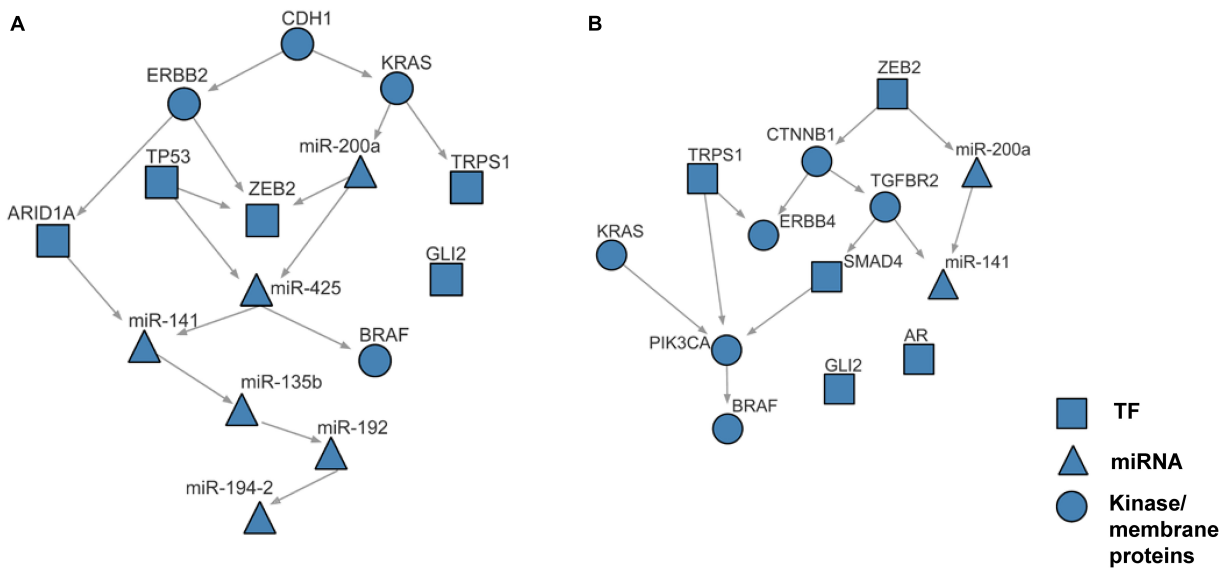


FIGURE 5 | The case studies of NEM-Tar on gastric cancer and colorectal cancer. **(A)** Reconstructed signaling network for the EMT subtype of gastric cancer. **(B)** Reconstructed signaling network for the CMS4 subtype of colorectal cancer.

TABLE 1 | WIGs assessing the impact of single perturbations (kinase only) on EMT in GC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (100,000 sampling, BH-adjusted *P*)
CDH1	591	57	66.15	<1e-05
ERBB2	491	44	51.65	<1e-05
KRAS	229	20	20.05	<1e-05
BRAF	14	1	2.71	3.21e-01

TABLE 2 | Double perturbations (kinase only) with top WIGs in GC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (50,000 sampling, BH-adjusted *P*)
CDH1/ERBB2	591	57	66.15	<5e-04
CDH1/KRAS	591	57	66.15	<5e-04
BRAF/CDH1	591	57	66.15	<5e-04
KRAS/ERBB2	558	51	59.12	<5e-04
BRAF/ERBB2	505	45	54.36	<5e-04
KRAS/BRAF	229	20	20.05	<5e-04

calculated WIGs for single and double perturbations to signaling elements in order to prioritize potential drug targets.

To select regulatory elements, we incorporated the signature genes of the MAP-kinase pathway (KRAS, BRAF, PIK3CA), and the TFs significantly upregulated in CMS4 ($\log_2$FC $>$ 1, BH-adjusted $P < 0.01$) as well as the miRNAs significantly downregulated in CMS4 ($\log_2$FC < -0.5, BH-adjusted $P < 0.01$). The regulatory elements were filtered through the integration with the somatic mutation profiles. More specifically, the kinases and TFs with the mutation frequency $>$ 5% were left, resulting in 11 kinases/TFs in 212 patient samples for analysis. The perturbations to miRNAs were measured by DNA methylation in the promoters, and two miRNAs were selected with highly significant hypermethylation (delta-beta $>$ 0.1, BH-adjusted $P < 0.001$) in the samples of the CMS4 subtype. The copy number variations (CNVs) profiles were also preprocessed, but the frequency of copy number gain was too low (less than 5%) to integrate. Finally, after integration with downstream E-genes ($\log_2$FC = 1, BH-adjusted $P = 0.01$) that are differentially expressed between CMS4 and non-CMS4 samples, we obtained a 212 $\times$ 1337 E-gene observation matrix for the following analysis.

The whole discretization analysis of E-genes is similar to what we did in gastric cancer. In CRC, the positive controls are patients belonging to the CMS4 subtype without any aberrations in any S-genes, while the negative controls are patients assigned to Non-CMS4 subtypes without aberrations in any S-genes. We set $\sigma = 0.6$ for the discretization, and the estimated error rates were $\alpha = 0.22$ and $\beta = 0.18$. Using the discretized E-gene data, we inferred the S-genes network regulating the CMS4 subtype of CRC (**Figure 5B**). Based on the WIG calculation (**Table 3**, **4**), we found that the perturbation on KRAS has the highest impact on the EMT pathway, though the influence is not specific to EMT, which is reasonable as KRAS is a frequently mutated oncogene in cancer. Currently, a variety of methods to inhibit KRAS for the treatment of metastatic CRC have been proposed (Porru et al., 2018). Besides, CTNNB1, which encodes β-catenin, has the second highest impact on the EMT pathway. CTNNB1 is involved in the Wnt-β-catenin signaling pathway, which often drives a transcriptional program that is reminiscent of EMT (Anastas and Moon, 2013). Particularly, the role of Wnt-β-catenin signaling in CRC and its potential as a therapeutic target for CRC has been extensively explored. Existing drugs targeting β-catenin, such as Aspirin, are already available, and several small molecules are under clinical trials (Cheng et al., 2019). Furthermore, the combinatorial perturbations to KRAS and CTNNB1, as well as KRAS and TGFBR2, enhanced the causal impact on the EMT pathway compared to their single perturbations, suggesting potential combination therapies for the specific CMS4 subtype of CRC.

DISCUSSION

Although quite a few computational approaches have been developed for the identification of cancer therapeutic targets, they differ in the types of input data, the design of models/algorithms, the output of the results and the angles of biological interpretations. The unique strength of our NEM-Tar lies in its capability to prioritize not only individual therapeutic targets but also combinational therapies, which has not been realized before as far as we know. As a result, it is very difficult to quantitatively compare NEM-Tar with other computational approaches directly. However, we tried to make a rough comparison with two widely used methods, DawnRank (Hou and Ma, 2014) and DriverNet (Bashashati et al., 2012), which were proposed to discover cancer driver genes. Using DawnRank, we found that for the CMS4 subtype in CRC, AR, and GLI2, two TFs in our regulatory network, were also ranked among the top 5% (**Supplementary Table 3**). More excitingly, CDH1 and TP53 were ranked as the top two drivers for the EMT subtype in GC (**Supplementary Table 3**). When it comes to the result of DriverNet, only CDH1 and TP53 were prioritized as the 2nd and 3rd for EMT subtype in GC (**Supplementary Table 4**). However,

TABLE 3 | WIGs assessing the impact of single perturbations (kinase only) on EMT in CRC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (100,000 sampling, BH-adjusted *P*)
KRAS	525	49	38.61	1.48e-01
CTNNB1	151	14	23.26	< 1e-05
TGFBR2	85	10	16.67	< 1e-05
PIK3CA	26	3	5.98	3.14e-02
BRAF	15	2	3.94	5.21e-01
ERBB4	23	2	2.95	5.25e-01

TABLE 4 | Double perturbations (kinase only) with top WIGs in CRC.

S-genes	Total No. of downstream E-genes within the regulon	No. of E-genes (EMT related) within the regulon	WIG	Significance of WIG (50,000 sampling, BH-adjusted *P*)
KRAS/CTNNB1	650	60	55.88	<5e-04
KRAS/TGFBR2	584	56	49.29	2.73e-04
KRAS/ERBB4	548	51	41.56	1.06e-01
KRAS/BRAF	525	49	38.61	2.62e-01
KRAS/PIK3CA	525	49	38.61	1.71e-01
BRAF/CTNNB1	151	14	23.26	<5e-04
PIK3CA/CTNNB1	151	14	23.26	<5e-04
TGFBR2/CTNNB1	151	14	23.26	<5e-04
ERBB4/CTNNB1	151	14	23.26	< 5e-04
TGFBR2/ERBB4	108	12	19.62	4.20e-05

no driver genes were found consistent between NEM-Tar and DriverNet for the CMS4 subtype of CRC (**Supplementary Table 5**). It should be noted that DawnRank and DriverNet could dissect the driver genes only based on the modeling of association networks, which lack the inference of causal relationships and cannot measure double or multiple therapeutic targets. Furthermore, neither DriverNet nor DawnRank were designed to distinguish TFs and kinases and could not incorporate perturbation information at other levels of gene expression regulations except for gene mutations. Instead, NEM-Tar was developed to prioritize potential therapeutic targets using regulatory network inference based on nested effects models.

The hierarchical causal relationship between signaling components is not only central for understanding the regulatory mechanism of cancers but also critical for developing potential drug targets to overcome the pervasive genetic redundancies. Inspired by NEMs encoding subset relations between observed downstream effects of experimental perturbations in signaling genes, we proposed NEM-Tar to infer signaling networks from various genetic and epigenetic perturbations to regulatory elements such as kinases, transcriptional factors, and miRNAs. The marginal likelihood function of NEM-Tar is similar to the original likelihood function of NEM, except the state matrix of regulators (S-genes) in our model. Based on NEM-Tar, a new score named weighted information gain (WIG) was defined to assess the causal impact of S-genes on downstream reporter genes.

Colorectal cancer and GC are two major malignancies of the gastrointestinal tract, for which molecular subtyping has been well studied. To exemplify the usefulness of NEM-Tar, we performed two case studies to infer signaling networks that drive the poor prognosis subtypes of GC and CRC, respectively. In GC, we found that among the top significant signaling genes with high WIGs, CDH1, and ERBB2 are particularly attractive. Indeed, the FDA-approved drug Trastuzumab targeting ERBB2 has already been established with chemotherapy as a first-line treatment for HER2-positive metastatic advanced GC patients. Our further evaluation of combinatorial perturbations suggested that simultaneous inhibition of CDH1 and ERBB2/KRAS/BRAF, ERBB2, and KRAS/BRAF, as well as KRAS and BRAF may be potential combination therapies. For CMS4 CRC, except for KRAS, a representative oncogene employed as a therapeutic target, the kinase CTNNB1 with the second highest WIG may be a potential alternative therapeutic target to CRC, and combinatorial inhibition of KRAS and CTNNB1 may provide a potential combination therapy.

Within the inferred signaling networks, we noticed many interesting interactions between the S-gene regulators. First, in the signaling network inferred for the EMT subtype in GC (**Figure 5A**), CDH1 and ERBB2 were prioritized as potential therapeutic targets (**Table 1**). A signal flow was inferred between them, which could be explained by the direct interaction (PPI) between them (Guo et al., 2014) or their PPIs via β-catenin (CTNNB1) (Schroeder et al., 2002; Tang et al., 2008). The signal flow miR-200a→ZEB2 could be strongly supported by the previous finding that miR-200a can regulate the expression of ZEB2 by directly binding the 3′UTR (Cong et al., 2013). Furthermore, the signal flow KRAS→miR-200a was also supported by the previous finding that oncogenic KRAS activation can suppress the expression of miR-200s

(Zhong et al., 2016), and TP53→ZEB2 could be verified by their interactions with the miR-200 family (Rokavec et al., 2014). Second, in the signaling network inferred for the CMS4 subtype in CRC (**Figure 5B**), the advantages of our work were demonstrated more explicitly. The signal flow KRAS→PIK3CA→BRAF, supported by the MAP-kinase pathway (Dhillon et al., 2007), i.e., the PPIs between KRAS and PIK3CA (Hart et al., 2015) and between PIK3CA and BRAF (Shen et al., 2017), which is known as a typical signaling pathway driving EMT. The interaction between TGFBR2 and SMAD4 is involved in the TGFβ signaling pathway (Zhang et al., 1996). The signal flow CTNNB1→TGFBR2 is involved in the crosstalk between Wnt/β-catenin and TGFβ signaling pathways (Tian and Phillips, 2002). Together, the literature supports the effectiveness of NEM-Tar in predicting the regulatory hierarchy involving multiple redundant pathways driving EMT. Moreover, we also found signal flows between miRNAs, like the links miR-200a→miR-425, miR-141→miR-135b (**Figure 5A**) and miR-200a→miR-141 (**Figure 5B**), which are interesting but have not been previously reported yet. The miRNAs may interact indirectly via intermediate regulators, which were not included in the regulatory network inference based on our criteria for the selection of S-genes. The crosstalk between the miRNAs might also indicate their synergistic relationship on co-regulating downstream targets, which is frequently reported in the literature [reviewed in Xu et al. (2016)]. Integrating the computational prediction with experimental validation will be more convincing in revealing the crosstalks between the miRNAs, which will be an interesting direction to explore in our future work.

NEM-Tar can be improved in multiple ways in our future work. First, known signaling pathway structures can be incorporated into the model as prior knowledge to strengthen the accuracy of inference. Second, NEM-Tar proposed in this article is designed for binary effects and treats E-genes as independent random variables. However, we can possibly model log odds ratios like the methods in Tresch and Markowetz (2008), where alternative and null distribution are both normal, to decrease the information loss. Third, in this work, we focused on the S-genes with subtype-specificity or with functional relations reported to key pathways (e.g., MAP-kinase) or biological processes (e.g., EMT), and therefore the number of S-genes was limited. The limitation of scalability to a larger perturbation scale could be one future direction to improve our method. In our simulation study, greedy hill-climbing demonstrated high and robust performance in signaling networks with up to 30 S-genes, which meets the regular need for signaling network inference and drug targets prioritization. Many techniques may improve the performance of MCMC sampling (Andrieu et al., 2003), which warrants further exploration in our future work. Last but not least, we can also change the modeling framework radically using graph embedding based methods (Yue et al., 2020), as the observation of S-genes and E-genes are all high-dimensional vectors. However, the question of how to preserve the assumption of nested subset structures in the embedding space needs to be conquered tactfully.

In conclusion, NEM-Tar presents a useful computational framework for dissecting the regulatory architecture underlying specific cancer subtypes and prioritizing potential drug targets. With the explosive increase of high-throughput sequencing data, NEM-Tar warrants further evaluation using large-scale multi-omics data cohorts in the future.

DATA AVAILABILITY STATEMENT

The datasets analyzed during the current study are available in the TCGA repository (https://cancergenome.nih.gov/).

AUTHOR CONTRIBUTIONS

XW contributed to study concept and design. YZ and LZ contributed to data collection, analysis, and interpretation. XW contributed to critical revision of the manuscript for important intellectual content. LZ provided important advice and assistance for manuscript drafting. XW supervised the study. All authors read and approved the final manuscript.

FUNDING

This work was supported by a grant from Guangdong Basic and Applied Basic Research Foundation (Project No. 2019B030302012), a grant supported by the Young Scientists Fund of the National Natural Science Foundation of China (81802384), a grant by the Science Technology and Innovation Commission of Shenzhen Municipality (Project No. JCYJ20200109120425045), and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21101115, 11102317, 11103718, 11103619, R4017-18, C4041-17GF) awarded to XW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.608042/full#supplementary-material

REFERENCES

Anastas, J. N., and Moon, R. T. (2013). WNT signalling pathways as therapeutic targets in cancer. *Nat. Rev. Cancer* 13, 11–26. doi: 10.1038/nrc3419

Anchang, B., Sadeh, M. J., Jacob, J., Tresch, A., Vlad, M. O., Oefner, P. J., et al. (2009). Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proc. Natl. Acad. Sci. U. S. A.* 106, 6447–6452. doi: 10.1073/pnas.0809822106

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Mach. Learn.* 50, 5–43. doi: 10.1023/a:1020281327116

Bang, Y.-J., Van Cutsem, E., Feyereislova, A., Chung, H. C., Shen, L., Sawaki, A., et al. (2010). Trastuzumab in combination with chemotherapy versus

chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* 376, 687–697. doi: 10.1016/S0140-6736(10)61121-X

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., et al. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* 13:R124. doi: 10.1186/gb-2012-13-12-r124

Bernards, R. (2012). A missing link in genotype-directed cancer therapy. *Cell* 151, 465–468. doi: 10.1016/j.cell.2012.10.014

Caggiari, L., Miolo, G., Buonadonna, A., Basile, D., Santeufemia, D. A., Cossu, A., et al. (2017). Characterizing metastatic HER2-positive gastric cancer at the CDH1 haplotype. *Int. J. Mol. Sci.* 19:19010047. doi: 10.3390/ijms19010047

Cardner, M., Meyer-Schaller, N., Christofori, G., and Beerenwinkel, N. (2019). Inferring signalling dynamics by integrating interventional with observational data. *Bioinformatics* 35, i577–i585. doi: 10.1093/bioinformatics/btz325

Cheng, X., Xu, X., Chen, D., Zhao, F., and Wang, W. (2019). Therapeutic potential of targeting the Wnt/β-catenin signaling pathway in colorectal cancer. *Biomed. Pharmacother.* 110, 473–481. doi: 10.1016/j.biopha.2018.11.082

Cong, N., Du, P., Zhang, A., Shen, F., Su, J., Pu, P., et al. (2013). Downregulated microRNA-200a promotes EMT and tumor growth through the wnt/β-catenin pathway by targeting the E-cadherin repressors ZEB1/ZEB2 in gastric adenocarcinoma. *Oncol. Rep.* 29, 1579–1587. doi: 10.3892/or.2013.2267

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

Dai, Y.-H., Tang, Y.-P., Zhu, H.-Y., Lv, L., Chu, Y., Zhou, Y.-Q., et al. (2012). ZEB2 promotes the metastasis of gastric cancer and modulates epithelial mesenchymal transition of gastric cancer cells. *Dig. Dis. Sci.* 57, 1253–1260. doi: 10.1007/s10620-012-2042-6

Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* 26, 3279–3290. doi: 10.1038/sj.onc.1210421

Fessler, E., Jansen, M., De Sousa, E., Melo, F., Zhao, L., Prasetyanti, P. R., et al. (2016). A multidimensional network approach reveals microRNAs as determinants of the mesenchymal colorectal cancer subtype. *Oncogene* 35, 6026–6037. doi: 10.1038/onc.2016.134

Fröhlich, H., Praveen, P., and Tresch, A. (2011). Fast and efficient dynamic nested effects models. *Bioinformatics* 27, 238–244. doi: 10.1093/bioinformatics/btq631

Fröhlich, H., Sahin, O., Arlt, D., Bender, C., and Beissbarth, T. (2009). Deterministic Effects Propagation Networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics* 10:322. doi: 10.1186/1471-2105-10-322

Gastric Adenocarcinoma (2020). *My Cancer Genome*. Available online at: https://www.mycancergenome.org/content/disease/gastric-adenocarcinoma/ (accessed September 18, 2020).

Green, M. R. (2004). Targeting targeted therapy. *N. Engl. J. Med.* 350, 2191–2193. doi: 10.1056/NEJMe048101

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nat. Med.* 21, 1350–1356. doi: 10.1038/nm.3967

Guo, W.-F., Zhang, S.-W., Liu, L.-L., Liu, F., Shi, Q.-Q., Zhang, L., et al. (2018). Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics* 34, 1893–1903. doi: 10.1093/bioinformatics/bty006

Guo, Z., Neilson, L. J., Zhong, H., Murray, P. S., Zanivan, S., and Zaidel-Bar, R. (2014). E-cadherin interactome complexity and robustness resolved by quantitative proteomics. *Sci. Signal.* 7:rs7. doi: 10.1126/scisignal.2005473

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 163, 1515–1526. doi: 10.1016/j.cell.2015.11.015

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

Kiyozumi, Y., Iwatsuki, M., Yamashita, K., Koga, Y., Yoshida, N., and Baba, H. (2018). Update on targeted therapy and immune therapy for gastric cancer, 2018. *J. Cancer Metastasis. Treat* 4:31.

Kotiang, S., and Eslami, A. (2020). A probabilistic graphical model for system-wide analysis of gene regulatory networks. *Bioinformatics* 36, 3192–3199. doi: 10.1093/bioinformatics/btaa122

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Stat.* 22, 79–86.

Lambert, A. W., Pattabiraman, D. R., and Weinberg, R. A. (2017). Emerging Biological Principles of Metastasis. *Cell* 168, 670–691. doi: 10.1016/j.cell.2016.11.037

Lin, Y., Wu, W., Sun, Z., Shen, L., and Shen, B. (2018). MiRNA-BD: an evidence-based bioinformatics model and software tool for microRNA biomarker discovery. *RNA Biol.* 15, 1093–1105. doi: 10.1080/15476286.2018.1502590

MacNeil, L. T., Pons, C., Arda, H. E., Giese, G. E., Myers, C. L., and Walhout, A. J. M. (2015). Transcription Factor Activity Mapping of a Tissue-Specific in vivo Gene Regulatory Network. *Cell Syst* 1, 152–162. doi: 10.1016/j.cels.2015.08.003

Manatakis, D. V., Raghu, V. K., and Benos, P. V. (2018). piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks. *Bioinformatics* 34, i848–i856. doi: 10.1093/bioinformatics/bty591

Markowetz, F., Bloch, J., and Spang, R. (2005). Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* 21, 4026–4032. doi: 10.1093/bioinformatics/bti662

Markowetz, F., Kostka, D., Troyanskaya, O. G., and Spang, R. (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* 23, i305–i312. doi: 10.1093/bioinformatics/btm178

Markowetz, F., and Spang, R. (2007). Inferring cellular networks–a review. *BMC Bioinformatics* 8:S5. doi: 10.1186/1471-2105-8-S6-S5

Mezlini, A. M., and Goldenberg, A. (2017). Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases. *PLoS Comput. Biol.* 13:e1005580. doi: 10.1371/journal.pcbi.1005580

Niederberger, T., Etzold, S., Lidschreiber, M., Maier, K. C., Martin, D. E., Fröhlich, H., et al. (2012). MC EMiNEM maps the interaction landscape of the Mediator. *PLoS Comput. Biol.* 8:e1002568. doi: 10.1371/journal.pcbi.1002568

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., and Thiery, J. P. (2016). EMT: 2016. *Cell* 166, 21–45. doi: 10.1016/j.cell.2016.06.028

Pham, V. V. H., Liu, L., Bracken, C. P., Goodall, G. J., Long, Q., Li, J., et al. (2019). CBNA: A control theory based method for identifying coding and non-coding cancer drivers. *PLoS Comput. Biol.* 15:e1007538. doi: 10.1371/journal.pcbi.1007538

Pirkl, M., and Beerenwinkel, N. (2018). Single cell network analysis with a mixture of Nested Effects Models. *Bioinformatics* 34, i964–i971. doi: 10.1093/bioinformatics/bty602

Pirkl, M., Diekmann, M., van der Wees, M., Beerenwinkel, N., Fröhlich, H., and Markowetz, F. (2017). Inferring modulators of genetic interactions with epistatic nested effects models. *PLoS Comput. Biol.* 13:e1005496. doi: 10.1371/journal.pcbi.1005496

Pirkl, M., Hand, E., Kube, D., and Spang, R. (2016). Analyzing synergistic and non-synergistic interactions in signalling pathways using Boolean Nested Effect Models. *Bioinformatics* 32, 893–900. doi: 10.1093/bioinformatics/btv680

Polyak, K., and Garber, J. (2011). Targeting the missing links for cancer therapy. *Nat. Med.* 17, 283–284. doi: 10.1038/nm0311-283

Porru, M., Pompili, L., Caruso, C., Biroccio, A., and Leonetti, C. (2018). Targeting KRAS in metastatic colorectal cancer: current strategies and emerging opportunities. *J. Exp. Clin. Cancer Res.* 37, 719–711. doi: 10.1186/s13046-018-0719-1

Rokavec, M., Li, H., Jiang, L., and Hermeking, H. (2014). The p53/microRNA connection in gastrointestinal cancer. *Clin. Exp. Gastroenterol.* 7, 395–413. doi: 10.2147/CEG.S43738

Russell, S., and Norvig, P. (2016). *Artificial intelligence: A modern approach, global edition*, 3rd Edn. London: Pearson Education.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529. doi: 10.1126/science.1105809

Schroeder, J. A., Adriance, M. C., McConnell, E. J., Thompson, M. C., Pockaj, B., and Gendler, S. J. (2002). ErbB-beta-catenin complexes are associated with human infiltrating ductal breast and murine mammary tumor virus (MMTV)-Wnt-1 and MMTV-c-Neu transgenic carcinomas. *J. Biol. Chem.* 277, 22692–22698. doi: 10.1074/jbc.M201975200

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., et al. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* 14, 573–576. doi: 10.1038/nmeth.4225

Siebourg-Polster, J., Mudrak, D., Emmenlauer, M., Rämö, P., Dehio, C., Greber, U., et al. (2015). NEMix: single-cell nested effects models for probabilistic pathway stimulation. *PLoS Comput. Biol.* 11:e1004078. doi: 10.1371/journal.pcbi.1004078

Tang, Y., Liu, Z., Zhao, L., Clemens, T. L., and Cao, X. (2008). Smad7 stabilizes beta-catenin binding to E-cadherin complex and promotes cell-cell adhesion. *J. Biol. Chem.* 283, 23956–23963. doi: 10.1074/jbc.M800351200

Tian, Y. C., and Phillips, A. O. (2002). Interaction between the transforming growth factor-beta type II receptor/Smad pathway and beta-catenin during transforming growth factor-beta1-mediated adherens junction disassembly. *Am. J. Pathol.* 160, 1619–1628. doi: 10.1016/s0002-9440(10)61109-1

Tresch, A., and Markowetz, F. (2008). Structure learning in Nested Effects Models. *Stat. Appl. Genet. Mol. Biol.* 7:Article9. doi: 10.2202/1544-6115.1332

Wang, X., Yuan, K., Hellmayr, C., Liu, W., and Markowetz, F. (2014). Reconstructing evolving signalling networks by hidden Markov nested effects models. *Ann. Appl. Stat.* 8, 448–480. doi: 10.1214/13-AOAS696

Xie, Y.-H., Chen, Y.-X., and Fang, J.-Y. (2020). Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduct. Target. Ther.* 5:22. doi: 10.1038/s41392-020-0116-z

Xu, J., Shao, T., Ding, N., Li, Y., and Li, X. (2016). miRNA–miRNA crosstalk: from genomics to phenomics. *Brief. Bioinform.* 2016:bbw073. doi: 10.1093/bib/bbw073

Yamaguchi, H., Chang, S.-S., Hsu, J. L., and Hung, M.-C. (2014). Signaling cross-talk in the resistance to HER family receptor targeted therapy. *Oncogene* 33, 1073–1081. doi: 10.1038/onc.2013.74

Yoon, C., Till, J., Cho, S.-J., Chang, K. K., Lin, J.-X., Huang, C.-M., et al. (2019). KRAS activation in gastric adenocarcinoma stimulates epithelial-to-mesenchymal transition to cancer stem-like cells and promotes metastasis. *Mol. Cancer Res.* 17, 1945–1957. doi: 10.1158/1541-7786.MCR-19-0077

Yu, C., Qi, X., Lin, Y., Li, Y., and Shen, B. (2020). iODA: An integrated tool for analysis of cancer pathway consistency from heterogeneous multi-omics data. *J. Biomed. Inform.* 112, 103605. doi: 10.1016/j.jbi.2020.103605

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2020). Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 36, 1241–1251. doi: 10.1093/bioinformatics/btz718

Zhang, Y., Feng, X., We, R., and Derynck, R. (1996). Receptor-associated Mad homologues synergize as effectors of the TGF-beta response. *Nature* 383, 168–172. doi: 10.1038/383168a0

Zhao, M., Liu, Y., Zheng, C., and Qu, H. (2019). dbEMT 2.0: An updated database for epithelial-mesenchymal transition genes with experimentally verified information and precalculated regulation information for cancer metastasis. *J. Genet. Genomics* 46, 595–597. doi: 10.1016/j.jgg.2019.11.010

Zhong, X., Zheng, L., Shen, J., Zhang, D., Xiong, M., Zhang, Y., et al. (2016). Suppression of MicroRNA 200 family expression by oncogenic KRAS activation promotes cell survival and epithelial-mesenchymal transition in KRAS-driven cancer. *Mol. Cell. Biol.* 36, 2742–2754. doi: 10.1128/mcb.00079-16

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fgene.2021.595436

Association of CLDN6 and CLDN10 With Immune Microenvironment in Ovarian Cancer: A Study of the Claudin Family

Peipei Gao[1], Ting Peng[1], Canhui Cao[1], Shitong Lin[1], Ping Wu[1], Xiaoyuan Huang[1,2], Juncheng Wei[1,2], Ling Xi[1,2], Qin Yang[3] and Peng Wu[1,2]**

[1] Cancer Biology Research Center (Key Laboratory of the Ministry of Education), Department of Obstetrics and Gynecology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2] Department of Obstetrics and Gynecology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [3] Institute of Pathology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

OPEN ACCESS

Edited by:
Bairong Shen,
Sichuan University, China

Reviewed by:
Rajeev K. Singla,
Sichuan University, China
Zhongqiu Xie,
University of Virginia, United States

****Correspondence:***
Peng Wu
pengwu8626@tjh.tjmu.edu.cn
Qin Yang
70918404@qq.com

Specialty section:
This article was submitted to Systems Biology Archive, a section of the journal Frontiers in Genetics

Received: *16 August 2020*
Accepted: *05 May 2021*
Published: *23 June 2021*

Citation:
Gao P, Peng T, Cao C, Lin S, Wu P, Huang X, Wei J, Xi L, Yang Q and Wu P (2021) Association of CLDN6 and CLDN10 With Immune Microenvironment in Ovarian Cancer: A Study of the Claudin Family. Front. Genet. 12:595436. doi: 10.3389/fgene.2021.595436

Background: The claudin family is a group of transmembrane proteins related to tight junctions. While their involvement in cancer has been studied extensively, their relationship with the tumor immune microenvironment remains poorly understood. In this research, we focused on genes related to the prognosis of ovarian cancer and explored their relationship with the tumor immune microenvironment.

Methods: The cBioPortal for Cancer Genomics database was used to obtain the genetic variation pattern of the claudin family in ovarian cancer. The ONCOMINE and Gene Expression Profiling Interactive Analysis (GEPIA) databases were used to explore the mRNA expression of claudins in cancers. The prognostic potential of these genes was examined via the Kaplan-Meier plotter. The enrichment of immunological signatures was determined by gene set enrichment analysis (GSEA). The correlations between claudins and the tumor immune microenvironment in ovarian cancer were investigated via the Tumor Immune Estimation Resource (TIMER).

Results: Claudin genes were altered in 363 (62%) of queried patients/samples. Abnormal expression levels of claudins were observed in various cancers. Among them, CLDN3, CLDN4, CLDN6, CLDN10, CLDN15, and CLDN16 were significantly correlated with overall survival in patients with ovarian cancer. GSEA revealed that CLDN6 and CLDN10 were significantly enriched in immunological signatures of B cell, CD4 T cell, and CD8 T cell. Furthermore, CLDN6 and CLDN10 were negatively correlated and positively correlated, respectively, with immune cell infiltration in ovarian cancer. The expression levels of CLDN6 and CLDN10 were also negatively correlated and positively correlated, respectively, with various gene markers of immune cells in ovarian cancer. Thus, CLDN6 and CLDN10 may participate in

immune cell infiltration in ovarian cancer, and these mechanisms may be the reason for poor prognosis.

Conclusion: Our study showed that CLDN6 and CLDN10 were prognostic biomarkers correlated with the immune microenvironment in ovarian cancer. These results reveal new roles for CLDN6 and CLDN10 as potential therapeutic targets in the treatment of ovarian cancer.

Keywords: ovarian cancer, CLDN6, CLDN10, prognosis, immune microenvironment

INTRODUCTION

Ovarian cancer is the most lethal gynecological cancer among women (Siegel et al., 2020). Although surgical techniques and combined chemotherapy applications have progressed since the 1970s, the 5 year survival rate of advanced ovarian cancer is only 40–45% (Henderson et al., 2018). Therefore, improved treatment of ovarian cancer remains an urgent issue. Immunotherapy is an emerging treatment for several solid tumors, which shows improved outcomes in patients. With the application of various immune-based interventions in ovarian cancer, immunotherapy has been proven useful in advanced disease (Bogani et al., 2020).

The claudin (CLDN) family consists of more than 20 transmembrane proteins, which are major components of tight junctions. They serve as a physical barrier to prevent molecules from passing freely through the paracellular space between epithelial or endothelial cell sheets and also play critical roles in maintaining cell polarity and signal transductions (Weinstein et al., 1976; Wodarz, 2000; Tsukita et al., 2001; Kirschner et al., 2013). Previous research has recognized various claudin gene expression patterns and identified several genes dysregulated in cancers (Hewitt et al., 2006). These genes play roles in the tumorigenesis of solid tumors (Swisshelm et al., 2005; Hagen, 2019) and represent promising targets for cancer detection, prognosis, and therapy (Morin, 2005). However, the relationship between claudins and the tumor immune microenvironment has not yet been elucidated. This study comprehensively analyzed claudin expression in ovarian cancer and further explored the relationship between claudins and the immune microenvironment.

MATERIALS AND METHODS

cBioPortal

The cBioPortal database[1] (Cerami et al., 2012; Gao et al., 2013) is an open platform for cancer genomics analysis. In total, 585 samples of ovarian serous cystadenocarcinoma (The Cancer Genome Atlas (TCGA), Pan-Cancer Atlas) were used for genetic variation analyses through the cBioPortal.

ONCOMINE Database Analysis

Claudin expression levels in various cancers were analyzed via the ONCOMINE database[2] (Rhodes et al., 2007), which includes more than 35 types of cancer and normal samples.

Gene Expression Profiling Interactive Analysis (GEPIA)

GEPIA v2[3] (Tang et al., 2017) is used to analyze the RNA sequencing expression data of 9736 tumors and 8587 normal samples from the TCGA and GTEx projects using a standard processing pipeline. The expression profile of the claudins in ovarian cancer was explored via GEPIA v2. The p-value cutoff was 0.05 and $|\log_2 FC|$ cutoff was 1.5.

Kaplan-Meier Plotter Database Analysis

The Kaplan-Meier plotter[4] (Gyorffy et al., 2012) assesses the effects of 54,000 genes on survival in 21 cancer types. The largest datasets include breast (n = 6234), ovarian (n = 2190), lung (n = 3452), and gastric (n = 1440) cancer. The system includes gene chip and RNA-seq data-sources from the Gene Expression Omnibus (GEO), European Genome-Phenome Archive (EGA), and TCGA databases. The prognostic significance of claudins in ovarian cancer was analyzed via the online database.

Tumor Immune Estimation Resource (TIMER)

TIMER[5] (Li et al., 2017) allows comprehensive analysis of tumor-infiltrating immune cells. The correlation between claudin expression and immune cell infiltration was analyzed using this database. TIMER v2, an updated and enhanced version of TIMER, was used to analyze immune infiltration across diverse cancer types.

Statistical Analyses

The expression levels of claudins are presented as mean ± standard deviation (SD). Kaplan-Meier survival curves were established based on the log-rank test. The hazard

Abbreviations: GEPIA, Gene Expression Profiling Interactive Analysis; TIMER, Tumor Immune Estimation Resource; GSEA, gene set enrichment analyses; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; EGA, European Genome-Phenome Archive; FPKM, Fragments per kilobase per million; TPM, transcripts per million; TILs, tumor-infiltrating lymphocytes; NK, natural killer cells; Tregs, regulatory T cells; CAFs, cancer associated fibroblasts; TAMs, Tumor-associated macrophages; CPE, clostridium perfringens enterotoxin. OS, overall survival; PFS, progression free survival; PPS, post progression survival.

[1]https://www.cbioportal.org/

[2]https://www.ONCOMINE.org/resource/login.html

[3]http://gepia2.cancer-pku.cn/

[4]http://kmplot.com/analysis/index.php?p=background

[5]https://cistrome.shinyapps.io/timer/

ratio (HR) was determined using the Cox model. Spearman correlation was used for correlation analysis. A p-value of < 0.05 was considered to be significant.

RESULTS

Gene Variation of Claudins in Ovarian Cancer

Twenty-four reviewed proteins of the claudin family were obtained from the UniProt Knowledgebase (UniProtKB)[6] (**Table 1**) [an additional file shows this in more detail (see **Table 1**)]. Firstly, we investigated the genetic variation of the claudin family in ovarian cancer using the cBioPortal for Cancer Genomics. Twenty-four genes were queried in 585 samples of ovarian serous cystadenocarcinoma (TCGA, Pan-Cancer Atlas). **Figure 1A** shows the alteration frequency of genetic variation in serous ovarian cancer. As shown in **Figure 1B**, the queried genes were altered in 363 (62%) queried patients/samples. The top three gene variations were *CLDN11* (24%), *CLDN16* (22%), and *CLDN1* (16%). Differences in overall survival (OS) between the altered and unaltered groups were compared using the Kruskal-Wallis test. We found that OS was reduced in the altered group compared to the unaltered group ($p = 7.981e\text{-}3$) (**Figure 1C**). Previous studies have shown that the claudin family is dysregulated in a variety of tumors and is involved in diagnosis, tumorigenesis, and prognosis (Zhang et al., 2013; Barros-Filho et al., 2015; Zhou et al., 2018). Thus, the claudin family is worthy of further research in ovarian cancer.

[6]https://www.uniprot.org/

Expression of Claudins Is Dysregulated in Various Cancers

To explore the mRNA expression of the claudin family, we investigated the expression profiles of claudins in various cancers via the ONCOMINE database. The thresholds were: p-value of 0.05, fold change of 1.5, and gene rank of all. Significant analyses are shown in **Supplementary Figure 1** (those with < 3 significant analyses were not considered). Results showed that most claudins were dysregulated in various cancers. To verify the expression of claudins in ovarian cancer, GEPIA2 was used to analyze mRNA expression in TCGA and GTEx samples. The $|\ Log_2FC|$ cutoff was set to 1.5 and the p-value cutoff was set to 0.05. As shown in **Figure 2**, eight genes were overexpressed in ovarian cancer samples compared with normal tissue samples and included CLDN1, CLDN3, CLDN4, CLDN6, CLDN7, CLDN9, CLDN10, and CLDN16. Furthermore, three genes showed low expression in the ovarian cancer samples compared with normal tissue samples and included CLDN5, CLDN11, and CLDN15.

Correlation of Claudin Expression With Ovarian Cancer Prognosis

To identify genes with clinical significance, we studied the relationship between differentially expressed genes (DEGs) and ovarian cancer patient prognosis using the Kaplan-Meier plotter. As shown in **Figure 3**, overexpressed genes CLDN3, CLDN4, CLDN6, and CLDN16 were significantly correlated

TABLE 1 | Twenty-four reviewed proteins of claudin family from the UniProtKB.

Entry	Status	Gene names	Protein names	Organism
O95832	Reviewed	CLDN1	Claudin-1 (Senescence-associated epithelial membrane protein)	Homo sapiens
P78369	Reviewed	CLDN10	Claudin-10 (Oligodendrocyte-specific protein-like) (OSP-like)	Homo sapiens
O75508	Reviewed	CLDN11	Claudin-11 (Oligodendrocyte-specific protein)	Homo sapiens
P56749	Reviewed	CLDN12	Claudin-12	Homo sapiens
O95500	Reviewed	CLDN14	Claudin-14	Homo sapiens
P56746	Reviewed	CLDN15	Claudin-15	Homo sapiens
Q9Y5I7	Reviewed	CLDN16	Claudin-16 (Paracellin-1) (PCLN-1)	Homo sapiens
P56750	Reviewed	CLDN17	Claudin-17	Homo sapiens
P56856	Reviewed	CLDN18	Claudin-18	Homo sapiens
Q8N6F1	Reviewed	CLDN19	Claudin-19	Homo sapiens
P57739	Reviewed	CLDN2	Claudin-2 (SP82)	Homo sapiens
P56880	Reviewed	CLDN20	Claudin-20	Homo sapiens
Q8N7P3	Reviewed	CLDN22	Claudin-22	Homo sapiens
Q96B33	Reviewed	CLDN23	Claudin-23	Homo sapiens
A6NM45	Reviewed	CLDN24/CLDN21	Putative claudin-24 (Claudin-21)	Homo sapiens
C9JDP6	Reviewed	CLDN25	Putative claudin-25	Homo sapiens
O15551	Reviewed	CLDN3	Claudin-3 (CPE-receptor 2)	Homo sapiens
H7C241	Reviewed	CLDN34	Claudin-34	Homo sapiens
O14493	Reviewed	CLDN4	Claudin-4 (CPE-receptor)	Homo sapiens
O00501	Reviewed	CLDN5	Claudin-5 (Transmembrane protein deleted in VCFS) (TMDVCF)	Homo sapiens
P56747	Reviewed	CLDN6	Claudin-6 (Skullin)	Homo sapiens
O95471	Reviewed	CLDN7	Claudin-7	Homo sapiens
P56748	Reviewed	CLDN8	Claudin-8	Homo sapiens
O95484	Reviewed	CLDN9	Claudin-9	Homo sapiens

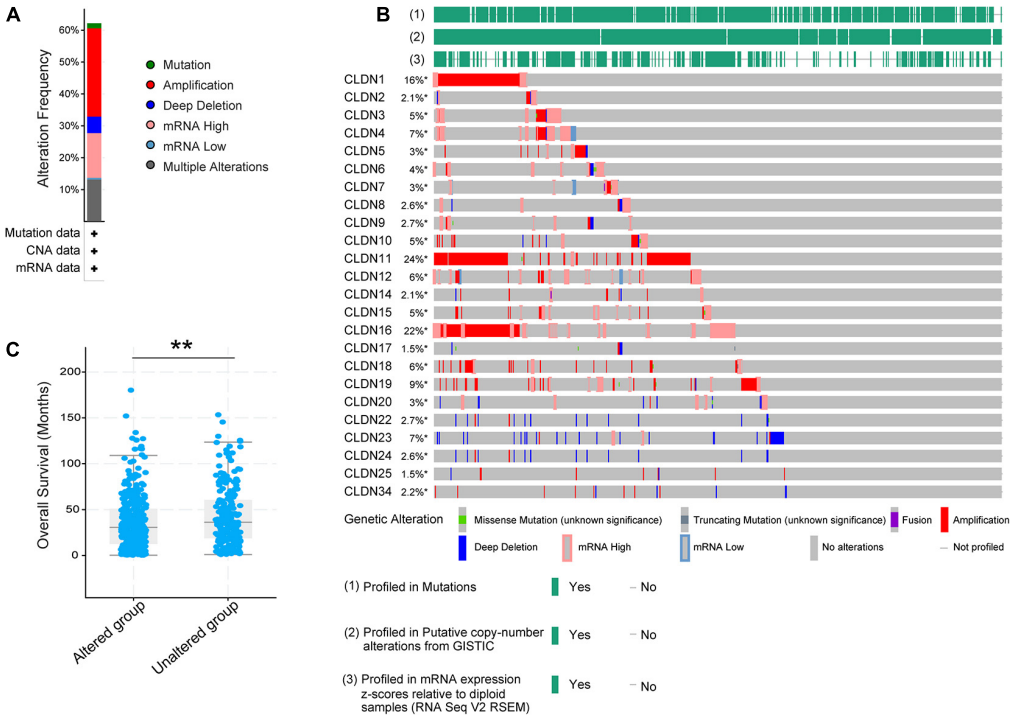


FIGURE 1 | The genetic variation of the claudin family in ovarian cancer through the cBioPortal. **(A)** The alteration frequency of the claudin family in serous ovarian cancer. **(B)** The oncoprint of the claudin family in serous ovarian cancer. **(C)** The overall survival difference of serous ovarian cancer between the altered and unaltered group (**$p < 0.01$).

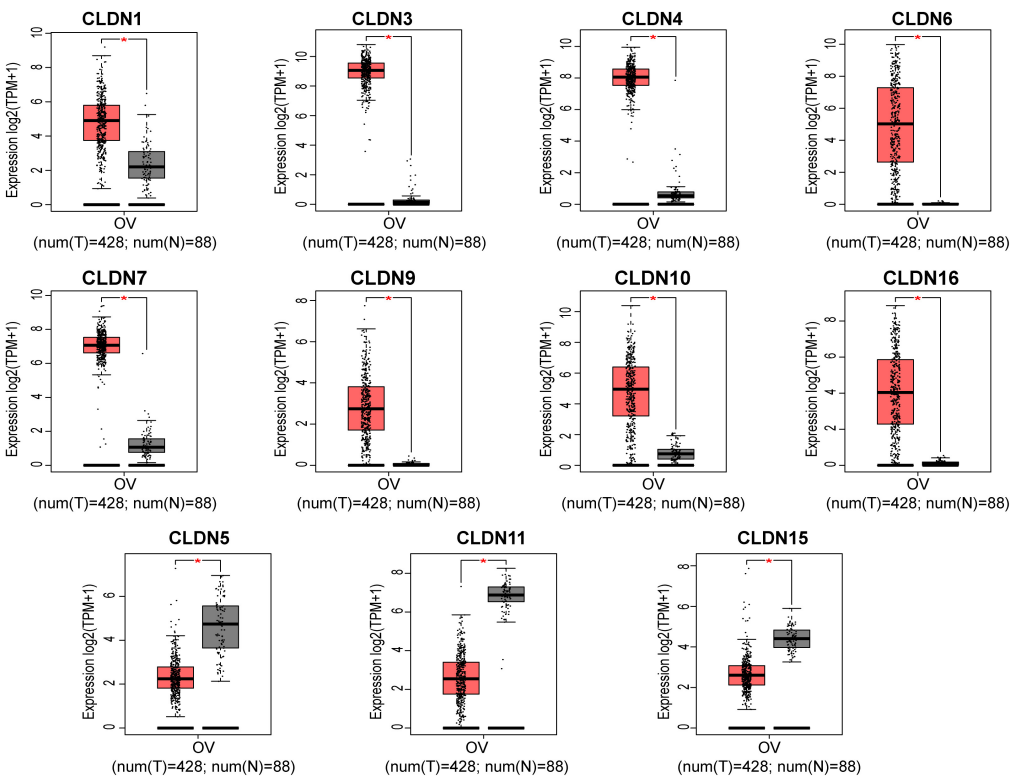


FIGURE 2 | The mRNA expression of claudins in TCGA samples and the GTEx normal samples via GEPIA2. (*$p < 0.01$).

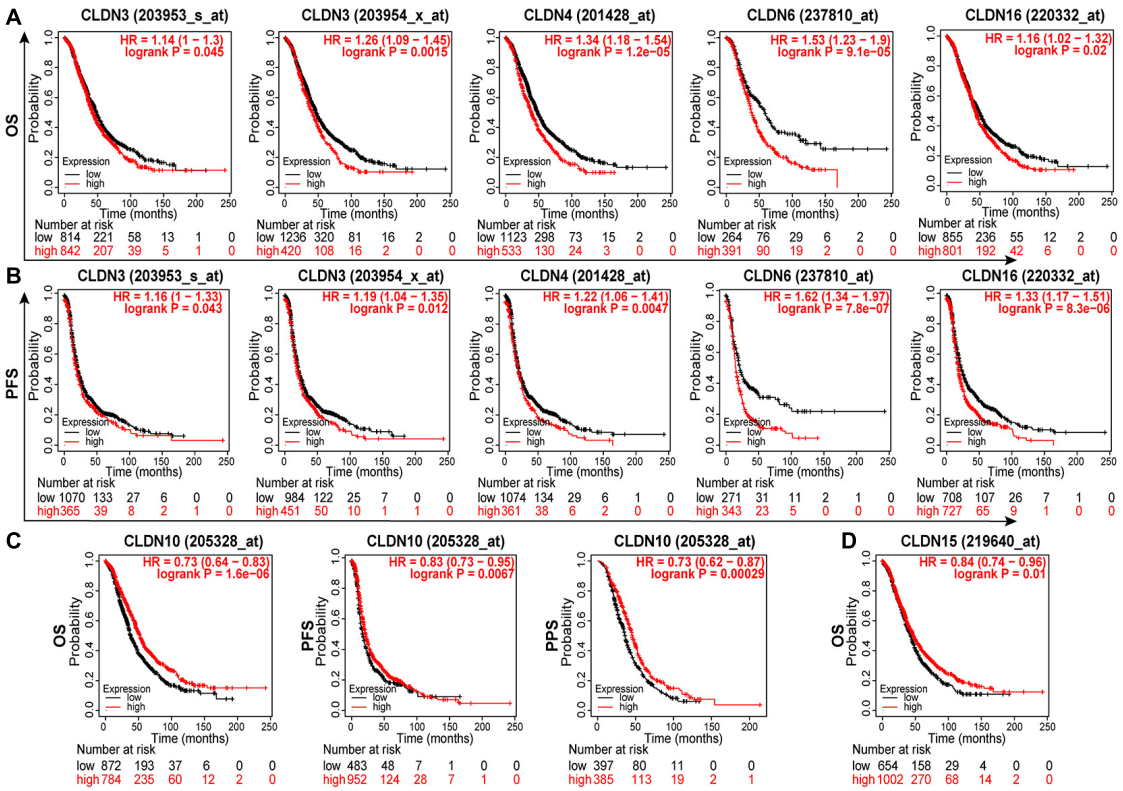


FIGURE 3 | The relationship between claudin expression and the prognosis of ovarian cancer patients through Kaplan-Meier plotter. The overexpression of CLDN3, CLDN4, CLDN6, and CLDN16 were significantly correlated with poor OS **(A)** and PFS **(B)**. **(C)** The overexpression of CLDN10 predicted good OS, PFS, and PPS. **(D)** The low expression of CLDN15 predicted poor OS in ovarian cancer. OS, overall survival; PFS, progression free survival; PPS, post progression survival.

with poor OS (**Figure 3A**) and progression free survival (PFS) (**Figure 3B**). In addition, high expression of CLDN10 and CLDN15 were predictive of a good prognosis in ovarian cancer patients (**Figures 3C,D**). Surprisingly, CLDN10 was overexpressed in cancer, but patients with high expression of CLDN10 showed good OS (HR = 0.73, logrank P = 1.6e-06), PFS (HR = 0.83, logrank P = 0.0067), and post progression survival (PPS, HR = 0.73, logrank P = 0.00029). These results are somewhat counterintuitive, and the underlying mechanism requires further exploration.

TCGA projects have identified four molecular subtypes of high-grade serous ovarian carcinoma (HGSOC) (Cancer Genome Atlas Research Network, 2011): (i) the differentiated subtype; (ii) the immunoreactive subtype; (iii) the mesenchymal subtype; and (iv) the proliferative subtype. Among them, T-cell chemokine ligands, CXCL11 and CXCL10, and the receptor, CXCR3, characterized the immunoreactive subtype. Then, Thorsson et al. (2018) developed a global immune classification of solid tumors based on the transcriptomic profiles of 33 cancer types. They identified six distinct immune subtypes: C1 (Wound healing); C2 (IFN-γ dominant); C3 (Inflammatory); C4 (Lymphocyte depleted); C5 (Immunologically quiet); C6 (TGF-β dominant). These six categories represent features of the tumor microenvironment (Charoentong et al., 2017). In this research, we explored the relationships between the expression of differentially expressed genes related to prognosis and molecular subtypes or immune subtypes of ovarian cancer via the TISIDB (Ru et al., 2019). The Kruskal-Wallis test was used. As **Supplementary Figure 2** shows, claudins including CLDN3, CLDN6, CLDN10, and CLDN15 are differentially expressed in different immune subtypes. And, claudins including CLDN3, CLDN4, CLDN6, CLDN10, and CLDN16 are differentially expressed in different molecular subtypes (**Supplementary Figure 3**). Among them, CLDN6 is relatively low expression, and CLDN10 is relatively high expression in the immunoreactive subtype.

GSEA of Immunological Signature Gene Sets

To characterize the potential function of claudins, GSEA was performed using gene expression data from TCGA ovarian cancer patients. Immunological signature gene sets were used. As shown in **Figure 4**, CLDN6 and CLDN10 were related to the effector differentiation of B cell, CD4 T cell, and CD8 T cell.

Correlation Analyses Between Claudins and Tumor Immune Microenvironment

To understand the role of claudins in immunity, we downloaded 379 RNA-seq FPKM (Fragments per kilobase per million) data

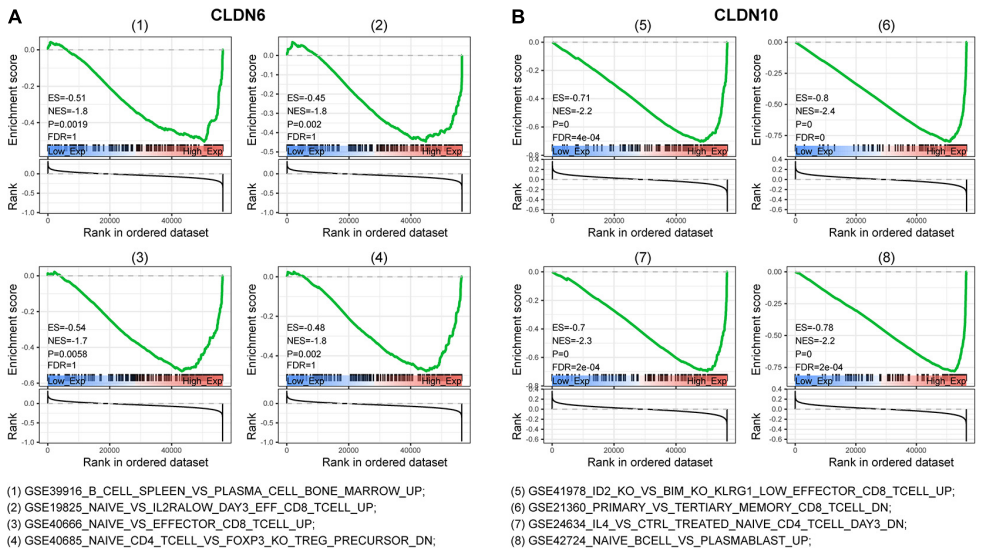


FIGURE 4 | Gene set enrichment analysis (GSEA) of c7 (immunologic signatures) for CLDN6 and CLDN10. CLDN6 **(A)** and CLDN10 **(B)** were related to effector differentiation of B cell, CD8 T cell, and CD4 T cell.

of ovarian cancer from TCGA. Subsequently, the FPKM was converted to TPM (transcripts per million) (Li et al., 2010). The ESTIMATE algorithm (Yoshihara et al., 2013) was used to predict tumor purity based on TCGA ovarian cancer samples. Then, the relationship between claudin expression and the immune microenvironment was explored. As shown in **Figure 5A**, a significant negative correlation between CLDN6 expression and the immune score was observed (Spearman correlation = −0.23, $p < 0.001$). A significant positive correlation between CLDN10 expression and immune score (spearman correlation = 0.21, $p < 0.001$) was observed (**Figure 5B**). However, the expression levels of CLDN6 and CLDN10 were not correlated with the stromal score.

We next examined the relationship between immune cell infiltration and claudin expression. RNA-seq TPM data ($n = 379$) from TCGA ovarian cancer were used to assess 22 immune cells subtype concentrations with the CIBERSORT algorithm (Newman et al., 2019). TCGA samples were grouped by the median values of CLDN6 and CLDN10, respectively. Activated dendritic cells differed significantly between the CLDN6_high and CLDN6_low groups. Several cell types were significantly different between the CLDN10_high and CLDN10_low group, including naïve B cells, memory B cells, naïve CD4 T cells, CD4 memory-activated T cells, monocytes, M1 macrophages, and activated dendritic cells (**Figure 5C**).

The microarray expression values of ovarian cancer were used to calculate the abundances of six immune infiltrates (B cells, $CD4^+$ T cells, $CD8^+$ T cells, Neutrophils, Macrophages, and Dendritic cells) via the TIMER algorithm (Yoshihara et al., 2013). The gene expression levels correlated with tumor purity are displayed in the left-most panel (**Figures 6A,B**). Results showed that CLDN6 expression was negatively correlated with infiltration of B cell (partial correlation = −0.284, p = 2.21e-10), $CD8^+$ T cells (partial correlation = −0.254, p = 1.64e-08), neutrophils (partial correlation = −0.152, p = 8.29e-04), and dendritic cells (partial correlation = −0.182, p = 6.31e-05) (**Figure 6A**). In contrast, there was a small but significant positive correlation between CLDN10 expression and infiltration of neutrophils (partial correlation = 0.185, p = 4.66e-05), and dendritic cells (partial correlation = 0.153, p = 7.74e-04) (**Figure 6B**).

To more accurately describe the relationship between gene expression and immune cell infiltration, we used the TIMER, CIBERSORT, quanTIseq, xCell, MCP-counter, and EPIC algorithms to assess the immune infiltration in tumor tissue (Sturm et al., 2019). TIMER2 provides a platform to analyze immune infiltrates across diverse cancer types based on available TCGA RNA-seq data (Li et al., 2016; Li T. et al., 2020). The correlations between claudin expression (CLDN6 and CLDN10) and immune cell infiltration in ovarian cancer are shown in **Table 2**. As seen in **Figure 6C**, CLDN6 was negatively correlated with immune cell infiltration, including that of B cells, $CD8^+$ T cells, effector memory $CD4^+$ T cells, M1 macrophages, and myeloid dendritic cells. In contrast, CLDN10 was positively correlated with immune cell infiltration, including that of B cells, $CD8^+$ T cells, effector memory $CD4^+$ T cells, M1 macrophages, and myeloid dendritic cells (**Figure 6D**). Relevant evidence suggests that cancer-associated fibroblasts (CAFs) play an important role in the progression of ovarian cancer (Mhawech-Fauceglia et al., 2014; Leung et al., 2018). Interestingly, here, CAFs also showed a positive correlation with CLDN6 expression, but a negative correlation with CLDN10 expression. In ovarian cancer, increased infiltration of tumor-infiltrating lymphocytes (TILs), and more specifically $CD8^+$ T cells, have been proven to be associated with improved clinical outcomes (Sato et al., 2005; Hamanishi et al., 2007; Ovarian Tumor Tissue Analysis et al., 2017). These results suggest that CLDN6 and CLDN10 may participate in immune cell infiltration in ovarian cancer, and these mechanisms may be the reason for poor prognosis.

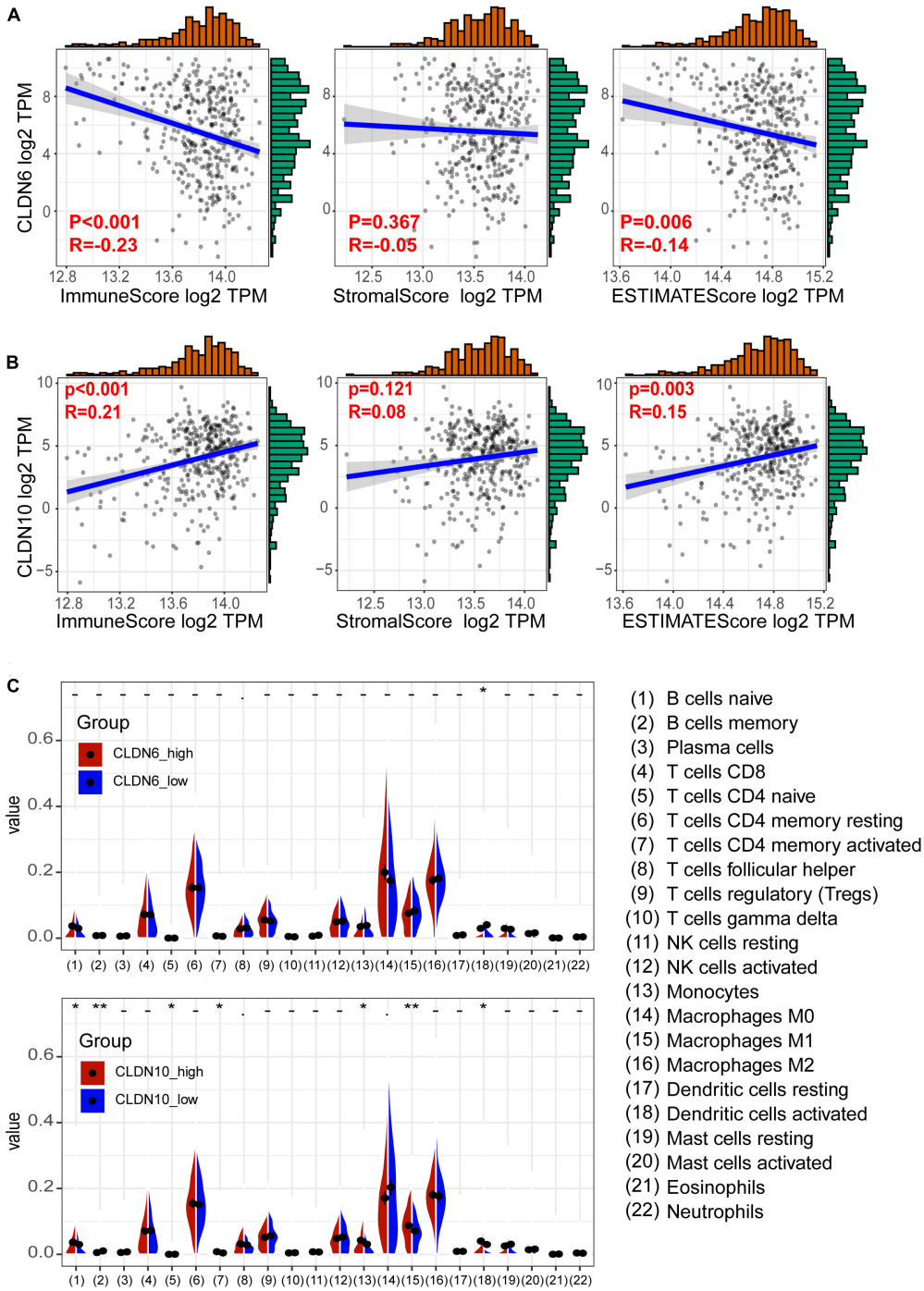


FIGURE 5 | Relationship between claudins expression and tumor immune microenvironment. **(A)** The expression of CLDN6 has a negative correlation with immune score and ESTIMATE score. **(B)** The expression of CLDN10 has a positive correlation with immune score and ESTIMATE score. **(C)** The difference of 22 immune cells between the claudin-high group and claudin-low group (*$p < 0.05$, **$p < 0.01$).

Relationship Between Claudin Expression and Gene Markers of Immune Cells

To further illustrate the correlation between claudins and the immune microenvironment, we analyzed the relationship between CLDN6 and CLDN10 expression and gene markers of various immune cells in ovarian cancer (TIMER2 database), including B cells, T cells (general), $CD8^+$ T cells, macrophages, dendritic cells, neutrophils, monocytes, natural killer (NK) cells, and regulatory T cells (Tregs) (**Table 3**). Purity-adjusted

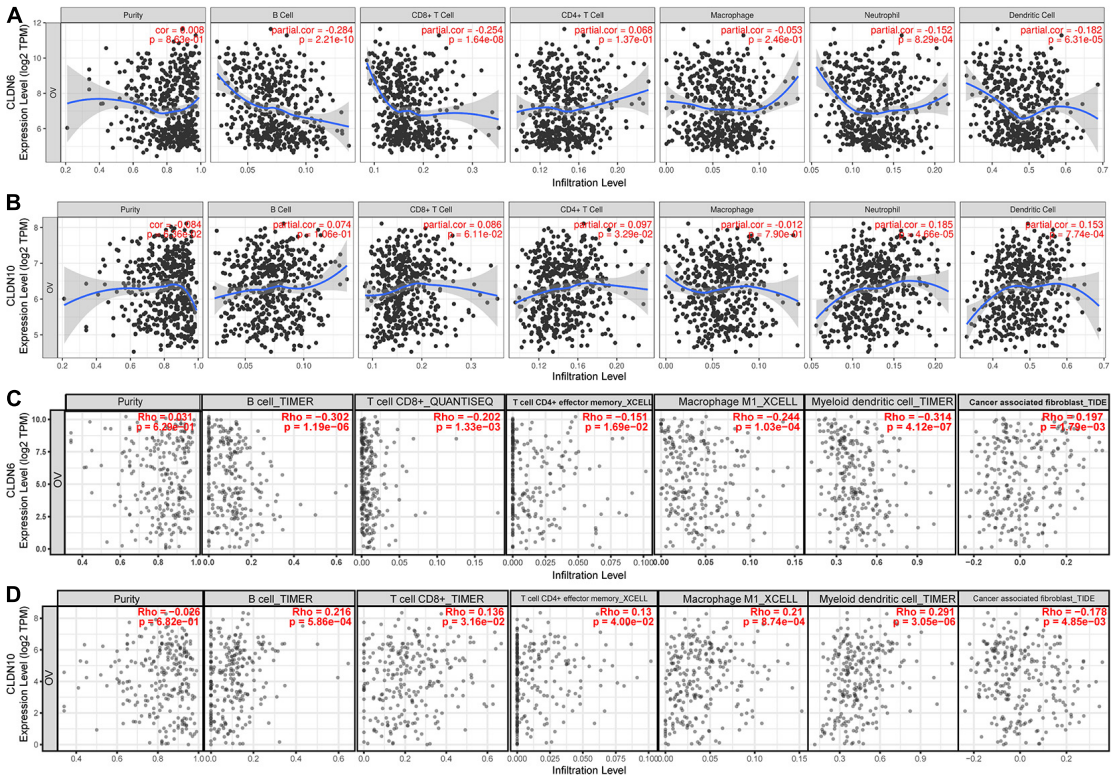


FIGURE 6 | The relationship between immune cell infiltration and claudin expression. Correlation analysis of immune cell infiltration and CLDN6 expression **(A)**, and CLDN10 expression **(B)** based on the microarray expression values of ovarian cancer through TIMER. Correlation analysis of immune cell infiltration and CLDN6 expression **(C)**, and CLDN10 expression **(D)** based on available TCGA RNA-seq data of ovarian cancer via TIMER2.

correlation heatmaps are shown in **Supplementary Figure 4**. After correlation adjustment by purity, CLDN6 expression was negatively correlated with most gene markers of dendritic cells, M1 macrophages, monocytes, NK cells, and tumor-associated macrophages (TAMs) in ovarian cancer. In contrast, CLDN10 expression was positively correlated with gene markers of dendritic cells, T cells (general), and TAMs in ovarian cancer.

Studies have shown that the tumor-infiltrating immune cells mentioned above are related to the tumor immunotherapy response (Rodriguez et al., 2018). Immune cell-based immunotherapy (Baci et al., 2020), including NK Cells (Nersesian et al., 2019) and dendritic cells (Stiff et al., 2013), play important roles in the treatment of ovarian cancer. Taken together, these analyses and our research indicate that CLDN6 and CLDN10 may play important roles in immunotherapy in the future.

DISCUSSION

CLDN6 and CLDN10 are important components of the claudin family related to tight junctions. Claudins were considered promising targets for diagnosis and therapy since they were involved in uncontrolled cancer growth and metastasis (Martin and Jiang, 2001; Morin, 2005; Bose and Mukhopadhyay, 2010). Moreover, studies have shown that they not only play a vital role in tumorigenesis (Swisshelm et al., 2005; Arabzadeh et al., 2007; Hagen, 2019), but also drug resistance (Gao et al., 2017).

CLDN6 had been demonstrated abnormal expression and can be a prognostic marker in cancers including ovarian cancer (Wang et al., 2013), endometrial cancer (Kojima et al., 2020), gastric cancer (Kohmoto et al., 2020), breast carcinoma (Liu et al., 2016; Jia et al., 2019), and lung cancer (Micke et al., 2014). Bioinformatic analysis has revealed that CLDN6 is regulated by a diverse set of transcription factors and promotes cancer cell behavior via the ASK1-p38/JNK MAPK secretory signaling pathway (Lin et al., 2017). A study revealed that CLDN6 may be a novel targeted therapy for ovarian cancer as a receptor for clostridium perfringens enterotoxin (Lal-Nag et al., 2012). In addition, 6PHU3, a T-cell-engaging bispecific single chain antibody with anti-CD3/anti-CLDN6 specificities, upregulated the cytotoxicity of T cells and made T cells acquire an effector phenotype (Stadler et al., 2016). Another recent study showed that CLDN6 as a chimeric antigen receptor target in solid tumors can be a strategy to overcome inefficient CAR-T cell stimulation *in vivo* (Reinhard et al., 2020). These studies suggested that CLDN6 has important research value in the treatment of cancer.

CLDN10, a glandular epithelial marker in epithelial ovarian cancer (Seo et al., 2010), was reported to be a key immune-related

TABLE 2 | Correlation analysis between claudins and immune infiltration in ovarian cancer via TIMER2.0.

		CLDN6			CLDN10		
Cancer	**Infiltrates**	**rho**	***p***	**adj.p**	**rho**	***p***	**adj.p**
OV (*n* = 303)	B cell memory_CIBERSORT	−0.018	0.777	0.9214	−0.1938	**	*
OV (*n* = 303)	B cell memory_CIBERSORT-ABS	−0.0185	0.7713	0.9214	−0.1795	**	*
OV (*n* = 303)	B cell memory_XCELL	−0.0386	0.5446	0.7855	0.091	0.1521	0.3381
OV (*n* = 303)	B cell naive_CIBERSORT	0.0053	0.9343	0.9895	0.255	***	***
OV (*n* = 303)	B cell naive_CIBERSORT-ABS	−0.0058	0.9272	0.9895	0.2577	***	***
OV (*n* = 303)	B cell naive_XCELL	0.0915	0.15	0.4803	−0.142	*	0.0952
OV (*n* = 303)	B cell plasma_CIBERSORT	0.1164	0.0666	0.3075	−0.0337	0.5963	0.7755
OV (*n* = 303)	B cell plasma_CIBERSORT-ABS	0.0741	0.2443	0.5768	0.0036	0.9552	0.9837
OV (*n* = 303)	B cell plasma_XCELL	0.04	0.5302	0.7759	−0.12	0.0587	0.1821
OV (*n* = 303)	B cell_EPIC	0.045	0.4801	0.7541	−0.149	*	0.0782
OV (*n* = 303)	B cell_MCPCOUNTER	0.2482	***	**	−0.0836	0.1888	0.3814
OV (*n* = 303)	B cell_QUANTISEQ	0.1153	0.0694	0.3139	−0.1177	0.0636	0.1866
OV (*n* = 303)	B cell_TIMER	−0.3021	***	***	0.2164	***	**
OV (*n* = 303)	B cell_XCELL	−0.1283	*	0.2616	0.0756	0.2345	0.4401
OV (*n* = 303)	Cancer associated fibroblast_EPIC	0.1377	*	0.1353	−0.0907	0.1537	0.4081
OV (*n* = 303)	Cancer associated fibroblast_MCPCOUNTER	0.1594	*	0.0746	−0.0955	0.133	0.3766
OV (*n* = 303)	Cancer associated fibroblast_TIDE	0.197	**	*	−0.178	**	*
OV (*n* = 303)	Cancer associated fibroblast_XCELL	0.1913	**	*	−0.1201	0.0585	0.2122
OV (*n* = 303)	Class-switched memory B cell_XCELL	−0.1073	0.091	0.3747	0.1094	0.085	0.2267
OV (*n* = 303)	Common lymphoid progenitor_XCELL	−0.0628	0.3235	0.6596	0.0795	0.2112	0.4607
OV (*n* = 303)	Common myeloid progenitor_XCELL	−0.1444	*	0.139	0.0333	0.6009	0.8165
OV (*n* = 303)	Endothelial cell_EPIC	0.092	0.1478	0.4554	−0.1135	0.0738	0.2627
OV (*n* = 303)	Endothelial cell_MCPCOUNTER	0.15	*	0.1218	−0.1109	0.0807	0.2771
OV (*n* = 303)	Endothelial cell_XCELL	0.0923	0.1466	0.4554	−0.0893	0.16	0.403
OV (*n* = 303)	Eosinophil_CIBERSORT	0.1312	*	0.1921	−0.006	0.9255	0.9687
OV (*n* = 303)	Eosinophil_CIBERSORT-ABS	0.1299	*	0.1983	−0.0054	0.9323	0.9707
OV (*n* = 303)	Eosinophil_XCELL	0.0472	0.4588	0.7698	−0.0908	0.1531	0.3919
OV (*n* = 303)	Granulocyte-monocyte progenitor_XCELL	0.0423	0.5061	0.7873	0.0061	0.9236	0.9687
OV (*n* = 303)	Hematopoietic stem cell_XCELL	0.0704	0.2685	0.6192	−0.1648	**	0.0568
OV (*n* = 303)	Macrophage M0_CIBERSORT	0.12	0.0586	0.2045	−0.1693	**	*
OV (*n* = 303)	Macrophage M0_CIBERSORT-ABS	0.0854	0.1791	0.431	−0.1219	0.0546	0.168
OV (*n* = 303)	Macrophage M1_CIBERSORT	−0.1565	*	0.0812	0.1868	**	*
OV (*n* = 303)	Macrophage M1_CIBERSORT-ABS	−0.1201	0.0585	0.2045	0.1764	**	*
OV (*n* = 303)	Macrophage M1_QUANTISEQ	−0.1115	0.0792	0.2541	0.1631	**	*
OV (*n* = 303)	Macrophage M1_XCELL	−0.2436	***	**	0.2096	***	**
OV (*n* = 303)	Macrophage M2_CIBERSORT	−0.1332	*	0.1481	0.0946	0.1366	0.3176
OV (*n* = 303)	Macrophage M2_CIBERSORT-ABS	−0.1201	0.0585	0.2045	0.1292	*	0.1388
OV (*n* = 303)	Macrophage M2_QUANTISEQ	−0.0632	0.3207	0.6029	0.1233	0.0521	0.1619
OV (*n* = 303)	Macrophage M2_TIDE	0.3074	***	***	−0.2819	***	***
OV (*n* = 303)	Macrophage M2_XCELL	−0.2827	***	***	0.0992	0.1183	0.2886
OV (*n* = 303)	Macrophage/Monocyte_MCPCOUNTER	−0.1563	*	0.0812	0.0675	0.2884	0.5842
OV (*n* = 303)	Macrophage/Monocyte_MCPCOUNTER	−0.1563	*	0.1115	0.0675	0.2884	0.5244
OV (*n* = 303)	Macrophage_EPIC	−0.1983	**	*	0.1515	*	0.0698
OV (*n* = 303)	Macrophage_TIMER	0.0371	0.5602	0.7984	−0.1785	**	*
OV (*n* = 303)	Macrophage_XCELL	−0.2767	***	***	0.1879	**	*
OV (*n* = 303)	Mast cell activated_CIBERSORT	0.0135	0.8325	0.9299	−0.0271	0.6699	0.8355
OV (*n* = 303)	Mast cell activated_CIBERSORT-ABS	0.0118	0.8527	0.9352	−0.0284	0.6555	0.8323
OV (*n* = 303)	Mast cell resting_CIBERSORT	−0.0645	0.3106	0.65	0.0765	0.2289	0.4775
OV (*n* = 303)	Mast cell resting_CIBERSORT-ABS	−0.0775	0.223	0.5626	0.0979	0.1233	0.3433
OV (*n* = 303)	Mast cell_XCELL	−0.1516	*	0.1157	−0.0698	0.2723	0.5282
OV (*n* = 303)	MDSC_TIDE	0.3588	***	***	−0.1393	*	0.1339
OV (*n* = 303)	Monocyte_CIBERSORT	0.0449	0.481	0.7776	0.0739	0.2454	0.5578
OV (*n* = 303)	Monocyte_CIBERSORT-ABS	−0.0003	0.9966	0.9966	0.124	0.0507	0.2355
OV (*n* = 303)	Monocyte_MCPCOUNTER	−0.1563	*	0.1115	0.0675	0.2884	0.5842
OV (*n* = 303)	Monocyte_QUANTISEQ	−0.3974	***	***	0.1651	**	0.0626

(Continued)

TABLE 2 | Continued

		CLDN6			CLDN10		
Cancer	**Infiltrates**	**rho**	***p***	**adj.p**	**rho**	***p***	**adj.p**
OV (*n* = 303)	Monocyte_XCELL	−0.1109	0.0807	0.3318	0.0824	0.195	0.5043
OV (*n* = 303)	Myeloid dendritic cell activated_CIBERSORT	−0.1643	**	0.0559	0.1554	*	0.069
OV (*n* = 303)	Myeloid dendritic cell activated_CIBERSORT-ABS	−0.1626	*	0.0573	0.1618	*	0.0564
OV (*n* = 303)	Myeloid dendritic cell activated_XCELL	−0.2327	***	**	0.1691	**	*
OV (*n* = 303)	Myeloid dendritic cell resting_CIBERSORT	−0.0371	0.5605	0.7955	−0.0546	0.3908	0.635
OV (*n* = 303)	Myeloid dendritic cell resting_CIBERSORT-ABS	−0.0367	0.5642	0.7962	−0.0475	0.4551	0.6843
OV (*n* = 303)	Myeloid dendritic cell_MCPCOUNTER	−0.1032	0.1044	0.2989	0.0276	0.6652	0.8057
OV (*n* = 303)	Myeloid dendritic cell_QUANTISEQ	0.363	***	***	−0.1552	*	0.0693
OV (*n* = 303)	Myeloid dendritic cell_TIMER	−0.3143	***	***	0.2908	***	***
OV (*n* = 303)	Myeloid dendritic cell_XCELL	−0.1196	0.0595	0.2138	0.1565	*	0.0675
OV (*n* = 303)	Neutrophil_CIBERSORT	−0.1029	0.1053	0.4127	0.1114	0.0793	0.2453
OV (*n* = 303)	Neutrophil_CIBERSORT-ABS	−0.0951	0.1345	0.4605	0.1072	0.0913	0.2681
OV (*n* = 303)	Neutrophil_MCPCOUNTER	−0.0017	0.9786	0.9929	−0.0367	0.5639	0.7514
OV (*n* = 303)	Neutrophil_QUANTISEQ	0.1785	**	0.0595	−0.0207	0.7447	0.863
OV (*n* = 303)	Neutrophil_TIMER	−0.0724	0.2552	0.61	0.0614	0.3348	0.5858
OV (*n* = 303)	Neutrophil_XCELL	−0.0869	0.1714	0.5122	0.0842	0.1851	0.418
OV (*n* = 303)	NK cell activated_CIBERSORT	−0.0263	0.6796	0.8663	0.0296	0.6423	0.8424
OV (*n* = 303)	NK cell activated_CIBERSORT-ABS	−0.0404	0.5256	0.7786	0.12	0.0587	0.2122
OV (*n* = 303)	NK cell resting_CIBERSORT	−0.1009	0.1124	0.3225	−0.0246	0.6989	0.8788
OV (*n* = 303)	NK cell resting_CIBERSORT-ABS	−0.1109	0.0808	0.266	−0.0226	0.7224	0.8908
OV (*n* = 303)	NK cell_EPIC	−0.1815	**	*	0.1149	0.0703	0.2474
OV (*n* = 303)	NK cell_MCPCOUNTER	−0.1553	*	0.0848	0.1402	*	0.12
OV (*n* = 303)	NK cell_QUANTISEQ	−0.0556	0.3821	0.6781	0.0411	0.519	0.7789
OV (*n* = 303)	NK cell_XCELL	−0.0824	0.1951	0.4491	0.0799	0.2087	0.4765
OV (*n* = 303)	Plasmacytoid dendritic cell_XCELL	−0.208	***	*	0.2213	***	**
OV (*n* = 303)	T cell $CD4^+$ (non-regulatory)_QUANTISEQ	−0.0536	0.3998	0.7259	−0.0638	0.3156	0.5912
OV (*n* = 303)	T cell $CD4^+$ (non-regulatory)_XCELL	0.0077	0.9032	0.9663	−0.0723	0.2555	0.5347
OV (*n* = 303)	T cell $CD4^+$ central memory_XCELL	0.0456	0.4736	0.7811	0.0344	0.5892	0.8122
OV (*n* = 303)	T cell $CD4^+$ effector memory_XCELL	−0.1513	*	0.1109	0.1302	*	0.1625
OV (*n* = 303)	T cell $CD4^+$ memory activated_CIBERSORT	−0.0047	0.9411	0.9798	0.0538	0.3982	0.6743
OV (*n* = 303)	T cell $CD4^+$ memory activated_CIBERSORT-ABS	−0.0041	0.9485	0.9798	0.0526	0.409	0.6835
OV (*n* = 303)	T cell $CD4^+$ memory resting_CIBERSORT	0.1047	0.0994	0.329	0.015	0.8141	0.9242
OV (*n* = 303)	T cell $CD4^+$ memory resting_CIBERSORT-ABS	0.0014	0.9827	0.992	0.0943	0.1378	0.3757
OV (*n* = 303)	T cell $CD4^+$ memory_XCELL	0.0253	0.6916	0.897	0.0693	0.2762	0.5595
OV (*n* = 303)	T cell $CD4^+$ naive_CIBERSORT	0.1349	*	0.1741	−0.1428	*	0.1147
OV (*n* = 303)	T cell $CD4^+$ naive_CIBERSORT-ABS	0.1349	*	0.1741	−0.1428	*	0.1147
OV (*n* = 303)	T cell $CD4^+$ naive_XCELL	−0.1611	*	0.0828	0.1101	0.083	0.2652
OV (*n* = 303)	T cell $CD4^+$ Th1_XCELL	−0.1385	*	0.1608	0.0499	0.4328	0.7009
OV (*n* = 303)	T cell $CD4^+$ Th2_XCELL	0.0625	0.3263	0.6506	0.0766	0.2287	0.522
OV (*n* = 303)	T cell $CD4^+$ _EPIC	0.0428	0.5014	0.8099	−0.0148	0.8168	0.9242
OV (*n* = 303)	T cell $CD4^+$ _TIMER	0.1149	0.0703	0.2735	−0.0058	0.9273	0.9753
OV (*n* = 303)	T cell $CD8^+$ central memory_XCELL	−0.1749	**	*	0.1568	*	0.0801
OV (*n* = 303)	T cell $CD8^+$ effector memory_XCELL	0.0858	0.177	0.4688	0.0796	0.2107	0.4441
OV (*n* = 303)	T cell $CD8^+$ naive_XCELL	0.1924	**	*	−0.1191	0.0606	0.2145
OV (*n* = 303)	T cell $CD8^+$ _CIBERSORT	−0.0534	0.4012	0.6829	0.0301	0.6366	0.8318
OV (*n* = 303)	T cell $CD8^+$ _CIBERSORT-ABS	−0.0453	0.4765	0.7086	0.0702	0.2695	0.5033
OV (*n* = 303)	T cell $CD8^+$ _EPIC	0.0434	0.4951	0.7166	−0.0542	0.3944	0.6552
OV (*n* = 303)	T cell $CD8^+$ _MCPCOUNTER	−0.0322	0.613	0.7909	0.0925	0.1455	0.3528
OV (*n* = 303)	T cell $CD8^+$ _QUANTISEQ	−0.2023	**	*	0.1851	**	*
OV (*n* = 303)	T cell $CD8^+$ _TIMER	−0.1707	**	*	0.1363	*	0.139
OV (*n* = 303)	T cell $CD8^+$ _XCELL	−0.0544	0.3923	0.6765	−0.0078	0.9028	0.9629
OV (*n* = 303)	T cell follicular helper_CIBERSORT	−0.036	0.5716	0.8255	0.0032	0.9605	0.9889
OV (*n* = 303)	T cell follicular helper_CIBERSORT-ABS	−0.0618	0.3316	0.7046	0.058	0.3618	0.6466
OV (*n* = 303)	T cell gamma delta_CIBERSORT	−0.0281	0.6591	0.8771	−0.0738	0.2458	0.5578
OV (*n* = 303)	T cell gamma delta_CIBERSORT-ABS	−0.0276	0.6642	0.8771	−0.0735	0.2481	0.5578
OV (*n* = 303)	T cell gamma delta_XCELL	−0.0545	0.3918	0.7431	0.03	0.6374	0.8533
OV (*n* = 303)	T cell NK_XCELL	−0.1745	**	0.064	−0.001	0.9869	0.9937
OV (*n* = 303)	T cell regulatory (Tregs)_CIBERSORT	−0.0056	0.9299	0.9886	−0.0417	0.5123	0.7769
OV (*n* = 303)	T cell regulatory (Tregs)_CIBERSORT-ABS	−0.0278	0.6622	0.8546	−0.006	0.9248	0.971
OV (*n* = 303)	T cell regulatory (Tregs)_QUANTISEQ	−0.001	0.9873	0.9998	0.1678	**	*
OV (*n* = 303)	T cell regulatory (Tregs)_XCELL	0.0683	0.283	0.575	0.0402	0.5276	0.783

*$P < 0.05$;**$P < 0.01$;***$P < 0.001$.

TABLE 3 | Correlation analysis between claudins and markers of immune cells in ovarian cancer via TIMER2.0.

			CLDN6			CLDN10		
Cancer	**Immune cells**	**Gene markers**	**rho**	***p***	**adj.p**	**rho**	***p***	**adj.p**
OV (*n* = 303)	B cell	CD19	0.122	0.052	0.189	−0.075	0.268	0.477
OV (*n* = 303)	B cell	CD79A	0.022	0.692	0.853	−0.063	0.307	0.521
OV (*n* = 303)	$CD8^+$ T cell	CD8A	−0.103	0.103	0.305	0.0977	0.121	0.298
OV (*n* = 303)	$CD8^+$ T cell	CD8B	−0.032	0.613	0.798	0.0925	0.145	0.336
OV (*n* = 303)	DC	CD1C	−0.158	*	0.098	0.0864	0.172	0.467
OV (*n* = 303)	DC	HLA-DPA1	−0.253	***	**	0.2298	***	**
OV (*n* = 303)	DC	HLA-DPB1	−0.3	***	***	0.2535	***	***
OV (*n* = 303)	DC	HLA-DQB1	−0.224	***	**	0.2259	***	**
OV (*n* = 303)	DC	HLA-DRA	−0.325	***	***	0.2428	***	**
OV (*n* = 303)	DC	ITGAX	−0.182	**	*	0.0859	0.178	0.469
OV (*n* = 303)	DC	NRP1	0.1252	*	0.235	−0.004	0.996	0.997
OV (*n* = 303)	M1 Macrophage	IRF5	−0.186	**	*	0.0896	0.157	0.342
OV (*n* = 303)	M1 Macrophage	NOS2	0.1436	*	0.106	−0.033	0.545	0.757
OV (*n* = 303)	M1 Macrophage	PTGS2	0.0961	0.135	0.347	0.0093	0.886	0.942
OV (*n* = 303)	M2 Macrophage	CD163	−0.106	0.096	0.289	0.0646	0.31	0.529
OV (*n* = 303)	M2 Macrophage	MS4A4A	−0.112	0.072	0.238	0.1147	0.077	0.206
OV (*n* = 303)	M2 Macrophage	VSIG4	−0.152	*	0.086	0.0768	0.224	0.433
OV (*n* = 303)	Monocyte	CD86	−0.222	***	**	0.1457	*	0.084
OV (*n* = 303)	Monocyte	CSF1R	−0.196	**	*	0.0717	0.256	0.473
OV (*n* = 303)	NK cell	KIR2DL1	−0.001	0.924	0.986	0.0991	0.117	0.388
OV (*n* = 303)	NK cell	KIR2DL3	−0.226	***	**	0.1527	*	0.096
OV (*n* = 303)	NK cell	KIR2DL4	−0.258	***	**	0.1563	*	0.08
OV (*n* = 303)	NK cell	KIR2DS4	−0.097	0.121	0.391	0.0847	0.185	0.475
OV (*n* = 303)	NK cell	KIR3DL1	0.019	0.764	0.936	0.1037	0.105	0.348
OV (*n* = 303)	NK cell	KIR3DL2	−0.063	0.318	0.631	0.1495	*	0.107
OV (*n* = 303)	NK cell	KIR3DL3	−0.044	0.466	0.751	0.0571	0.368	0.682
OV (*n* = 303)	Neutrophil	CCR7	−0.068	0.324	0.633	0.0943	0.138	0.421
OV (*n* = 303)	Neutrophil	CEACAM8	−0.065	0.344	0.658	−0.034	0.619	0.839
OV (*n* = 303)	Neutrophil	ITGAM	−0.185	**	*	0.0575	0.367	0.682
OV (*n* = 303)	T cell (general)	CD2	−0.157	*	0.065	0.1651	**	*
OV (*n* = 303)	T cell (general)	CD3D	−0.142	*	0.106	0.1524	*	0.077
OV (*n* = 303)	T cell (general)	CD3E	−0.126	*	0.175	0.1581	*	0.051
OV (*n* = 303)	TAM	CCL2	−0.171	**	*	0.1709	**	*
OV (*n* = 303)	TAM	CD68	−0.203	**	*	0.105	0.093	0.258
OV (*n* = 303)	TAM	IL10	0.046	0.432	0.707	−0.007	0.948	0.979
OV (*n* = 303)	Tfh	IL21	−0.128	*	0.164	−0.016	0.844	0.934
OV (*n* = 303)	Tfh	BCL6	−0.195	**	*	0.1285	*	0.158
OV (*n* = 303)	Th1	IFNG	−0.088	0.186	0.438	0.1323	*	0.146
OV (*n* = 303)	Th1	STAT1	−0.077	0.229	0.489	0.0894	0.158	0.384
OV (*n* = 303)	Th1	STAT4	−0.009	0.873	0.959	0.0768	0.225	0.468
OV (*n* = 303)	Th1	TBX21	−0.159	*	0.078	0.1587	*	0.063
OV (*n* = 303)	Th1	TNF	−0.038	0.568	0.778	0.02	0.759	0.882
OV (*n* = 303)	Th17	IL17A	−0.073	0.265	0.528	0.0043	0.943	0.981
OV (*n* = 303)	Th17	STAT3	−0.042	0.488	0.731	0.0117	0.857	0.938
OV (*n* = 303)	Th2	GATA3	−0.061	0.305	0.562	−0.084	0.206	0.436
OV (*n* = 303)	Th2	IL13	−0.017	0.761	0.893	0.0647	0.303	0.575
OV (*n* = 303)	Th2	STAT5A	−0.115	0.062	0.215	−0.051	0.423	0.681
OV (*n* = 303)	Th2	STAT6	−0.046	0.458	0.711	0.0869	0.176	0.396
OV (*n* = 303)	Treg	CCR8	−0.004	0.893	0.968	0.0211	0.741	0.882
OV (*n* = 303)	Treg	FOXP3	−0.059	0.415	0.675	0.0635	0.317	0.588
OV (*n* = 303)	Treg	STAT5B	0.152	*	0.081	−0.1677	**	*
OV (*n* = 303)	Treg	TGFB1	−0.127	0.052	0.181	0.0153	0.815	0.924

*DC, Dendritic cell; NK cel, Natural killer cell; TAM, Tumor-associated macrophage; Tfh, Follicular helper T cell; Treg, Regulatory T cell; *P < 0.05; **P < 0.01; ***P < 0.001.*

gene in clear cell renal cell carcinoma (Yang et al., 2021) and papillary thyroid carcinoma (Xiang et al., 2020). Furthermore, CLDN10 expression has proved to be a prognostic marker for ovarian cancer (Li Z. et al., 2020).

The present study combined and analyzed the prognostic potential of CLDN6 and CLDN10 with the tumor immune microenvironment. Consistent with previous reports, both CLDN6 and CLDN10 showed high expression in ovarian cancer. Prognostic analysis showed that the overexpression of CLDN6 was related to a poor prognosis for patients with ovarian cancer. However, CLDN10 overexpression predicted a better prognosis compared to the low CLDN10 expression group. We also found that CLDN6 overexpression was negatively related to immune cell infiltration, whereas CLDN10 overexpression was positively correlated with immune cell infiltration. Moreover, we found that CLDN6 and CLDN10 were related to gene markers of dendritic cells, NK cells, and TAMs. These results may explain why the overexpression of CLDN6 and low expression of CLDN10 predict poor OS in ovarian cancer. This study revealed that the prognostic potential of CLDN6 and CLDN10 is related to the tumor immune microenvironment in ovarian cancer.

Relevant evidence has emerged that immune-related gene expression and TILs are related to the prognosis, recurrence (Ojalvo et al., 2018), and chemotherapeutic response (Choi et al., 2020) of ovarian cancer. Furthermore, the presence of TILs may improve clinical outcomes in ovarian cancer patients (Odunsi, 2017). Immune cell-based immunotherapy (Baci et al., 2020), including NK Cells (Nersesian et al., 2019) and dendritic cells (Stiff et al., 2013), play an important role in the treatment of ovarian cancer. Previous studies and our analyses suggest that CLDN6 may be involved in immune evasion and that they could be an ideal candidate for immunotherapy in ovarian cancer. Future studies on the combined application of claudin-based molecular targeted therapy and immunotherapy are necessary.

CONCLUSION

CLDN6 and CLDN10 were identified as potential prognostic biomarkers and were correlated with immune cell infiltration in ovarian cancer. Our results revealed new roles for CLDN6 and CLDN10 in ovarian cancer and their potential as therapeutic targets in cancer treatment.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

PeW was responsible for the study conception and design. PG, TP, CC, and SL were involved in data acquisition, data analysis, and interpretation. PG drafted the manuscript and took charge of supervising the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by the Natural Science Foundation of China (81772775 to JW and 82072895 to PeW).

ACKNOWLEDGMENTS

This manuscript was released as a pre-print at Research Square (https://www.researchsquare.com/article/rs-40048/v1) (DOI: 10.21203/rs.3.rs-40048/v1) (Gao et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.595436/full#supplementary-material

Supplementary Figure 1 | The significant unique analyses of claudins expression in ONCOMINE. (Red: overexpression; Blue: low expression).

Supplementary Figure 2 | Associations between the expression of claudins and immune subtypes of ovarian cancer.

Supplementary Figure 3 | Associations between the expression of claudins and molecular subtypes of ovarian cancer.

Supplementary Figure 4 | The correlations between claudins (CLDN6 and CLDN10) expression and gene markers of immune cells across gynecologic oncology (Red: positive correlation; Blue: negative correlation).

REFERENCES

Arabzadeh, A., Troy, T. C., and Turksen, K. (2007). Changes in the distribution pattern of Claudin tight junction proteins during the progression of mouse skin tumorigenesis. *BMC Cancer* 7:196. doi: 10.1186/1471-2407-7-196

Baci, D., Bosi, A., Gallazzi, M., Rizzi, M., Noonan, D. M., Poggi, A., et al. (2020). The Ovarian Cancer Tumor Immune Microenvironment (TIME) as Target for Therapy: a Focus on Innate Immunity Cells as Therapeutic Effectors. *Int. J. Mol. Sci.* 21:3125. doi: 10.3390/ijms21093125

Barros-Filho, M. C., Marchi, F. A., Pinto, C. A., Rogatto, S. R., and Kowalski, L. P. (2015). High Diagnostic Accuracy Based on CLDN10, HMGA2, and LAMB3 Transcripts in Papillary Thyroid Carcinoma. *J. Clin. Endocrinol. Metab.* 100, E890–E899.

Bogani, G., Lopez, S., Mantiero, M., Ducceschi, M., Bosio, S., Ruisi, S., et al. (2020). Immunotherapy for platinum-resistant ovarian cancer. *Gynecol. Oncol.* 158, 484–488.

Bose, C. K., and Mukhopadhyay, A. (2010). Claudin and ovarian cancer. *J. Turk. Ger. Gynecol. Assoc.* 11, 48–54.

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

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. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.cd-12-0095

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Choi, K. U., Kim, A., Kim, J. Y., Kim, K. H., Hwang, C., Lee, S. J., et al. (2020). Differences in immune-related gene expressions and tumor-infiltrating lymphocytes according to chemotherapeutic response in ovarian high-grade serous carcinoma. *J. Ovarian Res.* 13:65.

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.

Gao, P., Peng, T., Cao, C., Lin, S., Wu, P., Huang, X., et al. (2020). CLDN6 and CLDN10 are Associated with Immune Infiltration of Ovarian Cancer: a Study of Claudin Family. *Front. Genet.* doi: 10.21203/rs.3.rs-40048/v1 [Preprint].

Gao, Y., Liu, X., Li, T., Wei, L., Yang, A., Lu, Y., et al. (2017). Cross-validation of genes potentially associated with overall survival and drug resistance in ovarian cancer. *Oncol. Rep.* 37, 3084–3092. doi: 10.3892/or.2017.5534

Gyorffy, B., Lanczky, A., and Szallasi, Z. (2012). Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer* 19, 197–208. doi: 10.1530/erc-11-0329

Hagen, S. J. (2019). Unraveling a New Role for Claudins in Gastric Tumorigenesis. *Cell. Mol. Gastroenterol. Hepatol.* 8, 151–152. doi: 10.1016/j.jcmgh.2019.04.004

Hamanishi, J., Mandai, M., Iwasaki, M., Okazaki, T., Tanaka, Y., Yamaguchi, K., et al. (2007). Programmed cell death 1 ligand 1 and tumor-infiltrating CD8+ T lymphocytes are prognostic factors of human ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* 104, 3360–3365. doi: 10.1073/pnas.0611533104

Henderson, J. T., Webber, E. M., and Sawaya, G. F. (2018). Screening for Ovarian Cancer: updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* 319, 595–606. doi: 10.1001/jama.2017.21421

Hewitt, K. J., Agarwal, R., and Morin, P. J. (2006). The claudin gene family: expression in normal and neoplastic tissues. *BMC Cancer* 6:186. doi: 10.1186/1471-2407-6-186

Jia, H., Chai, X., Li, S., Wu, D., and Fan, Z. (2019). Identification of claudin-2, -6, -11 and -14 as prognostic markers in human breast carcinoma. *Int. J. Clin. Exp. Pathol.* 12, 2195–2204.

Kirschner, N., Rosenthal, R., Furuse, M., Moll, I., Fromm, M., Brandner, J. M., et al. (2013). Contribution of tight junction proteins to ion, macromolecule, and water barrier in keratinocytes. *J. Invest. Dermatol.* 133, 1161–1169. doi: 10.1038/jid.2012.507

Kohmoto, T., Masuda, K., Shoda, K., Takahashi, R., Ujiro, S., Tange, S., et al. (2020). Claudin-6 is a single prognostic marker and functions as a tumor-promoting gene in a subgroup of intestinal type gastric cancer. *Gastric Cancer* 23, 403–417. doi: 10.1007/s10120-019-01014-x

Kojima, M., Sugimoto, K., Tanaka, M., Endo, Y., Kato, H., Honda, T., et al. (2020). Prognostic Significance of Aberrant Claudin-6 Expression in Endometrial Cancer. *Cancers* 12:2748. doi: 10.3390/cancers12102748

Lal-Nag, M., Battis, M., Santin, A. D., and Morin, P. J. (2012). Claudin-6: a novel receptor for CPE-mediated cytotoxicity in ovarian cancer. *Oncogenesis* 1:e33. doi: 10.1038/oncsis.2012.32

Leung, C. S., Yeung, T. L., Yip, K. P., Wong, K. K., Ho, S. Y., Mangala, L. S., et al. (2018). Cancer-associated fibroblasts regulate endothelial adhesion protein LPP to promote ovarian cancer chemoresistance. *J. Clin. Invest.* 128, 589–606. doi: 10.1172/jci95200

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26, 493–500. doi: 10.1093/bioinformatics/btp692

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17:174.

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77, e108–e110.

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48, W509–W514.

Li, Z., Xuan, W., Huang, L., Chen, N., Hou, Z., Lu, B., et al. (2020). Claudin 10 acts as a novel biomarker for the prognosis of patients with ovarian cancer. *Oncol. Lett.* 20, 373–381.

Lin, D., Guo, Y., Li, Y., Ruan, Y., Zhang, M., Jin, X., et al. (2017). Bioinformatic analysis reveals potential properties of human Claudin-6 regulation and functions. *Oncol. Rep.* 38, 875–885. doi: 10.3892/or.2017.5756

Liu, Y., Jin, X., Li, Y., Ruan, Y., Lu, Y., Yang, M., et al. (2016). DNA methylation of claudin-6 promotes breast cancer cell migration and invasion by recruiting MeCP2 and deacetylating H3Ac and H4Ac. *J. Exp. Clin. Cancer Res.* 35:120.

Martin, T. A., and Jiang, W. G. (2001). Tight junctions and their role in cancer metastasis. *Histol. Histopathol.* 16, 1183–1195.

Mhawech-Fauceglia, P., Wang, D., Samrao, D., Kim, G., Lawrenson, K., Meneses, T., et al. (2014). Clinical Implications of Marker Expression of Carcinoma-Associated Fibroblasts (CAFs) in Patients with Epithelial Ovarian Carcinoma After Treatment with Neoadjuvant Chemotherapy. *Cancer Microenviron* 7, 33–39. doi: 10.1007/s12307-013-0140-4

Micke, P., Mattsson, J. S., Edlund, K., Lohr, M., Jirström, K., Berglund, A., et al. (2014). Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. *Int. J. Cancer* 135, 2206–2214. doi: 10.1002/ijc.28857

Morin, P. J. (2005). Claudin proteins in human cancer: promising new targets for diagnosis and therapy. *Cancer Res.* 65, 9603–9606. doi: 10.1158/0008-5472.can-05-2782

Nersesian, S., Glazebrook, H., Toulany, J., Grantham, S. R., and Boudreau, J. E. (2019). Naturally Killing the Silent Killer: NK Cell-Based Immunotherapy for Ovarian Cancer. *Front. Immunol.* 10:1782. doi: 10.3389/fimmu.2019.01782

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37, 773–782. doi: 10.1038/s41587-019-0114-2

Odunsi, K. (2017). Immunotherapy in ovarian cancer. *Ann. Oncol.* 28, viii1–viii7.

Ojalvo, L. S., Thompson, E. D., Wang, T. L., Meeker, A. K., Shih, I. M., Fader, A. N., et al. (2018). Tumor-associated macrophages and the tumor immune microenvironment of primary and recurrent epithelial ovarian cancer. *Hum. Pathol.* 74, 135–147. doi: 10.1016/j.humpath.2017.12.010

Ovarian Tumor Tissue Analysis, C., Goode, E. L., Block, M. S., Kalli, K. R., Vierkant, R. A., Chen, W., et al. (2017). Dose-Response Association of CD8+ Tumor-Infiltrating Lymphocytes and Survival Time in High-Grade Serous Ovarian Cancer. *JAMA Oncol.* 3:e173290.

Reinhard, K., Rengstl, B., Oehm, P., Michel, K., Billmeier, A., Hayduk, N., et al. (2020). An RNA vaccine drives expansion and efficacy of claudin-CAR-T cells against solid tumors. *Science* 367, 446–453. doi: 10.1126/science.aay5967

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9, 166–180. doi: 10.1593/neo.07112

Rodriguez, G. M., Galpin, K. J. C., Mccloskey, C. W., and Vanderhyden, B. C. (2018). The Tumor Microenvironment of Epithelial Ovarian Cancer and Its Influence on Response to Immunotherapy. *Cancers* 10:242. doi: 10.3390/cancers10080242

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an integrated repository portal for tumor-immune system interactions. *Bioinformatics* 35, 4200–4202. doi: 10.1093/bioinformatics/btz210

Sato, E., Olson, S. H., Ahn, J., Bundy, B., Nishikawa, H., Qian, F., et al. (2005). Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* 102, 18538–18543. doi: 10.1073/pnas.0509182102

Seo, H. W., Rengaraj, D., Choi, J. W., Ahn, S. E., Song, Y. S., Song, G., et al. (2010). Claudin 10 is a glandular epithelial marker in the chicken model as human epithelial ovarian cancer. *Int. J. Gynecol. Cancer* 20, 1465–1473.

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30.

Stadler, C. R., Bähr-Mahmud, H., Plum, L. M., Schmoldt, K., Kölsch, A. C., Türeci, Ö, et al. (2016). Characterization of the first-in-class T-cell-engaging bispecific single-chain antibody for targeted immunotherapy of solid tumors expressing the oncofetal protein claudin 6. *Oncoimmunology* 5:e1091555. doi: 10.1080/2162402x.2015.1091555

Stiff, P. J., Czerlanis, C., and Drakes, M. L. (2013). Dendritic cell immunotherapy in ovarian cancer. *Expert Rev. Anticancer Ther.* 13, 43–53.

Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., et al. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* 35, i436–i445.

Swisshelm, K., Macek, R., and Kubbies, M. (2005). Role of claudins in tumorigenesis. *Adv. Drug Deliv. Rev.* 57, 919–928. doi: 10.1016/j.addr.2005.01.006

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., Zhang, Z., et al. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102.

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, 812–830.e14.

Tsukita, S., Furuse, M., and Itoh, M. (2001). Multifunctional strands in tight junctions. *Nat. Rev. Mol. Cell Biol.* 2, 285–293. doi: 10.1038/35067088

Wang, L., Jin, X., Lin, D., Liu, Z., Zhang, X., Lu, Y., et al. (2013). Clinicopathologic significance of claudin-6, occludin, and matrix metalloproteinases -2 expression in ovarian carcinoma. *Diagn. Pathol.* 8:190.

Weinstein, R. S., Merk, F. B., and Alroy, J. (1976). The structure and function of intercellular junctions in cancer. *Adv. Cancer Res.* 23, 23–89. doi: 10.1016/s0065-230x(08)60543-6

Wodarz, A. (2000). Tumor suppressors: linking cell polarity and growth control. *Curr. Biol.* 10, R624–R626.

Xiang, Z., Zhong, C., Chang, A., Ling, J., Zhao, H., Zhou, W., et al. (2020). Immune-related key gene CLDN10 correlates with lymph node metastasis but predicts favorable prognosis in papillary thyroid carcinoma. *Aging* 12, 2825–2839. doi: 10.18632/aging.102780

Yang, W., Li, L., Zhang, K., Ma, K., Gong, Y., Zhou, J., et al. (2021). CLDN10 associated with immune infiltration is a novel prognostic biomarker for clear cell renal cell carcinoma. *Epigenomics* 13, 31–45. doi: 10.2217/epi-2020-0256

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.

Zhang, Z., Wang, A., Sun, B., Zhan, Z., Chen, K., Wang, C., et al. (2013). Expression of CLDN1 and CLDN10 in lung adenocarcinoma in situ and invasive lepidic predominant adenocarcinoma. *J. Cardiothorac. Surg* 8:95.

Zhou, Y., Xiang, J., Bhandari, A., Guan, Y., Xia, E., Zhou, X., et al. (2018). CLDN10 is Associated with Papillary Thyroid Cancer Progression. *J Cancer* 9, 4712–4717. doi: 10.7150/jca.28636

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2021
doi: 10.3389/fgene.2021.607817

Multi-Omics Data Fusion for Cancer Molecular Subtyping Using Sparse Canonical Correlation Analysis

*Lin Qi[1], Wei Wang[1], Tan Wu[1], Lina Zhu[1], Lingli He[1] and Xin Wang[1,2]**

[1] Department of Biomedical Sciences, City University of Hong Kong, Shenzhen, China, [2] Key Laboratory of Biochip Technology, Biotech and Health Centre, Shenzhen Research Institute, City University of Hong Kong, Shenzhen, China

It is now clear that major malignancies are heterogeneous diseases associated with diverse molecular properties and clinical outcomes, posing a great challenge for more individualized therapy. In the last decade, cancer molecular subtyping studies were mostly based on transcriptomic profiles, ignoring heterogeneity at other (epi-)genetic levels of gene regulation. Integrating multiple types of (epi)genomic data generates a more comprehensive landscape of biological processes, providing an opportunity to better dissect cancer heterogeneity. Here, we propose sparse canonical correlation analysis for cancer classification (SCCA-CC), which projects each type of single-omics data onto a unified space for data fusion, followed by clustering and classification analysis. Without loss of generality, as case studies, we integrated two types of omics data, mRNA and miRNA profiles, for molecular classification of ovarian cancer (n = 462), and breast cancer (n = 451). The two types of omics data were projected onto a unified space using SCCA, followed by data fusion to identify cancer subtypes. The subtypes we identified recapitulated subtypes previously recognized by other groups (all P- values < 0.001), but display more significant clinical associations. Especially in ovarian cancer, the four subtypes we identified were significantly associated with overall survival, while the taxonomy previously established by TCGA did not (P- values: 0.039 vs. 0.12). The multi-omics classifiers we established can not only classify individual types of data but also demonstrated higher accuracies on the fused data. Compared with iCluster, SCCA-CC demonstrated its superiority by identifying subtypes of higher coherence, clinical relevance, and time efficiency. In conclusion, we developed an integrated bioinformatic framework SCCA-CC for cancer molecular subtyping. Using two case studies in breast and ovarian cancer, we demonstrated its effectiveness in identifying biologically meaningful and clinically relevant subtypes. SCCA-CC presented a unique advantage in its ability to classify both single-omics data and multi-omics data, which significantly extends the applicability to various data types, and making more efficient use of published omics resources.

Keywords: multi-omics, data fusion, cancer subtyping, canonical correlation analysis, ovarian cancer, breast cancer

OPEN ACCESS

Edited by:
Bairong Shen,
Sichuan University, China

Reviewed by:
Alessandro Giuliani,
National Institute of Health (ISS), Italy
Chao Xu,
University of Oklahoma Health Sciences Center, United States

****Correspondence:***
Xin Wang
xin.wang@cityu.edu.hk

Specialty section:
This article was submitted to Systems Biology Archive, a section of the journal Frontiers in Genetics

Received: *18 September 2020*
Accepted: *05 July 2021*
Published: *22 July 2021*

Citation:
Qi L, Wang W, Wu T, Zhu L, He L and Wang X (2021) Multi-Omics Data Fusion for Cancer Molecular Subtyping Using Sparse Canonical Correlation Analysis. Front. Genet. 12:607817. doi: 10.3389/fgene.2021.607817

INTRODUCTION

It has been recognized that cancers are heterogeneous diseases comprising multiple subtypes with distinct molecular properties associated with discrepant clinical outcomes. In the last decade, tremendous efforts have been made in identifying cancer molecular subgroups (Zhao et al., 2019). Unlike traditional cancer classification based on histopathological characteristics or individual mutations, these studies employed unsupervised classification to identify biologically coherent subgroups. However, the pre-existing studies were mostly based on single-omics data, especially transcriptomic data, ignoring molecular heterogeneity occurring at other (epi-)genetic levels of gene regulation such as copy number variation, and DNA methylation. Recent advances in high-throughput biotechnologies, especially next-generation sequencing technologies, made it possible to generate (epi)genomic profiles at a significantly reduced cost, providing an opportunity for integrative analysis of multiple types of omics data. Genome-wide, multi-omics profiles of tissue samples from large-scale patient cohorts enabled a more comprehensive dissection of cancer molecular heterogeneity. International consortia such as the cancer genome atlas (TCGA) have assembled multiple cancer data types from 1,000 patients, making integrative methods essential for a better understanding of cancer biology. However, due to the difference in data scale, the complexity of dimensionality, effective integration of multi-omics data for cancer subtyping remains a significant challenge (Bersanelli et al., 2016).

To address the challenge, several computational models have been proposed, which showed promising performance. For instance, non-negative matrix factor (NMF) can be used to project multi-omics data onto dimension-reduced space for integration based on non-negative matrix decomposition (Zhang et al., 2011, 2012). However, the prerequisite of non-negative matrices needs to be satisfied, and proper normalization of the input data is crucial. Joint and individual variation explained (or JIVE) can also be used for integrative analysis of multi-omics data by quantifying the joint variation between data types followed by decomposition to reduce the dimensionality. The application of JIVE to glioblastoma showed better characterizations of different subtypes, but the robustness remains a concern due to potential outliers affecting the factorization based on principal component analysis (PCA) (Lock et al., 2013). iCluster (Shen et al., 2009) and its extensions iClusterPlus (Kirk et al., 2012) learned a joint latent variable model for integrative clustering on multiple types of data. Despite the widely demonstrated usefulness, the scalability of iCluster and its related methods to a genome-wide scale was questionable (Shen et al., 2009). Wang et al. (2014) developed a novel bioinformatic approach named "similarity network fusion (SNF)", which iteratively fused similarity networks constructed from each type of single-omics data into a similarity network by a nonlinear combination method. SNF showed better performance than single-omics methods in cancer subtyping, as demonstrated in multiple case studies (Wang et al., 2014). However, iCluster and SNF do not provide a classification framework, and they both rely on a complete dataset of multi-omics profiles for the clustering of new samples, which is often not available, and significantly limiting their general applicability (Wang et al., 2014).

To overcome the above-mentioned challenges, we propose to fuse different types of omics data for clustering and classification of tumor samples by canonical correlation analysis (CCA) (Hotelling, 1936), a classical statistical analysis method used in multi-views biometric identification. The CCA algorithm measured the correlation between two sets of multi-dimensional data and projected onto a unified space in which the transformed vectors are maximally correlated. However, the classical CCA could not be easily applied to analyze high-throughput data in which the number of variables is much larger than the number of samples. PCA was commonly used to reduce dimensions but may discard important information of correlation and discrimination for 1,000 of variables (Witten et al., 2009). Sparse CCA solved the problem by employing singular value decomposition, seeking sparsity in both sets of variables simultaneously (Witten et al., 2009). The efficiency of SCCA (Sparse CCA) had been demonstrated in simulated genomic data in previous studies (Witten et al., 2009), providing a rationale for us to employ SCCA for cancer subtyping analysis.

In this study, we propose to project single-omics data onto a unified space by SCCA for data fusion, followed by clustering analysis on the fused data to identify cancer subtypes (**Figure 1A**). The trained projection matrices, combined with a trained classifier, can be subsequently used to either single-omics or multi-omics classifications (**Figure 1B**). Using two case studies in ovarian cancer and breast cancer, we demonstrated the usefulness of sparse canonical correlation analysis for cancer classification (SCCA-CC) in cancer classification using multi-omics profiles in the TCGA database[1] as well as single-omics datasets from other independent datasets. Furthermore, we demonstrated that SCCA-CC is superior to other popular methods such as iCluster in the coherence and clinical relevance of identified cancer subtypes, and the running time consumed.

MATERIALS AND METHODS

Data Collection and Curation

We collected mRNA and miRNA expression profiles for 462 ovarian cancer patients and 451 breast cancer patients from the TCGA database. Single-omics (mRNA or miRNA) datasets were collected from gene expression omnibus (GEO). More specifically, we downloaded one mRNA dataset (Tothill dataset, GSE9891, and n = 285) (Tothill et al., 2008), and three miRNA datasets: OC133 (GSE73582, n = 133), OC179 (GSE73581, n = 179), and Bagnoli (GSE25204, n = 130) datasets (Bagnoli et al., 2016) in the ovarian cancer case study. In the breast cancer study, we downloaded the GSE22220 series (Buffa et al., 2011), which includes a mRNA dataset (GSE22219, n = 216) and a

[1] https://cancergenome.nih.gov/

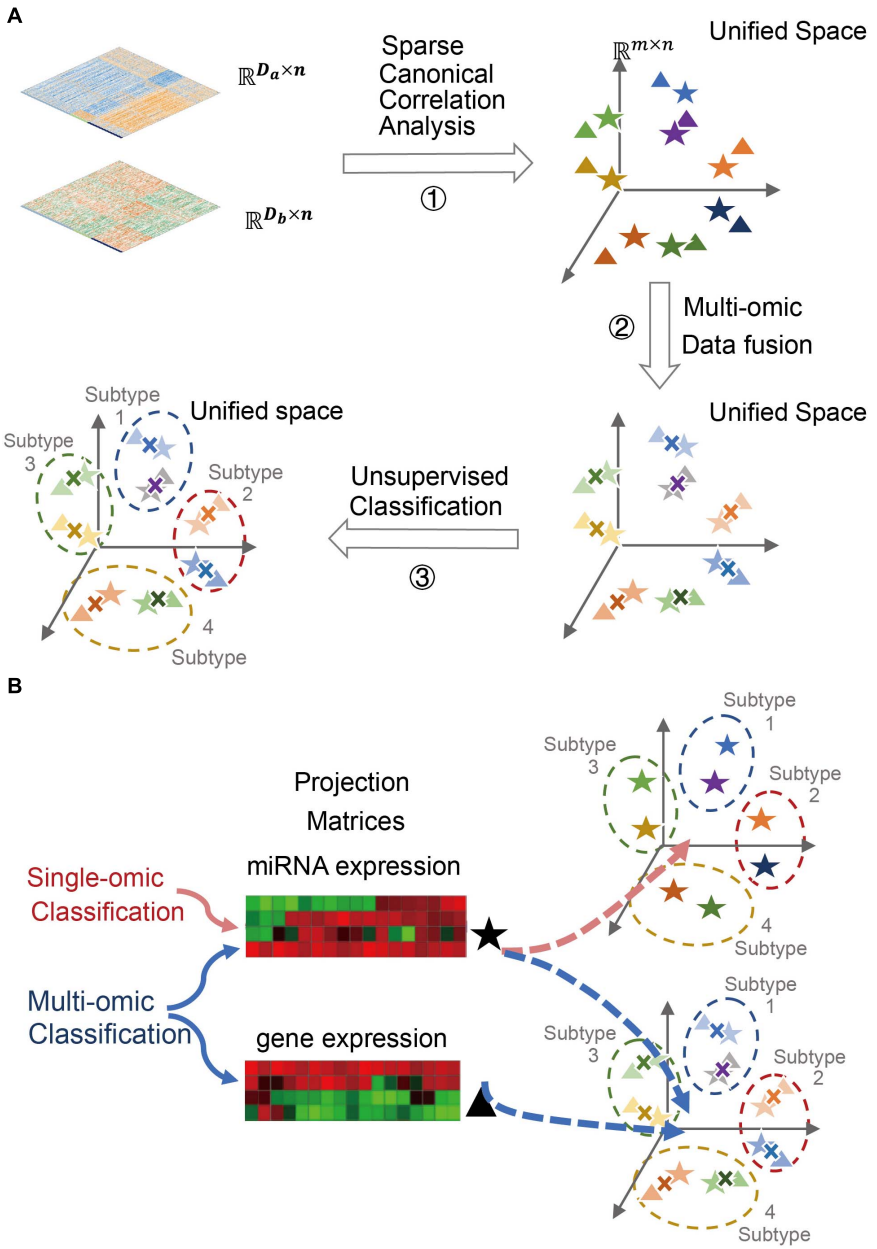


FIGURE 1 | Cancer subtyping and classification using SCCA-CC. **(A)** A schematic figure illustrating the three major steps for multi-omics cancer subtyping. A toy example is used to illustrate the projection of mRNA and miRNA expression data of the same set of patient samples onto lower-dimensional unified space by sparse canonical correlation analysis (SCCA), followed by data fusion and unsupervised classification. **(B)** A schematic figure illustrating the versatile classifier can not only classify fused multi-omics data, but also individual single-omics data.

miRNA dataset (GSE22216, $n = 210$), of which 207 samples have both types of data.

Penalized Canonical Correlation Analysis and Data Fusion

Canonical correlation analysis was proposed in 1936, which was aimed to use fewer combinatorial variables to reflect the correlation between the original two variable groups (Hotelling, 1936). The measurement of the correlation between the two groups of variables makes it possible to fuse different biometrics. CCA projected the two groups of variables onto a unified space in which the transformed vectors were maximally correlated. As the classical CCA could not handle high-dimensional data with small sample size, sparse CCA introduced convex penalty functions to overcome the challenge (Witten et al., 2009). Given two sets of zero-mean random vectors, $A = [a_1, a_2, ..., a_n] \in R^{D_a \times n}$, $B = [b_1, b_2, ..., b_n] \in R^{D_b \times n}$, we can

obtain the objective projection matrices $P_a \in R^{D_a \times m}$ and $P_b \in R^{D_b \times m}$ corresponding to A and B, respectively by SCCA to maximize the correlation coefficient ρ:

$$\rho = \frac{P_a^T A B^T P_b}{\sqrt{(P_a^T A A^T P_a)(P_b^T B B^T P_b)}}$$

Feature-level fusion meant the aggregation of features obtained from various methods of feature extraction. As the features were compressed and extracted to some extent compared with the raw data, the complexity is much lower, and the computation is much more efficient. Much more importantly, feature-level fusion is more tolerant to specific data types, enabling the effective fusion of various omics data. Sparse CCA was implemented using the "CCA" function in R package "PMA", which performs sparse CCA using the penalized matrix decomposition. Lasso penalty was used to obtain the corresponding canonical vector to enforce sparsity by setting the parameters "typex" and "typez" to "standard". The sparsity was determined by the penalties applied to the input matrix. The penalties were set to the default value of 0.3 in the analyses. The number of canonical vectors was determined by the lower number of dimensions of the preprocessed mRNA and miRNA data as mentioned in Methods by the parameter "K". The other parameters were kept by default in the function.

After projecting two types of omics data to the same space, $A^P = P_a^T A$, $A^P \in R^{m \times n}$ and $B^P = P_b^T B$, $B^P \in R^{m \times n}$, we can subsequently fuse them by a weighted averaging strategy:

$$Z = \alpha A^P + (1 - \alpha) B^P = \alpha P_a^T A + (1 - \alpha) P_b^T B$$

where, $\alpha \in [0, 1]$ represents the fusion coefficient. In our case studies, we set an equal weight (fusion coefficient) for each type of omics data, and the fused data Z is used for the following consensus clustering analysis.

Clustering and Classification Analysis

To identify molecular subtypes, we performed unsupervised classification on the fused TCGA data in the unified space. To ensure robustness, we employed the widely adopted consensus clustering method (Monti et al., 2003), with 500 iterations and 0.9 subsampling ratio, to assess the clustering stability. The consensus clustering was implemented by the "ConsensusClusterPlus" function of the R package "ConsensusClusterPlus" with k-means clustering algorithm using Euclidean distance (Monti et al., 2003). The fused TCGA data, together with the subtyping labels, were used to train a classifier. More specifically, we explored various classification methods such as random forests (RF) (R package "randomForest") (Breiman, 2001), support vector machine (SVM) (R package "e1071") (Cortes and Vapnik, 1995), k-nearest neighbors algorithm (KNN) (R package "class") (Venables and Ripley, 2021), minimum distance algorithm (Min-Dis), and Bayesian classifier (R package "e1071") (Cortes and Vapnik, 1995), and selected the one yielding the lowest error rate for the following analysis. More specifically, in the RF classification analysis of ovarian cancer, the number of trees was set to 1,000 and the other parameters were set by default. In the SVM classification analysis of breast carcinoma (BRCA), we used the radial basis kernel and set the cost of constraints violation to 10.

Statistical Analysis

Statistical analysis was conducted with R software (version 3.6.1[2]). SigClust (Huang et al., 2015), a statistical method for testing the significance of clustering results, was used to evaluate the subtypes we identified. Differential gene expression analysis was performed by comparing each subtype with the others using the R package "limma" (Ritchie et al., 2015). Biological characterizations of cancer subtypes were based on gene set enrichment analysis (GSEA) using R package "HTSanalyzeR2" (Wang et al., 2011). Cox regression analyses were performed by R package "survival"[3]. A *p*- value of less than 0.05 was considered statistically significant in all tests.

RESULTS

Molecular Subtyping of Ovarian Cancer Using SCCA-CC

Ovarian cancer is one of the most lethal malignancies in women. Although most ovarian cancer patients can be cured during the early stage, more than 80% of ovarian cancers are diagnosed at advanced stages. Similar to other major malignancies, ovarian cancer has been recognized as a molecularly heterogeneous disease underlying the diverse clinical outcomes. Recently, Tothill et al. (2008) performed unsupervised classification of gene expression profiles for 285 high-grade serous ovarian cancer (HGSOC) samples, resulting in the identification of four distinct subtypes: immunoreactive, differentiated, proliferative, and mesenchymal subgroups. TCGA network recapitulated these subtypes based on transcriptomic profiles of more than 500 OvCa cases (Cancer Genome Atlas Research Network, 2011). More recently, it was found that compared to other subtypes, the mesenchymal subtype displayed higher invasiveness and was associated with poor overall survival (Konecny et al., 2014). Despite the well-established taxonomy, the subtyping studies were based on transcriptomic profiles, ignoring potential heterogeneity at other levels of gene regulations. Furthermore, the classifiers based on gene expression signatures cannot be applied to other types of omics data, greatly limiting the applicability of these classification systems.

Unsupervised Classification of the Fused Multi-Omics Data Identified More Clinically Relevant Subtypes

In total, we obtained matched mRNA and miRNA expression profiles from 462 ovarian cancer samples in the TCGA cohort. To eliminate the impacts of magnitude scale and ensure the comparability of data, within each type of omics data we performed z-score normalization and filtered out genes or miRNAs with low between-sample variations (median absolute deviation, or MAD < 0.75). The preprocessed mRNA and

[2]http://www.Rproject.org

[3]https://github.com/therneau/survival

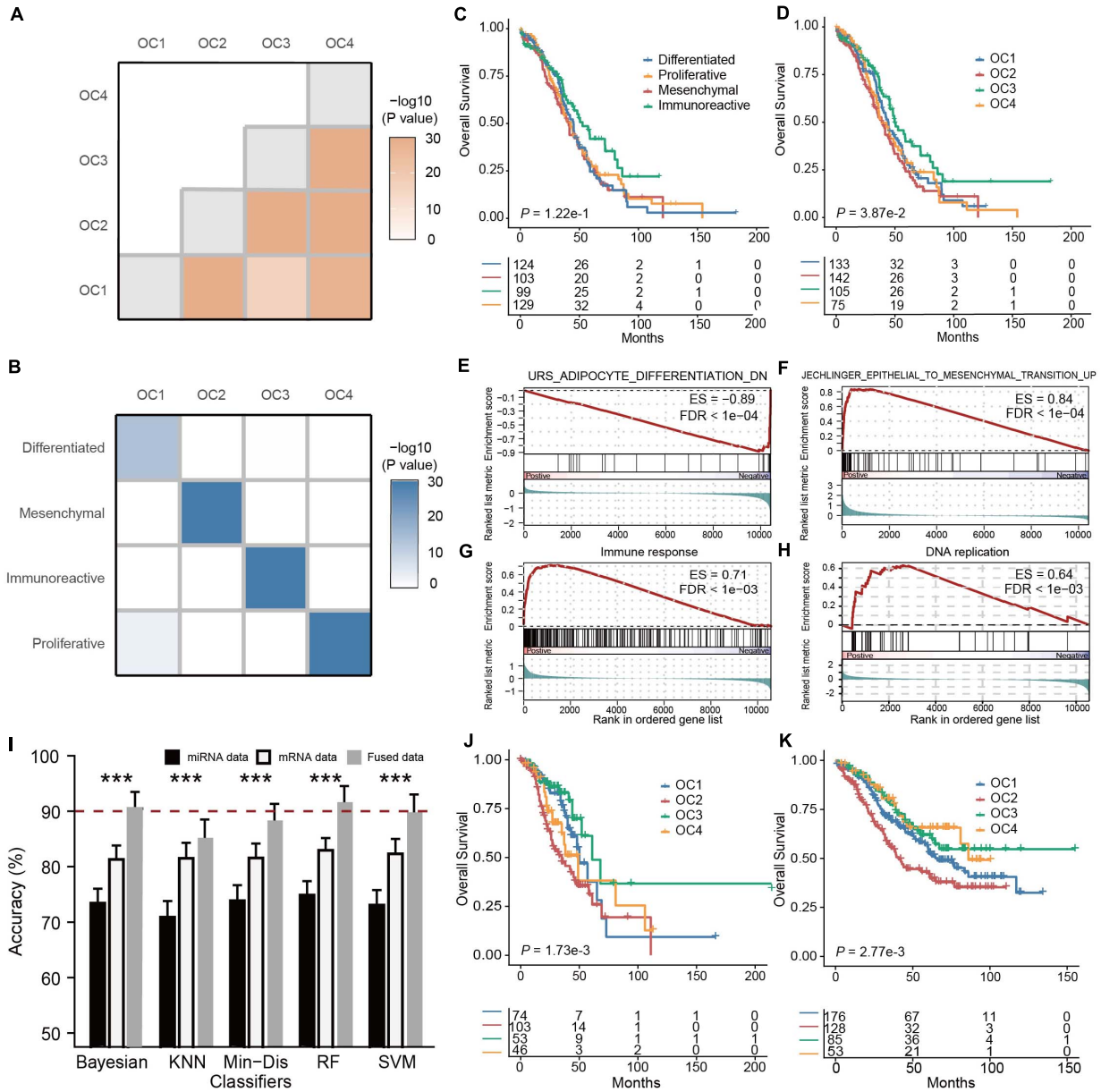


FIGURE 2 | Multi-omics subtyping of ovarian cancer using sparse canonical correlation analysis for cancer classification (SCCA-CC). **(A)** A heatmap showing the statistical significance of the differences between the identified OC subtypes. The color depth is proportionate to the -$\log_{10}$(*P*- values) derived from SigClust. **(B)** A heatmap illustrating the association between ovarian cancer subtypes identified by SCCA-CC and the cancer genome atlas (TCGA). The heatmap is colored in proportion to the -$\log_{10}$(*P*- values) derived from hypergeometric tests. **(C,D)** Kaplan-Meier plots showing the association of the subtypes identified by panel **(C)** TCGA and **(D)** SCCA-CC, respectively. *P*- values were calculated based on log–rank tests. **(E–H)** GSEA plots illustrating the representative pathways dysregulated in each molecular subtype identified by SCCA-CC. **(I)** A bar plot comparing the classification performance of the five classifiers on mRNA data, miRNA data, and the fused data, respectively. *P*- values were calculated based on Wilcoxon signed–rank tests. *** indicates $P < 0.001$. **(J,K)** Kaplan-Meier plots illustrating the association between the four subtypes identified by SCCA-CC with overall survival in this panel **(J)** the Tothill mRNA dataset and **(K)** the merged miRNA data. *P*- values were calculated based on log–rank tests.

miRNA data were subsequently projected onto a unified space using SCCA, followed by data fusion based on a weighted averaging strategy ($\alpha = 0.5$) (**Figure 1A**). Using the fused data, we performed consensus clustering and observed that subdivision into four clusters generated the most robust classification (**Supplementary Figure 1**), suggesting the existence of four major ovarian cancer subtypes (OC1-4). Using SigClust (Huang et al., 2015), a statistical method for testing the significance of clustering results, and we found that indeed the differences between subtypes were statistically significant (all $P < 0.001$, **Figure 2A**). To interpret the four OC subtypes we identified, we compared our clustering result with the TCGA taxonomy

(Cancer Genome Atlas Research Network, 2011; **Supplementary Figure 2**). Interestingly, we found that each OC subtype identified by SCCA-CC was significantly associated with one of the subtypes identified by TCGA (**Figure 2B**, all $P < 0.001$, hypergeometric tests; $P < 0.001$, McNemar–Bowker test), suggesting that SCCA-CC recapitulated the four subtypes previously defined, i.e., proliferative, immunoreactive, differentiated, and mesenchymal (Tothill et al., 2008). Notably, the four OC subtypes we identified are significantly associated with overall survival (**Figure 2D**, $P = 0.039$, log–rank test), while the four TCGA subtypes did not (**Figure 2C**, $P = 0.12$, log–rank test), supporting our hypothesis that incorporating different types of omics data may identify more clinically relevant subtypes than single-omics approaches.

To further elucidate the OC subtypes, we performed differential gene expression analysis by comparing each subtype with the others and then identified subtype-specific biological functions based on GSEA (**Supplementary Table 1**). We confirmed that OC1 is differentiated-like, featured with dysregulated cell differentiation signatures; OC2 is mesenchymal-like, displaying upregulated epithelial-to-mesenchymal transition (Jechlinger et al., 2003); OC3 is immunoreactive-like, characterized by activated immune responses; and OC4 is proliferative-like, characterized by upregulated DNA replication, which were all consistent with previous studies (Jechlinger et al., 2003; Verhaak et al., 2013; Wang et al., 2017; **Figures 2E–H**).

The Multi-Omics Classifier Was Able to Classify Both Single-Omics and Multi-Omics Data

A unique advantage of a multi-omics classifier lies in its ability to handle both single-omics data and multi-omics data, making more efficient use of different types of data potentially (**Figure 1B**). In our study, using the mRNA-miRNA fused data obtained by the projection and fusion from randomly selected 200 TCGA samples, we constructed multiple classifiers based on RF, SVM, KNN, Min-Dis, and Bayesian classifier. Using the clustering labels as the reference, we evaluated the performance of these classifiers on the fused mRNA-miRNA data, the mRNA and the miRNA data alone for the other 262 TCGA samples, respectively. To obtain a stable and robust estimation of the performance, we repeated the tests 100 times. Compared to miRNA-based classification results, all classifiers demonstrated higher accuracies on the mRNA data (**Figure 2I**, all $P < 0.001$, Wilcoxon Signed–rank tests), and, remarkably, achieved even higher accuracies on the fused data (**Figure 2I**, all $P < 0.001$, Wilcoxon Signed–rank tests). The results supported our hypothesis that SCCA-CC achieved higher classification performance when more information is incorporated.

Independent Validations Verified the General Applicability of Multi-Omics Classification

Among the various classifiers, RF showed relatively higher accuracy and lower volatility (standard deviation): 91.6% using the fused data, 83.0% using only the mRNA data, and 74.8% using only the miRNA data (**Figure 2I**). Therefore, we trained a multi-omics classifier based on RF using all the 462 TCGA samples. To evaluate the general applicability of the classifier to other independent datasets, we tested the Tothill mRNA dataset (Tothill et al., 2008) ($n = 279$) and a miRNA dataset ($n = 442$) merged from GSE73581, GSE73582, and Bagnoli miRNA (Bagnoli et al., 2015) datasets. In both datasets, the predicted OC subtypes showed a significant association with survival (**Figures 2J,K**, both $P < 0.01$, log–rank tests). More specifically, patients classified to OC2 (mesenchymal-like) had the worst overall survival, while those classified to OC3 (immunoreactive-like) had the best outcome, which was consistent with previous studies (Jechlinger et al., 2003; Verhaak et al., 2013; Wang et al., 2017). These results demonstrated the multi-omics classifier's potential to classify other independent datasets with different types of omics data. Notably, it was the first time ever that the three miRNA datasets (GSE73581, GSE73582, and Bagnoli) could be classified, since the previous classification method developed by TCGA only takes mRNA data as input (Cancer Genome Atlas Research Network, 2011).

Molecular Subtyping of Breast Cancer Using SCCA-CC

Breast carcinoma is the most common type of gynecological cancer, as it alone accounts for 24.2% of all new cancer incidences in women in 2018 (Bray et al., 2018). Over the past two decades, breast cancer mortality has been reduced remarkably since 1989 (Siegel et al., 2020), mainly attributed to population-wide screening based on mammography and improved therapeutics. However, since breast cancer is also a heterogeneous disease, a significant proportion of patients eventually died due to limited benefit from chemotherapy (Jemal et al., 2017). The intrinsic subtypes of breast cancer, including luminal A, luminal B, basal-like, Her2-enriched, and normal-like have been well characterized and widely adopted (Perou et al., 2000; Sorlie et al., 2001). Importantly, the five intrinsic subtypes are characterized by distinct molecular properties, associate with different clinical outcomes. In particular, patients classified to the Her2-positive subtype showed poor survival, while those assigned to the luminal A subtype displayed more favorable outcome (Sorlie et al., 2001, 2003; Yersal and Barutca, 2014; Dai et al., 2015). For breast cancer subtype prediction, PAM50 is the most popular classifier (Parker et al., 2009), but since the classification system was established based on transcriptomic profiles, it cannot be applied to other types of omics data.

Unsupervised Classification of the Fused Multi-Omics Data Recapitulated the Five Intrinsic Subtypes of Breast Cancer

Like our case study in ovarian cancer, we performed unsupervised classification on 451 patient samples of breast cancer with matched mRNA and miRNA expression profiles in the TCGA cohort. Z-score normalization was applied to each type of omics data, followed by the filtering of genes or miRNAs with low between-sample variations (MAD < 0.5). We projected the preprocessed data onto a lower-dimensional space by SCCA for data fusion using the weighted averaging method ($\alpha = 0.5$) subsequently. Based on consensus clustering of the fused data, we determined the optimal five breast cancer subtypes (**Supplementary Figure 3**). Pairwise comparisons between the subtypes showed significant differences, suggesting

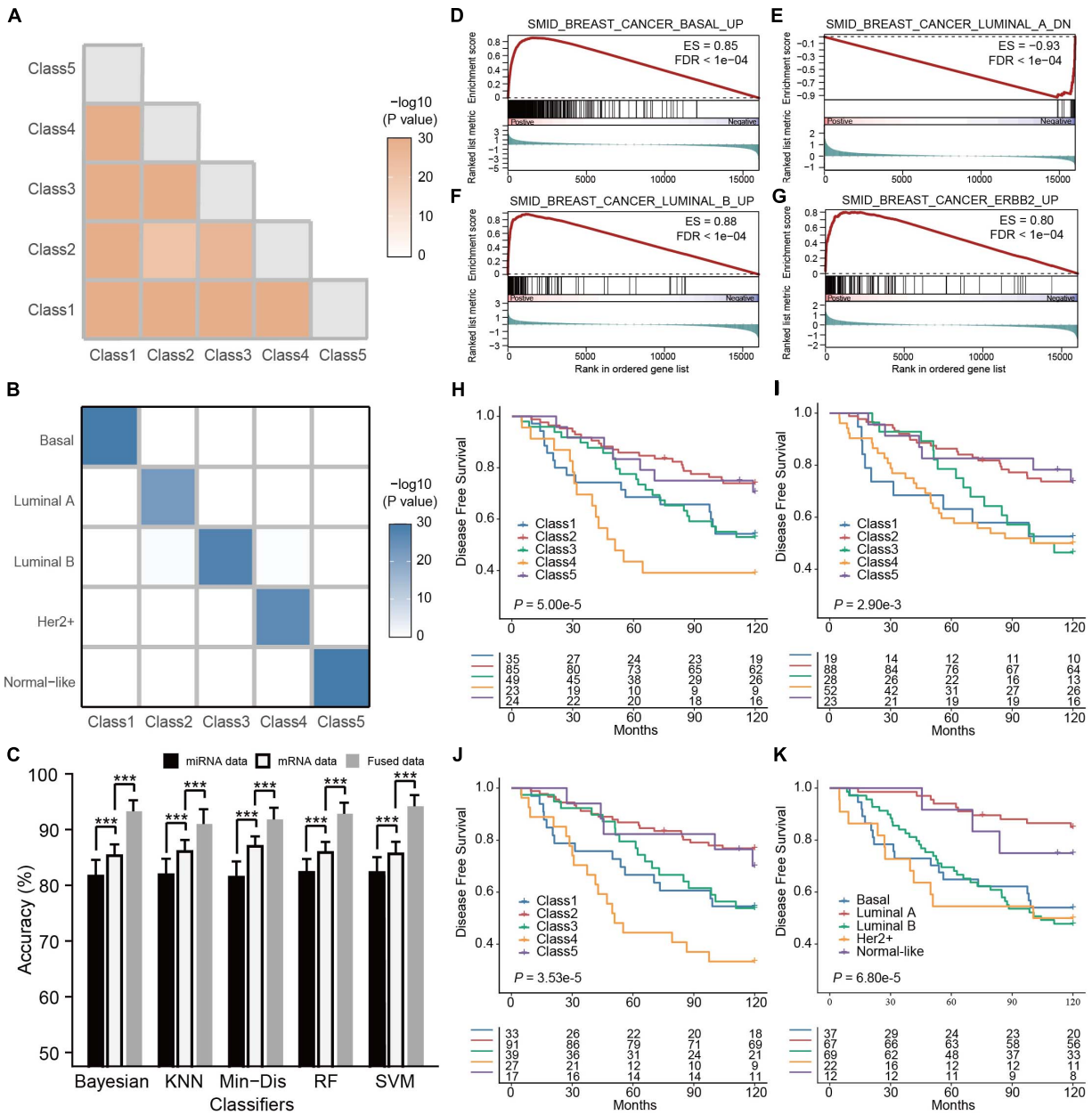


FIGURE 3 | Multi-omics subtyping of breast cancer using sparse canonical correlation analysis for cancer classification (SCCA-CC). **(A)** A heatmap showing the statistical significance of the differences between the identified breast cancer subtypes. The color depth is proportionate to the -$\log_{10}$(*P*- values) derived from SigClust. **(B)** A heatmap illustrating the association between the subtypes identified by SCCA-CC and PAM50. The heatmap is colored in proportion to the -$\log_{10}$(*P*-values) derived from hypergeometric tests. **(C)** A bar plot comparing the classification performance of the five classifiers on mRNA data, miRNA data and the fused data, respectively. *P*- values were calculated based on Wilcoxon signed–rank tests. *** indicates $P < 0.001$. **(D–G)** GSEA plots illustrating the representative pathways dysregulated in the Class 1 (Basal-like), Class 2 (Luminal A-like), Class 3 (Luminal B-like), and Class 4 (Her2+-like) subtypes identified by SCCA-CC. **(H–J)** Kaplan-Meier plots showing the association of the subtypes identified by SCCA-CC using **(H)** the GSE22219 mRNA dataset, **(I)** the GSE22216 miRNA dataset and **(J)** the fused mRNA and miRNA dataset with disease-free survival. **(K)** Kaplan-Meier plot showing the association of the subtypes identified by PAM50 with disease-free survival. *P*- values were calculated by log–rank tests.

the significance of the clustering (all $P < 0.001$, **Figure 3A**). Similar to the ovarian cancer study, each breast cancer subtype we identified was significantly associated with an intrinsic subtype classified by PAM50 (**Figure 3B** and **Supplementary Figure 4**, all $P < 0.001$, hypergeometric tests; $P < 0.001$, McNemar–Bowker test). To further elucidate the biological properties associated with identified subtypes, we performed differential gene expression analysis by comparing each subtype with the others, followed by GSEA to detect subtype-specific biological functions. The GSEA results (**Supplementary Table 1**)

suggested that Class 1, Class 2, Class 3, and Class 4 were enriched for the gene expression signatures representative of the basal, luminal A, luminal B, and Her2+ (ERBB2) subtypes, respectively (**Figures 3D–G**; Smid et al., 2008). Since Class 5 recapitulated the normal-like subtype, and therefore we did not notice any particular biological process representative of this subtype.

The Multi-Omics Classifier Was Able to Classify Both Single-Omics and Multi-Omics Data

For breast cancer, we also employed the five classification algorithms: RF, SVM, KNN, Min-Dis, and Bayesian classifier to build the classifiers. Using the labels obtained from the consensus clustering as the reference, we randomly selected 200 breast cancer samples from the TCGA cohort to construct classifiers based on the mRNA-miRNA fused data and evaluated the performance on other 251 samples. We repeated the tests 100 times and compared the results of all types of omics data for each classifier. Similar to the ovarian cancer case study, we also found that all the classifiers demonstrated higher accuracies on the mRNA data than on the miRNA data (**Figure 3C**, all $P < 0.001$, Wilcoxon Signed–rank tests), and they achieved even higher accuracies on the mRNA-miRNA fused data (**Figure 3C**, all $P < 0.001$, Wilcoxon Signed–rank tests). Consistent with ovarian cancer, our results in breast cancer further demonstrated the improved classification performance of SCCA-CC on multi-omics data.

Independent Validations Verified the General Applicability of Multi-Omics Classification

In this case study, SVM demonstrated the best performance and relatively low volatility: 94.8% using the fused data, 88.74% using only the mRNA data, and 80.4% using only the miRNA data. Therefore, we trained a multi-omics classifier based on SVM using all the 451 TCGA samples and evaluated the general applicability of the classifier to other independent datasets. The GSE22220 series, including a mRNA dataset ($n = 216$), a miRNA dataset ($n = 210$), of which 207 samples have both types of data, were used for validations (Buffa et al., 2011). Using either the mRNA or miRNA dataset alone, we found that the predicted subtypes by the multi-omics classifier showed a significant association with survival (**Figures 3H,I**, both $P < 0.01$, log–rank tests). More interestingly, a higher significance of prognosis was observed using the predicted subtypes based on the fused data (**Figure 3J**, $P < 0.001$, log–rank test). Regardless of single-omics or multi-omics classifications, the predicted Class 4 (Her2+ like) subtype always displayed the worst overall survival, while the Class 2 (Luminal A like) subtype showed more favorable clinical outcome. These results about clinical associations were consistent with previous studies, demonstrating the general applicability of the multi-omics classifier (Sorlie et al., 2001, 2003; Yersal and Barutca, 2014; Dai et al., 2015). Compared with the PAM50 classification on the same dataset, the SCCA-CC classification was more significantly associated with survival (**Figures 3J,K**, $P = 3.5e\text{-}5$ and $6.8e\text{-}5$ for SCCA-CC and PAM50 classifications, respectively). Together, our case study suggested that SCCA-CC was able to identify subtypes that are more clinically relevant, and again supported our hypothesis that incorporating different types of omics data may capture more comprehensive intrinsic characteristics of breast cancer than a single data type.

Benchmark Study

In order to demonstrate the superiority, we directly compared SCCA-CC with iCluster on the datasets we analyzed in the case studies based on three commonly used measures: (i) *P*- values derived from log–rank tests in the Kaplan-Meier analysis to show the association between subtypes and survival; (ii) Silhouette score evaluating the cluster coherence. A higher Silhouette score indicates that samples are more similar within subtypes; and (iii) The algorithm running time evaluating computational complexity. Using varying numbers of genes preselected based on MAD, we performed subtyping analysis using SCCA-CC and iCluster, respectively. As a result, we found SCCA-CC outperformed iCluster based on the three different clustering performance measures in almost all the different scenarios (**Figures 4A,B**). The algorithm running time is acceptable when a small number of genes were used for both methods, but the time iCluster spent increased exponentially with the number of genes, suggesting better scalability of SCCA-CC (**Figure 4C**).

To further compare SCCA-CC with other cancer taxonomies and clinical risk factors, we performed Cox regression analyses in ovarian cancer (TCGA dataset) and breast cancer (GSE22220), respectively. For both cancer types, we first employed iCluster to identify the subtypes with default parameters and evaluated the associations between the identified subtypes with the reference subtyping results based on TCGA or PAM50 (**Supplementary Figures 5, 6**). In ovarian cancer, we found that the SCCA-CC taxonomy showed higher statistical association with patient survival than the classifications based on iCluster and TCGA in the univariate analysis (**Table 1**). After adjusting for other clinical factors such as age and stage, the SCCA-CC classification did not show significant prognostic power. The lack of significance in the survival difference is not surprising, since the TCGA cohort includes HGSOC patients only, who showed very poor overall survival in general. HGSOC patients are diagnosed at advanced stages, and the 5-year overall survival rate (20–30%) has not significantly improved over the last 20–30 years (Davidson et al., 2014; Konstantinopoulos et al., 2015). These patients were difficult to stratify by other pre-existing classifiers such as the TCGA taxonomy itself (**Figure 2C**). Even with our SCCA-CC classifier, the survival difference is marginally significant ($P = 0.031$, Univariate Cox regression in **Table 1**; $P = 0.0387$, and log–rank test in **Figure 2D**) in the TCGA cohort. However, the independent validation dataset, with only miRNA expression profiles for ovarian cancer, includes not only high-grade serous tumors but also other ovarian cancer histotypes that are less aggressive. Therefore, in the miRNA cohort, the overall survival of the patients showed much higher diversity (**Figures 2J,K**). Based on the univariate and multivariate analysis, we found the SCCA-CC classification showed significant prognostic power, even after adjusting for age and stage information (**Supplementary Table 3**). Furthermore, in breast cancer SCCA-CC also outperformed iCluster classification and the PAM classification, and the prognostic power remains after adjusting age and grade factors (**Table 2**). Together, using

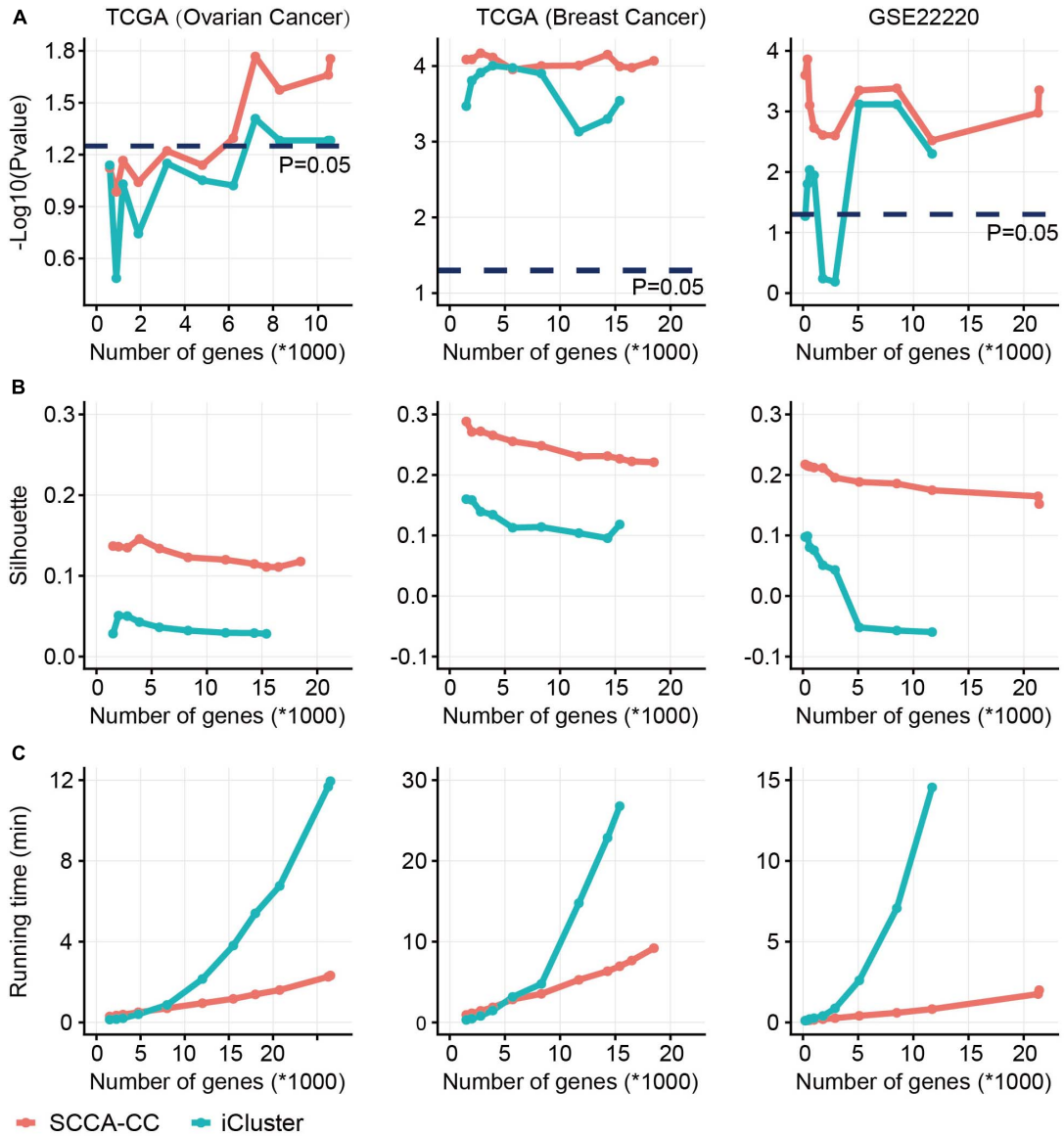


FIGURE 4 | A comparison of sparse canonical correlation analysis for cancer classification (SCCA-CC) with iCluster. Using varying numbers of genes preselected based on MAD, we compared the classification performance between SCCA-CC and iCluster based on **(A)** *P*- values indicative of association with survival calculated by log–rank tests, **(B)** Silhouette score representing the coherence of clusters, and **(C)** the algorithm running time evaluating the computational complexity.

both the ovarian and breast cancer case studies, we demonstrated the better performance of SCCA-CC in identifying molecular subtypes that are more coherent and clinically relevant.

Interpretations of the Canonical Variate Pairs

Sparse CCA provides sets of variables with sparse loadings, which is consistent with the belief that only a small number of genes are expressed under specific conditions (Parkhomenko et al., 2007). Previous studies have used sparse CCA to investigate the associations between different types of omics data, e.g., identification of sets of genes that are correlated with sets of SNPs and copy number variations (Parkhomenko et al., 2007, 2009; Waaijenborg et al., 2008). For better understanding the biology underlying the CCA we further analyzed the pairwise correlations of mRNAs and miRNAs, and build miRNA-mRNA regulatory networks in our case studies.

In ovarian cancer, we checked the first canonical variate pair of mRNAs and miRNAs and found 105 non-zero mRNA variables and 12 non-zero miRNA variables. Pairwise correlation coefficients (n = 1260) were calculated between these variables using their original expression data. Interestingly, we found apparent correlation (negative or positive) between the expression levels of mRNAs and miRNAs, suggesting their potential interactions (**Figure 5A**). As a comparison, we

TABLE 1 | Univariate and multivariate Cox regression analyses in ovarian cancer using the TCGA dataset.

	Univariate		Multivariate	
	HR (95% CI)	***P*- value**	**HR (95% CI)**	***P*- value**
Age (≥65 vs. < 65)	1.37 (1.07–1.76)	0.014	1.32 (1.02–1.71)	0.034
Stage (Late vs. Early)	2.33 (1.10∼4.94)	0.028	2.21 (1.04–4.70)	0.039
SCCA-CC (multinomial)	1.04 (0.85–1.08)	0.466		
TCGA labels (multinomial)	0.94 (0.95–1.19)	0.279		
SCCA-CC (OC2 vs. OC1/3/4)	1.33 (1.03∼1.73)	0.031	1.21 (0.93–1.58)	0.161
iCluster (iCluster 1 vs. iCluster 2–4)	1.25 (0.94∼1.67)	0.12		
TCGA labels (Mesenchymal vs. Others)	0.82 (0.91∼1.63)	0.18		

TABLE 2 | Univariate and multivariate Cox regression analyses in breast cancer using the GSE22220 series dataset.

	Univariate		Multivariate	
	HR (95% CI)	***P*- value**	**HR (95% CI)**	***P*- value**
Age (≥65 vs. < 65)	2.28 (1.42∼3.68)	0.0007	1.81 (1.09–2.99)	0.021
Grade (2–3 vs. 1)	1.82 (1.06∼3.11)	0.030	1.55 (0.89–2.68)	0.118
ER status (1 vs. 0)	0.80 (0.51∼1.26)	0.33		
SCCA-CC (multinomial)	0.86 (0.96–1.39)	0.127		
PAM50 (multinomial)	1.06 (0.78–1.15)	0.575		
SCCA-CC (Class 4 vs. Classes 1–3, 5)	2.95 (1.73∼5.02)	< 0.0001	1.95 (1.07–3.55)	0.030
iCluster (iCluster 5 vs. iCluster 1–4)	2.49 (1.54∼4.02)	0.0002	1.68 (0.97–2.91)	0.063
PAM50 (Her2+ vs. others)	1.77 (0.93∼3.35)	0.08		

generated a background distribution of correlation coefficients based on random sampling of 1260 pairs of mRNAs and miRNAs from all the input data, repeating for 1,000 times. As a result, the randomly selected mRNAs and miRNAs showed lack of association (**Figure 5A**), suggesting the functional relevance of the non-zero mRNAs and miRNAs variables. Based on the interesting correlation observed, we hypothesize that physical interactions may underlie the expression associations between these mRNAs and miRNAs selected by sparse CCA. To test the hypothesis, we built a miRNA-mRNA regulatory network by collecting both experimentally validated miRNA-target interactions (from miRecords (Xiao et al., 2009), miRTarBase (Huang et al., 2020), and TarBase (Karagkouni et al., 2018) and predicted miRNA-target interactions with evolutionary conservation (from TargetScan (Agarwal et al., 2015), PITA (Kertesz et al., 2007) and miRanda (Enright et al., 2003). As expected, most of the mRNAs (99 out of the total 105) and all the miRNAs are interconnected (**Figure 5C**), suggesting that these miRNAs have intense physical interactions with the mRNAs. As an example, CDR2L, with the second largest weight, is the target of hsa-miR-125b, hsa-miR-142-3p, hsa-miR-142-5p and hsa-miR-222 based on targetScan, and/or PITA predictions (Kertesz et al., 2007; Agarwal et al., 2015; **Supplementary Table 2**). Similarly, MMD, with the third largest weight, is the target of hsa-miR-142-3p, hsa-miR-142-5p, hsa-miR-223, hsa-miR-224, and hsa-miR-335 (Kertesz et al., 2007; Agarwal et al., 2015; **Supplementary Table 2**).

Similarly, in BRCA we checked the first canonical variate pair of genes and miRNAs dataset and found 204 non-zero mRNA variables and 57 non-zero miRNA variables. Pairwise correlation coefficients (n = 11628) were calculated between these variables using their original expression data. As a result, we also found strong correlation (negative or positive) between the expression levels of mRNAs and miRNAs (**Figure 5B**). A background distribution of correlation coefficients based on random sampling of 11,628 pairs of mRNAs and miRNAs from all the input data, repeating for 1,000 times. As expected, the randomly selected mRNAs and miRNAs also showed lack of association (**Figure 5B**). We further built a miRNA-mRNA regulatory network to investigate whether physical interactions may also underlie the expression associations of mRNAs and miRNAs selected by sparse CCA in BRCA. Indeed, most of the genes (195 out of the total 204) and the miRNAs (52 out of the total 57) are interconnected (**Figure 5D**), suggesting that these miRNAs have intense physical interactions with the mRNAs. For instance, Gene UBE2T with the highest weight is the target of hsa-mir-96, hsa-mir-200c, has-miR-182 based on targetScan, and/or PITA predictions (Kertesz et al., 2007; Agarwal et al., 2015; **Supplementary Table 2**). Gene CKS2, with the second highest weight, is the target of hsa-mir-200c, has-miR-429, and has-miR-33b (Kertesz et al., 2007; Agarwal et al., 2015; **Supplementary Table 2**).

Together, these results provide compelling evidence that the sparse CCA selected biologically relevant genes and miRNAs, which explains their strong expression correlation enabling multi-omic data fusion in the projected space.

DISCUSSION

Cancer molecular heterogeneity hampers the selection of patients for more optimized clinical management and the design of targeted agents. During the last decade, tremendous efforts have been made to dissecting the inter-tumor heterogeneity in an overwhelming number of studies based on unsupervised classification of high-throughput omics profiles. These studies gained novel insights into cancer biology with important clinical implications, which laid a solid foundation for precision medicine. However, most of these studies were based on single-omics data, especially transcriptomic data, which ignored other genetic and epigenetic levels of gene regulation, and resulting in only partial understanding of cancer heterogeneity. Recent studies have seen a growing interest in integrating multiple types of omics data for more comprehensive cancer subtyping, but few

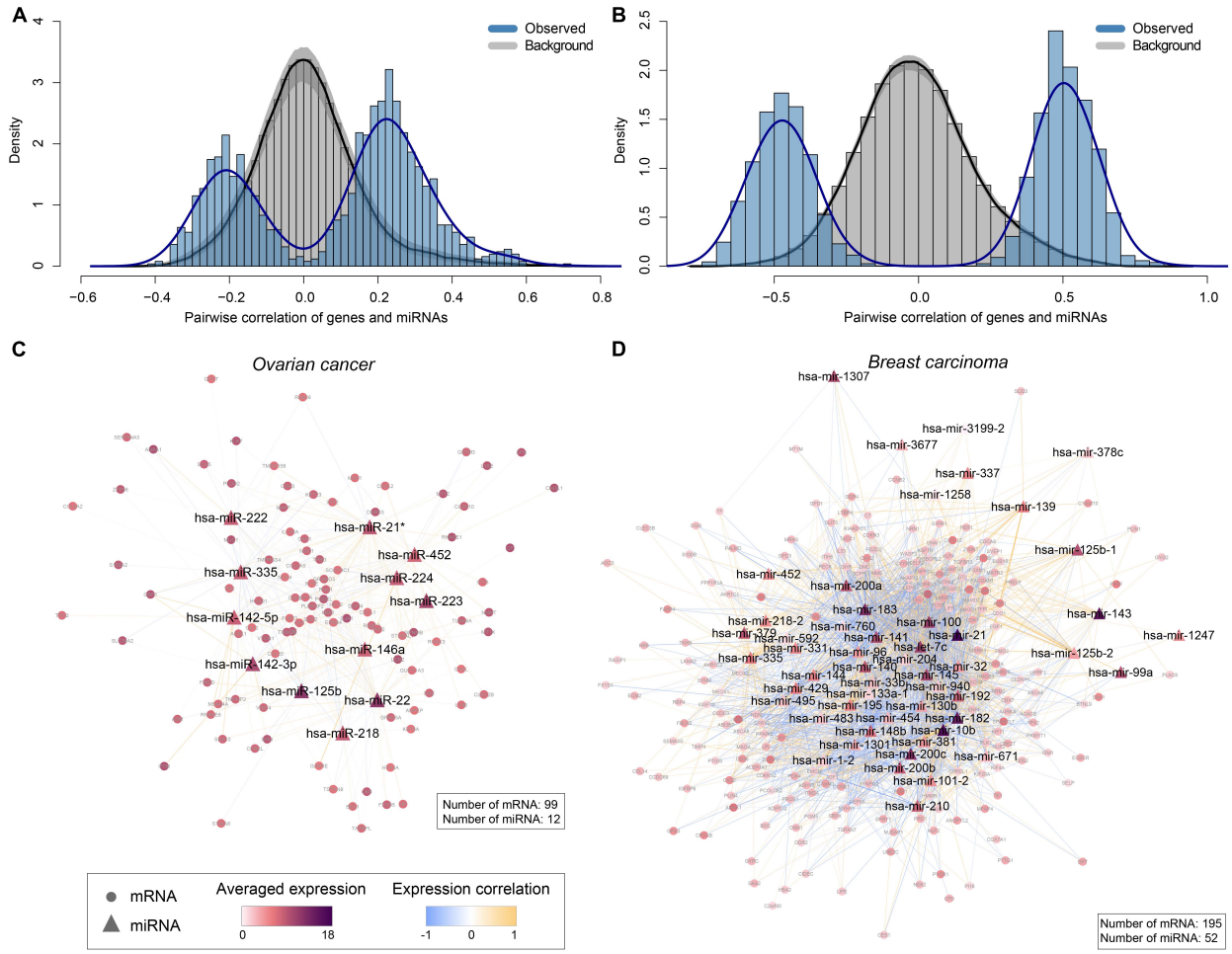


FIGURE 5 | Interpretations of the canonical variate pairs. **(A,B)** The distributions of observed pairwise correlation coefficients between non-zero mRNA variables and miRNA variables in the first canonical variates, compared to the corresponding background distributions in this panel **(A)** ovarian cancer and **(B)** breast carcinoma (BRCA), respectively. The gray area represents 95% confidence intervals. **(C,D)** MiRNA-mRNA regulatory networks constructed based on non-zero mRNA and miRNA variables selected by sparse CCA in this panel **(C)** ovarian cancer and **(D)** breast carcinoma, respectively. Triangles and circles represent miRNAs and mRNAs, respectively. Edges represent the interactions between miRNAs and mRNAs experimentally validated and/or predicted by databases including miRecords, miRTarBase, TarBase, TargetScan, PITA, and miRanda. Nodes are colored in proportion to the averaged log2 transformed expression levels of mRNAs/miRNAs across all TCGA samples, and edges are colored based on Pearson correlation coefficients between the expression levels of miRNAs and mRNAs.

existing methods can classify both single-omics and multi-omics data. In this study, we developed SCCA-CC, a robust and efficient framework for cancer subtyping and classifications based on data fusion using sparse CCA followed by unsupervised classification. Using two case studies on multiple independent cohorts, we demonstrated that SCCA-CC was able to identify biologically meaningful and clinically more relevant taxonomies.

Conventional CCA may suffered from the high dimensionality of genomic data where the number of observations greatly exceeds the number of samples, leading to high risk of potential collinearity and unstable estimates (Waaijenborg et al., 2008; Parkhomenko et al., 2009; Boutte and Liu, 2010). PCA is a powerful dimension reduction method, which has been used prior to CCA in some applications. However, in our study, we did not perform PCA prior to CCA due to the following considerations: (1) We employed sparse CCA but not the conventional CCA in our study. In the sparse CCA (Witten et al., 2009), a penalized matrix decomposition is introduced using a LASSO penalty to compute a rank-K approximation of a matrix (Witten et al., 2009; Lin et al., 2013). This is inspired by several penalization methods presented in the regression context (Zou et al., 2006; Wright et al., 2009; Wu et al., 2009). As reported before, the problem of multicollinearity can be mitigated by the use of sparse loadings in the CCA algorithm (Waaijenborg and Zwinderman, 2009; Witten et al., 2009; Boutte and Liu, 2010; Lin et al., 2013). (2) Cross-platform applicability. In practice, the strategy to perform PCA before CCA may be difficult to be applied to other datasets based on different gene expression profiling platforms. For instance, a lot of newly identified genes in RNA-seq data were often missing in early gene expression microarrays published many years ago. (3) After performing PCA prior to CCA, the generalized eigenvector problem is changed into the eigen-system computation of a nonsymmetric matrix which is unstable as previously reported (Swets and Weng, 1996).

(4) Performing PCA prior to CCA may discard dimensions that contain important discriminative information (Xing et al., 2016).

For both ovarian and breast cancers, the subtypes identified by SCCA-CC recapitulated the taxonomies previously established, as suggested by the pairwise statistical tests and GSEA. However, SCCA-CC derived subtypes showed a more significant association with clinical outcomes. Notably, ovarian cancer subtypes identified by SCCA-CC were significantly associated with overall survival in the TCGA cohort, while the four subtypes previously defined by TCGA did not show any significant clinical association. On independent mRNA, miRNA, and fused datasets, SCCA-CC demonstrated consistent clinical associations in both our ovarian and breast cancer studies. These results demonstrated that SCCA-CC is able to detect the biologically coherent subgroups in different types of single-omics data, and incorporating multiple types of omics data can further improve the prediction performance.

More importantly, a number of published studies with only miRNA expression profiles cannot be classified using existing subtyping systems based on mRNA expression signatures. SCCA-CC presented a unique advantage in its ability to classify both single-omics data and multi-omics data, which significantly extends the general applicability to make efficient use of public resources. More specifically, we constructed multi-omics classifiers using the fused data with the consensus clustering labels as the reference and evaluated the robustness of the classification performance by iterative testing. The validity of multi-omics classifiers was verified by the observed significant prognostic power on both the mRNA dataset and the miRNA dataset. Notably, it was the first time ever that the miRNA datasets could be classified since the previous classification assays only take mRNA data as input.

In a benchmark study against iCluster, SCCA-CC also demonstrated its superiority in the higher coherence and clinical relevance of identified cancer subtypes, and lower computational complexity. Furthermore, the strength of SCCA-CC also lies in the biological interpretability. The non-zero mRNAs and miRNAs selected by sparse CCA had strong correlation in their expression levels, which can be explained by their intense physical interactions. These results provide compelling evidence that the sparse CCA selected biologically relevant genes and miRNAs.

Despite the demonstrated usefulness, the major limitation of SCCA-CC lies in the limited types of omics data we used in the study. Only mRNA and miRNA data were fused for classification, likely missing heterogeneity occurring at other omics levels. Thus, our future work will focus on integrating more types of omics data to dissect the heterogeneity more comprehensively. Considering the differences in the dimensionalities and data scales between various types of omics data, how to properly preprocess the data for effective data fusion remains a significant challenge.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

XW contributed to study concept and design. LQ, WW, and TW contributed to data collection, analysis, and interpretation. XW contributed to critical revision of the manuscript for important intellectual content. LZ and LH provided important advice and assistance for manuscript drafting. XW supervised the study. All authors read and approved the final manuscript.

FUNDING

This work was supported by a grant from Guangdong Basic and Applied Basic Research Foundation (Project No. 2019B030302012), a grant supported by the Young Scientists Fund of the National Natural Science Foundation of China (81802384), and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CityU 21101115, 11102317, 11103718, 11103619, R4017-18, C4041-17GF, and AoE/M-401/20) awarded to XW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.607817/full#supplementary-material

Supplementary Figure 1 | Consensus clustering based on the fused data in ovarian cancer. **(A)** Heatmap illustrating the consensus matrices for k = 4, 5, and 6. **(B)** Consensus cumulative distribution function (CDF) plot for k varying from 2 to 6. **(C)** Delta area plot shows the relative change in the area under the consensus cumulative distribution function (CDF) curve comparing k and k–1. At k = 4, there is no appreciable increase (Delta area $<$ 0.1). **(D)** Gap statistic suggesting the optimal number of clusters is four in the TCGA dataset. Error bars indicate SEM.

Supplementary Figure 2 | A heatmap illustrating the pairwise comparison between the subtypes identified by SCCA-CC and those defined by the cancer genome atlas (TCGA) in the TCGA ovarian cancer dataset.

Supplementary Figure 3 | Consensus clustering based on the fused data in breast cancer. **(A)** Heatmap illustrating the consensus matrices for k = 5, 6, and 7. **(B)** Consensus cumulative distribution function (CDF) plot for k varying from 2 to 7. **(C)** Delta area plot shows the relative change in the area under the consensus cumulative distribution function (CDF) curve comparing k and k–1. At k = 5, there is no appreciable increase (Delta area $<$ 0.1). **(D)** Gap statistic suggesting the optimal number of clusters is five in the TCGA dataset. Error bars indicate SEM.

Supplementary Figure 4 | A heatmap illustrating the pairwise comparison between the subtypes identified by SCCA-CC and those defined by PAM50 in the TCGA breast cancer dataset.

Supplementary Figure 5 | Heatmaps showing the pairwise comparison between the subtypes identified by iCluster and the cancer genome atlas (TCGA) in the TCGA ovarian cancer dataset. **(A)** The statistical significance of association quantified by hypergeometric tests. **(B)** The detailed confusion matrix.

Supplementary Figure 6 | Heatmaps showing the pairwise comparison between the subtypes identified by iCluster and PAM50 in the GSE22219 breast cancer dataset. **(A)** The statistical significance of association quantified by hypergeometric tests. **(B)** The detailed confusion matrix.

Supplementary Table 1 | The result of gene set enrichment analysis (xlsx).

Supplementary Table 2 | The weights of canonical variates (xlsx).

Supplementary Table 3 | Univariate and multivariate Cox regression analyses in the independent miRNA validation datasets (xlsx).

REFERENCES

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005. doi: 10.7554/eLife.05005

Bagnoli, M., Canevari, S., Califano, D., Losito, S., Maio, M. D., Raspagliesi, F., et al. (2016). Development and validation of a microRNA-based signature (MiROvaR) to predict early relapse or progression of epithelial ovarian cancer: a cohort study. *Lancet Oncol.* 17, 1137–1146. doi: 10.1016/s1470-2045(16)30108-5

Bagnoli, M., De Cecco, L., Granata, A., Nicoletti, R., Marchesi, E., Alberti, P., et al. (2015). Identification of a chrXq27.3 microRNA cluster associated with early relapse in advanced stage ovarian cancer patients. *Oncotarget* 6:9643. doi: 10.18632/oncotarget.3998

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinform.* 17(Suppl. 2):15.

Boutte, D., and Liu, J. (2010). "Sparse canonical correlation analysis applied to fMRI and genetic data fusion", in *Proceedings of the 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Hong Kong, 422–426.

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

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Buffa, F. M., Camps, C., Winchester, L., Snell, C. E., Gee, H. E., Sheldon, H., et al. (2011). microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Res.* 71, 5635–5645. doi: 10.1158/0008-5472.can-11-0489

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., et al. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.* 5, 2929–2943.

Davidson, B., Rosenfeld, Y. B. Z., Holth, A., Hellesylt, E., Tropé, C. G., Reich, R., et al. (2014). VICKZ2 protein expression in ovarian serous carcinoma effusions is associated with poor survival. *Hum. Pathol.* 45, 1520–1528. doi: 10.1016/j.humpath.2014.03.005

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in *Drosophila*. *Genome Biol.* 5:R1.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28:321. doi: 10.2307/2333955

Huang, H., Liu, Y., Yuan, M., and Marron, J. S. (2015). Statistical significance of clustering using Soft thresholding. *J. Comput. Graph. Stat.* 24, 975–993. doi: 10.1080/10618600.2014.948179

Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* 48, D148–D154.

Jechlinger, M., Grunert, S., Tamir, I. H., Janda, E., Lüdemann, S., Waerner, T., et al. (2003). Expression profiling of epithelial plasticity in tumor progression. *Oncogene* 22, 7155–7169. doi: 10.1038/sj.onc.1206887

Jemal, A., Ward, E. M., Johnson, C. J., Cronin, K. A., Ma, J., Ryerson, B., et al. (2017). Annual report to the nation on the status of cancer, 1975-2014, featuring survival. *J. Natl. Cancer Inst.* 109:djx030. doi: 10.1093/jnci/djx030

Karagkouni, D., Paraskevopoulou, M. D., Chatzopoulos, S., Vlachos, I. S., Tastsoglou, S., Kanellos, I., et al. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.* 46, D239–D245.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39, 1278–1284. doi: 10.1038/ng2135

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28, 3290–3297. doi: 10.1093/bioinformatics/bts595

Konecny, G. E., Wang, C., Hamidi, H., Winterhoff, B., Kalli, K. R., Dering, J., et al. (2014). Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J. Natl. Cancer Inst.* 106:dju249. doi: 10.1093/jnci/dju249

Konstantinopoulos, P. A., Ceccaldi, R., Shapiro, G. I., and D'Andrea, A. D. (2015). Homologous recombination deficiency: exploiting the fundamental vulnerability of ovarian cancer. *Cancer Discov.* 5, 1137–1154. doi: 10.1158/2159-8290.cd-15-0714

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H.-W., and Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinform.* 14:245. doi: 10.1186/1471-2105-14-245

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat.* 7, 523–542. doi: 10.1214/12-AOAS597

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* 52:118.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi: 10.1200/jco.2008.18.1370

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.* 1(Suppl. 1):S119.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.* 8:1. doi: 10.2202/1544-6115.1406

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25, 2906–2912. doi: 10.1093/bioinformatics/btp543

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G. M., et al. (2008). Subtypes of breast cancer show preferential site of relapse. *Cancer Res.* 68, 3108–3114. doi: 10.1158/0008-5472.can-07-5644

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10869–10874. doi: 10.1073/pnas.191367098

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8418–8423. doi: 10.1073/pnas.0932692100

Swets, D. L., and Weng, J. J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Transact. Patt. Anal. Mach. Intel.* 18, 831–836. doi: 10.1109/34.531802

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* 14, 5198–5208. doi: 10.1158/1078-0432.ccr-08-0196

Venables, W. N., and Ripley, B. D. (2021). *Modern Applied Statistics With S*, 4th Edn. Available online at: https://www.stats.ox.ac.uk/pub/MASS4/ (accessed June 23, 2021).

Verhaak, R. G. W., Tamayo, P., Yang, J.-Y., Hubbard, D., Zhang, H., Creighton, C. J., et al. (2013). Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* 123, 517–525.

Waaijenborg, S., and Zwinderman, A. H. (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinform.* 10:315.

Waaijenborg, S., Verselewel de Witt Hamer, P. C., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat. Appl. Genet. Mol. Biol.* 7:3.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wang, C., Armasu, S. M., Kalli, K. R., Maurer, M. J., Heinzen, E. P., Keeney, G. L., et al. (2017). Pooled clustering of high-grade serous ovarian cancer gene expression leads to novel consensus subtypes associated with survival and surgical outcomes. *Clin. Cancer Res.* 23, 4077–4085. doi: 10.1158/1078-0432.ccr-17-0246

Wang, X., Terfve, C., Rose, J. C., and Markowetz, F. (2011). HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics* 27, 879–880. doi: 10.1093/bioinformatics/btr028

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534. doi: 10.1093/biostatistics/kxp008

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 210–227.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25, 714–721. doi: 10.1093/bioinformatics/btp041

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* 37, D105–D110.

Xing, X., Wang, K., Yan, T., and Lv, Z. (2016). Complete canonical correlation analysis with application to multi-view gait recognition. *Pattern Recogn.* 50, 107–117. doi: 10.1016/j.patcog.2015.08.011

Yersal, O., and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J. Clin. Oncol.* 5, 412–424. doi: 10.5306/wjco.v5.i3.412

Zhang, S., Li, Q., Liu, J., and Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* 27, i401–i409.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

Zhao, L., Lee, V. H. F., Ng, M. K., Yan, H., and Bijlsma, M. F. (2019). Molecular subtyping of cancer: current status and moving toward clinical applications. *Brief. Bioinform.* 20, 572–584. doi: 10.1093/bib/bby026

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.* 15, 265–286.

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.

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