frontiers | Research Topics

Machine learning-based methods for RNA data analysis - volume III

Edited by
Lihong Peng, Minxian Wallace Wang, Jialiang Yang and Liqian Zhou

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-83251-490-0
DOI 10.3389/978-2-83251-490-0

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

Machine learning-based methods for RNA data analysis - volume III

Topic editors

Lihong Peng — Hunan University of Technology, China

Minxian Wallace Wang — Beijing Institute of Genomics, Chinese Academy of Sciences (CAS), China

Jialiang Yang — Geneis (Beijing) Co. Ltd, China

Liqian Zhou — Hunan University of Technology, China

Citation

Peng, L., Wang, M. W., Yang, J., Zhou, L., eds. (2023). *Machine learning-based methods for RNA data analysis - volume III*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-490-0

Table of contents

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 01 July 2022
doi: 10.3389/fgene.2022.912711

D3K: The Dissimilarity-Density-Dynamic Radius K-means Clustering Algorithm for scRNA-Seq Data

Guoyun Liu[1], Manzhi Li[1,2], Hongtao Wang[1], Shijun Lin[1], Junlin Xu[3], Ruixi Li[4], Min Tang[5] and Chun Li[1]*

[1]School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2]Key Laboratory of Data Science and Smart Education, Ministry of Education, Hainan Normal University, Haikou, China, [3]College of Information Science and Engineering, Hunan University, Changsha, China, [4]Geneis Beijing Co., Ltd., Beijing, China, [5]School of Life Sciences, Jiangsu University, Zhenjiang, China

A single-cell sequencing data set has always been a challenge for clustering because of its high dimension and multi-noise points. The traditional K-means algorithm is not suitable for this type of data. Therefore, this study proposes a Dissimilarity-Density-Dynamic Radius-K-means clustering algorithm. The algorithm adds the dynamic radius parameter to the calculation. It flexibly adjusts the active radius according to the data characteristics, which can eliminate the influence of noise points and optimize the clustering results. At the same time, the algorithm calculates the weight through the dissimilarity density of the data set, the average contrast of candidate clusters, and the dissimilarity of candidate clusters. It obtains a set of high-quality initial center points, which solves the randomness of the K-means algorithm in selecting the center points. Finally, compared with similar algorithms, this algorithm shows a better clustering effect on single-cell data. Each clustering index is higher than other single-cell clustering algorithms, which overcomes the shortcomings of the traditional K-means algorithm.

Keywords: Dissimilarity matrix, density, dynamic radius, ScRNA-seq, K-means

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Xiangtao Li,
Jilin University, China
Jujuan Zhuang,
Dalian Maritime University, China
Qi Ren,
Tianjin University, China

****Correspondence:***
Manzhi Li
lmz20031979@163.com

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

Received: *04 April 2022*
Accepted: *25 April 2022*
Published: *01 July 2022*

Citation:
Liu G, Li M, Wang H, Lin S, Xu J, Li R, Tang M and Li C (2022) D3K: The Dissimilarity-Density-Dynamic Radius K-means Clustering Algorithm for scRNA-Seq Data. Front. Genet. 13:912711. doi: 10.3389/fgene.2022.912711

1 INTRODUCTION

Since the start of genome Project, genome sequencing has been carried out rapidly, and a large amount of genome data has been mined. In order to obtain the information needed by people, bioinformatics emerges as The Times require (Li and Wong, 2019; Liu et al., 2021). It is an interdisciplinary subject composed of life science and computer science, which can dig out the biological significance contained in the chaotic biological data (Sun et al., 2022). Transcriptome is an important research field in bioinformatics, which can study gene function and gene structure from an overall level, and reveal specific biological processes and molecular mechanisms in the process of disease occurrence (Qi et al., 2021; Tang et al., 2020). In order to study the transcriptome, it must be sequenced first, but traditional sequencing techniques ignore the critical differences of individual cells, which will mask the heterogeneous expression between cells and make it difficult to detect subtle potential changes (Huang et al., 2017; Liu et al., 2020). To solve this problem, the single cell RNA sequencing (scrNA-SEQ) technology was developed (Qiao et al., 2017).

scRNA-seq is a powerful method for analyzing gene expression patterns and quickly determining the correct gene expression patterns of thousands of single cells (Potter, 2018). By analyzing scRNA-seq data, we can identify rare cell populations, find subgroup types with different functions, and reveal the regulatory relationship between genes. scRNA-seq can not only show the complexity of single-cell horizontal structure but also improve biomedical research and solve various problems in biology (Yang et al., 2019).

Although the research prospect of scRNA-seq is comprehensive, it also brings new problems and challenges (Kiselev et al., 2019). The scRNA-seq data are high-dimensional and noisy (Xu et al., 2020). Therefore, many clustering methods have been proposed to deal with high-dimensional data structures and noise distribution (Jiang et al., 2018; Zhang et al., 2021; Zhuang et al., 2021). Most of the existing scRNA-seq clustering methods can be divided into unsupervised or semi-supervised clustering (Chen et al., 2016). Zhang et al., (2018) et al. proposed an improved K-means algorithm based on density canopy to find the appropriate center point by calculating the density of the sample data set; Li et al. proposed a new improved algorithm based on T-SNE and density canopy algorithm, called density-canopy-K-means (Li et al., 2019). Compared with similar methods, this clustering algorithm shows stable and efficient clustering performance on single-cell data, thus overcoming the shortcomings of traditional methods; Dong and Zhu, (2020) et al. calculated the dissimilarity parameter between each model by calculating the dissimilarity function between samples and selected the maximum dissimilarity parameter value as the initial clustering center point; Zhu (Zhuang et al., 2021) et al. proposed a new sparse subspace clustering method, which can describe the relationship between cells in a subspace; Ruiqing (Zheng et al., 2019) et al. proposed a method for detecting scRNA-seq cell types based on similarity learning. Wang et al., (2022) propose the scHFC, which is a hybrid fuzzy clustering method optimized by natural computation based on Fuzzy C Mean (FCM) and Gath-Geva (GG) algorithms. The FCM algorithm is optimized by simulated annealing algorithm, and the genetic algorithm is applied to cluster the data to output a membership matrix. Gan et al., (2022). propose a new deep structural clustering method for sc RNA-seq data, named scDSC, which integrates the structural information into deep clustering of single cells. The study by Gan et al., (2022) not only explained the cell typing method behaviors under different experimental settings but also provided a general guideline for the choice of the method according to the scientific goal and dataset properties. Li et al., (2019) Surrogate-Assisted Evolutionary Deep Imputation Model (SEDIM) is proposed to automatically design the architectures of deep neural networks for imputing gene expression levels in scRNA-seq data without any manual tuning. Yu et al., (2022)propose a single-cell model-based deep graph embedding clustering (scTAG) method, which simultaneously learns cell–cell topology representations and identifies cell clusters based on a deep graph convolutional network. Li et al., (2021) propose a multiobjective evolutionary clustering based on adaptive non-negative matrix factorization (MCANMF) for multiobjective single-cell RNA-seq data clustering. Peng et al., (2020) compared 12 single-cell clustering methods and found that most of them improved based on the K-means algorithm.

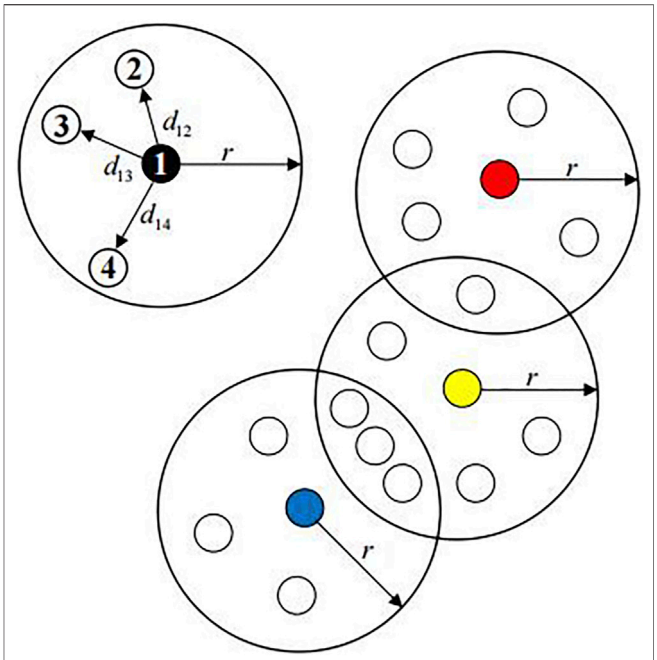


FIGURE 1 | Dissimilarity density ρ.

The K-means algorithm (Macqueen, 1966; Lloyd, 1982) was first proposed by Steinhaus in 1955, Lloyd in 1957, Ball and Hall in 1965, and McQueen in 1967 in different scientific fields. Once the algorithm is put forward, it is widely used in various areas because of its simple principle and easy implementation. At the same time, it is also commonly used in scRNA-seq clustering. However, the K-means algorithm still has some problems. Including that the value of K is difficult to determine, the clustering result depends on the selection of the initial center point, and it is easy to fall into the optimal local solution. In addition, the K-means algorithm is sensitive to noise points and outliers, and it is not practical for nonconvex data sets or data with too significant differences in category size. These problems will have a particular impact on the clustering results. To solve this problem, many workers have carried out a lot of research.

Due to the high-dimensional characteristics of single cells, we reduce the dimension of data sets and then cluster them, which can not only improve the clustering effect but also visually analyze the clustering results. This technology has been widely used in scRNA-seq clustering. Common dimensionality reduction algorithms include Principal Components Analysis (PCA), Locality Preserving Projections (LPP), t-distributed Stochastic Neighbor Embedding (t-SNE), Multidimensional Scaling (MDS), Isometric feature mapping (Isomap), and Locally Linear Embedding (LLE).

Based on dimension reduction, we propose a scRNA-seq clustering method: The dissimilarity-Density-Dynamic Radius-K-means algorithm. The algorithm obtains a set of initial center points by calculating the product of dissimilarity density ρ, average dissimilarity of candidate clusters α, and disparity of candidate clusters s. At the same time, the algorithm can optimize the clustering results by adjusting the dynamic radius parameters.

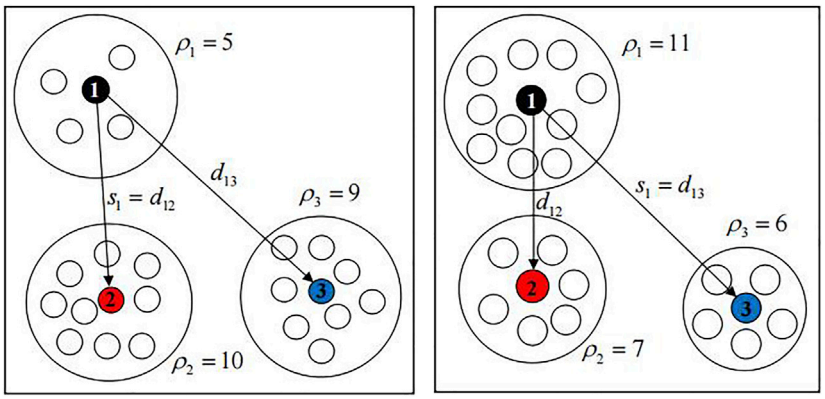


FIGURE 2 | Dissimilarity of candidate clusters s_i.

We apply this algorithm to single-cell data sets, and the obtained indicators (NMI, FMeasure_node, Accuracy, and RandIndex) are superior to those of other algorithms. They can be used as an effective tool for scRNA-seq clustering.

The main significance of this study lies in the establishment of a clustering model based on single-cell sequencing data, which can be used to cluster cells with similar gene expression patterns into the same cell type so as to infer cell functions and understand the correlation between diseases and genomic characteristics. A more precise and unbiased classification of cells would have a huge impact in oncology, genetics, immunology, and other research fields.

2 MATERIALS AND METHODS

2.1 Theoretical Presentation

The K-means algorithm will randomly select K points as initial center points when clustering, which will make the algorithm fall into optimal local solution, and the obtained clustering distribution is not optimal. It is possible to divide a smaller group into one cluster and a large cluster into several small groups. Therefore, the initial center point of the optimal group should meet the following requirements: the difference between the initial center point and other sample points in the group should be as slight as possible; The difference to sample points between the groups is as large as possible.

In thisarticle , the concept of dissimilarity is used when selecting the center point. The so-called dissimilarity is the dissimilarity between two objects, and its expression form is a $n \times n$ matrix

$$\begin{bmatrix} \boldsymbol{d}(\boldsymbol{a}_1, \boldsymbol{a}_1) & \boldsymbol{d}(\boldsymbol{a}_1, \boldsymbol{a}_2) & \cdots & \boldsymbol{d}(\boldsymbol{a}_1, \boldsymbol{a}_2) \\ \boldsymbol{d}(\boldsymbol{a}_2, \boldsymbol{a}_1) & \boldsymbol{d}(\boldsymbol{a}_2, \boldsymbol{a}_2) & \cdots & \boldsymbol{d}(\boldsymbol{a}_2, \boldsymbol{a}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{d}(\boldsymbol{a}_n, \boldsymbol{a}_1) & \boldsymbol{d}(\boldsymbol{a}_n, \boldsymbol{a}_2) & \cdots & \boldsymbol{d}(\boldsymbol{a}_n, \boldsymbol{a}_n) \end{bmatrix},$$

where $d(a_i, a_j)$ represents the degree of dissimilarity between objects a_i and a_j, which is usually a non-negative value. The more similar the two things are, the closer the case is to 0; Otherwise, the closer the matter is to 1. We find that if the dissimilarity density ρ of a point is greater, the fact is more likely to become the initial center point.

The dissimilarity density ρ of sample points x_i is the number of samples whose dissimilarity with sample objects x_i less than the dynamic radius r. Because there are often some noise points in single-cell data sets, if the average dissimilarity is taken as the radius, this fixed-radius algorithm will make the dissimilarity density ρ inaccurate and affect the selection of the initial center point. At the same time, the fixed radius will also cause the number of clusters to be unsatisfactory. Therefore, the traditional fixed-radius method is no longer suitable for single-cell clustering. If it is set to the dynamic radius r, it can effectively solve this problem and is more conducive to single-cell data clustering. The dynamic radius here is the ratio of the average dissimilarity between samples to the dynamic parameters T. The degree of dissimilarity is a model which fully considers the comprehensive distance and dynamic radius, constructed the dissimilarity matrix, and converts the single-cell data into a phase dissimilarity matrix. It can be used to better judge the differences between cells, not just by the distance between them.

2.2 Basic Definitions

$X = \{x_1, x_2, \cdots, x_n\}$is set as the sample data set to be clustered, where $x = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}, i \in \{1, 2, \cdots, n\}$, and ρ is the number of attributes.

DEFINITION 1. Dissimilarity d_{ij} between sample points x_i and x_j:

$$\boldsymbol{d}_{ij} = \sum_{s=1}^{p} \boldsymbol{d}_{ij}^{(s)}, \tag{1}$$

among them

$$\boldsymbol{d}_{ij}^{(s)} = \frac{\left|\boldsymbol{x}_{is} - \boldsymbol{x}_{js}\right|}{\boldsymbol{max}\{\boldsymbol{x}_{rs}\} - \boldsymbol{min}\{\boldsymbol{x}_{rs}\}} \tag{2}$$

represents the dissimilarity of the s th attribute between the sample point and, x_{rs} is all the values of the s th attribute.

DEFINITION 2. Constructing dissimilarity matrix d:

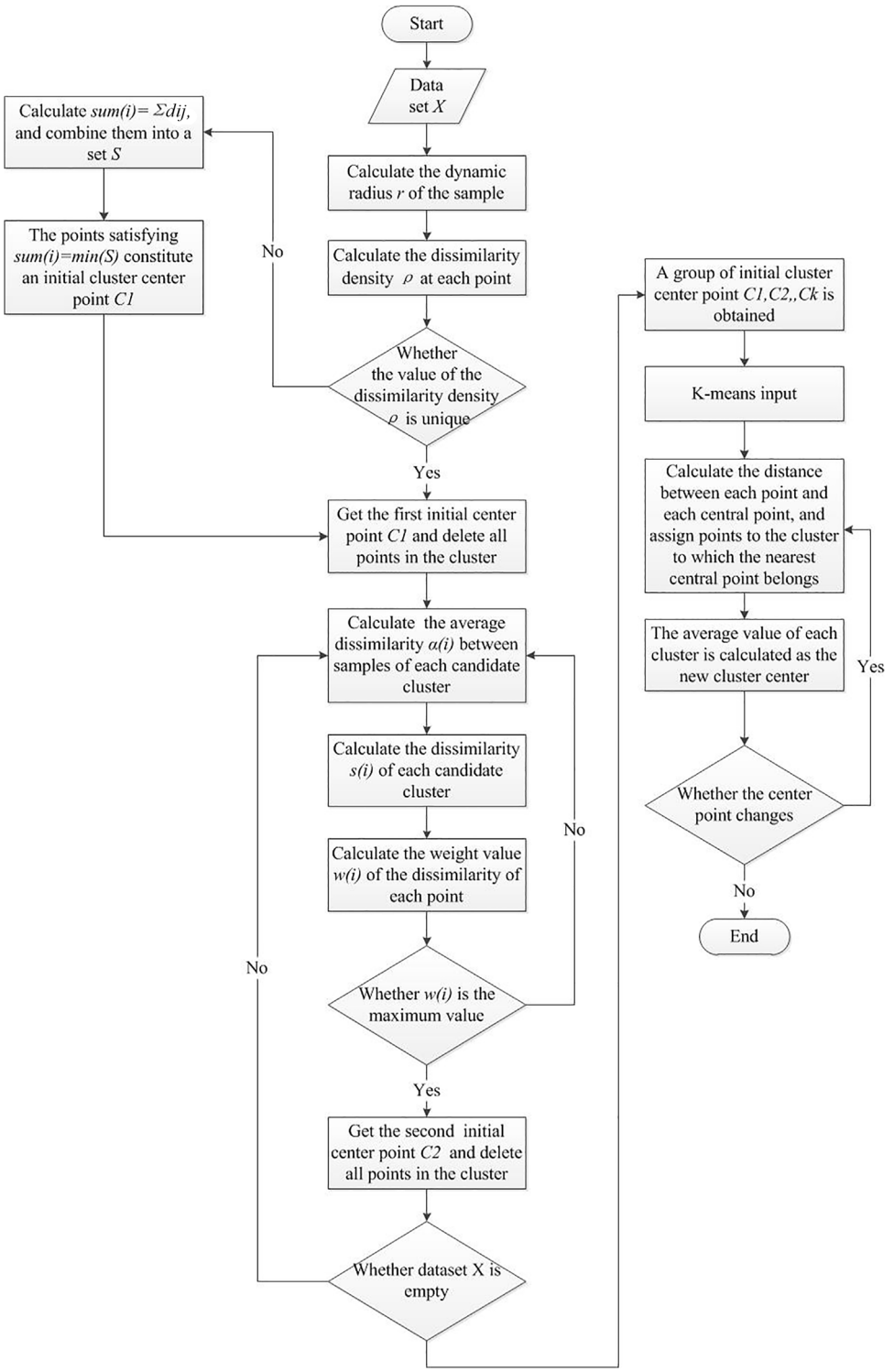


FIGURE 3 | Algorithm block diagram.

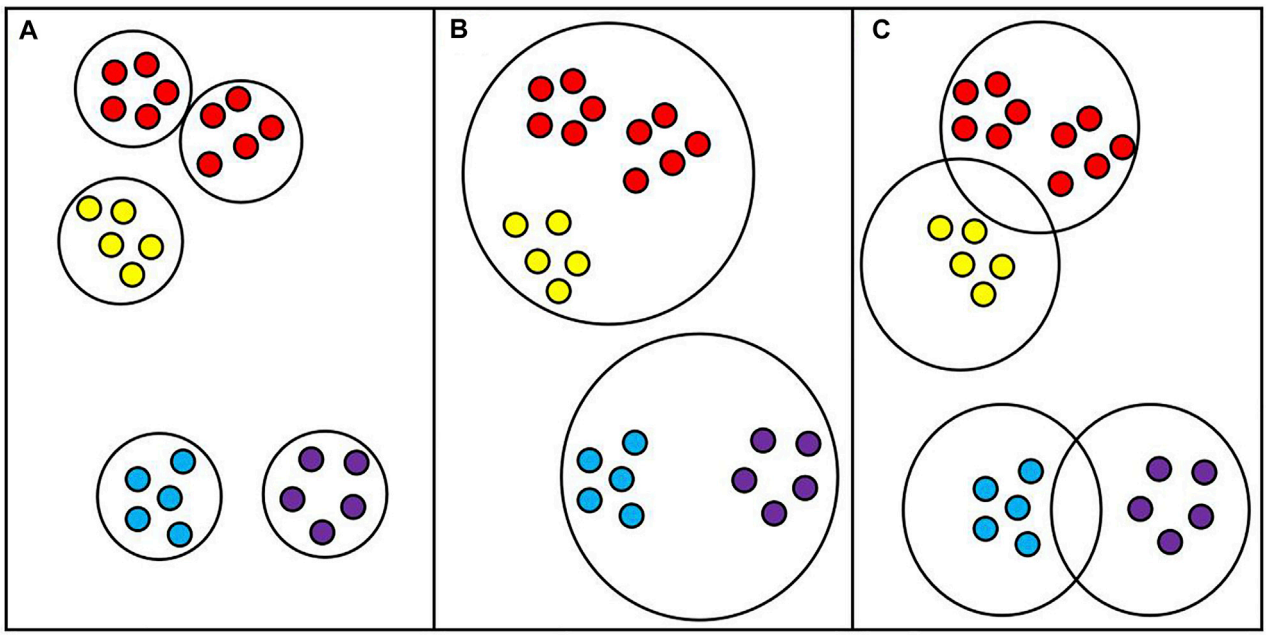


FIGURE 4 | Clustering at a fixed radius. **(A)**: The radius is too small; **(B)**: The radius is too large; **(C)**: The radius is appropriate.

$$d = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & 0 & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 0 \end{bmatrix}, \tag{3}$$

where d_{ij} represents the dissimilarity between the sample points x_i and x_j.

DEFINITION 3. Dynamic radius r of data set X:

$$r = \frac{Mean\_r(d)}{T}, \tag{4}$$

among them

$$Mean\_r(d) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}, \tag{5}$$

T is the dynamic radius parameter, and the value is as follows:

$$T = -0.423 + 0.328K - 1.211mead(d) + 0.662\max(d) + 1.631\min(d), \tag{6}$$

where K represents the number of data categories; mean means the average of dissimilarity; max represents the maximum phase dissimilarity; and min represents the minimum phase dissimilarity.

DEFINITION 4. Sample dissimilarity density ρ:

$$\rho = \sum_{j=1}^{n} \delta(d_{ij} - r), \tag{7}$$

where $\delta(z) = \begin{cases} 1, & z \le 0 \\ 0, & others \end{cases}$, ρ_i represents the dissimilarity density of the sample object x_i, which is the number of points satisfied $d_{1i} < r$.

The sample point dissimilarity density is the number of points that satisfy $d_{1i} < r$. As shown in **Figure 1**, the conditions are $d_{11}, d_{12}, d_{13}, d_{14}$. so the dissimilarity density of sample point 1 is 4. Similarly, the dissimilarity density of red dots is 6; the dissimilarity density of yellow dots is 8; the dissimilarity density of blue dots is 7. It is to be noted that the points in the intersection of two great circles can be calculated repeatedly.

DEFINITION 5. According to Definition 4, ρ is the number of samples whose dissimilarity with the sample object x_i is less than the dynamic radius r. Samples meeting the conditions form a candidate cluster, where the average dissimilarity between the samples of the candidate cluster is

$$\alpha(i) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}, \tag{8}$$

DEFINITION 6. The dissimilarity s_i of candidate clusters represents the dissimilarity between sample x_i objects x_j, which satisfies the following formula

$$s_i = \begin{cases} \min(d_{ij}), & \exists p(j) > p(i) \\ \max(d_{ij}), & \exists p(j) \le p(i) \end{cases}, \tag{9}$$

As shown on the left of **Figure 2**, the dissimilarity density of sample point 1 is 5, and there is a dissimilarity density larger than it, so the smallest dissimilarity is selected as the candidate cluster

TABLE 1 | Summary of six scRNA-seq data sets used in this study.

Data set	The number of cells	The number of genes	The number of clusters
Kolod	704	10685	3
Pollen	249	14805	11
Ting	114	11405	5
Ioh	429	18087	8
Goolam	124	16384	5
Usoskin	622	17772	4
Xin	1600	39851	8
Zeisel	3005	4412	48
Macosko	6418	12822	39

TABLE 2 | Clustering indexes after dimensionality reduction.

		Kolod	Pollen	Usoskin	Ting	Ioh	Goolam	Xin	Zeisel	Macosko
Original Data	NMI	0.5202	0.8533	0.3139	0.7262	0.5512	0.6218	0.5338	0.5262	0.4772
	FM	0.8207	0.7837	0.5923	0.534	0.6013	0.7605	0.5468	0.3260	0.3726
	Accuracy	0.6960	0.7807	0.5907	0.7746	0.5734	0.8097	0.8744	0.4985	0.4399
	RandIndex	0.7080	0.9323	0.7011	0.8370	0.7924	0.8140	0.6971	0.9230	0.9092
t-SNE	NMI	**0.8344**	**0.9169**	**0.7197**	**0.8402**	**0.8296**	**0.7298**	**0.6087**	**0.5741**	**0.6954**
	FM	**0.9025**	**0.8682**	**0.8032**	**0.9494**	**0.8540**	**0.9363**	**0.5456**	**0.3564**	**0.5339**
	Accuracy	**0.9071**	**0.9149**	**0.6521**	**0.9033**	**0.8748**	**0.8952**	**0.9306**	**0.5784**	**0.6790**
	RandIndex	**0.9005**	**0.5335**	**0.8804**	**0.9197**	**0.9459**	**0.8937**	**0.7088**	**0.9297**	**0.9488**
PCA	NMI	0.5557	0.8190	0.3435	0.8318	0.6398	0.6674	0.5821	0.4031	0.3433
	FM	0.7710	0.8013	0.5486	0.9077	0.6616	0.7779	0.5873	0.2254	0.2456
	Accuracy	0.7685	0.8233	0.5723	0.8947	0.6727	0.8653	0.9175	0.4254	0.3398
	RandIndex	0.7905	0.9475	0.6837	0.9127	0.8648	0.8679	0.7119	0.9188	0.9185
MDS	NMI	0.5519	0.8123	0.3438	0.8228	0.6444	0.7202	0.5960	0.4033	0.3441
	FM	0.7679	0.5588	0.5588	0.8429	0.6674	0.7927	0.6046	0.2255	0.2420
	Accuracy	0.7648	0.5723	0.5723	0.8596	0.6681	0.8871	0.9219	0.4252	0.3363
	RandIndex	0.7883	0.6845	0.6845	0.9067	0.8647	0.9078	0.7288	0.9189	0.9163
Isomap	NMI	0.4574	0.7350	0.3686	0.9173	0.7812	0.6535	0.6002	0.5338	0.5063
	FM	0.7797	0.6632	0.6709	0.8064	0.8292	0.7295	0.5852	0.3307	0.4182
	Accuracy	0.7741	0.6908	0.6672	0.8684	0.8436	0.8734	0.9207	0.5196	0.4634
	RandIndex	0.7590	0.9070	0.7372	0.9104	0.9355	0.8173	0.7240	0.9251	0.9133
LLE	NMI	0.5358	0.8941	0.4951	0.8172	0.7867	0.7205	0.5831	0.5719	0.6020
	FM	0.8006	0.8931	0.7353	0.8458	0.7843	0.3620	0.6042	0.3620	0.5398
	Accuracy	0.7955	0.9076	0.7267	0.8772	0.8462	0.5237	0.8882	0.5237	0.5734
	RandIndex	0.7897	0.9695	0.7841	0.8763	0.9225	0.8978	0.7240	0.8978	0.9405
LPP	NMI	0.7105	0.8875	0.6887	0.7869	0.7709	0.7056	0.5543	0.4819	0.4517
	FM	0.7977	0.8460	0.8559	0.8351	0.7449	0.7991	0.5506	0.2664	0.3275
	Accuracy	0.7979	0.8594	0.8376	0.8509	0.8089	0.8790	0.9006	0.4516	0.4020
	RandIndex	0.7925	0.9620	0.8680	0.8572	0.8693	0.8996	0.7036	0.9197	0.9232

dissimilarity of sample point 1; as shown in the right of **Figure 2**, the dissimilarity density of sample point 1 is 11, and there is no dissimilarity density larger than it. Therefore, the biggest dissimilarity is selected as the candidate cluster dissimilarity of sample point 1.

By analyzing Definitions 5, 6, when the candidate cluster is formed with x_i as the center point, if the average dissimilarity value $\alpha(i)$ between samples of the candidate cluster is smaller, the dissimilarity of the cluster is very small, and the similarity is very high; similarly, the greater the value of s_i, the greater the dissimilarity between samples. Therefore, the dissimilarity density ρ, the average dissimilarity $\alpha(i)$, and the dissimilarity value s_i of candidate clusters can be taken as the standard to measure the initial center point, which is specifically defined as follows:

DEFINITION 7. The dissimilarity weight formula for selecting the cluster center point is as follows:

$$\omega_i = \rho_i * \frac{1}{\alpha_i} * s_i, \tag{10}$$

among them, the point with the most significant weight of dissimilarity is the initial center point.

2.3 Algorithm Flow and Block Diagram

2.3.1 Algorithm Flow

The Dissimilarity-Density-Dynamic Radius-K-means algorithm calculates the dissimilarity density ρ of sample points, the average dissimilarity α of candidate clusters, and the dissimilarity value s of candidate clusters to obtain the dissimilarity weight ω of sample points and determine a group

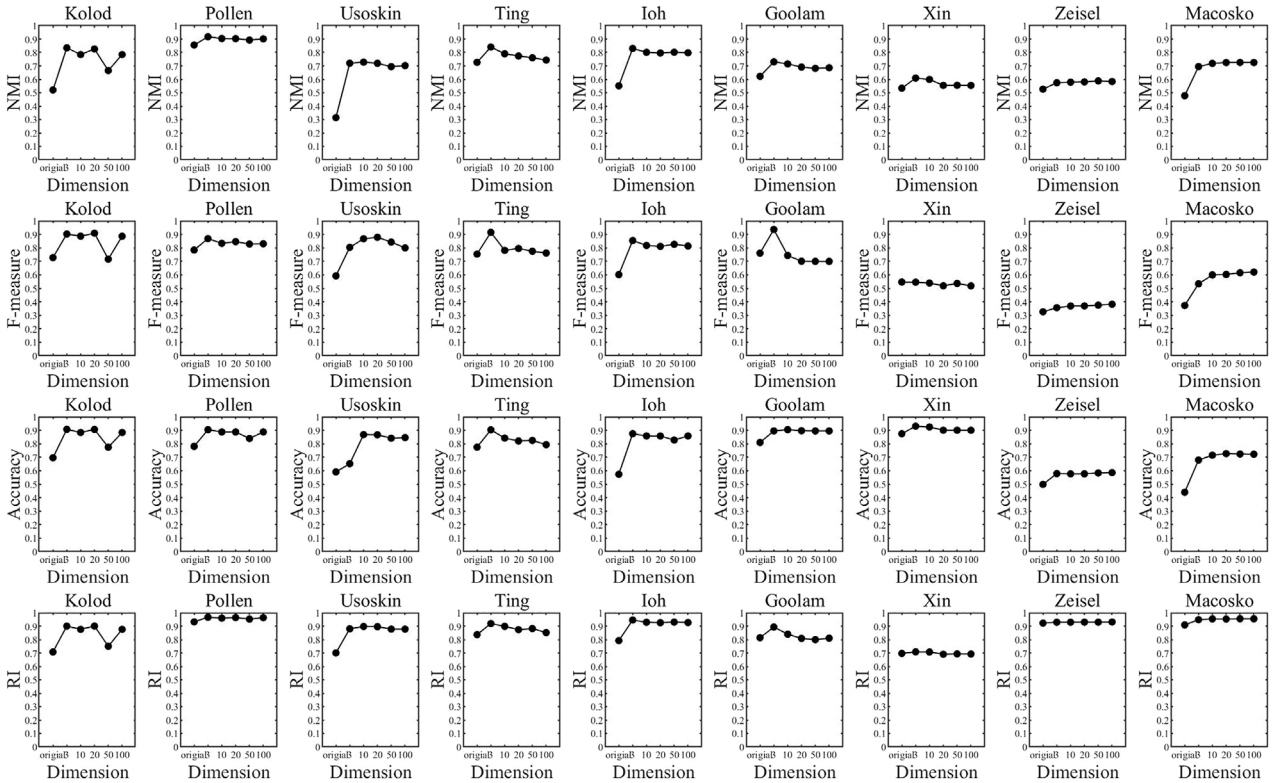


FIGURE 5 | Clustering index values of different dimension clustering.

of initial center points. Then, the obtained center point is used as the initial center point of K-means for clustering. The flow of the Dissimilarity-Density-Dynamic Radius-K-means algorithm is as follows:

1) Giving a data set $X = \{x_1, x_2, \cdots, x_n\}$;
2) Calculating the dissimilarity density of all points in x is in accordance with the definition and form a set ρ;
3) Finding that point corresponding to the maximum value from the dissimilarity set ρ; if the number of the value is 1, the point is taken as the first initial clustering center point; if the number of the maximum value is not 1, the calculated $sum(i) = \sum_{j=1}^{n} d_{ij}$, wherein $d_{ij} \le r$, $j = 1, 2, \cdots, n$, and form the set S, that satisfying $sum(i) = \min(S)$ point is taken as the first initial center point;
4) Obtaining a first initial clustering center point at this time, recording $C1$, and putting it in the set C at that time $C = \{C1\}$. Then, points satisfying $d_{1i} < r$ are then removed from the data set X;
5) Calculating the weight value ω_i of the dissimilarity of the remaining point according to the definition, wherein the second initial center point is the point with the maxumun weight value of the distinction and is recorded $C2$ and put in set C at that time $C = \{C1, C2\}$. Then, deleting the points that meet the criteria;
6) Repeating step 5 until that data set is empty, $C = \{C1, C2, \cdots, Ck\}$;
7) At this time, a group of initial center points C and the number K of clustering have been obtained, and the parameters are brought into the k-means algorithm for clustering;
8) Calculating the distance between each point in the sample and the initial center point, classifying the space into the cluster where the center point with the smallest distance between each other is located, and calculating the new center points of each group;
9) Repeating the step 8 until the division condition of all sample points remain unchanged or the central point does not change;
10) Output that clustering result.

2.3.2 Algorithm Block Diagram

The algorithm block diagram is shown in **Figure 3**.

2.4 The Necessity of Setting the Dynamic Radius Parameter T

When introducing the D3K algorithm, we put forward the definition of dynamic radius R; the so-called dynamic radius is the ratio of average dissimilarity and dynamic radius parameter T. The distribution of the data set is not uniform. If the distribution of the data set is too scattered or too close, the average dissimilarity will be too large or too small. If the average dissimilarity is taken as the radius, the clustering result will be inaccurate, which will affect

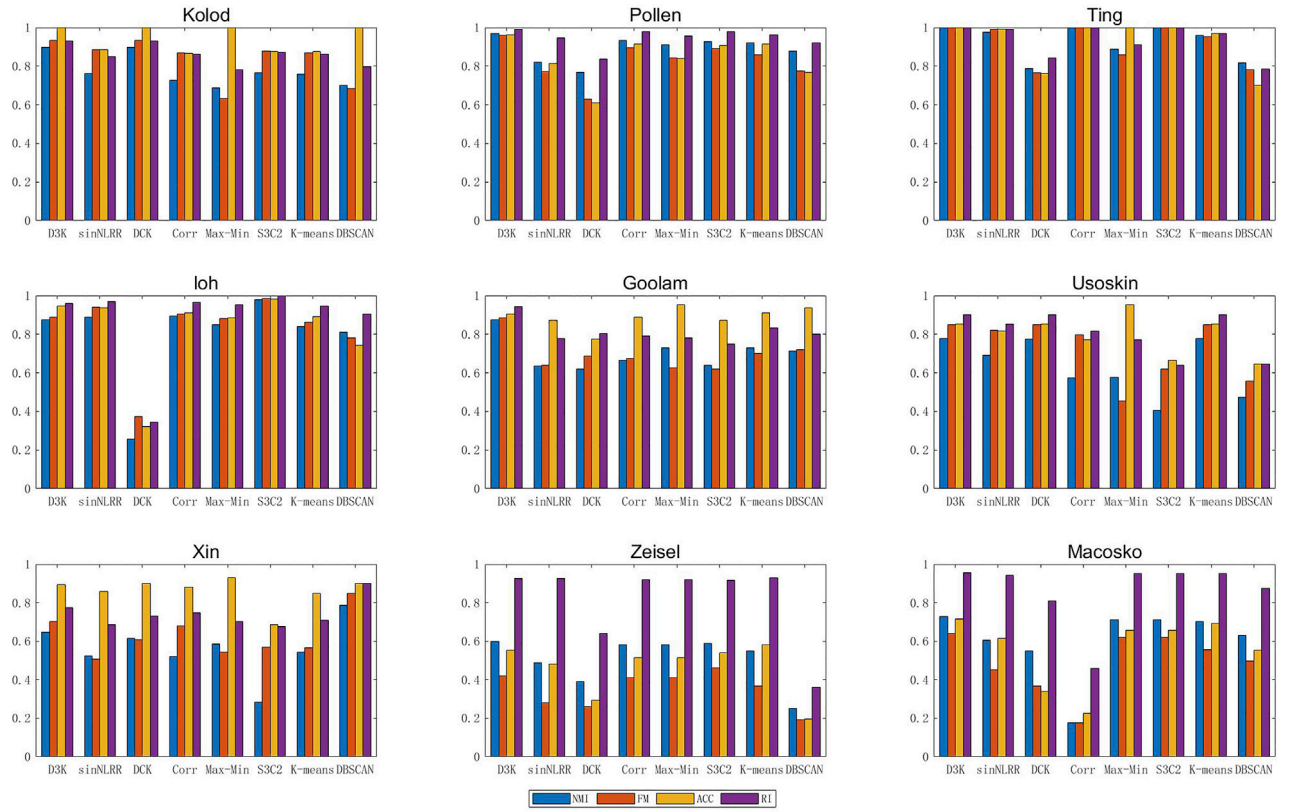


FIGURE 6 | Index of the D3K algorithm in single-cell data aggregation class.

the selection of the initial center point and result in an inaccurate clustering result. If the dynamic radius parameter is added, the radius can be adjusted flexibly according to the data characteristics so as to optimize the clustering result. As shown in **Figure 4** below:

As shown in **Figure 4C**, for clustering results under ideal conditions, appropriate radii are set and clusters are divided reasonably. However, if the average dissimilarity is taken as the radius, the average dissimilarity will be too small for some overly tight data sets, which will make the radius smaller, and the original cluster will be divided into two or more clusters, as shown in **Figure 4A**. For some data sets that are too scattered or have noise points, the average dissimilarity will be very large. In this case, taking the average dissimilarity as the radius will make the radius very large so that originally different clusters can be divided into one cluster, as shown in **Figure 4B**. Therefore, adding the dynamic radius parameter T into the model can reasonably adjust the radius size according to the data characteristics and optimize the result of cluster division.

The dynamic radius parameter T is considered from multiple perspectives, including the maximum, minimum, average, and the number of clusters K. Considering many aspects, we get the optimal solution through the greedy algorithm and then fit the equation of the dynamic radius parameter T through a large amount of data. Among them, the dissimilarity between each point and itself is 0, so the dissimilarity between each point and itself should be removed when selecting the minimum value of phase dissimilarity, that is, the value with the smallest foreign phase dissimilarity except 0. By observing the equation of dynamic radius parameter T, it is found that the coefficient of K value of the number of clusters is only 0.328, indicating that although the dynamic radius parameter T is related to the number of clusters, it does not account for the main factor, and the optimal solution of T is in an interval, so the equation can be satisfied without a particularly accurate K value.

3 RESULT

To verify the algorithm, we selected nine groups of single-cell data sets for experiments, namely, Kolod, Pollen, Ting, Ioh, Goolam, Usoskin, Xin, Zeisel, and Macosko data sets. **Table 1** shows the details of the data set.

Clustering the data in **Table 1** after dimension reduction can improve the clustering effect and visually analyze the clustering results. We compare the effects of six dimensionality reduction methods on single-cell data and visually examine the clustering results and find out an algorithm suitable for dimensionality reduction of single-cell data. At the same time, to verify the quality of the algorithm, we compare it with other single-cell clustering algorithms and finally confirm the selection of parameter T in this study.

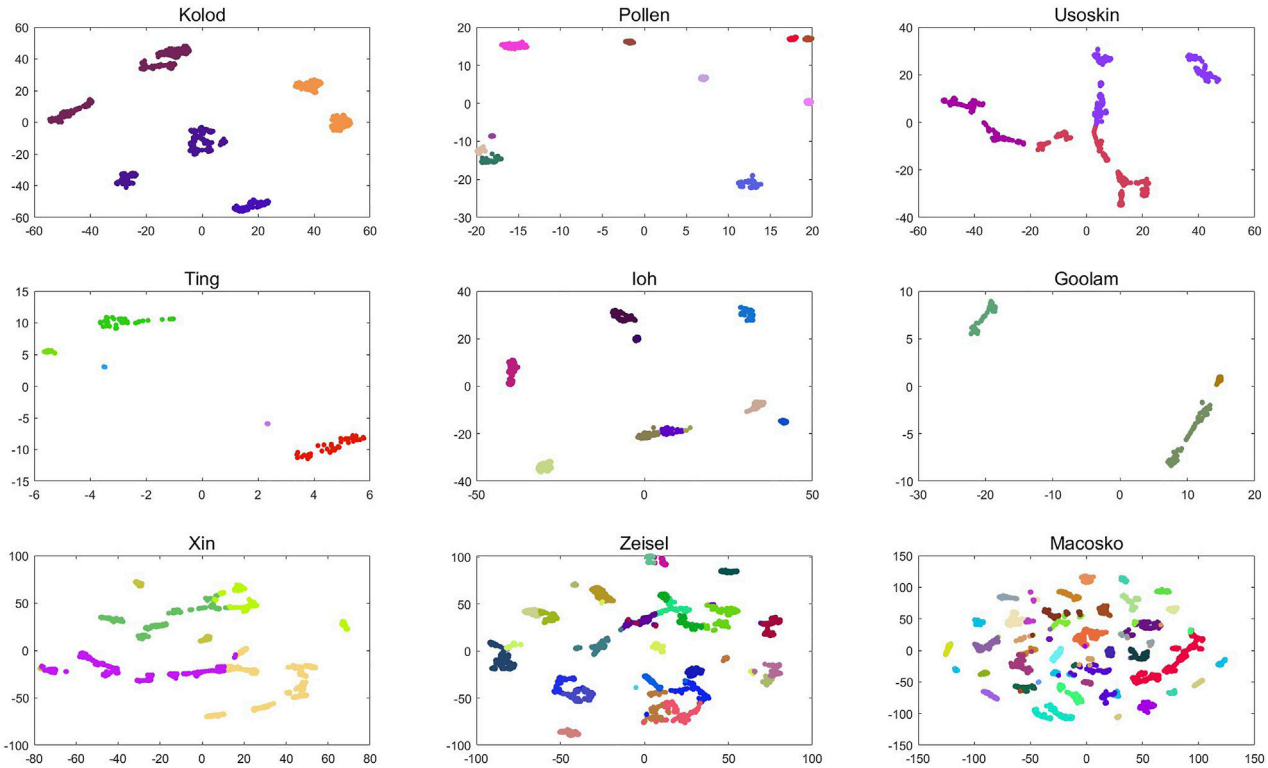


FIGURE 7 | D3K algorithm visualization analysis.

3.1 Dimension Reduction

To find a dimension reduction algorithm suitable for single-cell data sets, we preprocess single-cell data with different dimension reduction algorithms and then cluster the reduced data to obtain clustering results. Here, we compare six dimensionality reduction algorithms: T-SNE, PCA, MDS, LPP, and LLE Isomap. By reducing dimensions in clustering, we obtain the data in **Table 2**:

By analyzing the data in **Table 2**, it can be found that after dimensionality reduction is used, the values of each index of clustering have been significantly improved, indicating that dimensionality reduction is very important for clustering, which can not only greatly increase the accuracy of clustering but also reduce the calculation time. At the same time, it can be found that in most of the data, the t-SNE algorithm has the best improvement effect. Therefore, the T-SNE algorithm can be used as an effective tool for single-cell clustering.

In the previous experiments, we have concluded that the t-SNE algorithm is more suitable for single-cell data dimension reduction, but how many dimensions to reduce the dimension is more suitable for clustering is still a problem to be discussed. To this end, we set up the following experiments: The t-SNE algorithm with the best dimensional reduction effect for single-cell data was selected, and six groups of single-cell data were reduced to 3, 10, 20, 50, and 100 dimensions for K-means clustering, and the clustering index results in different dimensions were analyzed. In order to compare the differences between different dimensions more clearly, the results are presented in a broken line graph. As shown in **Figure 5**:

Through the analysis of **Figure 5**, it is found that each data set has an inflection point in three dimensions, that is to say, the data will be reduced to three-dimension clusterings, and the clustering result will be significantly improved. Although some data still improve after three-dimension clustering, the increase is very small and can be almost ignored. Therefore, we can conclude that the t-SNE algorithm has the best clustering effect when the data are reduced to three dimensional ones. Therefore, in the following experiments, we uniformly used the t-SNE algorithm to reduce single-cell data to three dimensional ones for clustering.

3.2 Comparison With Other Clustering Algorithms

To verify the effectiveness of the D3K algorithm, we selected seven single-cell clustering algorithms to compare with it, namely, DCK (Zhang et al., 2018), S3C2 (Zhuang et al., 2021), sinNLRR (Zheng et al., 2019), Corr (Dong et al., 2018), Max-Min (Sen et al., 2018), K-means, and DBSCAN algorithm.

The nine groups of single-cell data in **Table 1** were clustered by the single-cell clustering algorithm described above, and each index (NMI, FMeasure_node, Accuracy, RandIndex) of the clustering result was obtained to obtain **Figure 6** as follows:

Compared with other clustering algorithms, the D3K algorithm is obviously higher than different algorithms in

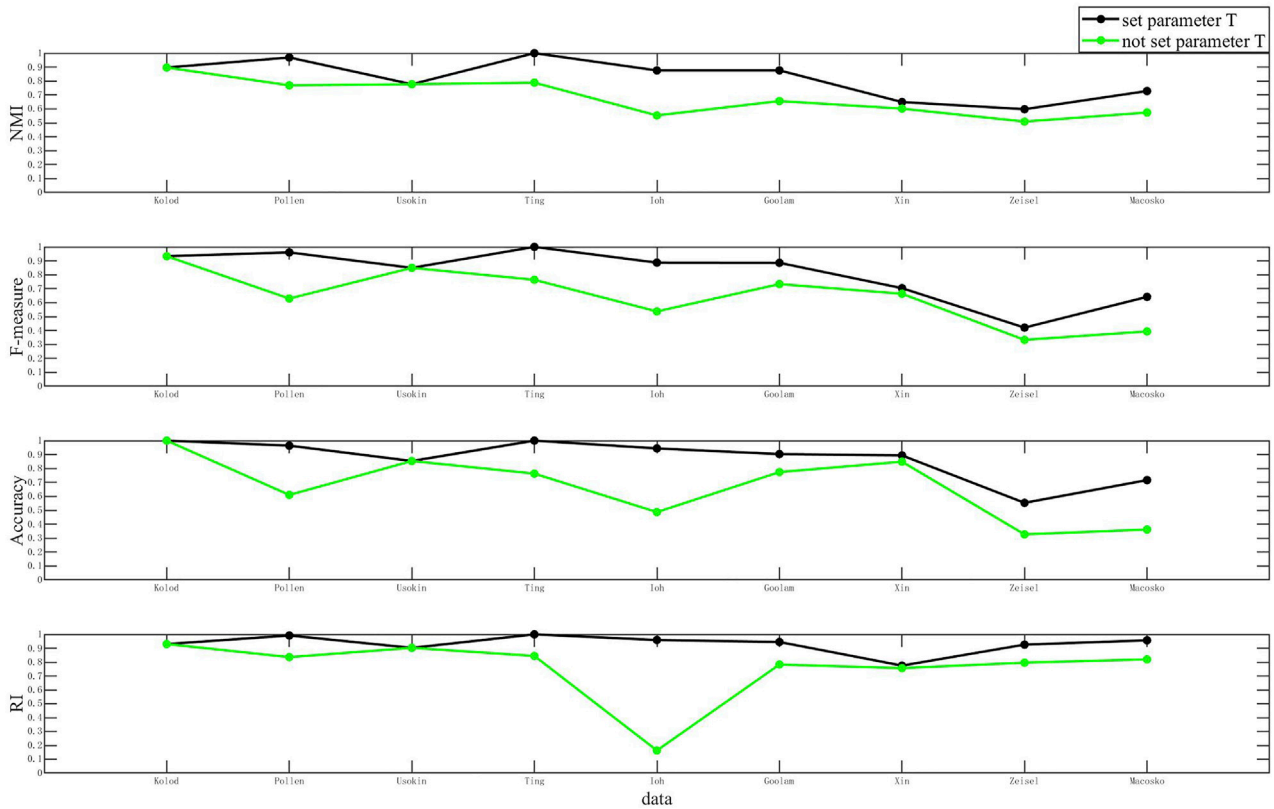


FIGURE 8 | Clustering results when T and not T are set.

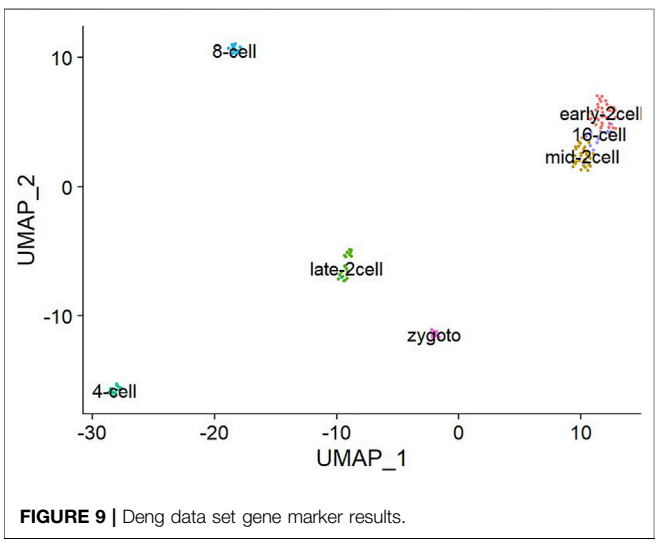


FIGURE 9 | Deng data set gene marker results.

various indexes, and the results of multiple indexes are basically above 0.8, among which the effects of multiple indexes of the Pollen data set can reach above 0.95, especially Ting data set, and the results all are 1. It can be seen that the D3K algorithm can achieve ideal clustering results for both small and large data sets and can be used as a clustering model for single-cell data.

Visual analysis of clustering results can not only clearly display complex data in the form of images but also intuitively observe the differences between clusters and the size of differences within clusters. For single-cell data, this study first constructs its dissimilarity degree matrix and then obtains the cluster label of single-cell data through clustering. According to the cluster label, visual analysis of the dissimilarity matrix can not only show the clustering results of single cells after clustering but also make the distance within the same cluster smaller and the distance between different clusters larger. The following **Figure 7** is a visual analysis of the clustering results of six groups of single-cell data, and the clustering results of the D3K algorithm are displayed in the form of images.

As shown in **Figure 7**, the visualization results of the D3K algorithm after clustering 10 groups of single-cell data are shown. It can be seen that the D3K algorithm can perfectly divide these data into different cell types according to the labels after clustering and make the differences within clusters after clustering very small, but the differences between clusters are very large.

3.3 Validation of Parameter T

When introducing the D3K algorithm, we propose the definition of dynamic radius r, and the so-called dynamic radius is the ratio of the average degree of difference to the dynamic radius parameter. The distribution of the dataset is not uniform, and if the dataset

distribution is too scattered or too tight, it will cause the average difference to be too large or too small. If the radius is based on the average degree of difference, it will affect the selection of the initial center point, resulting in inaccurate clustering results. By adding the dynamic radius parameter, you can find the right radius for each set of data to optimize clustering results.

In order to explain the necessity of the dynamic radius parameter T more rigorously, we set up the following experiment and ninesets of single-cell data were taken and clustered using D3K. The dynamic radius parameter T is not added to the first cluster, and the dynamic radius parameter T is added to the second cluster to compare the difference between the results. The result is shown in **Figure 8**:

As shown in **Figure 8**, the comparison of clustering results of the D3K algorithm when T is set and not set is shown. The abscissa of each of these plots represents the dataset, and the ordinate coordinate represents the values of each metric. The black polyline represents the clustering result when T is set, and the green polyline represents the clustering result when T is not set. The analysis found that the clustering results when setting T were better than the clustering results when T was not set. It is to be noted that setting the T value can optimize the clustering results and make the clustering results more accurate.

3.4 Genetic Markers

The task of single-cell scrNA-SEQ sequencing is not only to cluster single-cell sequencing data but also to cluster cells with similar gene expression patterns into the same cell type. Extraction of gene markers from the single-cell level of single-cell RNA-SEQ and cell identification is also an important part because it can assist in subsequent analysis of gene interactions. As shown in **Figure 9**, after annotation of the Deng data cluster class, its marker genes can be determined. The Deng marker genes include Early-2cell, mid-2cell, late-2cell, 4cell, 8cell, 16cell, and Zygoto. By clustering single-cell data, gene markers can be realized more effectively, which is convenient for further research on a single cell.

4 DISCUSSION

scRNA-seq can quickly determine the precise gene expression patterns of thousands of single cells and reveal the complexity of the horizontal structure of single cells, thus improving biomedical research and solving various problems in biology. However, due to the high dimension and multi-noise characteristics of single-cell sequencing data sets, it brings significant challenges to the traditional clustering algorithm. In this study, we propose a Dissimilarity-Density-Dynamic Radius-K-means clustering algorithm. By selecting the dynamic radius, the algorithm effectively calculates the dissimilarity density ρ of the data set, the average dissimilarity α of candidate clusters, and the dissimilarity s of candidate clusters, finds a group of high-quality initial center points, and achieves the purpose of improving the K-means algorithm.

We use the Dissimilarity-Density-Dynamic Radius-K-means clustering algorithm to cluster some single-cell data sets and evaluate the clustering results. Experiments show that the Dissimilarity-Density-Dynamic Radius-K-means clustering algorithm has good performance for single-cell data clusters. At the same time, we also compared with other single-cell clustering algorithms. Experiments show that the Dissimilarity-Density-Dynamic Radius-K-means clustering algorithm is superior to other single-cell clustering algorithms.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These datasets can be found here: This study used data from the GEO database and can be found below: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

GL and ML proposed the algorithm, wrote the code, and wrote the manuscript. HW revised the original manuscript. SL ran other single-cell clustering algorithms. JX, RL, and MT discussed the proposed algorithm and carried out further research.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61903106), the Hainan Province Natural Science Foundation (No. 621MS0773, No. 118QN231), Key Laboratory of Data Science and Smart Education, Ministry of Education, the Project of Hainan Key Laboratory for Computational Science and Application, and Hainan Normal University for the funding of the Ph.D.

ACKNOWLEDGMENTS

The authors appreciate Binbin Ji, Lingyu Cui, and others for useful discussion.

REFERENCES

Belkin, M., and Niyogi, P. (2001). Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Adv. Neural Inf. Process. Syst.* 14, 585–591. doi:10.7551/mitpress/2F1120.003.0080

Chen, L., Xu, Z., Wang, H., and Liu, S. (2016). An Ordered Clustering Algorithm Based on K-Means and the Promethee Method. *Int. J. Mach. Learn. Cyber.* 9, 917–926. doi:10.1007/s13042-016-0617-9

Dong, J., Hu, Y., Fan, X., Wu, X., Mao, Y., Hu, B., et al. (2018). Single-cell RNA-Seq Analysis Unveils a Prevalent Epithelial/mesenchymal Hybrid State during Mouse Organogenesis. *Genome Biol.* 19, 31. doi:10.1186/s13059-018-1416-2

Dong, Q., and Zhu, Z. (2020). A New K-Means Algorithm for Selecting Initial Clustering Center. *Statistics Decis.* 36, 32–35. doi:10.13546/j.cnki.tjyjc.16.007

Gan, Y., Huang, X., Zou, G., Zhou, S., and Guan, J. (2022). Deep Structural Clustering for Single-Cell RNA-Seq Data Jointly through Autoencoder and Graph Neural Network. *Briefings Bioinforma.* 23 (2), 1467–5463. doi:10.1093/bib/bbac018

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix Completion with Side Information and its Applications in Predicting the Antigenicity of Influenza Viruses. *Bioinformatics* 33, 3195–3201. doi:10.1093/bioinformatics/btx390

Jiang, H., Sohn, L. L., Huang, H., and Chen, L. (2018). Single Cell Clustering Based on Cell-Pair Differentiability Correlation and Variance Analysis. *Bioinformatics* 34, 3684–3694. doi:10.1093/bioinformatics/bty390

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data. *Nat. Rev. Genet.* 20, 273–282. doi:10.1038/s41576-018-0088-9

Li, M., Wang, H., Long, H., Xiang, J., and Yang, J. (2019). Community Detection and Visualization in Complex Network by the Density-Canopy-Kmeans Algorithm and MDS Embedding, *IEEE Access*, 7, 120616–120625. doi:10.1109/ACCESS.2936248

Li, X., Li, S., Huang, L., Zhang, S., and Wong, K.-c. (2021). High-throughput Single-Cell RNA-Seq Data Imputation and Characterization with Surrogate-Assisted Automated Deep Learning. *Briefings Bioinforma.* 23 (1), 1. doi:10.1093/bib/bbab368

Li, X., and Wong, K.-C. (2019). Single-Cell RNA Sequencing Data Interpretation by Evolutionary Multiobjective Clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 17, 1773–1784. doi:10.1109/TCBB.2019.2906601

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-Of-Origin. *Front. Cell Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Lloyd, S. (1982). Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* 28, 129–137. doi:10.1109/TIT.1982.1056489

Macqueen, J. B. (1966). *Some Methods for Classification and Analysis of Multivariate Observations* 1 (14), 281–297.

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020). Single-cell RNA-Seq Clustering: Datasets, Models, and Algorithms. *RNA Biol.* 17, 765–783. doi:10.1080/15476286.2020.1728961

Potter, S. S. (2018). Single-cell RNA Sequencing for the Study of Development, Physiology and Disease. *Nat. Rev. Nephrol.* 14, 479–492. doi:10.1038/s41581-018-0021-7

Qi, R., Wu, J., Guo, F., Xu, L., and Zou, Q. (2021). A Spectral Clustering with Self-Weighted Multiple Kernel Learning Method for Single-Cell RNA-Seq Data. *Brief. Bioinform* 22 (4), bbaa216. doi:10.1093/bib/bbaa216

Qiao, L., Efatmaneshnik, M., Ryan, M., and Shoval, S. (2017). Product Modular Analysis with Design Structure Matrix Using a Hybrid Approach Based on MDS and Clustering. *J. Eng. Des.* 28, 433–456. doi:10.1080/09544828.2017.1325858

Sen, X. U., Hua, X., Jing, X. U., Xiufang, X. U., Gao, J., and Jing, A. N. (2018). Cluster Ensemble Approach Based on T-Distributed Stochastic Neighbor Embedding. *J. Electron. Inf. Technol.* 40 (6), 1316–1322. doi:10.11999/JEIT170937

Sun, X., Lin, X., Li, Z., and Wu, H. (2022). A Comprehensive Comparison of Supervised and Unsupervised Methods for Cell Type Identification in Single-Cell RNA-Seq. *Briefings Bioinforma.* 23. doi:10.1093/bib/bbab567

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2020). Indicator Regularized Non-negative Matrix Factorization Method-Based Drug Repurposing for COVID-19. *Front. Immunol.* 11, 603615. doi:10.3389/fimmu.2020.603615

Wang, J., Xia, J., Tan, D., Lin, R., Su, Y., and Zheng, C.-H. (2022). scHFC: a Hybrid Fuzzy Clustering Method for Single-Cell RNA-Seq Data Optimized by Natural Computation. *Briefings Bioinforma.* 23 (2), bbaa588. doi:10.1093/bib/bbab588

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: an Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinformatics* 36, 3139–3147. doi:10.1093/bioinformatics/btaa109

Yang, J., Liao, B., Zhang, T., and Xu, Y. (2019). Editorial: Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine. *Front. Genet.* 10, 1358. doi:10.3389/fgene.2019.01358

Yu, Z., Lu, Y., and Wang, Y. (2022). ZINB-based Graph Embedding Autoencoder for Single-Cell RNA-Seq Interpretations.

Zhang, G., Zhang, C., and Zhang, H. (2018). Improved K-Means Algorithm Based on Density Canopy. *Knowledge-Based Syst.* 145, 289–297. doi:10.1016/j.knosys.2018.01.031

Zhang, Z., Cui, F., Wang, C., Zhao, L., and Zou, Q. (2021). Goals and Approaches for Each Processing Step for Single-Cell RNA Sequencing Data. *Briefings Bioinforma.* 22, bbaa314. doi:10.1093/bib/bbaa314

Zheng, R., Li, M., Liang, Z., Wu, F.-X., Pan, Y., and Wang, J. (2019). SinNLRR: a Robust Subspace Clustering Method for Cell Type Detection by Non-negative and Low-Rank Representation. *Bioinformatics* 35, 3642–3650. doi:10.1093/bioinformatics/btz139

Zhuang, J., Cui, L., Qu, T., Ren, C., Xu, J., Li, T., et al. (2021). A Streamlined scRNA-Seq Data Analysis Framework Based on Improved Sparse Subspace Clustering. *IEEE Access* 9, 9719–9727. doi:10.1109/ACCESS.2021.3049807

Conflict of Interest: RL was employeed by Geneis Beijing Co., Ltd.

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

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

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 18 July 2022
doi: 10.3389/fgene.2022.908826

Transcriptome Analysis Reveals Hub Genes Regulating Autophagy in Patients With Severe COVID-19

Jinfeng Huang[1,2†], Yimeng Wang[1†], Yawen Zha[3], Xin Zeng[1], Wenxing Li[4] and Meijuan Zhou[1,2]*

[1]Department of Radiation Medicine, Guangdong Provincial Key Laboratory of Tropical Disease Research, School of Public Health, Southern Medical University, Guangzhou, China, [2]Jiangmen Central Hospital, Affiliated Jiangmen Hospital of Sun Yat-sen University, Jiangmen, China, [3]Department of Radiation Oncology II, Zhongshan People's Hospital, Zhongshan, China, [4]Department of Biochemistry and Molecular Biology, School of Basic Medicine, Southern Medical University, Guangzhou, China

Background: The COVID-19 pandemic has currently developed into a worldwide threat to humankind. Importantly, patients with severe COVID-19 are believed to have a higher mortality risk than those with mild conditions. However, despite the urgent need to develop novel therapeutic strategies, the biological features and pathogenic mechanisms of severe COVID-19 are poorly understood.

Methods: Here, peripheral blood mononuclear cells (PBMCs) from four patients with severe COVID-19, four patients with mild COVID-19, and four healthy controls were examined by RNA sequencing (RNA-Seq). We conducted gene expression analysis and Venn diagrams to detect specific differentially expressed genes (DEGs) in patients with severe disease compared with those with mild conditions. Gene Ontology (GO) enrichment analysis was performed to identify the significant biological processes, and protein–protein interaction networks were constructed to extract hub genes. These hub genes were then subjected to regulatory signatures and protein–chemical interaction analysis for certain regulatory checkpoints and identification of potent chemical agents. Finally, to demonstrate the cell type-specific expression of these genes, we performed single-cell RNA-Seq analyses using an online platform.

Results: A total of 144 DEGs were specifically expressed in severe COVID-19, and GO enrichment analysis revealed a significant association of these specific DEGs with autophagy. Hub genes such as *MVB12A*, *CHMP6*, *STAM*, and *VPS37B* were then found to be most significantly involved in the biological processes of autophagy at the transcriptome level. In addition, six transcription factors, including SRF, YY1, CREB1, PPARG, NFIC, and GATA2, as well as miRNAs, namely, hsa-mir-1-3p, and potent chemical agents such as copper sulfate and cobalt chloride, may cooperate in regulating the autophagy hub genes. Furthermore, classical monocytes may play a central role in severe COVID-19.

Conclusion: We suggest that autophagy plays a crucial role in severe COVID-19. This study might facilitate a more profound knowledge of the biological characteristics and

OPEN ACCESS

Edited by:
Jialiang Yang,
Genesis Beijing Co., Ltd., China

Reviewed by:
Yanfa Sun,
Longyan University, China
Haifang Zhang,
Second Affiliated Hospital of Soochow University, China

****Correspondence:***
Meijuan Zhou
fyzmj@163.com

†These authors have contributed equally to this work

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

Received: *31 March 2022*
Accepted: *06 June 2022*
Published: *18 July 2022*

Citation:
Huang J, Wang Y, Zha Y, Zeng X, Li W and Zhou M (2022) Transcriptome Analysis Reveals Hub Genes Regulating Autophagy in Patients With Severe COVID-19. Front. Genet. 13:908826. doi: 10.3389/fgene.2022.908826

progression of COVID-19 and the development of novel therapeutic approaches to achieve a breakthrough in the current COVID-19 pandemic.

Keywords: severe COVID-19, differentially expressed genes, protein ubiquitination, RNA sequencing, peripheral blood mononuclear cells

INTRODUCTION

The current COVID-19 pandemic, caused by novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to urgent healthcare issues worldwide. According to the World Health Organization, 223 countries or regions had reported 456,797,217 confirmed cases of COVID-19 by 14 March 2022, including 6,043,094 deaths. The manifestations of COVID-19 vary, and most infected individuals have only mild symptoms similar to typical pneumonia or even no symptoms (Wu and McGoogan, 2020). Furthermore, mortality is mainly observed in patients with severe COVID-19 with severe respiratory failure associated with interstitial lung pneumonia and acute respiratory distress syndrome (Berlin et al., 2020). In countries that did not implement active control measures, the case fatality rate of COVID-19 was as high as ~10% (Iype and Gulati, 2020). However, treatment options are limited to symptomatic treatment to reduce the severity of symptoms, and no curative treatment is available. Moreover, in COVID-19, especially in the severe forms, the characteristics and effects of biological reactions are still poorly understood, which prompts researchers to search for better predictors of clinical outcomes and tools to provide information for developing new therapeutic targets and appropriate therapeutic measures. Transcriptome profiling by RNA sequencing offers sufficient gene expression analysis for characterizing COVID-19 and explains biological pathways and key genes that are not yet targeted by current therapies. In this way, Mahmud et al. (2021) identified transcription factor–gene interactions, protein–drug interactions, and DEG-miRNA coregulatory networks with differentially expressed genes (DEGs) for effective treatment of COVID-19. Auwul et al. (2021) identified that common gene signatures and pathways between COVID-19 and chronic kidney disease (CKD) could be therapeutic targets in COVID-19 patients with CKD as a comorbidity using the RNA sequencing (RNA-Seq) transcriptomic dataset of peripheral blood mononuclear cells (PBMCs) infected with SARS-CoV-2.

Autophagy refers to the process of sealing a part of the cytoplasm in the double-membrane autophagosome and delivering it to the lysosome for degradation; it is an essential cellular mechanism to cope with various stress conditions (such as starvation, energy deprivation, and pathogen invasion) and maintain a steady-state balance (Feng et al., 2014). As a monitoring mechanism, autophagy is also involved in resisting the foreign invasion of viruses. In response to viral infection, the autophagic activity is activated by host cells through virus-encoded activators, cellular stresses provoked by infection, and sensing of viral constituents mediated by Toll-like receptors (TLRs) (Viret et al., 2018). As a defense mechanism during viral infection, the autophagic activity could deliver the virus or viral protein to the lysosome for degradation, transport viral nucleic acids and antigens to endolysosomal compartments for innate and adaptive immune responses, and regulate virus-induced cell death (Levine et al., 2011). SARS-CoV-2 is an enveloped, approximately 30 kb single-stranded RNA β-coronavirus (Wu et al., 2020). Several studies have demonstrated that infection with SARS-CoV-2 may be associated with autophagy. Miao et al. (2020) demonstrated that SARS-CoV-2 virus infection would block autophagy, resulting in the accumulation of autophagosomes and causing late endosomal sequestration of the homotypic fusion and protein sorting (HOPS) component VPS39. In contrast, Hui et al. (2021) provided evidence that SARS-CoV-2 promotes autophagy to suppress type I interferon response.

Here, to explore the biological characteristics and progression in patients with severe COVID-19 as opposed to those with mild COVID-19, we first pre-processed raw data on GSE167930 and screened specific DEGs for severe COVID-19. Gene Ontology analysis of these DEGs was performed to gain knowledge regarding their biological processes. Subsequently, we examined the most significant term and performed protein–protein interaction (PPI) network analysis to extract the hub genes regulating autophagy. Furthermore, we identified transcription factors (TFs) and microRNAs (miRNAs) at regulatory checkpoints using these hub genes. We then analyzed the protein–chemical interaction network, to determine potent chemical agents. Finally, to determine the cell type-specific expression of these genes, we performed single-cell RNA-Seq analysis using an online dataset. The sequential workflow of the processes in this study is shown in **Figure 1**.

MATERIALS AND METHODS

Sample Collection, Data Processing, and Differential Expression Analysis

The study cohort comprised peripheral blood mononuclear cells from four mild COVID-19 patients, four severe COVID-19 patients, and four healthy controls. All samples were subjected to RNA-Seq analysis, and the results could be obtained from the GEO database of the National Center for Biotechnology Information (NCBI). The GEO accession ID of the dataset was GSE167930, which was already deposited for early published article by our team (Zhou et al., 2021). The limma R package was used for RNA-Seq to identify significant DEGs (the cut-off value of fold change >2 and fold change <0.5; p-value < 0.05). To screen specific DEGs for severe COVID-19, we set two gene clusters. In cluster 1, DEGs were significantly expressed in severe COVID-19 patients compared with healthy controls. In cluster 2, DEGs were significantly expressed in mild COVID-19 patients compared

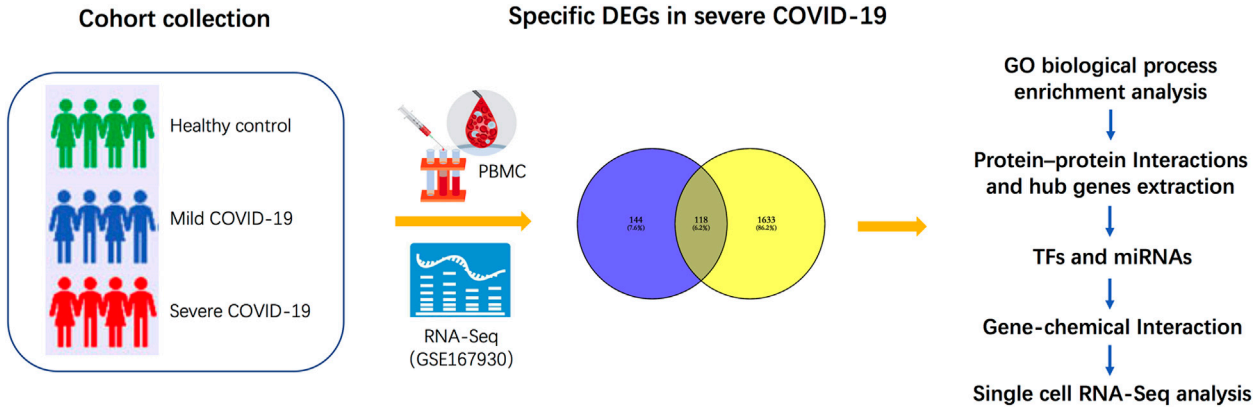


FIGURE 1 | Schematic illustration of the overall general workflow of this study.

with severe COVID-19 patients. The Venn diagram of cluster 1 and cluster 2 was used to specifically distinguish DEGs associated with severe COVID-19 patients. This differential expression analysis was performed and figures were obtained using the SangerBox tools, a free online platform for data analysis (http://vip.sangerbox.com/).

Gene Ontology Enrichment Analysis

Gene Ontology (http://geneontology.org/) stores a database of gene annotations that participate in biological processes; it calculates the probability of obtaining at least as many genes with the observed annotations. DAVID (https://david.ncifcrf.gov) was used as a data source for Gene Ontology enrichment analysis of the 144 DEGs specifically expressed in severe COVID-19, and the significant enrichments were filtered based on p-value < 0.05 and FDR (q-value) < 0.05. The enrichment analysis was performed, and figures mentioned earlier were obtained using the SangerBox tools (http://vip.sangerbox.com/).

Protein–Protein Interaction Network Analysis and Hub Gene Cluster Identification

The genes involved in the crucial biological process of severe COVID-19 were included in the STRING database (https://string-db.org/) (version 11.0) to construct a PPI network. Cytoscape v.3.7.1 was then used for the visual presentation of the results from STRING. Hub gene cluster analysis was conducted using the Molecular Complex Detection (MCODE) plugin.

Transcriptional and Post-transcriptional Network Analysis

The hub genes involved in the crucial biological processes of severe COVID-19 were used to recognize its TF gene and gene-miRNA network using the JASPAR database and TarBase database v8.0 on the NetworkAnalyst platform, respectively. Subsequently, Cytoscape v.3.7.1 was used to obtain a visual presentation of the TF gene and gene-miRNA interaction network.

Gene–Chemical Interaction Network Analysis

The Comparative Toxicogenomics Database in the NetworkAnalyst tool was further used to recognize the relationship of potential chemical agents and hub genes in the crucial biological process of severe COVID-19. For a visual presentation of the gene–chemical interaction network, Cytoscape v.3.7.1 was used.

Single Cell RNA-Seq Analysis

To indicate the cell type-specific expression of hub genes involved in autophagy in this study, we performed single-cell RNA-Seq analysis on a free online database platform: COVID-19 Cell Atlas Data Mining Site (http://www.covidcellatlas.com/) (Unterman et al., 2022). We set the comparison as COVID-19 Stable versus Progressive on the website, where "Stable" refers to patients hospitalized in internal medicine wards who eventually recovered and were discharged, that is, mild patients in our study, and "Progressive" refers to severe patients who required admission to the ICU and eventually succumbed to the disease. To show the identified cell types that express the gene, the UMAP Explorer was used for plotting the gene expression. We then imported the five most specifically expressed cell types into interactive connectome to explore the intercellular ligand–receptor pair interactions. All the figures were obtained from this online platform.

RESULTS

Identification of Differentially Expressed Genes Specific for Severe COVID-19 Patients

To identify the DEGs specific for severe COVID-19, we first compared genes expressed in severe COVID-19 to those expressed in healthy controls and set these 262 DEGs in cluster 1 (the cut-off value of fold change >2 and fold

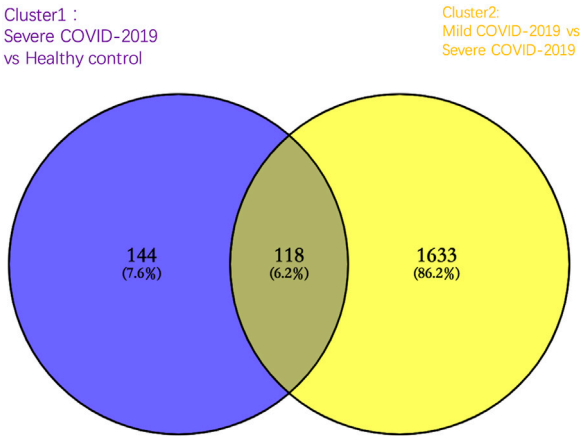


FIGURE 2 | Venn diagram of significantly altered genes in cluster 1 [severe COVID-19 group compared with healthy control group (purple)] and cluster 2 [mild COVID-19 group compared with the severe COVID-19 group (yellow)]. A total of 144 genes were significantly altered specifically in severe COVID-19 patients.

change <0.5; *p*-value < 0.05). Following this, in cluster 2, a total of 1751 DEGs were significantly changed in mild COVID-19 patients compared with severe COVID-19 patients. As shown in the Venn diagram of cluster 1 and cluster 2, there were 144 DEGs specifically expressed in severe COVID-19, and these 144 DEGs were employed to accomplish further determination of biological process enrichment analysis (**Figure 2**).

Gene Ontology Enrichment Analysis

To gain insight into the regulation of genes and the transmission of signals that occur during the progression of severe COVID-19, we performed Gene Ontology enrichment analysis. In biological process enrichment with DAVID at *p*-value < 0.05 and FDR (q-value) < 0.05, we found a total of 11 GO terms enriched significantly for the specific 144 DEGs in severe COVID-19 patients (**Figure 3**). **Table 1** presents the 11 GO terms based on the number of genes included. The biological process terms "Autophagy" (GO:0006914, *p*-value 0.0001, FDR 0.0327) and "Process utilizing autophagic mechanism" (GO:0006919, *p*-value 0.0001, FDR 0.0327) were considered the crucial biological process as they contained maximum 12 genes including *ATM*, *CHMP6*, *EP300*, *RIPK2*, *ATP6V0E1*, *VPS37B*, *ATP6V1E2*, *PLEKHM1*, *ZC3H12A*, *STAM*, *MVB12A,* and *RALB*. In **Figure 4A**, the stacked bar chart visualized these gene expressions in patients with severe COVID-19, mild COVID-19, and healthy control. Also, a total of six genes (*EP300*, *RIPK2*, *STK4*, *TRAF2*, *RFNG,* and *RALB*) were contained in "Positive regulation of protein binding" (GO:0032092, *p*-value 0.0001, FDR 0.0294). The GO term "Virion assembly" (GO:0019068, *p*-value 0.0001, FDR 0.0133) contained *TBC1D20*, *CHMP6*, *RPS27A*, *VPS37B,* and *MVB12A*. The biological process terms of "Multi-organism transport" (GO:0044766, *p*-value 0.0001, FDR 0.0365) and "Multi-organism localization" (GO:0050706, *p*-value 0.0001, FDR 0.0365) contained five genes including *THOC7*, *RPS27A*, *VPS37B*, *XPO1*, and *MVB12A*. A total of four genes (*RIPK2*, *PML*, *ZC3H12A,* and *PANX1*) were contained in "Regulation of interleukin-1 beta secretion" (GO:0050706, *p*-value 0.0002, FDR 0.0474). Then, four genes including *UBE2J2*, *LEPROT*, *RHOU,* and *HSPA1L* consist in "Regulation of protein targeting to mitochondrion" (GO:1903214, *p*-value 0.0002, FDR 0.0474) and "Positive regulation of protein targeting to mitochondrion" (GO:1903955, *p*-value 0.0001, FDR 0.0294). Furthermore, four genes (*CHMP6*, *VPS37B*, *STAM,* and *MVB12A*) were contained in terms of "Multivesicular body assembly" (GO:0036258, *p*-value 0.0001, FDR 0.0294) and "Multivesicular body organization" (GO:0036257, *p*-value 0.0001, FDR 0.0294).

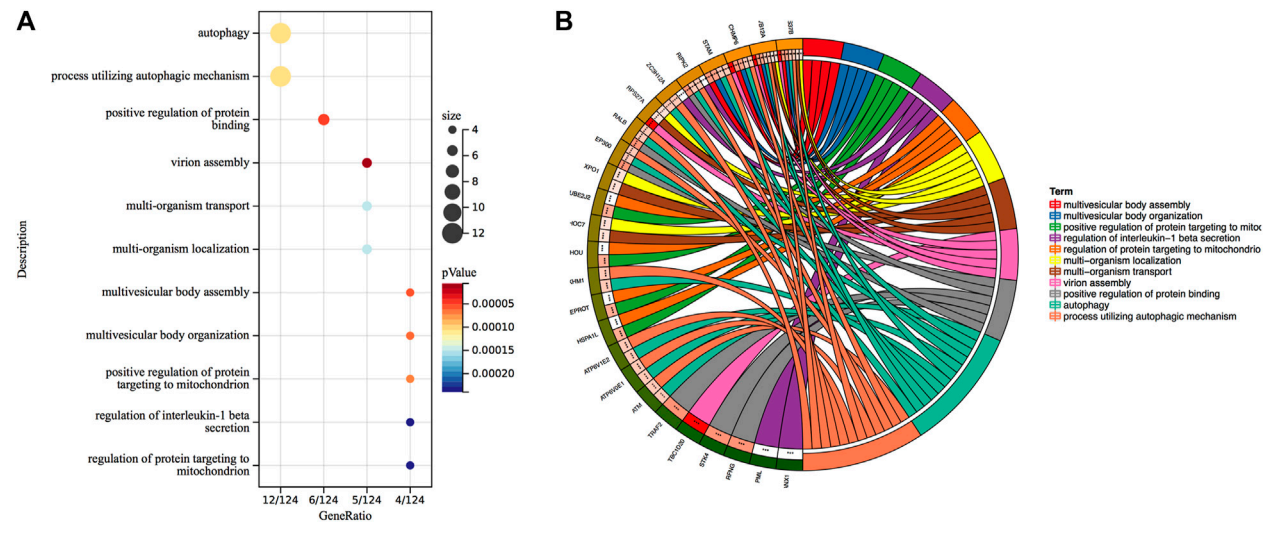


FIGURE 3 | (A) Gene Ontology biological process enrichment analysis of 144 specific DEGs in severe COVID-19. Overall, 11 terms are presented. The size of the circle indicates the relative contribution of the genes to the activity of the term. Colors show the *p*-value. **(B)** Circles show genes contained in each term.

TABLE 1 | Gene Ontology biological process enrichment analysis of 144 specific DEGs in severe COVID-19 cases.

ID	Term	Size	Official Gene Symbol	*p*-value	FDR
GO: 0006914	Autophagy	12	*ATM, CHMP6, EP300, RIPK2, ATP6V0E1, VPS37B, ATP6V1E2, PLEKHM1, ZC3H12A, STAM, MVB12A, RALB*	0.0001	0.0327
GO: 0061919	Process utilizing the autophagic mechanism	12	*ATM, CHMP6, EP300, RIPK2, ATP6V0E1, VPS37B, ATP6V1E2, PLEKHM1, ZC3H12A, STAM, MVB12A, RALB*	0.0001	0.0327
GO: 0032092	Positive regulation of protein binding	6	*EP300, RIPK2, STK4, TRAF2, RFNG, RALB*	0.0001	0.0294
GO: 0019068	Virion assembly	5	*TBC1D20, CHMP6, RPS27A, VPS37B, MVB12A*	0.0001	0.0133
GO: 0044766	Multi-organism transport	5	*THOC7, RPS27A, VPS37B, XPO1, MVB12A*	0.0001	0.0365
GO: 1902579	Multi-organism localization	5	*THOC7, RPS27A, VPS37B, XPO1, MVB12A*	0.0001	0.0365
GO: 0050706	Regulation of interleukin-1 beta secretion	4	*RIPK2, PML, ZC3H12A, PANX1*	0.0002	0.0474
GO: 1903214	Regulation of protein targeting the mitochondrion	4	*UBE2J2, LEPROT, RHOU, HSPA1L*	0.0002	0.0474
GO: 1903955	Positive regulation of protein targeting the mitochondrion	4	*UBE2J2, LEPROT, RHOU, HSPA1L*	0.0001	0.0294
GO: 0036258	Multivesicular body assembly	4	*CHMP6, VPS37B, STAM, MVB12A*	0.0001	0.0294
GO: 0036257	Multivesicular body organization	4	*CHMP6, VPS37B, STAM, MVB12A*	0.0001	0.0294

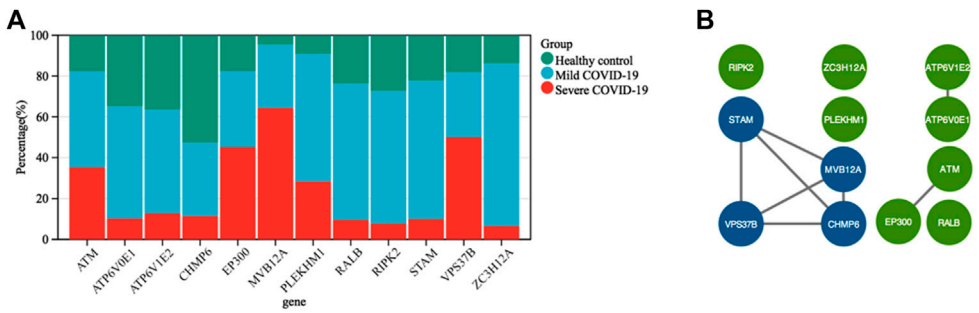


FIGURE 4 | (A) Stacked bar chart shows the expression value percentage of the 12 genes involved in the autophagy process in patients with severe COVID-19, mild COVID-19, and healthy controls. **(B)** Protein–protein interaction network was created using STRING and visualized in Cytoscape. The core cluster analyzed by MCODE is indicated with blue color.

Protein–Protein Interaction and Identification of Hub Genes Involved in Autophagy

To further acquire the core genes associated with autophagy in severe COVID-19, we conducted a PPI network of the total 12 genes involved in GO terms "Autophagy" and "Process utilizing autophagic mechanism" using STRING. In this manner, 12 nodes and eight edges were obtained with a local clustering coefficient of 0.667 and a PPI enrichment *p*-value of 0.00032 (**Figure 4B**). The data file was calculated by MCODE, a novel Cytoscape plugin to identify significant gene clusters. Subsequently, we obtained only one gene cluster, which consisted of four nodes (*MVB12A*, *CHMP6*, *STAM*, and *VPS37B*) and six edges. Therefore, *MVB12A*, *CHMP6*, *STAM*, and *VPS37B* were regarded as hub genes regulating autophagy in severe COVID-19.

Determination of Regulatory Signatures

The selected hub genes as mentioned earlier (*MVB12A*, *CHMP6*, *STAM*, and *VPS37B*) were evaluated with TF gene and gene-miRNA interaction network analysis to detect transcriptional signatures and post-transcriptional regulatory signatures. **Figure 5** shows the TF gene interaction network drawn by Cytoscape. The transcription factors, namely, SRF, YY1, CREB1, PPARG, and NFIC, linked with *VPS37B* and *MVB12A* and GATA2 connected with *MVB12A* and *STAM*. The gene-miRNA interaction network has 66 nodes and 72 edges (**Figure 6**). Among these miRNAs,

FIGURE 5 | Regulatory interaction network of TF gene was identified using the NetworkAnalyst tool. Herein, the circle nodes are genes (blue); the square nodes are TFs (green); TFs targeting more than two genes simultaneously are shown in orange.

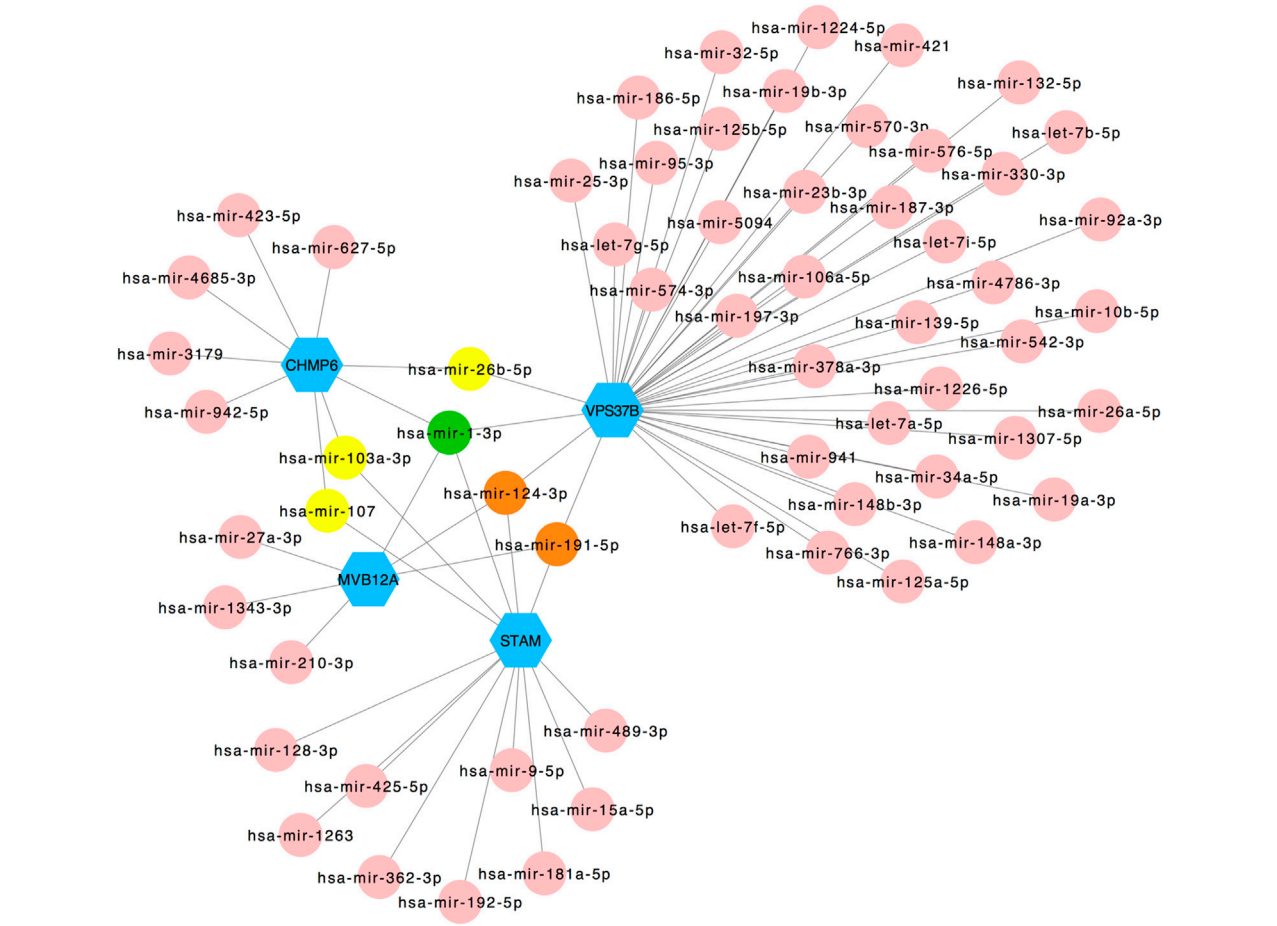


FIGURE 6 | Regulatory interaction network of gene-miRNA was identified using the NetworkAnalyst tool. Genes are shown in blue; miRNAs are shown in pink; miRNAs targeting two genes are shown in yellow; miRNAs targeting three genes are shown in orange; miRNAs targeting four genes are shown in green.

hsa-mir-26b-5p connected with *CHMP6* and *VPS37B*, and hsa-mir-103a-3p and hsa-mir-107 were associated with *CHMP6* and *STAM*. Subsequently, hsa-mir-124-3p and hsa-mir-191-5p were commonly linked with *VPS37B*, *CHMP6*, and *STAM*. Furthermore, we considered hsa-mir-1-3p to be the most pivotal miRNA as it was the common post-transcriptional factor for all the four hub genes.

Construction of the Gene–Chemical Interaction Network and Identification of Potent Chemical Agents

We used the four hub genes to analyze their interactions with different chemical agents in addition to using them for identifying potential antiviral agents associated with autophagy. Using the

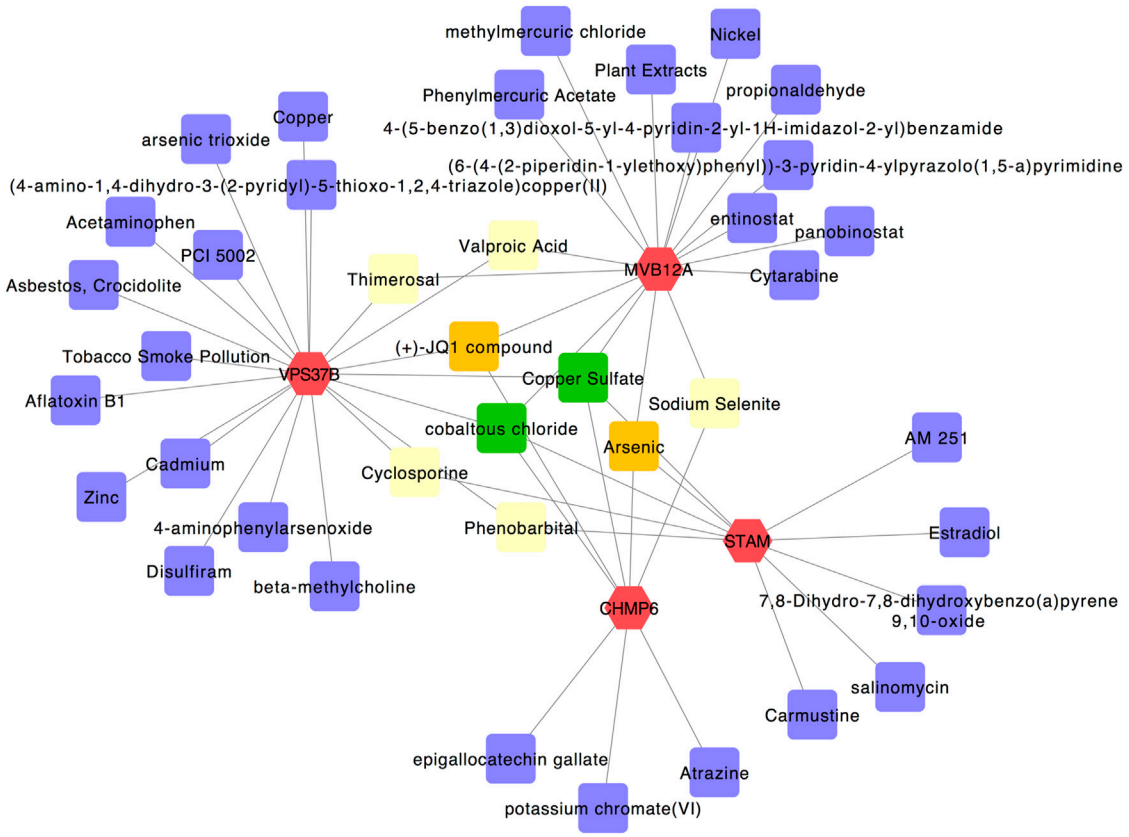


FIGURE 7 | Gene–chemical interaction network. Genes are colored in red; chemical agents are colored in purple; chemical agents targeting two genes simultaneously are shown in yellow; chemical agents targeting three genes simultaneously are shown in orange; chemical agents targeting all four genes simultaneously are shown in green.

Comparative Toxicogenomics Database in the NetworkAnalyst tool, 40 chemical agents were predicted (**Figure 7**). Among these, thimerosal and valproic acid were linked with *VPS37B* and *MVB12A*, cyclosporine and phenobarbital were connected with *VPS37B* and *CHMP6*, and sodium selenite was associated with *MVB12A* and *CHMP6*. In addition, (+)-JQ1 compound was connected with *VPS37B*, *CHMP6*, and *MVB12A*, and arsenic was linked with *STAM*, *CHMP6*, and *MVB12A*. Most importantly, copper sulfate and cobaltous chloride were associated with all the four hub genes and could be considered to be the most relevant potent chemical agents.

Single-Cell RNA-Seq Analysis

An online single-cell RNA-Seq platform was used to assess the cell type-specific expressions of hub genes (*MVB12A*, *CHMP6*, *STAM*, and *VPS37B*). *MVB12A* was highly expressed in classical monocytes and effector T cells (**Figure 8B**). CHMP6 was highly enriched in plates and effector T cells (**Figure 8C**). *STAM* was highly expressed in Tregs and dying T & NK cells (**Figure 8D**). *VPS37B* was highly expressed in dying T & NK cells and effector T cells (**Figure 8E**). Subsequently, the interactive connectome tool in this platform was used to explore the intercellular ligand–receptor pair interactions between classical monocytes, effector T cells, plates, Tregs, and dying T & NK cells. **Figure 8F** shows that the ligands of effector T cells, plates, Tregs, and dying T & NK cells among PBMCs of severe COVID-19 patients are targeted to match the receptors of classical monocytes compared with the ligands of those cells of mild COVID-19 patients, and classical monocytes also secrete ligands to induce a cellular response through cognate receptors.

DISCUSSION

The COVID-19 pandemic is a major threat to a safe and healthy living environment of human beings and has resulted in more than 4 million deaths worldwide. It is a type of pneumonia, an infection with a virus named SARS-CoV-2 in the lungs. However, the condition is more complex. Severe COVID-19 strains display more aggressive symptoms, consequently resulting in a high mortality rate, especially the delta variant initially discovered in India in December 2020. Furthermore, the Omicron variant was discovered in South Africa a few months ago. Mild COVID-19 patients show moderate or even no symptoms. Therefore, identifying the characteristics of severe COVID-19 by comparison with mild COVID-19 would be more helpful for

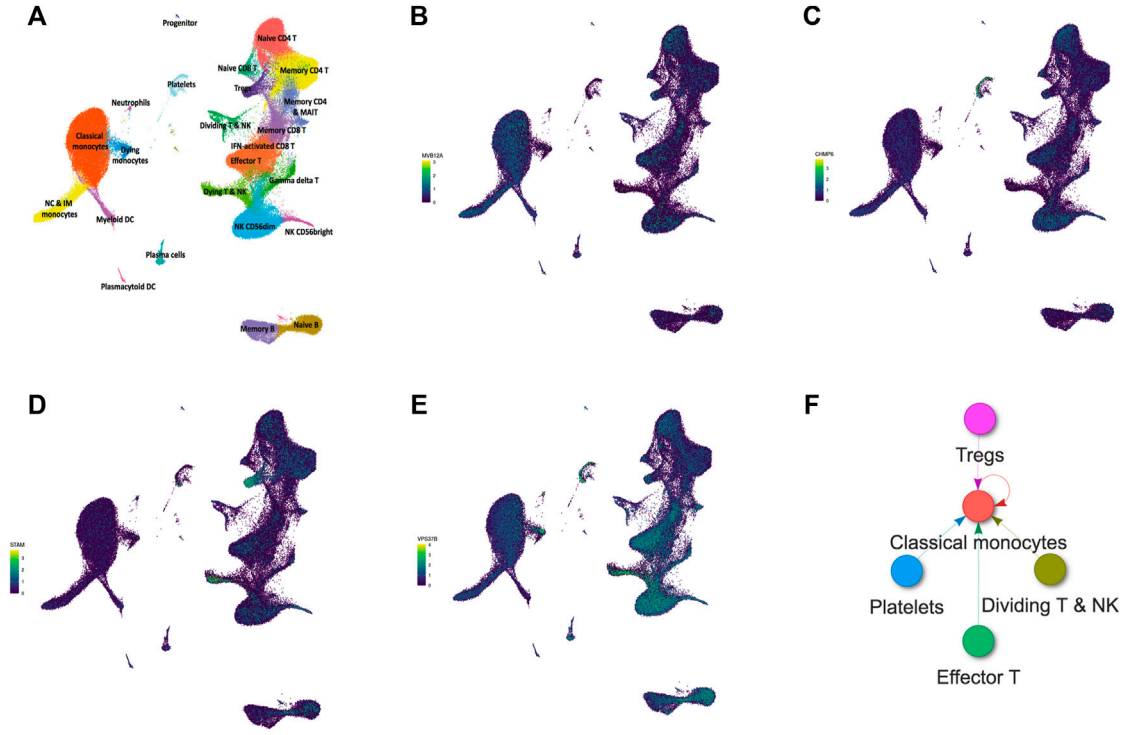


FIGURE 8 | Single-cell RNA-seq analysis of PBMCs from COVID-19 patients on an online database platform. **(A)** UMAP embedding of single-cell transcriptomes from 153,554 cells color-coded for the indicated cell type. **(B–E)** UMAP plots showing the expressions of MVB12A, CHMP6, STAM, and VPS37B in PBMCs. **(F)** Intercellular ligand–receptor pair interactions of PBMCs from COVID-19 progressive patients.

developing potential biomarkers and even new therapeutic targets. In this study, we first identified 144 specific DEGs in severe COVID-19 cases and performed the GO biological process enrichment analysis to acquire insight into the biological characteristics. Following this, we suggested that autophagy plays a key role in severe COVID-19, which corresponded to provide evidence in several studies. SARS-CoV-2 virus could block autophagy by infection or expression of ORF3a to sequestrate the HOPS component *VPS39* and impaired the assembly of the STX17-SNAP29-VAMP8 SNARE complex (Miao et al., 2020). Hui et al. (2021) reported that SARS-CoV-2 M protein induced mitophagy to block the downstream innate immunity signaling for inhibiting the type I IFN response. Thus, autophagy may crucially contribute to the SARS-CoV-2 viral lifecycle.

Assessment of the PPI network is considered a key pattern of protein affiliation and interaction. A total of 12 genes in the biological process of autophagy were involved in PPIs and the determination of hub genes. Here, four hub genes including *MVB12A*, *CHMP6*, *STAM*, and *VPS37B* were considered to be involved in the core regulation of autophagy in severe COVID-19 cases. Among these, Multivesicular Body Subunit 12A (*MVB12A*) is a component of the endosomal sorting required for transport I (ESCRT-I) complex. Its depletion and overexpression inhibit HIV-1 infectivity by inducing aberrant virion morphologies and altering viral Gag protein processing (Morita et al., 2007). The charged multivesicular body protein 6 (*CHMP6*) gene is the core component of endosomal sorting required for the transport III (ESCRT-III) complex, which is considered essential for viral-like particles (VLPs) and virion release (Kumar et al., 2016). The signal transducing adapter molecule (*STAM*) gene forms the endosomal sorting complex required for transport-0 (ESCRT-0), and vacuolar protein sorting-associated protein 37B (*VPS37B*) is a component of the ESCRT-I complex. All the four genes are components of the ESCRT complex, and several viruses take advantage of the ESCRT system for proliferation, budding, and transmission in infected cells (Ju et al., 2021; Meng et al., 2021). Consequently, we supposed the infection of SARS-CoV-2 may be allied to the ESCRT system.

Subsequently, the hub genes specialized for autophagy in severe COVID-19 were selected to predict their potential function at transcriptional and post-transcriptional levels. A number of transcription factors were detected. The serum response factor (SRF) has been demonstrated to modulate asymmetrical cardiac myocyte hypertrophy by constituting an epigenomic switch balancing the growth of adult ventricular myocytes in width versus length (Li et al., 2020). The transcription factor Yin-Yang 1 (YY1) played an essential role in apoptosis and angiogenesis, and its cardioprotective effects were associated with T helper 2 cytokine production and M2 macrophage polarization (Huang et al., 2021). In addition, cAMP responsive element binding protein 1 (CREB1) and its target genes identified by the recombinant canarypox vector ALVAC + Alum could augment immunogenicity and reduce

the HIV-1 infection rate (Tomalka et al., 2021). Furthermore, NFIC is related to digestive system carcinoma (Fang et al., 2021) (Liang et al., 2021) and regulates renal inflammation and renal fibrosis in patients with diabetic nephropathy (Zhang et al., 2021). Eventually, hepatitis B virus x protein can interact with GATA binding protein 2 (GATA2) to influence the activity of the ST2 promoter. We detected several significant miRNAs as latent post-transcriptional factors. We believed hsa-mir-1-3p to be the most pivotal miRNA in the process of autophagy in severe COVID-19 as it was targeted by four hub genes, and hsa-mir-1-3p has been identified to have a relationship with COVID-19 in several studies (Sardar et al., 2020; Sarma et al., 2020). It can inhibit influenza A virus replication by targeting the supportive host factor ATP6V1A (Peng et al., 2018). In addition, hsa-mir-1-3p was related to tumors such as endometrial cancer, (Czerwiński et al., 2021), metastatic prostate cancer (Mukherjee and Sudandiradoss, 2021), and breast cancer (Yan et al., 2021). In addition, hsa-mir-124-3p and hsa-mir-191-5p were commonly linked with *VPS37B*, *CHMP6*, and *STAM*.

Hsa-miR-124-3p was considered a potential candidate for treating COVID-19 (Prasad et al., 2021) and regulating ACE2 networks (Wicik et al., 2020). Then, hsa-miR-191-5p showed an inhibitory effect on HIV-1 replication (Zheng et al., 2021); it is associated with cervical lesions and can serve as a non-invasive biomarker (Ning et al., 2021). Next, the chemical agents that may target the common hub genes have been detected using the Comparative Toxicogenomics Database. Among significant chemical agents, copper sulfate has been proposed as a locally applied fungicide, bactericide, and astringent in medical practice (https://go.drugbank.com/drugs/DB06778). It may induce pulmonary fibrosis through EMT activation induced by the TGF-β1/Smad pathway and MAPK pathways (Guo et al., 2021). Moreover, copper sulfate has also been identified as a potential chemical agent in pathogenetic profiling of COVID-19 (Nain et al., 2021). Furthermore, cobaltous chloride is a chemical agent that has been found to have an application in certain insecticides and fungicides (https://www.britannica.com/science/cobaltous-chloride). As the evidence of potent chemical agents in severe COVID-19 is indirective, their roles need to be further studied to be confirmed. However, although these critical factors lack experimental verification, the correlations to autophagy suggest that they play a role in the prognosis of severe COVID-19.

Eventually, for determining whether the localization of these genes regulates autophagy, we assessed the cell type-specific expressions of *MVB12A*, *CHMP6*, *STAM*, and *VPS37B* using an online single cell RNA-Seq database platform. The results showed that classical monocytes, effector T cells, plates, Tregs, and dying T & NK cells play roles in autophagy. Furthermore, classical monocytes exhibit a central role among the five cell types that constitute cellular communication because all ligands match their receptors. Monocytes are phagocytic innate immune cells in blood circulation and depending on their respective expressions of CD14 and CD16 are traditionally divided into classical monocytes ($CD14^{++}CD16^{-}$), non-classical monocytes ($CD14^{+}CD16^{++}$), and intermediate monocytes ($CD14^{++}CD16^{+}$). In acute patients with severe COVID-19, the number of non-classical and intermediate monocytes is found to be significantly reduced, whereas circulating classical monocytes display clear signs of activation (Knoll et al., 2021). Vanderbeke et al. (2021)have demonstrated that classical pro-inflammatory monocytes (based on the expressions of S100A8, S100A9, and S100A12 markers) dominate COVID-19 immunopathology in most critical cases. The results also indicated that classical monocytes were the primary source of major COVID-19-mediating cytokines, including the monocyte chemoattractant CCL2 and its receptor CCR2, the neutrophil chemoattractant CXCL8, and TNF-α (Vanderbeke et al., 2021). In addition, the expression level of the monocyte chemoattractant CCR2, which is a classical monocyte, was higher than that of non-classical monocytes, and anti-CCR2 treatment improved the course of the disease in preclinical trials (Channappanavar et al., 2016). Thus, these results demonstrate a correlation between classical monocytes and COVID-19, which could contribute to the design of novel therapeutics for this pandemic. However, because the samples used in this experiment were already used before, our conclusions may be limited by direct experimental validation. There are also few studies reported on the relationship between COVID-19 and autophagy. To the best of our knowledge, this is the first study to propose that *MVB12A*, *CHMP6*, *STAM*, and *VPS37B* are crucial genes associated with autophagy of PBMCs in patients with severe COVID-19 as opposed to those with a mild condition. Classical monocytes may play a central role in this disease; accordingly, subsequent studies should deeply explore the insight into the relationship between autophagy and classical monocytes in severe COVID-19.

CONCLUSION

The present study highlights the potential specific pathogenic processes in severe COVID-19 relative to mild COVID-19 and identifies hub genes, regulatory components, and chemical agents that may help develop novel and efficacious clinical therapeutic targets. We first identified 144 specific DEGs in severe COVID-19 patients. Subsequently, using these DEGs, we identified autophagy as a critical biological process. Next, based on the PPI network, we identified the most significant gene cluster involving the hub genes of *MVB12A*, *CHMP6*, *STAM*, and *VPS37B*. Consequently, we determined that the most pivotal miRNA hsa-miR-1-3p may play a role at the regulatory level. Copper sulfate and cobaltous chloride were considered relevant potent chemical agents. Eventually, we reported that classical monocytes may play a central role in genes regulating autophagy in severe COVID-19 cases compared with mild ones. Overall, our findings will shed light on the knowledge regarding biological characteristics of severe COVID-19 cases, as well as help find novel therapeutic strategies enabling us to achieve breakthroughs in the current pandemic.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Conceptualization, JH and MZ; methodology, JH; software, JH; validation, YZ and MZ; formal analysis, MZ; investigation, JH; resources, MZ; data curation, MZ; writing—original draft preparation, JH; writing—review and editing, YW, XZ, WL, and MZ; visualization, JH; supervision, MZ; project administration, JH; funding acquisition, MZ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the Natural Science Foundation of China (NSFC) (Nos. 81673105 and 31971167).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.908826/full#supplementary-material

REFERENCES

Auwul, M. R., Zhang, C., Rahman, M. R., Shahjaman, M., Alyami, S. A., and Moni, M. A. (2021). Network-Based Transcriptomic Analysis Identifies the Genetic Effect of COVID-19 to Chronic Kidney Disease Patients: A Bioinformatics Approach. *Saudi J. Biol. Sci.* 28, 5647–5656. doi:10.1016/j.sjbs.2021.06.015

Berlin, D. A., Gulick, R. M., and Martinez, F. J. (2020). Severe Covid-19. *N. Engl. J. Med.* 383 (25), 2451–2460. doi:10.1056/NEJMcp2009575

Channappanavar, R., Fehr, A. R., Vijay, R., Mack, M., Zhao, J., Meyerholz, D. K., et al. (2016). Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. *Cell host microbe* 19 (2), 181–193. doi:10.1016/j.chom.2016.01.007

Czerwiński, M., Bednarska-Czerwińska, A., Ordon, P., Gradzik, M., Oplawski, M., Boroń, D., et al. (2021). Variances in the Expression of mRNAs and miRNAs Related to the Histaminergic System in Endometrioid Endometrial Cancer. *Biomedicines* 9 (11), 1535. doi:10.3390/biomedicines9111535

Fang, G., Fan, J., Ding, Z., Li, R., Lin, K., Fu, J., et al. (2021). Prognostic and Predictive Value of Transcription Factors Panel for Digestive System Carcinoma. *Front. Oncol.* 11, 670129. doi:10.3389/fonc.2021.670129

Feng, Y., He, D., Yao, Z., and Klionsky, D. J. (2014). The Machinery of Macroautophagy. *Cell Res.* 24 (1), 24–41. doi:10.1038/cr.2013.168

Guo, H., Jian, Z., Liu, H., Cui, H., Deng, H., Fang, J., et al. (2021). TGF-β1-induced EMT Activation via Both Smad-dependent and MAPK Signaling Pathways in Cu-Induced Pulmonary Fibrosis. *Toxicol. Appl. Pharmacol.* 418, 115500. doi:10.1016/j.taap.2021.115500

Huang, Y., Li, L., Chen, H., Liao, Q., Yang, X., Yang, D., et al. (2021). The Protective Role of Yin-Yang 1 in Cardiac Injury and Remodeling After Myocardial Infarction. *J. Am. Heart Assoc.* 10, e021895. doi:10.1161/jaha.121.021895

Hui, X., Zhang, L., Cao, L., Huang, K., Zhao, Y., Zhang, Y., et al. (2021). SARS-CoV-2 Promote Autophagy to Suppress Type I Interferon Response. *Sig Transduct. Target Ther.* 6 (1), 180. doi:10.1038/s41392-021-00574-8

Iype, E., and Gulati, S. (2020). Understanding the Asymmetric Spread and Case Fatality Rate (CFR) for COVID-19 Among Countries. *medRxiv* 2020, 20073791. doi:10.1101/2020.04.21.20073791

Ju, Y., Bai, H., Ren, L., and Zhang, L. (2021). The Role of Exosome and the ESCRT Pathway on Enveloped Virus Infection. *Int. J. Mol. Sci.* 22 (16), 9060. doi:10.3390/ijms22169060

Knoll, R., Schultze, J. L., and Schulte-Schrepping, J. (2021). Monocytes and Macrophages in COVID-19. *Front. Immunol.* 12, 720109. doi:10.3389/fimmu.2021.720109

Kumar, B., Dutta, D., Iqbal, J., Ansari, M. A., Roy, A., Chikoti, L., et al. (2016). ESCRT-I Protein Tsg101 Plays a Role in the Post-macropinocytic Trafficking and Infection of Endothelial Cells by Kaposi's Sarcoma-Associated Herpesvirus. *PLoS Pathog.* 12 (10), e1005960. doi:10.1371/journal.ppat.1005960

Levine, B., Mizushima, N., and Virgin, H. W. (2011). Autophagy in Immunity and Inflammation. *Nature* 469 (7330), 323–335. doi:10.1038/nature09782

Li, J., Tan, Y., Passariello, C. L., Martinez, E. C., Kritzer, M. D., Li, X., et al. (2020). Signalosome-Regulated Serum Response Factor Phosphorylation Determining Myocyte Growth in Width Versus Length as a Therapeutic Target for Heart Failure. *Circulation* 142 (22), 2138–2154. doi:10.1161/circulationaha.119.044805

Liang, X., Zhang, Z., Wang, L., Zhang, S., Ren, L., Li, S., et al. (2021). Mechanism of Methyltransferase like 3 in Epithelial-Mesenchymal Transition Process, Invasion, and Metastasis in Esophageal Cancer. *Bioengineered* 12, 10023–10036. doi:10.1080/21655979.2021.1994721

Mahmud, S. M. H., Al-Mustanjid, M., Akter, F., Rahman, M. S., Ahmed, K., Rahman, M. H., et al. (2021). Bioinformatics and System Biology Approach to Identify the Influences of SARS-CoV-2 Infections to Idiopathic Pulmonary Fibrosis and Chronic Obstructive Pulmonary Disease Patients. *Briefings Bioinforma.* 22, bbab115. doi:10.1093/bib/bbab115

Meng, B., Vallejo Ramirez, P. P., Scherer, K. M., Bruggeman, E., Kenyon, J. C., Kaminski, C. F., et al. (2021). EAP45 Association with Budding HIV -1: Kinetics and Domain Requirements. *Traffic* 22, 439–453. doi:10.1111/tra.12820

Miao, G., Zhao, H., Li, Y., Ji, M., Chen, Y., Shi, Y., et al. (2020). ORF3a of the COVID-19 Virus SARS-CoV-2 Blocks HOPS Complex-Mediated Assembly of the SNARE Complex Required for Autolysosome Formation. *Dev. Cell* 56, 427–442 e5. doi:10.1016/j.devcel.2020.12.010

Morita, E., Sandrin, V., Alam, S. L., Eckert, D. M., Gygi, S. P., and Sundquist, W. I. (2007). Identification of Human MVB12 Proteins as ESCRT-I Subunits that Function in HIV Budding. *Cell host microbe* 2 (1), 41–53. doi:10.1016/j.chom.2007.06.003

Mukherjee, S., and Sudandiradoss, C. (2021). Transcriptomic Analysis of Castration, Chemo-Resistant and Metastatic Prostate Cancer Elucidates Complex Genetic Crosstalk Leading to Disease Progression. *Funct. Integr. Genomics* 21, 451–472. doi:10.1007/s10142-021-00789-6

Nain, Z., Rana, H. K., Liò, P., Islam, S. M. S., Summers, M. A., and Moni, M. A. (2021). Pathogenetic Profiling of COVID-19 and SARS-like Viruses. *Briefings Bioinforma.* 22 (2), 1175–1196. doi:10.1093/bib/bbaa173

Ning, R., Meng, S., Wang, L., Jia, Y., Tang, F., Sun, H., et al. (2021). 6 Circulating miRNAs Can Be Used as Non-invasive Biomarkers for the Detection of Cervical Lesions. *J. Cancer* 12 (17), 5106–5113. doi:10.7150/jca.51141

Peng, S., Wang, J., Wei, S., Li, C., Zhou, K., Hu, J., et al. (2018). Endogenous Cellular MicroRNAs Mediate Antiviral Defense against Influenza A Virus. *Mol. Ther. - Nucleic Acids* 10, 361–375. doi:10.1016/j.omtn.2017.12.016

Prasad, K., Alasmari, A. F., Ali, N., Khan, R., Alghamdi, A., and Kumar, V. (2021). Insights into the SARS-CoV-2-Mediated Alteration in the Stress Granule Protein Regulatory Networks in Humans. *Pathogens* 10 (11), 1459. doi:10.3390/pathogens10111459

Sardar, R., Satish, D., and Gupta, D. (2020). Identification of Novel SARS-CoV-2 Drug Targets by Host MicroRNAs and Transcription Factors Co-regulatory Interaction Network Analysis. *Front. Genet.* 11, 571274. doi:10.3389/fgene.2020.571274

Sarma, A., Phukan, H., Halder, N., and Madanan, M. G. (2020). An In-Silico Approach to Study the Possible Interactions of miRNA between Human and

SARS-CoV2. *Comput. Biol. Chem.* 88, 107352. doi:10.1016/j.compbiolchem.2020.107352

Tomalka, J. A., Pelletier, A. N., Fourati, S., Latif, M. B., Sharma, A., Furr, K., et al. (2021). The Transcription Factor CREB1 Is a Mechanistic Driver of Immunogenicity and Reduced HIV-1 Acquisition Following ALVAC Vaccination. *Nat. Immunol.* 22, 1294–1305. doi:10.1038/s41590-021-01026-9

Unterman, A., Sumida, T. S., Nouri, N., Yan, X., Zhao, A. Y., Gasque, V., et al. (2022). Single-cell Multi-Omics Reveals Dyssynchrony of the Innate and Adaptive Immune System in Progressive COVID-19. *Nat. Commun.* 13 (1), 440. doi:10.1038/s41467-021-27716-4

Vanderbeke, L., Van Mol, P., Van Herck, Y., De Smet, F., Humblet-Baron, S., Martinod, K., et al. (2021). Monocyte-driven Atypical Cytokine Storm and Aberrant Neutrophil Activation as Key Mediators of COVID-19 Disease Severity. *Nat. Commun.* 12 (1), 4117. doi:10.1038/s41467-021-24360-w

Viret, C., Rozières, A., and Faure, M. (2018). Autophagy during Early Virus-Host Cell Interactions. *J. Mol. Biol.* 430, 1696–1713. doi:10.1016/j.jmb.2018.04.018

Wicik, Z., Eyileten, C., Jakubik, D., Simões, S. N., Martins, D. C., Pavão, R., et al. (2020). ACE2 Interaction Networks in COVID-19: A Physiological Framework for Prediction of Outcome in Patients with Cardiovascular Risk Factors. *J. Clin. Med.* 9 (11), 3743. doi:10.3390/jcm9113743

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* 579 (7798), 265–269. doi:10.1038/s41586-020-2008-3

Wu, Z., and McGoogan, J. M. (2020). Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *Jama* 323 (13), 1239–1242. doi:10.1001/jama.2020.2648

Yan, L.-r., Wang, A., Lv, Z., Yuan, Y., and Xu, Q. (2021). Mitochondria-related Core Genes and TF-miRNA-Hub mrDEGs Network in Breast Cancer. *Biosci. Rep.* 41 (1), BSR20203481. doi:10.1042/bsr20203481

Zhang, L., Zhang, L., Li, S., Zhang, Q., Luo, Y., Zhang, C., et al. (2021). Overexpression of Mm9_circ_013935 Alleviates Renal Inflammation and Fibrosis in Diabetic Nephropathy via the miR-153-3p/NFIC axis. *Can. J. Physiol. Pharmacol.* 99, 1199–1206. doi:10.1139/cjpp-2021-0187

Zheng, Y., Yang, Z., Jin, C., Chen, C., and Wu, N. (2021). hsa-miR-191-5p Inhibits Replication of Human Immunodeficiency Virus Type 1 by Downregulating the Expression of NUP50. *Arch. Virol.* 166 (3), 755–766. doi:10.1007/s00705-020-04899-7

Zhou, Z., Zhou, X., Cheng, L., Wen, L., An, T., Gao, H., et al. (2021). Machine Learning Algorithms Utilizing Blood Parameters Enable Early Detection of Immunethrombotic Dysregulation in COVID-19. *Clin. Transl. Med* 11 (9), e523. doi:10.1002/ctm2.523

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

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

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fgene.2022.933009

Finding Lung-Cancer-Related lncRNAs Based on Laplacian Regularized Least Squares With Unbalanced Bi-Random Walk

*Zhifeng Guo, Yan Hui, Fanlong Kong and Xiaoxi Lin**

Department of Oncology, Chifeng Municipal Hospital, Chifeng, China

Lung cancer is one of the leading causes of cancer-related deaths. Thus, it is important to find its biomarkers. Furthermore, there is an increasing number of studies reporting that long noncoding RNAs (lncRNAs) demonstrate dense linkages with multiple human complex diseases. Inferring new lncRNA-disease associations help to identify potential biomarkers for lung cancer and further understand its pathogenesis, design new drugs, and formulate individualized therapeutic options for lung cancer patients. This study developed a computational method (LDA-RLSURW) by integrating Laplacian regularized least squares and unbalanced bi-random walk to discover possible lncRNA biomarkers for lung cancer. First, the lncRNA and disease similarities were computed. Second, unbalanced bi-random walk was, respectively, applied to the lncRNA and disease networks to score associations between diseases and lncRNAs. Third, Laplacian regularized least squares were further used to compute the association probability between each lncRNA-disease pair based on the computed random walk scores. LDA-RLSURW was compared using 10 classical LDA prediction methods, and the best AUC value of 0.9027 on the lncRNADisease database was obtained. We found the top 30 lncRNAs associated with lung cancers and inferred that lncRNAs TUG1, PTENP1, and UCA1 may be biomarkers of lung neoplasms, non-small–cell lung cancer, and LUAD, respectively.

Keywords: lung cancer, lncRNA, biomarker, lncRNA-disease association, laplacian regularized least squares, unbalanced bi-random walk

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Guanghui Li,
East China Jiaotong University, China
JunLin Xu,
Hunan University, China

****Correspondence:***
Xiaoxi Lin
L18047666059@163.com

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

Received: *30 April 2022*
Accepted: *03 June 2022*
Published: *22 July 2022*

Citation:
Guo Z, Hui Y, Kong F and Lin X (2022) Finding Lung-Cancer-Related lncRNAs Based on Laplacian Regularized Least Squares With Unbalanced Bi-Random Walk. Front. Genet. 13:933009. doi: 10.3389/fgene.2022.933009

1 INTRODUCTION

Cancers are posing threat for the health of humans (Yang et al., 2013; Liu et al., 2021). Lung cancer is the most common cancer worldwide and one of the leading causes of cancer-relevant deaths, and it has been so for many years. Thus, in 2008, the global statistical analysis demonstrated that approximately 1.6 million new lung cancer cases were diagnosed, and 1.4 million deaths were confirmed globally. In 2012, there were 1.8 million of new lung cancer diagnoses and 1.6 million deaths (de Groot et al., 2018; Howlader et al., 2020). In 2018, the number of new lung cancer cases exceeded 2 million and the number of deaths exceeded 1.7 million (Yuan et al., 2019). In the United States, approximately 234,000 cases of lung cancer were diagnosed the same year. This year, lung cancer diagnosis account for 14 and 13% of new cases in men and women, respectively. Estimation of mortality is 83,550 and 70,500 deaths in men and women, respectively. Lung

carcinoma is one of cancers with the lowest survival rate. It is usually not diagnosed until an advanced stage (de Groot et al., 2018; Howlader et al., 2020).

Despite the fast development of lung cancer therapy, high morbidity and mortality rates still pose a severe challenge for cancer researchers. The majority of patients with advanced-stage lung cancer have been ultimately poorly diagnosed. Thus, designing efficient therapy strategies is extremely important for lung cancer patients. However, existing techniques applied to diagnosis and therapies of lung cancer remain suboptimal. Thus, better strategies supplementing or replacing the existing techniques are urgent. Genome-wide association studies have found numerous genetic variants relevant to various cancers, one-third of which are densely linked to noncoding regions. The noncoding RNAs can be used as biomarkers of lung cancers. Therefore, accurate biomarker identification is urgently required to effectively diagnose lung cancer and boost the survival rate while decreasing its mortality and morbidity (Huang et al., 2017; Roointan et al., 2019; Yang et al., 2020).

Long noncoding RNAs (lncRNAs) are a type of noncoding RNAs that has over 200 nucleotides and post-transcriptional modifications including splicing, capping, and polyadenylation. lncRNAs can be used as a guide for protein-DNA interactions, protein-RNA interactions, and protein–protein interactions (Peng et al., 2020a). With the fast advancement of cancer genomics, many lncRNAs have been demonstrated to be aberrantly expressed in diverse cancers and play key action in the development of tumors through modulation of cancer-related signaling pathways. lncRNAs can regulate survival, metastasis, angiogenesis, and proliferation of tumor cells. Therefore, lncRNAs can be used as potential biomarkers and therapeutic targets in cancers by interacting with proteins (Chandra Gupta and Nandan Tripathi, 2017). For example, Peng et al. and her groups (Peng et al., 2021a; Zhou L. Q. et al., 2021; Peng et al., 2021b; Zhou L. et al., 2021; Tian et al., 2021; Peng et al., 2022) designed a series of state-of-the-art lncRNA-protein interaction prediction methods and significantly improved biomarker identification for various diseases. In addition, lncRNA SNHG14, BCRT1, DSCAM-AS1, MaTAR24, and HOTAIR have been validated to densely link to breast cancer (Niknafs et al., 2016; Dong et al., 2018; Chang et al., 2020; Liang et al., 2020; Yang et al., 2022; Xue et al., 2016). HOTAIR has been reported to be highly expressed in non-small–cell lung cancer (NSCLC) and affect NSCLC tumorigenesis and metastasis. In addition, many biomarkers (for example, CA125, NSE, CEA, VEGF, and EGFR (Khanmohammadi et al., 2020) have been validated to associate with lung cancer.

More importantly, many machine learning methods, especially deep-learning methods, have been applied to identify lncRNA biomarkers of various diseases through lncRNA-disease association prediction. Thus, Fan et al. (2022) designed an LDA prediction method (GCRFLDA) using the graph convolutional matrix completion. Ma Y (Ma, 2022) exploited a deep multi-network embedding-based LDA inference framework. Wu et al. (2021) integrated graph auto-encoder and random forest for LDA prediction. Sheng et al. (2021) developed an attentional multi-level representation encoding method to find new LDAs combining convolutional and variance autoencoders. Zhao et al. (2022) proposed a heterogeneous graph attention network-based LDA identification model. These methods significantly improved the LDA prediction.

With the development of single cell RNA sequencing technologies (Peng et al., 2020b), we can obtain numerous RNA data. These data can improve the analyses of RNA data, for example, SARS-CoV-2 (Xu et al., 2020; Li et al., 2021). By finding new lncRNA biomarkers, we can design corresponding therapeutic strategies for lung cancer based on drug repositioning (Peng et al., 2015; Liu et al., 2020; Meng et al., 2022; Shen et al., 2022).

Although experimental methods found a few biomarkers for lung cancer, they are time-consuming and waste of resources. Therefore, computational techniques have been exploited to infer potential biomarkers for lung cancer. However, the majority of computational approaches need to improve the inference performance. In this study, to analyze the diagnostic, prognostic, and therapeutical potential of lncRNAs in lung cancer patients, we exploit a computational model combining Laplacian regularized least square and unbalanced bi-random walk, LDA-RLSURW, to predict possible lncRNA biomarkers for lung cancer.

2 DATASETS

First, the lncRNA-disease association dataset was collected. The dataset can be obtained from the lncRNADisease database at http://www.cuilab.cn/lncrnadisease (Chen et al., 2012). We obtained 82 lncRNAs, 157 diseases, and 701 associations after excluding lncRNAs without record in the lncRNADisease database and diseases with inappropriate names or without MeSH tree numbers.

3 METHODS

This study developed an lncRNA-disease association prediction method LDA-RLSURW. First, LDA-RLSURW computed disease semantic similarity and lncRNA functional similarity. Second, LDA-RLSURW calculated the initial association probability of each lncRNA-disease pair using unbalanced bi-random walk based on disease similarity matrix and lncRNA similarity, respectively. In conclusion, the computed initial lncRNA-disease association probabilities were further updated Laplacian regularized least squares. The flowchart of LDA-RLSURW is presented in **Figure 1**.

3.1 Disease Semantic Similarity

Semantic similarity between diseases can be computed using the directed acyclic graph (DAGs) based on their MeSH descriptors (Fan et al., 2020). Given a disease A, let its DAG be represented as $DAG_A = \{T_A, E_A\}$, where T_A denotes the ancestor node set of A

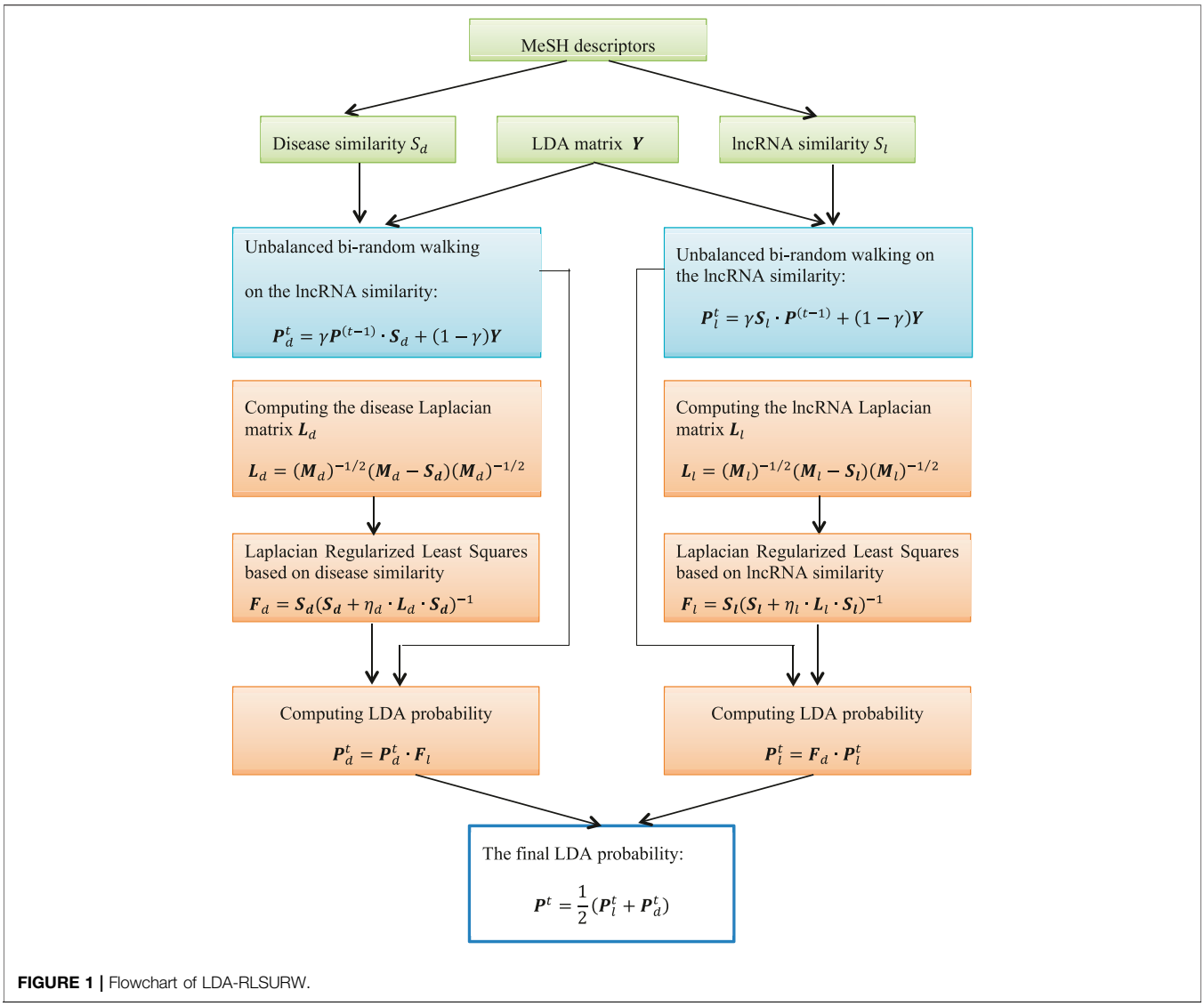


FIGURE 1 | Flowchart of LDA-RLSURW.

including A, and E_A denotes all edge set. For a disease term $t \in T_A$ in DAG_A, its semantic contribution to A can be computed by **Eq. 1** provided by LNCSIM1 (Chen et al., 2015):

$$SV_A^1(t) = \begin{cases} 1 & t = A \\ \max(\alpha \times SV_A^1(t') | t' \in C(t) & t \neq A \end{cases}, \quad (1)$$

where $C(t)$ denotes the children of t and α denotes a sematic contribution value of an edge linking t' to t in E_A.

In **Eq. 1**, we assume that terms at one identical layer from DAG_A have identical semantic contribution to A. However, when terms t_1 and t_2 are in the identical layer of DAG_A, and t_1 appears less than t_2 in DAG_A, the results from t_1 may be more specific than t_2. Thus, it could be more reasonable that $SV_A^1(t_1)$ is larger than $SV_A^1(t_2)$.

Considering this situation, we compute another semantic contribution value for disease A by **Eq. 2** provided by LNCSIM1 (Chen et al., 2015):

$$SV_A^2(t) = -log\frac{Dags(t)}{D}, \quad (2)$$

where D denotes the number of all diseases in the MeSH database and $Dags(t)$ denotes the number of DAG s, including the disease term t. In conclusion, the semantic contribution value of disease A in DAG_A can be computed by

$$SV_A^3(t) = \begin{cases} 1 & t = A \\ \max((\alpha + \beta)SV_A^3(t') | t' \in C(t) & t \neq A \end{cases}, \quad (3)$$

where β denotes the information content contribution factor, and

$$\beta = \frac{\max_{k \in K}(Dags(k)) - dags(t)}{D}, \quad (4)$$

where K denotes the disease set from the MeSH database.

Thus, the contribution of all diseases in DAG_A to A can be represented as

$$SV(A) = \sum_{t \in T_A} SV_A^3(t). \quad (5)$$

In summary, the semantic similarity between diseases A and B can be computed by **Eq. 6**:

$$S_d(A,B) = \frac{\sum_{t \in T_A \cap T_B} \left(SV_A^3(t) + SV_B^3(t)\right)}{SV(A) + SV(B)}. \quad (6)$$

3.2 lncRNA Functional Similarity

We calculate the lncRNA similarity using the approach provided by Fan et al. (2020). Assuming that *DG(u)*/*DG(v)* denotes diseases associated with lncRNA u/ v based on the LDA matrix, the lncRNA similarity between u and v was computed through semantic similarity between diseases involved in *DG(u)* and *DG(v)*. First, we construct a disease semantic similarity sub-matrix, where both rows and columns denote all diseases involved in *DG(u)*∪*DG(v)*, and the value of each element can be measured using the semantic similarity between corresponding diseases. Second, let d_u/ d_v denote one disease in *DG(u)*/*DG(v)*; the similarity between d_u/ d_v and *DG(v)*/*DG(u)* can be computed by **Eqs. 7** and **8**:

$$S(d_u, DG(v)) = \max_{d \in DG(v)} (S_d(d_u, d)), \quad (7)$$

$$S(d_v, DG(u)) = \max_{d \in DG(u)} (S_d(d_v, d)). \quad (8)$$

Third, the similarity between *DG(u)* to *DG(v)* and one between *DG(v)* to *DG(u)* can be calculated by **Eqs. 9** and **10**:

$$S_{u \to v} = \sum_{d \in DG(u)} S(d, DG(v)), \quad (9)$$

$$S_{v \to u} = \sum_{d \in DG(v)} S(d, DG(u)). \quad (10)$$

In conclusion, the similarity between two lncRNAs u and v can be computed by **Eq. 11**:

$$S_l(u,v) = \frac{S_{u \to v} + S_{v \to u}}{|DG(u)| + |DG(v)|}, \quad (11)$$

where $|DG(u)|/|DG(v)|$ indicates the number of diseases in $DG(u)/DG(v)$.

3.3 Unbalanced Bi-Random Walk

In this section, inspired by Shen et al. (2022), we consider that the lncRNA similarity network and the disease network and design an unbalance bi-random walk model to score lncRNA-disease pairs. The two networks exhibit different topological structures. Therefore, we use different optimal walking step sizes when randomly walking on these two networks. That is, we propose an unbalanced bi-random walk algorithm. First, we compute lncRNA-disease association scores by randomly walking with the maximal iteration number of n_l on the lncRNA network based on the lncRNA similarity by **Eq. 12**:

$$\boldsymbol{P}_l^t = \gamma \boldsymbol{S}_l \cdot \boldsymbol{P}^{(t-1)} + (1-\gamma)\boldsymbol{Y} \quad for \quad t = n_l. \quad (12)$$

In **Eq. 12**, at each step, the lncRNA similarity is fused with the random walk step by multiplying $\boldsymbol{S}_l$ on the left of the lncRNA-disease association probability matrix. $\gamma \in (0,1)$ is used to decrease the importance of circular bigraphs where the paths are longer during random walk and balance possible and known LDAs.

Second, we compute lncRNA-disease association scores by randomly walking with the maximal iteration number of n_d on the disease network based on the disease similarity by **Eq. 13**:

$$\boldsymbol{P}_d^t = \gamma \boldsymbol{P}^{(t-1)} \cdot \boldsymbol{S}_d + (1-\gamma)\boldsymbol{Y} \; for \; t = n_r. \quad (13)$$

In **Eq. 13**, at each step, disease similarity is fused with the random walk step by multiplying $\boldsymbol{S}_d$ on the right of the lncRNA-disease association probability matrix.

3.4 Laplacian Regularized Least Squares

In the last section, we compute the association probability for each lncRNA and disease using unbalanced bi-random walk method. However, for the algorithm, the jump condition is determined by known LDA data and the two similarity matrices. For a node n_i in an LDA network, if two other nodes n_j and n_k exhibit the same similarity with n_i, n_j and n_k may equally contribute to the jump. However, the node that has lower similarities with other nodes should have more contribution. Thus, we introduce Laplacian regularized least squares to solve the problem. First, the lncRNA Laplacian matrix $\boldsymbol{L}_l$ and the disease Laplacian matrix $\boldsymbol{L}_d$ are normalized to assess the jump probability for each node *via* **Eqs 14**, **15**.

$$\boldsymbol{L}_l = (\boldsymbol{M}_l)^{-1/2}(\boldsymbol{M}_l - \boldsymbol{S}_l)(\boldsymbol{M}_l)^{-1/2}, \quad (14)$$

$$\boldsymbol{L}_d = (\boldsymbol{M}_d)^{-1/2}(\boldsymbol{M}_d - \boldsymbol{S}_d)(\boldsymbol{M}_d)^{-1/2}, \quad (15)$$

where $\boldsymbol{M}_l/\boldsymbol{M}_d$ represent the diagonal matrices of lncRNAs/ diseases whose element $\boldsymbol{M}_l(\boldsymbol{i},\boldsymbol{i})/\boldsymbol{M}_d(\boldsymbol{j},\boldsymbol{j})$ denotes the summation of the i-th/ j-th row of $\boldsymbol{S}_l/\boldsymbol{S}_d$.

Second, to optimize the above minimum problems, the loss functions in the lncRNA and disease spaces are defined based on Laplacian matrices $\boldsymbol{L}_l$ and $\boldsymbol{L}_d$ *via* **Eqs. 11** and **12**, respectively:

$$\min_{F_l}\left[\|\boldsymbol{Y}^T - \boldsymbol{F}_l\|_F^2 + \eta_l \|\boldsymbol{F}_l \cdot \boldsymbol{L}_l \cdot (\boldsymbol{F}_l)^T\|_F^2\right], \quad (16)$$

$$\min_{F_d}\left[\|\boldsymbol{Y} - \boldsymbol{F}_d\|_F^2 + \eta_d \|\boldsymbol{F}_d \cdot \boldsymbol{L}_d \cdot (\boldsymbol{F}_d)^T\|_F^2\right], \quad (17)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $(\cdot)^T$ indicates the transpose, and η_v and η_d represent trade-off parameters. Models (11) and (12) can be solved *via* **Eqs. 13** and **14**, respectively:

$$\boldsymbol{F}_l^* = \boldsymbol{S}_l\left(\boldsymbol{S}_l + \eta_l \cdot \boldsymbol{L}_l \cdot \boldsymbol{S}_l\right)^{-1}\boldsymbol{Y}^T, \quad (18)$$

$$\boldsymbol{F}_d^* = \boldsymbol{S}_d\left(\boldsymbol{S}_d + \eta_d \cdot \boldsymbol{L}_d \cdot \boldsymbol{S}_d\right)^{-1}\boldsymbol{Y}. \quad (19)$$

To comprehensively detect the effect of unbalanced bi-random walk on the inference performance, we replace $\boldsymbol{Y}$ using LDA association probabilities computed by random walks. Assume that **Eqs. 20** and **21** can be defined as follows:

TABLE 1 | AUC values of LDA prediction methods on the lncRNADisease dataset.

	LNCSIM1/LNCSIM2	ILNCSIM	IDSSIM	RWRlncD	IIRWR
5-fold CV	0.8892/0.8881	0.8866	0.8966	0.6976	0.7781
	SIMCLDA	LRLSLDA	LLCPLDA	LDA-LNSUBRW	LDA-RLSURW
	0.7986	0.8174	0.8678	0.8874	0.9027

The LNCSIM1, LNCSIM2, LRLSLDA, and LDA-RLSURW are Laplacian regularized least square-based LDA methods, and the LDA-RLSURW can compute a better AUC. The results demonstrate that integrating unbalanced bi-random random walk can improve the performance. In addition, the IDSSIM and LDA-RLSURW computed the lncRNA similarity and disease similarity using the same method. The IDSSIM used the weighed K nearest known neighbor method to compute the lncRNA-disease association scores. The LDA-RLSURW outperforms IDSSIM, which show that the combination of Laplacian regularized least square and unbalanced bi-random walk can improve the LDA prediction performance compared to weighted K nearest known neighbor method. Both RWRlncD and IIRWR are random walk with restart-based LDA prediction methods. The SIMCLDA is an inductive matrix completion-based method. The LLCPLDA is a locality-constraint linear coding-based method. The LDA-RLSURW computes a better AUC than RWRlncD, IIRWR, SIMCLDA, and LLCPLDA, which further validates the powerful performance of LDA-RLSURW.

TABLE 2 | Inferred top 30 lncRNAs associated with LN.

Rank	lncRNAs	Evidence	Rank	lncRNAs	Evidence
1	MALAT1	Known	16	MINA	the MNDR database
2	HOTAIR	Known	17	PVT1	the MNDR database
3	MEG3	Known	18	**TUG1**	**Unconfirmed**
4	H19	Known	19	**PANDAR**	**Unconfirmed**
5	GAS5	Known	20	XIST	the MNDR database
6	UCA1	Known	21	**HULC**	**Unconfirmed**
7	CCAT2	Known	22	**HNF1A-AS1**	**Unconfirmed**
8	SPRY4-IT1	Known	23	**PTENP1**	**Unconfirmed**
9	CCAT1	Known	24	**KCNQ1OT1**	**Unconfirmed**
10	CDKN2B-AS1	Known	25	**HIF1A-AS2**	**Unconfirmed**
11	BANCR	Known	26	**DANCR**	**Unconfirmed**
12	BCYRN1	Known	27	**NPTN-IT1**	**Unconfirmed**
13	PCAT1	Known	28	**CRNDE**	**Unconfirmed**
14	SOX2-OT	Known	29	**CBR3-AS1**	**Unconfirmed**
15	CASC2	Known	30	**MIR31HG**	**Unconfirmed**

*The bold values denotes lncRNAs that were predicted to associate with LN and need to further validate in **Table 2**.*

$$\boldsymbol{F}_l = \boldsymbol{S}_l\left(\boldsymbol{S}_l + \eta_l \cdot \boldsymbol{L}_l \cdot \boldsymbol{S}_l\right)^{-1}, \tag{20}$$

$$\boldsymbol{F}_d = \boldsymbol{S}_d\left(\boldsymbol{S}_d + \eta_d \cdot \boldsymbol{L}_d \cdot \boldsymbol{S}_d\right)^{-1}. \tag{21}$$

At the t-th walking, **Eqs. 22** and **23** can be defined as

$$\boldsymbol{P}_l^t = \boldsymbol{F}_d \cdot \boldsymbol{P}_l^t, \tag{22}$$

$$\boldsymbol{P}_d^t = \boldsymbol{P}_d^t \cdot \boldsymbol{F}_l. \tag{23}$$

In conclusion, the LDA-RLSURW calculates the association score for each lncRNA-disease pair by combining association scores from the lncRNA and disease networks using **Eq. 24**:

$$\boldsymbol{P}^t = \frac{1}{2}\left(\boldsymbol{P}_l^t + \boldsymbol{P}_d^t\right). \tag{24}$$

4 EXPERIMENTS

4.1 Experimental Settings and Evaluation

The semantic contribution weight α is set as 0.5, the jump probability γ is set as 0.001, the maximal iteration number on the lncRNA network n_l is set as 31, the maximal iteration number on the disease network n_r is set as 1, and Laplacian regularized least square parameters η_l and η_d are set as 0.01. When the parameters are set as the above values, respectively, the LDA-RLSURW computes the best AUC on the lncRNADisease dataset. Therefore, we choose the parameters as the corresponding values. For other parameters, we set them as defaults provided by corresponding methods. The proposed LDA-RLSURW method and other comparative methods are evaluated using area under the receiver operating characteristic curve (AUC). Larger AUC values denote better performance.

4.2 Performance Comparison With Other Methods

To assess the performance of our proposed LDA-RLSURW method, we compare it with other 10 classical LDA prediction methods, that is, LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM (Fan W. et al., 2020). LNCSIM1 and LNCSIM2 measured the disease similarity separately using DAGs and the information content and computed association score for each lncRNA-disease pair by Laplacian regularized least squares. IDSSIM designed novel lncRNA functional similarity and disease semantic similarity computation approaches and computed the lncRNA-disease association scores using the computed similarity matrices and weighed K nearest known neighbor method. **Table 1** shows the AUC

TABLE 3 | Inferred top 30 lncRNAs associated with NSCLC.

Rank	lncRNAs	Evidence	Rank	lncRNAs	Evidence
1	MALAT1	Known	16	PANDAR	Known
2	HOTAIR	Known	17	HIF1A-AS1	Known
3	MEG3	Known	18	PCAT1	the MNDR database
4	GAS5	Known	19	CASC2	the MNDR database
5	H19	Known	20	SOX2-OT	the MNDR database
6	UCA1	Known	21	HULC	the MNDR database
7	CCAT2	Known	22	**MINA**	Unconfirmed
8	SPRY4-IT1	Known	23	**PTENP1**	Unconfirmed
9	CDKN2B-AS1	Known	24	HIF1A-AS2	the MNDR database
10	PVT1	Known	25	HNF1A-AS1	Known
11	CCAT1	Known	26	KCNQ1OT1	the MNDR database
12	TUG1	Known	27	CRNDE	the MNDR database
13	BANCR	Known	28	DANCR	the MNDR database
14	BCYRN1	Known	29	MIR31HG	the MNDR database
15	XIST	Known	30	NPTN-IT1	the MNDR database

*The bold values denotes lncRNAs that were predicted to associate with NSCLC and need to further validate in **Table 3**.*

values of LDA prediction methods on the lncRNADisease dataset. From **Table 1**, we can see that LDA-RLSURW computes the best AUC, which demonstrates the powerful LDA prediction performance of LDA-RLSURW.

4.3 Case Study

In this section, we conduct case studies to find potential lncRNA biomarkers for lung neoplasms, NSCLC, and adenocarcinoma of lung after confirming the performance of the proposed LDA-RLSURW method.

4.3.1 Finding Potential lncRNA Biomarkers for Lung Neoplasms

Lung neoplasms are one of the leading causes of death associated with malignant tumors in China (Khanmohammadi et al., 2020). Thus, Wang et al. (2020) investigated 14,528 lung cancer patients suffering from multiple primary malignant neoplasms (MPMN) and found 364 MPMN cases. In this section, we inferred the top 30 lncRNA biomarkers associated with lung neoplasms. The results are shown in **Table 2** and **Figure 2**. From **Table 2** and **Figure 2**, we can find that 15 lncRNAs are known to be associated with lung neoplasms in the lncRNADisease database, 3 lncRNAs (MINA, PVT1, and XIST) are unknown to be associated with lung neoplasms in the lncRNADisease database, which can be validated by the MNDR database (Cui et al., 2018). In addition, 12 lncRNAs are predicted to link to lung neoplasms and may be possible biomarkers of lung neoplasms.

More importantly, we predict that lncRNA taurine-upregulated gene 1 (TUG1) may be associated with lung neoplasms. TUG1 is one of lncRNAs that were first identified to associate with human disease. It is linked to diverse physiological processes, for example, gene regulation involved in translation, post-translation, transcription, and post-transcription. In this section, we infer that TUG1 may be the biomarker of lung neoplasms (Guo et al., 2020).

4.3.2 Finding Potential lncRNA Biomarkers for NSCLC

The NSCLC is a subtype of lung cancer. It is one of the leading causes of cancer death in the United States and accounts for 85% of lung cancers among all its subtypes. Although we have achieved important advancements in the NSCLC treatment, our understanding about the biology and mechanisms of NSCLC progression and early detection is still superficial. In this section, we aim to infer new lncRNA biomarkers for NSCLC after confirming the performance of LDA-RLSURW. The predicted top 30 lncRNAs associated with NSCLC are presented in **Table 3** and **Figure 3**. From **Table 3** and **Figure 3**, we can find that 18 lncRNAs associated with NSCLC are known in the lncRNADisease database, 10 lncRNAs associated with NSCLC have been validated in the MNDR database, and 2 lncRNAs (MINA and PTENP1) associated with NSCLC are unknown and require validation. The lncRNA PTENP1 has exerted the tumor-suppressive function through modulating PTEN expression in multiple malignancies. We predict that the

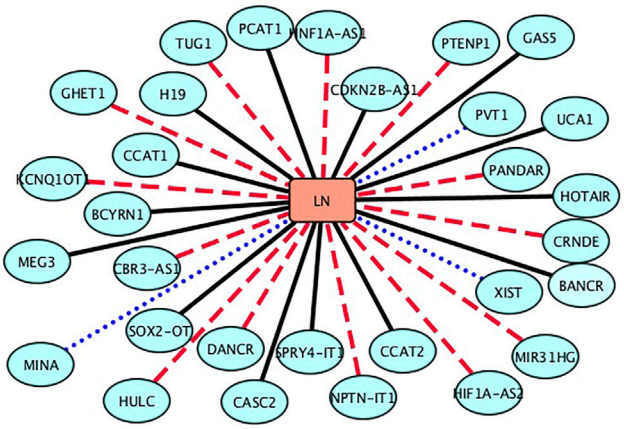


FIGURE 2 | Associations between the inferred top 30 lncRNAs and lung neoplasms (LN). Black solid lines represent known LDAs in the lncRNADisease database. Blue-dot lines represent LDAs that can be observed in the MNDR database. Red-dash lines represent LDAs predicted to be potential lncRNA biomarkers of LN.

TABLE 4 | Inferred top 30 lncRNAs associated with LUAD.

Rank	lncRNAs	Evidence	Rank	lncRNAs	Evidence
1	MALAT1	Known	16	**XIST**	Unconfirmed
2	HOTAIR	Known	17	**PANDAR**	Unconfirmed
3	MEG3	Known	18	**BCYRN1**	Unconfirmed
4	GAS5	Known	19	**PCAT1**	Unconfirmed
5	CCAT1	Known	20	**HULC**	Unconfirmed
6	HNF1A-AS1	the MNDR database	21	**CASC2**	Unconfirmed
7	MIAT	Known	22	**SOX2-OT**	Unconfirmed
8	H19	the MNDR database	23	**PTENP1**	Unconfirmed
9	**UCA1**	Unconfirmed	24	**MINA**	Unconfirmed
10	**CDKN2B-AS1**	Unconfirmed	25	**CRNDE**	Unconfirmed
11	**PVT1**	Unconfirmed	26	**DANCR**	Unconfirmed
12	**TUG1**	Unconfirmed	27	**WT1-AS**	Unconfirmed
13	**CCAT2**	Unconfirmed	28	**KCNQ1OT1**	Unconfirmed
14	**SPRY4-IT1**	Unconfirmed	29	**NPTN-IT1**	Unconfirmed
15	**BANCR**	Unconfirmed	30	**CCDC26**	Unconfirmed

*The bold values denotes lncRNAs that were predicted to associate with LUAD and need to further validate in **Table 4**.*

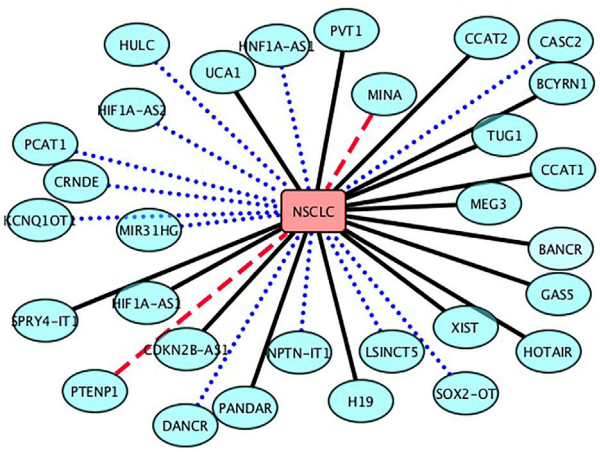


FIGURE 3 | Associations between the inferred top 30 lncRNAs and NSCLC. Black solid lines represent known LDAs in the lncRNADisease database. Blue-dot lines represent LDAs that can be observed in the MNDR database. Red-dash lines represent LDAs predicted to be potential lncRNA biomarkers of LN.

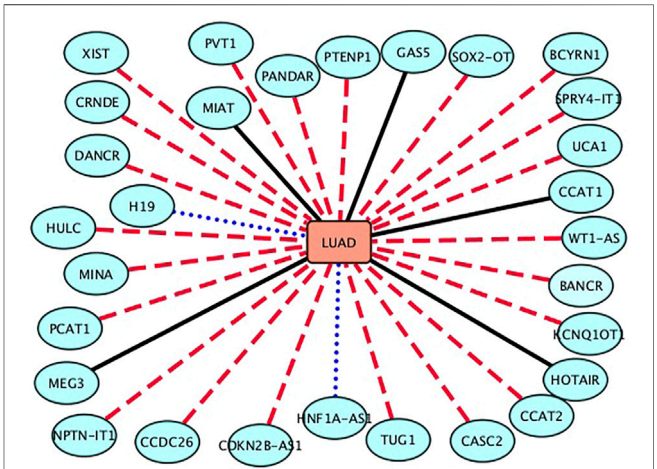


FIGURE 4 | Associations between the inferred top 30 lncRNAs and LUAD. Black solid lines represent known LDAs in the lncRNADisease database. Blue-dot lines represent LDAs that can be observed in the MNDR database. Red-dash lines represent LDAs predicted to be potential lncRNA biomarkers of adenocarcinoma of lung.

PTENP1 may be a potential biomarker of NSCLC (Herbst et al., 2018; Arbour and Riely, 2019; Fan et al., 2020; Leighl et al., 2019).

4.3.3 Finding Potential lncRNA Biomarkers for Lung Adenocarcinoma

The NSCLC is divided into three main subtypes: lung squamous cell carcinoma, large-cell lung cancer, and lung adenocarcinoma (LUAD), among which lung squamous cell carcinoma and LUAD are the most prevalent. In this section, we predict possible lncRNAs associated with LUAD. The results are shown in **Table 4** and **Figure 4**. From **Table 4** and **Figure 4**, we can find that 6 lncRNAs are known to associate with LUAD, 2 lncRNAs are not known to associate with LUAD in the lncRNADisease database, although they are known in the MNDR database, and 22 lncRNAs have not been confirmed to associate with LUAD.

Urothelial carcinoma associated 1 (UCA1) is an oncogenic lncRNA. It is highly expressed in many cancers. UCA1 can bind to tumor-suppressive microRNAs, activate a few pivotal signaling pathways, and alter epigenetic and transcriptional regulation. More importantly, its high expression is linked to poor clinicopathological characteristics. In this section, we predict that UCA1 may associate with LUAD and require validation (Yao et al., 2019).

5 DISCUSSION

LNCSIM1 and LNCSIM2 obtained better performance improvements based on cross-validation and case analyses. However, LNCSIM1 cannot effectively distinguish the

semantic contributions of various disease terms from the identical layer. LNCSIM2 computed the IC values only through integrating DAG information. ILNCSIM is an edge-based prediction model. It combined the concept of information content and the hierarchical structure of DAGs to compute disease semantic similarity.

The RWRlncD conducted random walk with restart on the lncRNA similarity network. However, the RWRlncD cannot be used to predict associated information for diseases without any associated lncRNAs. The IRWRLDA improved random walk-based method through setting an initial probability vector to reduce the disadvantages of random walk with restart. The SIMCLDA used an inductive matrix completion model to complement missing LDA information. The LRLSLDA utilized Laplacian regularized least square model to predict LDAs. The LLCLPLDA first applied a locality-constraint linear coding model to project the local-constraint characteristics of lncRNAs and diseases, and then propagated LDAs by the initial LDA. The LDA-LNSUBRW used linear neighborhood similarity measurement and unbalanced bi-random walk algorithm to find possible LDAs.

The LDA-RLSURW obtains better performance for lncRNA-disease association prediction. It has three advantages: First, it utilizes the biological features to compute the lncRNA and disease similarity. Second, it uses unbalanced bi-random walk to compute the lncRNA-disease association probability. In conclusion, it further computes the lncRNA-disease association probability combining Laplacian regularized least squares.

6 CONCLUSION

Lung cancer is one of the most threatening cancer forms worldwide. In this study, we designed a computational method, LDA-RLSURW, to find possible lncRNA biomarkers for lung cancer. LDA-RLSURW effectively combines unbalanced bi-random walk and Laplacian regularized least square. We predict that TUG1, PTENP1, and UCA1 may be the biomarkers of lung neoplasms, NSCLC and LUAD, respectively.

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

Conceptualization: ZG, YH, FK, and XL; methodology: ZG, YH, FK, and XL; project administration: XL; software: XL; writing original draft: ZG; writing review and editing: ZG and XL.

REFERENCES

Arbour, K. C., and Riely, G. J. (2019). Systemic Therapy for Locally Advanced and Metastatic Non-small Cell Lung Cancer. *Jama* 322 (8), 764–774. doi:10.1001/jama.2019.11058

Chandra Gupta, S., and Nandan Tripathi, Y. (2017). Potential of Long Non-coding RNAs in Cancer Patients: From Biomarkers to Therapeutic Targets. *Int. J. Cancer* 140 (9), 1955–1967. doi:10.1002/ijc.30546

Chang, K. C., Diermeier, S. D., Yu, A. T., Brine, L. D., Russo, S., Bhatia, S., et al. (2020). Matar25 Lncrna Regulates the Tensin1 Gene to Impact Breast Cancer Progression. *Nat. Commun.* 11, 1–19. doi:10.1038/s41467-020-20207-y

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNADisease: a Database for Long-Non-Coding RNA-Associated Diseases. *Nucleic acids Res.* 41 (D1), D983–D986. doi:10.1093/nar/gks1099

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA Functional Similarity Network Based on lncRNA-Disease Associations and Disease Semantic Similarity. *Sci. Rep.* 5 (1), 1–12. doi:10.1038/srep11338

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: an Updated Resource of ncRNA-Disease Associations in Mammals. *Nucleic Acids Res.* 46 (D1), D371–D374. doi:10.1093/nar/gkx1025

de Groot, P. M., Wu, C. C., Carter, B. W., and Munden, R. F. (2018). The Epidemiology of Lung Cancer. *Transl. Lung Cancer Res.* 7 (3), 220–233. doi:10.21037/tlcr.2018.05.06

Dong, H., Wang, W., Chen, R., Zhang, Y., Zou, K., Ye, M., et al. (2018). Exosome-mediated Transfer of lncRNA-SNHG14 P-romotes T-rastuzumab C-hemoresistance in B-reast C-ancer. *Int. J. Oncol.* 53, 1013–1026. doi:10.3892/ijo.2018.4467

Fan W, W., Shang, J., Li, F., Sun, Y., Yuan, S., and Liu, J. X. (2020). IDSSIM: an lncRNA Functional Similarity Calculation Model Based on an Improved Disease Semantic Similarity Method. *BMC Bioinforma.* 21 (1), 1–14. doi:10.1186/s12859-020-03699-9

Fan, Y., Chen, M., and Pan, X. (2022). GCRFLDA: Scoring lncRNA-Disease Associations Using Graph Convolution Matrix Completion with Conditional Random Field. *Brief. Bioinform* 23 (1), bbab361. doi:10.1093/bib/bbab361

Guo, C., Qi, Y., Qu, J., Gai, L., Shi, Y., and Yuan, C. (2020). Pathophysiological Functions of the lncRNA TUG1. *Curr. Pharm. Des.* 26 (6), 688–700. doi:10.2174/1381612826666191227154009

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The Biology and Management of Non-small Cell Lung Cancer. *Nature* 553 (7689), 446–454. doi:10.1038/nature25183

Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., et al. (2020). The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *N. Engl. J. Med.* 383 (7), 640–649. doi:10.1056/nejmoa1916623

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix Completion with Side Information and its Applications in Predicting the Antigenicity of Influenza Viruses. *Bioinformatics* 33 (20), 3195–3201. doi:10.1093/bioinformatics/btx390

Khanmohammadi, A., Aghaie, A., Vahedi, E., Qazvini, A., Ghanei, M., Afkhami, A., et al. (2020). Electrochemical Biosensors for the Detection of Lung Cancer Biomarkers: A Review. *Talanta* 206, 120251. doi:10.1016/j.talanta.2019.120251

Leighl, N. B., Page, R. D., Raymond, V. M., Daniel, D. B., Divers, S. G., Reckamp, K. L., et al. (2019). Clinical Utility of Comprehensive Cell-free DNA Analysis to Identify Genomic Biomarkers in Patients with Newly Diagnosed Metastatic Non-small Cell Lung Cancer. *Clin. Cancer Res.* 25 (15), 4691–4700. doi:10.1158/1078-0432.ccr-19-0624

Li, T., Huang, T., Guo, C., Wang, A., Shi, X., Mo, X., et al. (2021). Genomic Variation, Origin Tracing, and Vaccine Development of SARS-CoV-2: A Systematic Review. *Innovation* 2 (2), 100116. doi:10.1016/j.xinn.2021.100116

Liang, Y., Song, X., Li, Y., Chen, B., Zhao, W., Wang, L., et al. (2020). Lncrna Bcrt1 Promotes Breast Cancer Progression by Targeting Mir-1303/ptbp3 axis. *Mol. Cancer* 19, 85–20. doi:10.1186/s12943-020-01206-5

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor tissue-of-Origin[J]. *Front. Cell Dev. Biol.* 9, 886. doi:10.3389/fcell.2021.619330

Ma, Y. (2022). DeepMNE: Deep Multi-Network Embedding for lncRNA-Disease Association Prediction[J]. *IEEE J. Biomed. Health Inf.* 26 (7), 3539–3549. doi:10.1109/JBHI.2022.3152619

Meng, Y., Lu, C., Jin, M., Xu, J., Zeng, X., and Jang, J. (2022). A Weighted Bilinear Neural Collaborative Filtering Approach for Drug Repositioning[J]. *Briefings Bioinforma.* 23 (2), bbab581. doi:10.1093/bib/bbab581

Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., et al. (2016). The Lncrna Landscape of Breast Cancer Reveals a Role for Dscam-As1 in Breast Cancer Progression. *Nat. Commun.* 7, 1–14. doi:10.1038/ncomms12791

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2015). Predicting Drug-Target Interactions with Multi-Information Fusion. *IEEE J. Biomed. Health Inf.* 21 (2), 561–572. doi:10.1109/JBHI.2015.2513200

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022). EnANNDeep: An Ensemble-Based lncRNA–Protein Interaction Prediction Framework with Adaptive K-Nearest Neighbor Classifier and Deep Models[J]. *Interdiscip. Sci. Comput. Life Sci.* 14, 209–232. doi:10.1007/s12539-021-00483-y

Peng, L. H., Wang, C., Tian, X. F., Zhou, L. Q., and Li, K. Q. (2021b). Finding lncRNA-Protein Interactions Based on Deep Learning with Dual-Net Neural Architecture[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 2021, 3116232. doi:10.1109/TCBB.2021.3116232

Peng, L. H., Yuan, R. Y., Shen, L., Gao, P. F., and Zhou, L. Q. (2021a). LPI-EnEDT: an Ensemble Framework with Extra Tree and Decision Tree Classifiers for Imbalanced lncRNA-Protein Interaction Data Classification[J]. *BioData Min.* 14 (1), 1–22. doi:10.1186/s13040-021-00277-4

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2020a). Probing lncRNA-Protein Interactions: Data Repositories, Models, and Algorithms. *Front. Genet.* 10, 1346. doi:10.3389/fgene.2019.01346

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020b). Single-cell RNA-Seq Clustering: Datasets, Models, and Algorithms. *RNA Biol.* 17 (6), 765–783. doi:10.1080/15476286.2020.1728961

Roointan, A., Ahmad Mir, T., Wani, S. I., Mati-ur-Rehman, Hussain, K. K., Ahmed, B., et al. (2019). Early Detection of Lung Cancer Biomarkers through Biosensor Technology: A Review. *J. Pharm. Biomed. analysis* 164, 93–103. doi:10.1016/j.jpba.2018.10.017

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: An Anti-SARS-CoV-2 Drug Prioritizing Framework Combining an Unbalanced Bi-random Walk and Laplacian Regularized Least Squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Sheng, N., Cui, H., Zhang, T., and Xuan, P. (2021). Attentional Multi-Level Representation Encoding Based on Convolutional and Variance Autoencoders for lncRNA-Disease Association Prediction. *Brief. Bioinform* 22 (3), bbaa067. doi:10.1093/bib/bbaa067

Tian, X. F., Shen, L., Wang, Z. W., Zhou, L. Q., and Peng, L. H. (2021). A Novel lncRNA–Protein Interaction Prediction Method Based on Deep Forest with Cascade Forest Structure[J]. *Sci. Rep.* 11 (1), 1–15. doi:10.1038/s41598-021-98277-1

Wang, H., Hou, J., Zhang, G., et al. (2019). Clinical Characteristics and Prognostic Analysis of Multiple Primary Malignant neoplasms in patients with lung cancer [J]. *Cancer Gene Therapy* 26 (11), 419–426.

Wu, Q. W., Xia, J. F., Ni, J. C., and Zheng, C. H. (2021). GAERF: Predicting lncRNA-Disease Associations by Graph Auto-Encoder and Random Forest. *Brief. Bioinform* 22 (5), bbaa391. doi:10.1093/bib/bbaa391

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: an Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinformatics* 36 (10), 3139–3147. doi:10.1093/bioinformatics/btaa109

Xue, X., Yang, Y. A., Zhang, A., Fong, K.-W., Kim, J., Song, B., et al. (2016). LncRNA HOTAIR Enhances ER Signaling and Confers Tamoxifen Resistance in Breast Cancer. *Oncogene* 35 (21), 2746–2755. doi:10.1038/onc.2015.340

Yang, J., Grünewald, S., and Wan, X.-F. (2013). Quartet-Net: A Quartet-Based Method to Reconstruct Phylogenetic Networks. *Mol. Biol. Evol.* 30 (5), 1206–1217. doi:10.1093/molbev/mst040

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-Positive Breast Cancer Recurrence and Metastasis Risk from Histopathological Images and Clinical Information via Multimodal Deep Learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human Geroprotector Discovery by Targeting the Converging Subnetworks of Aging and Age-Related Diseases. *Geroscience* 42 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yao, F., Wang, Q., and Wu, Q. (2019). The Prognostic Value and Mechanisms of lncRNA UCA1 in Human Cancer. *Cancer Manag. Res.* 11, 7685–7696. doi:10.2147/cmar.s200436

Yuan, M., Huang, L. L., Chen, J. H., Wu, J., and Xu, Q. (2019). The Emerging Treatment Landscape of Targeted Therapy in Non-small-cell Lung Cancer. *Signal Transduct. Target Ther.* 4 (1), 61–14. doi:10.1038/s41392-019-0099-9

Zhao, X., Zhao, X., and Yin, M. (2022). Heterogeneous Graph Attention Network Based on Meta-Paths for lncRNA-Disease Association Prediction. *Brief. Bioinform* 23 (1), bbab407. doi:10.1093/bib/bbab407

Zhou, L. Q., Duan, Q., Tian, X. F., Tang, J. X., and Peng, L. H. (2021a). LPI-HyADBS: a Hybrid Framework for lncRNA-Protein Interaction Prediction Integrating Feature Selection and Classification[J]. *BMC Bioinforma.* 22 (1), 1–31. doi:10.1186/s12859-021-04485-x

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021b). LPI-deepGBDT: a Multiple-Layer Deep Framework Based on Gradient Boosting Decision Trees for lncRNA-Protein Interaction Identification. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8

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

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

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 19 August 2022
DOI 10.3389/fgene.2022.962028

OPEN ACCESS

EDITED BY
Jialiang Yang,
Geneis (Beijing) Co., Ltd., China

REVIEWED BY
Yaoxin Gao,
East China Normal University, China
Qiangfeng Yu,
Zhuhai People's Hospital, China

*CORRESPONDENCE
Shixin Huang,
254893284@qq.com
Zhijian Huang,
ehuang27@fjmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 05 June 2022
ACCEPTED 07 July 2022
PUBLISHED 19 August 2022

CITATION
Xiao C, Yang L, Jin L, Lin W, Zhang F, Huang S and Huang Z (2022), Prognostic and immunological role of cuproptosis-related protein FDX1 in pan-cancer.
Front. Genet. 13:962028.
doi: 10.3389/fgene.2022.962028

Prognostic and immunological role of cuproptosis-related protein FDX1 in pan-cancer

Chen Xiao[1†], Linhui Yang[2†], Liangzi Jin[3], Weiguo Lin[1], Faqin Zhang[1], Shixin Huang[4*] and Zhijian Huang[1*]

[1]Department of Gastroenterology, Fuzhou Second Hospital Affiliated to Xiamen University, The School of Clinical Medicine, Fujian Medical University, Fuzhou, China, [2]The Graduate School of Fujian Medical University, Fuzhou, China, [3]Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming, China, [4]Department of Ultrasound, Fujian Medical University Cancer Hospital, Fujian Cancer Hospital, Fuzhou, China

Background: Cancer is the second cause of death worldwide. Copperoptosis is a new mode of regulated cell death and is strongly associated with metabolic pathways. FDX1 is a key gene that promotes copperoptosis, and its impact on tumor pathogenesis and tumor immune response is indistinct and needs further exploration.

Methods: Data was mined from the Cancer Genome Atlas database, the Broad Institute Cancer Cell Line Encyclopedia database, and the International Cancer Genome Consortium. Survival analyses included the Kaplan–Meier method for calculating the cumulative incidence of survival events and the log-rank method for comparing survival curves between groups. Immune cell infiltration levels were calculated using the Spearman correlation test and correlated with FDX1 expression to assess significance. More correlation analyses between FDX1 expression and mutational markers, such as tumor mutational burden (TMB) and microsatellite instability (MSI), were also examined *via* Spearman assay to explore the relation between FDX1 expression and the sensitivity of common antitumor drugs.

Results: FDX1 expression was downregulated in most kinds of cancers, and this high expression indicated better overall survival and death-specific survival. For several cancer types, FDX1 expression had a positive correlation with immune cell infiltration, and FDX1 also had a positive correlation with TMB and MSI in some cancer types, linking its expression to the assessment of possible treatment responses.

Conclusion: The correlations between FDX1 expression and cancer in varioustissues, including clear links to cancer survival and prognosis, make FDX1 aninteresting biomarker and potential therapeutic target for cancer surveillance and futureresearch.

KEYWORDS

copperoptosis, pan-cancer, prognosis, biomarkers, immunological

Introduction

Cancer is the second cause of death worldwide. In 2020, approximately 19.3 million new cancer cases were found worldwide. Female breast cancer has become the commonest cancer diagnosed with approximately 2.3 million new cases (11.7%) exceeding lung cancer (11.4%), colorectal cancer (10.0%), prostate cancer (7.3%), and gastric cancer (5.6%) (Liu et al., 2021; Sung et al., 2021). Cancer is driven by genetic change, and the occurrence and development of cancer can be divided into three stages: transformation and growth of carcinogenic factors, promotion, and development of carcinogenesis. This is a multifactor, multistep complex process. Metabolism is significant in carcinogenesis, and recently, metabolism-targeted therapy has become an important part of tumor therapy. As tumorigenesis is complex, the conduction of a pan-cancer expression analysis of any gene of interest and the assessment of its correlation with clinical prognosis and potential molecular mechanisms are important. The publicly funded TCGA project contains functional genomics datasets of different tumors so pan-cancer analyses can be conducted (Tomczak et al., 2015; Blum et al., 2018; He et al., 2020a; He et al., 2020b; Zhao et al., 2021; Yang et al., 2022).

FDX1, also called adrenodoxin or hepatoredoxin, is a subunit of the augmin complex. The FDX1 gene is a small ferrithionein that transfers electrons from NADPH to mitochondrial cytochrome P450 *via* ferredoxin reductase, involved in the metabolism of steroids, vitamin D, and bile acids (Sheftel et al., 2010; Strushkevich et al., 2011). Diseases associated with FDX1 include cerebrotendinous xanthomatosis and xanthomatosis. The latest research shows that the FDX1 gene is a recently discovered important gene associated with copperoptosis (Tsvetkov et al., 2022). FDX1 positively regulates a specific metabolic pathway of copperoptosis, and FDX1 and Protein lipoylation are key regulators of copper ion carrier-induced cell death. FDX1 is associated with protein thioctanoylation, and FDX1 knockout results in loss of protein thioctanoylation. Protein thioctanoylation is a highly conserved posttranslational modification of lysine that occurs primarily on four enzymes that regulate the tricarboxylic acid cycle. Copper ions promote cell death by directly binding to thioctanoylated tricarboxylic acid cycle-related enzymes, and the knockout of FDX1 can save cell copperoptosis.

Copperoptosis is a new mode of regulating cell death (Tang et al., 2022). Copper ions are involved in cell death such as iron ions (Wang et al., 2022). Inhibiting mitochondrial respiration through drugs may be a strategy to fight diseases. In addition, some cancers express a large amount of thioctanoylated mitochondrial proteins, and with high respiration, the use of copper ion metal carriers to kill cancer cells may become a new method of treating cancer. FDX1 gene is a key gene that promotes copperoptosis, so the study of FDX1 is significant for tumorigenesis, progression, tumor prognosis, tumor treatment, and many other aspects in practice (Zhang et al., 2021). Here, bioinformatics analyses were conducted to evaluate different FDX1 expressions in tissues and their possible link with cancer. Its expression level was evidently associated with survival, immune cell function, and tumor mutation status. FDX1 can be used as a new prognostic marker for various malignancies and an indicator of cancer immunotherapy response.

Materials and methods

Data collection and processing

Pan-cancer sequencing data from the Cancer Genome Atlas (TCGA) database and the Broad Institute Cancer Cell Line Encyclopedia (CCLE) database (Illumina platform) and data related to hepatocellular carcinoma (LIHC) from the International Cancer Genome Consortium (ICGC)) databases were drawn through their portal for analysis (Hudson et al., 2010; Tomczak et al., 2015). The entire data set was screened, and missing and duplicate results were removed and converted by log2 (TPM + 1), using the rma function in the R package (R studio version: 1.2.1335, R version: 3.6.1). Relating information of clinic was also drawn through the portal, including the patient's age, gender, tumor stage, and clinical stage. In addition, the information that can only be downloaded from the TCGA database were tumor mutation load (TMB) and microsatellite instability (MSI). The calculation of TMB followed the total mutation incidence per million base pairs, and the calculation of MSI was from the amount of insertion or deletion events in a repeating genetic sequence. Data analysis was conducted using the Sangerbox tools (http://sangerbox.com/).

Cox regression analysis and survival analysis

In the ICGC and TCGA, Cox regression analysis was conducted to find out if FDX1 expression correlated with overall survival (OS) and disease-specific survival (DSS) for patients with different cancer types. Using the Kaplan–Meier method, the patients were grouped into high and low FDX1 expressions according to the optimal separation method, and the survival curve of patients with various cancer types was constructed. The analysis of specificity and time-dependent sensitivity of survival was conducted by deploying survival ROC and survival in R packages (rdocumentation.org/packages/survival). The difference between curves was checked via a log-rank test, and p values of less than 0.05 were regarded as important.

Immune cell infiltration and enrichment

Tumor Immune Estimation Resource (TIMER) (https://cistrome.shinyapps.io/timer/) is a computational network tool based on a database for immune cell infiltration that supplies infiltration scores for six common immune cell types, including

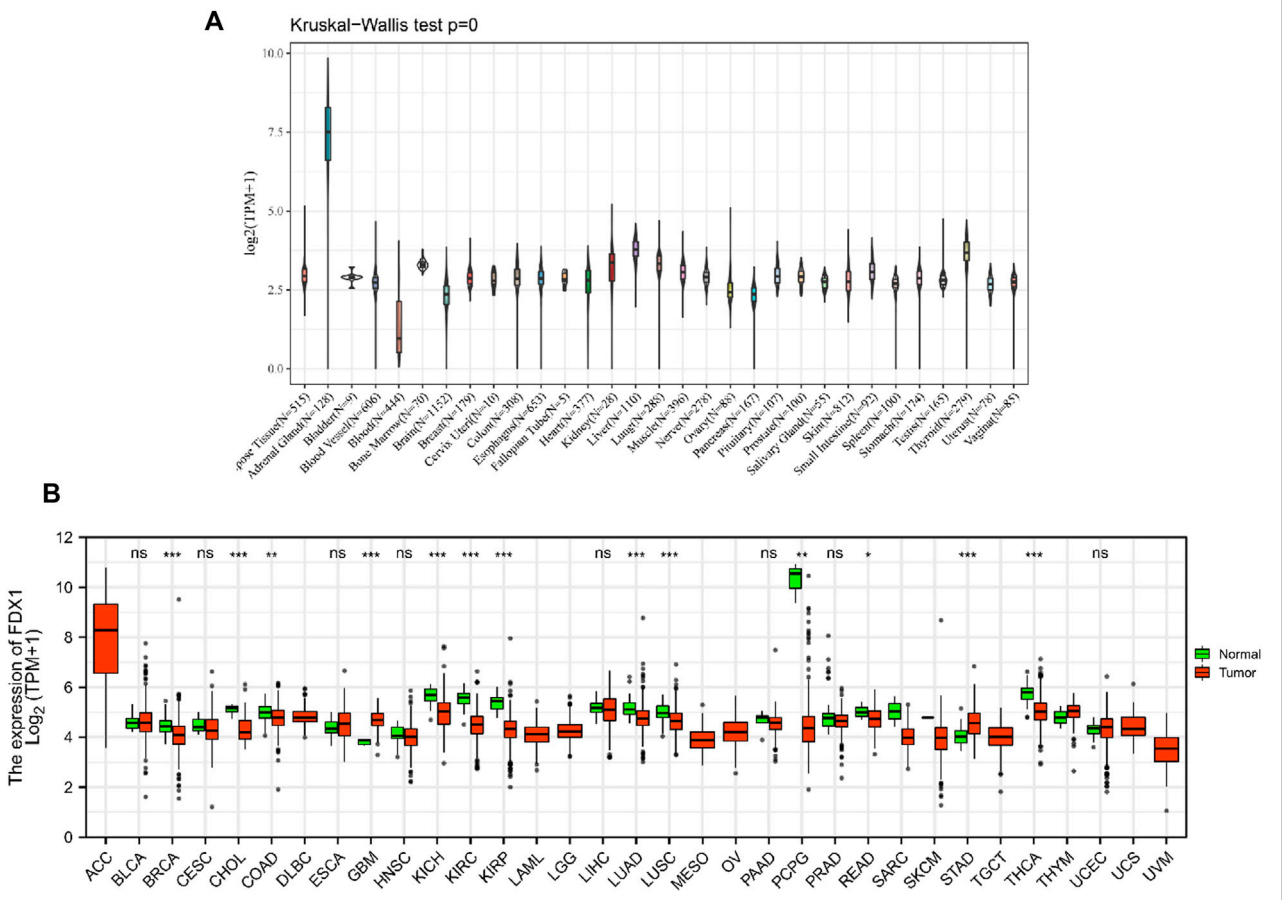


FIGURE 1
mRNA expression levels of FDX1 from different tissue sources and tumors. **(A)** normal mRNA expression levels of FDX1 in various tissues from the GTEx database. **(B)** differences in FDX1 mRNA expression between tumor and peritumoral samples from the Cancer Genome Atlas database. Abbreviations: BLCA, bladder urothelial carcinoma; CESC, cervical and cervical cancer; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid tumor diffuse large B-cell lymphoma; ESCA, esophageal cancer; HNSC, head and neck squamous cell carcinoma; KIRP, renal papillary cell carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PAAD, pancreatic cancer; PCPG, pheochromocytoma and paraneurysm; PRAD, prostate adenocarcinoma; reading, rectal adenocarcinoma; SARC, sarcoma; STAD, gastric adenocarcinoma; STES, gastric and esophageal cancer; TGCT, testicular germ cell tumor; THCA, thyroid cancer; THYM, thymoma; UCEC, endometrial cancer of the uterus; UCS, uterine carcinosarcoma.

B cells, CD4 + T cells, CD8 + T cells, macrophages, neutrophils, and dendritic cells (Li et al., 2016; Yang et al., 2017; Tavasolian et al., 2020; Hwu et al., 2021). The calculation of immune cell infiltration scores for pan-cancer data in the TCGA database was performed using TIMER and archived online. Here, correlation with FDX1 expression was tested with downloaded penetration data.

Correlation analysis of FDX1 expression in tumor microenvironment

The immune tumor microenvironment (TME) is a tumor cell–developing and –surviving microenvironment. It involves different elements surrounding tumor cells, stromal cells, etc. (Arneth, 2019; Liu et al., 2022a). The stromal and immune cell quantity in the tumor microenvironment affects the development and growth of cancer cells. The R package "ESTIMATE" is used to calculate StromalScore, ImmuneScore, and ESTIMATEScore, which is the sum of ImmuneScore and StromalScore (Yoshihara et al., 2013; Lv et al., 2021). Then, Spearman correlation analysis in R was used to analyze the association between FDX1 and stromal and immune scores.

Correlation analysis between FDX1 expression and immune infiltrating cell expression

TIMER is a database providing a platform for tumor immunoinfiltration analysis (Yang et al., 2017). In general

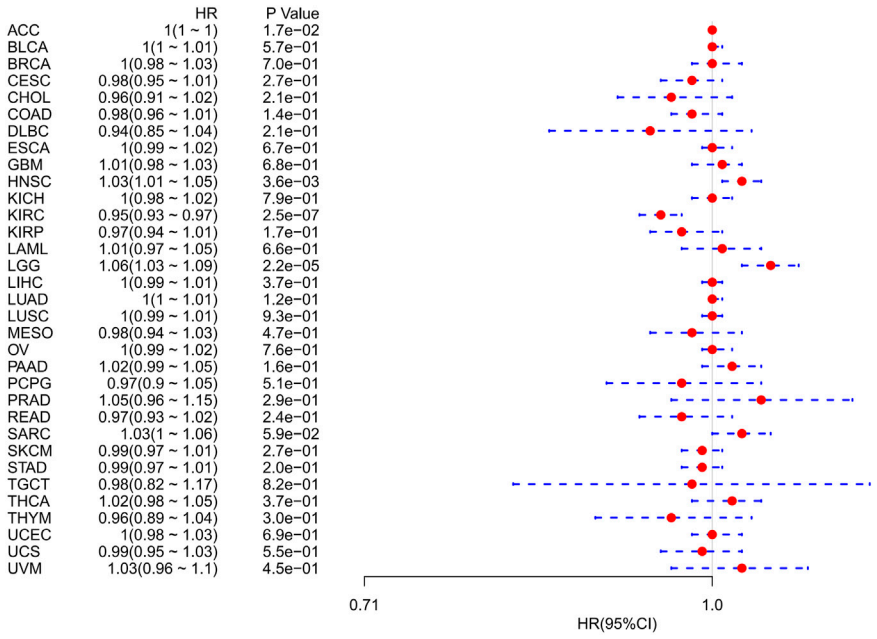


FIGURE 2
Association of FDX1 mRNA expression levels with overall survival in multiple tumors from the Cancer Genome Atlas database. Cox regression analysis, $p < 0.05$ was evident.

calculate the infiltration scores of six types of immune infiltrating cells: CD4 T cells, B cells, CD8 T cells, neutrophils, dendritic cells, and macrophages. The "gene" module in TIMER was used for the analysis of the correlation between FDX1 expression in the TCGA database and levels of immune infiltration across multiple cancer types.

Drug susceptibility analysis

A total of 60 cancer cells listed by the National Cancer Institute (NCI) Cancer Research Center are the basis of the CellMiner database. The NCI-60 cell line is the most popular cancer cell sample group for anticancer drug detection recently. Here, NCI-60 drug sensitivity data and RNA-seq gene expression data were downloaded, and the relation between genes and the sensitivity of common antitumor drugs was explored through correlation analysis.

Statistics

Correlations between FDX1 expression and target targets were assessed using Spearman correlation tests, including immune cell infiltration scores (as the description in the previous section for the six immune cell types), TMB, MSI, and mismatch repair (MMR) genes. According to whether the samples were paired, FDX1 expression levels were compared between groups or between tumors and normal tissue using paired *t*-test or *t*-test. *p* values below 0.05 are regarded as evident. All charts are generated from the R package of ggplot2 and forestplot.

Results

Expression levels of FDX1 in various normal and cancerous tissues

With the data of GTEx databases from different tissues in healthy individuals, it was determined that mRNA expression levels of FDX1 were similar in all tissues (Figure 1A), except for the adrenal gland. As an actively differentiated tissue, the higher expression levels of the adrenal gland were not unexpected. Further comparison of relatively normal tissues and respective tumors showed that FDX1 was lowly expressed in most tumors, except for GBM and STAD, showing that the opposite result was significant. Based on TCGA data, 13 of 33 cancer types (BRCA, CHOL, COAD, GBM, KICH, KIRC, KIRP, LUAD, LUSC, PCPG, READ, STAD, and THCA) showed significant differences in expression (Figure 1B).

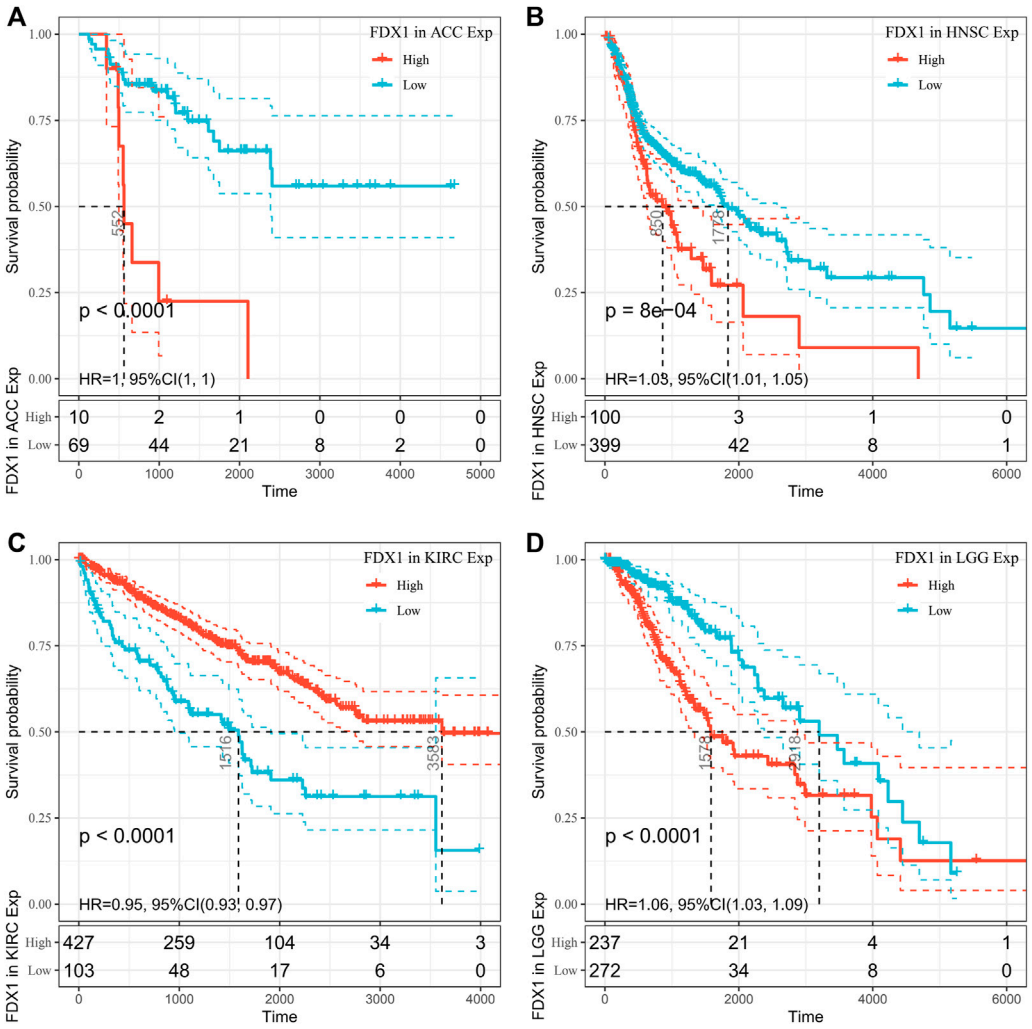


FIGURE 3
Overall survival (OS) difference of high and low FDX1 mRNA expression groups in significantly prognostically relevant tumors from the Cancer Genome Atlas database (by median expression dichotomy). **(A)** OS difference of ACC groups. **(B)** OS difference between HNSC groups. **(C)** OS difference between KIRC groups. **(D)** OS difference of LGG groups. $p < 0.05$ was regarded significant, with a dashed line of 95% CI.

Analysis of the relationship between FDX1 expression level and prognosis

Using univariate Cox regression analysis, we used data from the TCGA database to assess the correlation between the respective expression levels of FDX1 and OS in various cancers. The hazard ratios of FDX1 to ACC, HNSC, KIRC, and LGG were significant, with FDX1 having the highest risk in LGG, and being a tumor suppressor factor in KIRC (Figure 2). The survival analysis below, using patient data using the median expression value dichotomy for each cancer type (Figure 3), showed that survival differences were significant across OS-related cancer categories and that patients with high FDX1 expression had a better prognosis in some cancers.

However, OS may be influenced by noncancer-related deaths during follow-up. Therefore, the data on the correlation between DSS and FDX1 expression in various cancers were reanalyzed (Figure 4). The Cox regression analysis results were similar to those related to OS. Differences included the determination of a significant risk effect on THYM (except for the four cancers mentioned earlier, HNSCs were excluded for p greater than 0.05) and the calculation inability of the hazard ratio for FDX1 in LAML due to deficient relevant data. Cancer types with high FDX1 expression (KIRC, THYM) showed a favorable

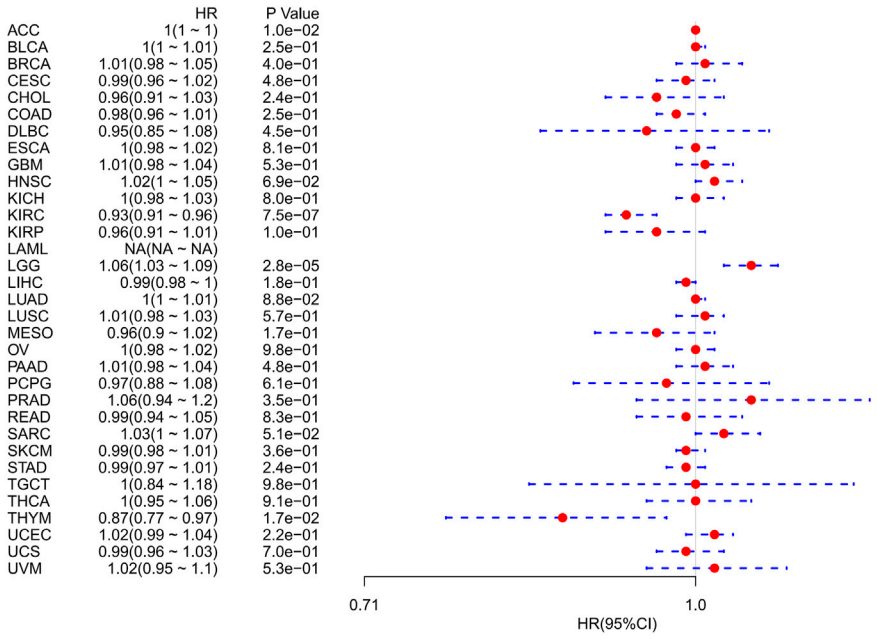


FIGURE 4
Correlation between FDX1 mRNA expression levels and disease-specific survival in multiple tumors from the Cancer Genome Atlas database. Cox regression analysis, $p < 0.05$ was evident.

prognosis compared with the low expression group as learned from the later survival analysis (Figure 5).

Correlation analysis of FDX1 with tumor microenvironment, immune infiltrating cells and immune-related cells in some immune pathways

The relation between FDX1 and immune and stromal scores was measured. We then visualize the remarkable results (Figure 6). As shown, immune scores in 11 of the 33 cancers were significantly associated with FDX1 expression, and ESTIMATEScore scores in 14 of the 33 cancers were significantly associated with FDX1. BRCA ($r = 0.169$, $p < 0.05$), LGG ($r = 0.423$, $p < 0.05$), PCPG ($R = 0.295$, $p < 0.05$), SARC ($r = 0.215$, p r = 0.223, $p < 0.05$) show a positively correlated. The highest correlation coefficient is LGG. In ACC ($r = -0.496$, $p < 0.05$), KIRC ($r = -0.17$, $p < 0.05$), THCA ($r = -0.395$, $p < 0.05$), THYM ($r = -0.232$, $p < 0.05$), UCEC ($r = -0.133$, $p < 0.05$) show a negative correlation. The highest correlation coefficient is ACC (Stepien et al., 2017). The lower the expression of FDX1, the higher the purity of tumor cells in some kinds of cancers.

The correlation between FDX1 and immunoinfiltrating cells in 33 kinds of cancers in the TIMER database was investigated.

FDX1 may modulate the tumor immune microenvironment by affecting immune infiltration in various cancer types

FDX1 expression and levels of immune cell infiltration in each cancer type were correlated to assess whether this pathway affects the tumor's immune microenvironment. Several tumors were found by using six immune cell types (B cells, CD4 + T cells, CD8 + T cells, neutrophils, macrophages, and dendritic cells) available in the TIMER database, derived from TCGA. There is indeed a significant correlation. We picked FDX1 with BRCA, HNSC, KIRC, LGG, STAD, and UCEC. Their corresponding linear regression plots showed that in most tumors, high FDX1 expression was correlated with potentially increased levels of immune cell infiltration. In particular, in STAD, FDX1 expression corresponded negatively with immune cell infiltration levels (Figure 7).

Correlation of FDX1 expression with certain immune checkpoint genes expression in some cancers

Several genes were now closely correlated to and considered checkpoint components in the immune response. The mRNA sequence database allowed assessing whether a link between FDX1 expression and the expression of such checkpoint genes

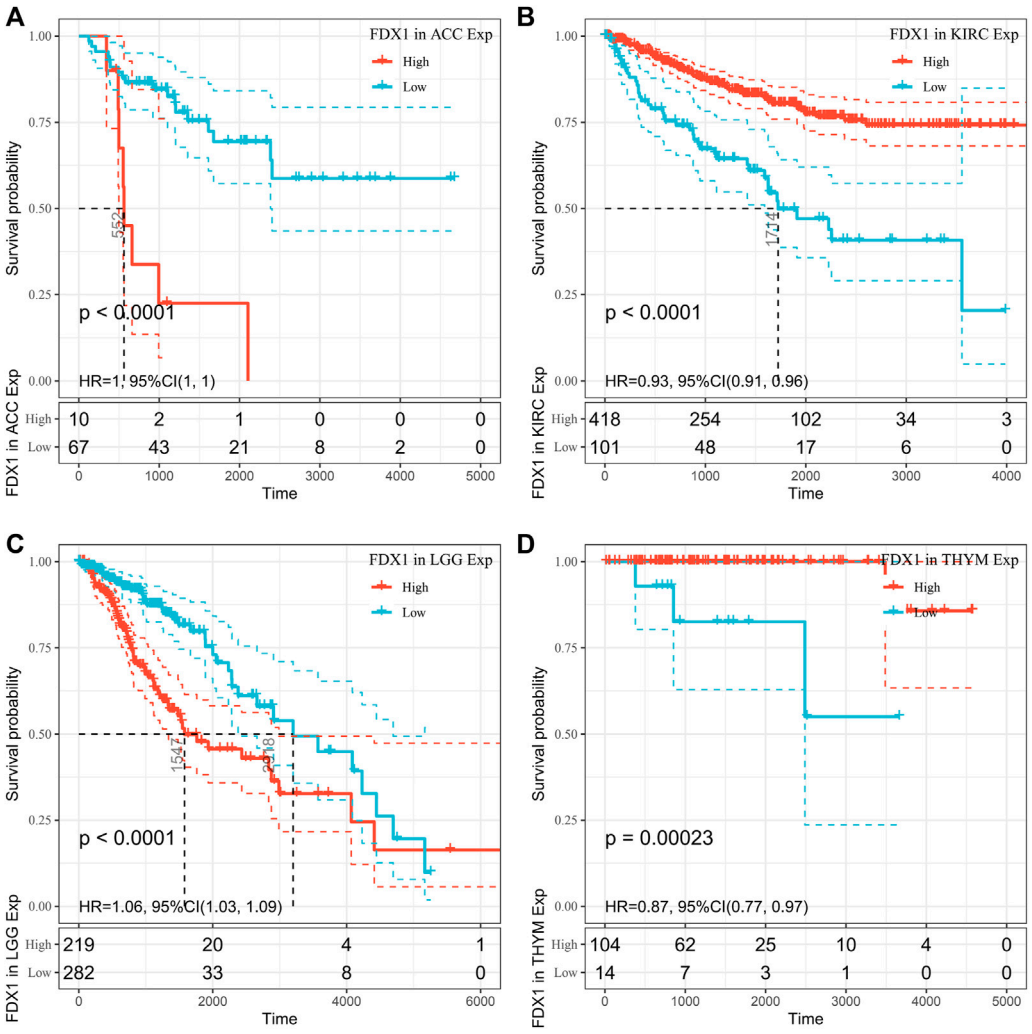


FIGURE 5
Disease-specific survival (DSS) difference between high and low FDX1 mRNA expression groups in significantly prognostically relevant tumors from the Cancer Genome Atlas database (by median expression dichotomy). **(A)** DSS differences between adrenal cortical carcinoma groups. **(B)** DSS differences between KIRC groups. **(C)** DSS differences between groups in LGG. **(D)** DSS differences between thymoma groups. $p < 0.05$ was considered a significant, 95% CI dashed line.

exists. Correlation analysis of FDX1 with checkpoint gene expression found a high correlation ($p < 0.05$) with tumor necrosis factor (TNF)–related immune genes (TNFRSF14, 15, 25) and CTLA4, PDCD1, CD274, NRP1, and VTCN1 in some kinds of cancers.

Moreover, in LGG and TGCT, THCA and THYM, important coexpressions of FDX1 with more immune checkpoint genes, were examined. The results, especially for LGG and TGCT, suggest that FDX1 modulates tumor immune responses by modulating immune checkpoint activity. In addition, in THCA and THYM, FDX1 expression was inversely related with most immune checkpoint molecules but not to a significant extent for some of them (Figure 8).

FDX1 is related to tumor mutational burden and microsatellite instability in some cancers

TMB and MSI are potent prognostic biomarkers and indicators of immunotherapy response in a variety of tumors. Their respective relationships to FDX1 expression in various cancers were examined to investigate the link between FDX1 activity and mutations in specific cancer types. The relation between FDX1 expression and TMB was significant ($p < 0.05$), and data were available for 10 of 32 cancer types (ESCA, HNSC, KIRC, LGG, LUAD, LUSC, STAD, THCA, THYM, and UCEC), among which ESCA, LGG, and STAD coefficients were the highest, whereas KIRC, LUAD, and

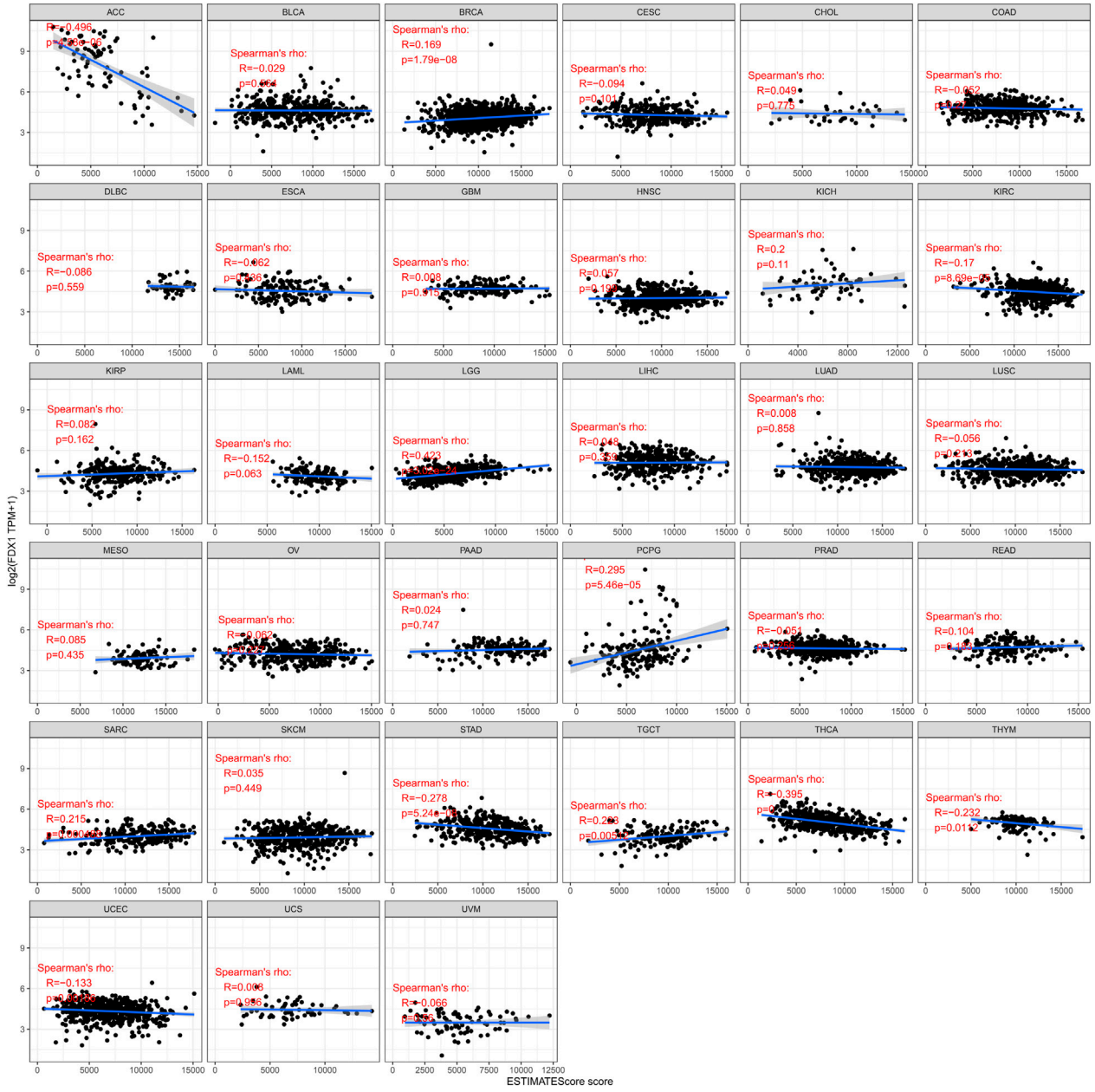


FIGURE 6
Correlation of FDX1 expression with ESTIMATEScore score in pan-cancer.

THCA coefficients were the lowest (Figure 9A). Coefficient values showed that FDX1 expression was positively associated with high mutation status in ESCA, LGG, and STAD but positively correlated to low mutation status in KIRC, LUAD, and THCA (especially THCA).

The relation between FDX1 expression and MSI was examined in 32 cancer types, and the correlation was statistically significant ($p < 0.05$) in nine cancer types (DLBC, HNSC, KIRC, LUAD, LUSC, PAAD, SKCM, STAD, and UCED) (Figure 9B). Among these cancer types, SKCM, PAAD, LUSC, LUAD, FDX1 expression, and MSI had a significant negative correlation, and the PAAD coefficient was the highest; conversely, in DLBC, HNSC, KIRC, STAD, and UCEC, FDX1 expression was positively correlated to MSI, and the DLBC coefficient was the highest. In particular, the STAD cohort had relatively high absolute coefficients associated with either TMB or MSI compared with other kinds of cancers; however, all the quantity of cancer

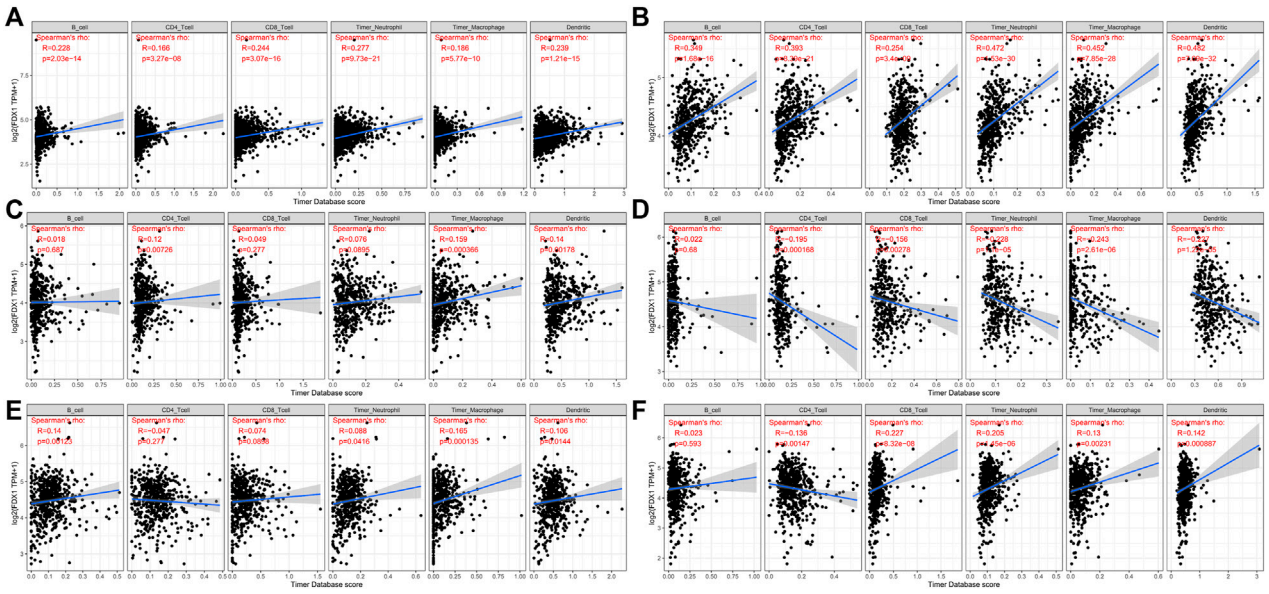


FIGURE 7
Correlation of six immune cell (B cells, CD4 + T cells, CD8 + T cells, neutrophils, macrophages, and dendritic cells) infiltration scores with FDX1 mRNA expression in six kinds of cancer [**(A)**: BRCA, **(B)**: IGG, **(C)**: HNSC, **(D)**: STAD, **(E)**: KIRC, and **(F)**: UCEC]. Spearman correlation test, $p < 0.05$ was significant.

categories showing evident associations with these mutational indicators was lower.

Pan-cancer expression and drug sensitivity

The CellMiner database was used to study the sensitivity of the FDX1 gene to common antitumor drugs and further calculate the correlation between gene expression and the drug IC50. Studies have shown that high expression of the FDX1 gene is associated with resistance to multiple antitumor drugs (Figure 10). Among them, FDX1 was negatively correlated with everolimus, JNJ-42756493, VE-821, AZD-8055, FDX1, MK-2206, avagacestat, and ENMD-2076 precursor and positively correlated with chelerythrine, ifosfamide, ribavirin, PX-316, nelarabine, vorinostat, and amonafide.

Discussion

Previous studies have shown that FDX1 is necessary for the synthesis of kinds of steroid hormones (Sheftel et al., 2010; Strushkevich et al., 2011). Mitochondrial cytochrome P450 is involved in reducing steroid production (Sheftel et al., 2010; Strushkevich et al., 2011). Its associated pathways include metabolic and inflammatory pathways. It has been reported that FDX1 can enhance the copper-dependent cell death induced by elesclomol and can offer new ideas to improve the efficacy of several cancer-targeted drugs. In addition, FDX1 can augment the copper-dependent cell death induced by elesclomol and can offer a new idea to promote the effect of some cancer-targeting agents (Tsvetkov et al., 2019). Current studies have shown that FDX1 is a key gene that promotes copperoptosis. Copperoptosis is a newly discovered mode of regulatory cell death, varying from other regulatory cell death characteristics such as pyroptosis, ferroptosis, and apoptosis. The relationship between copperoptosis and tumors: It has been found that patients with different cancers (such as breast cancer, thyroid cancer, cervical cancer, ovarian cancer, lung cancer, pancreatic cancer, prostate cancer, breast cancer, oral cancer, and bladder cancer) have serum and tumor tissue copper content that is significantly changed (Baltaci et al., 2017) (Stepien et al., 2017). Copper also promotes angiogenesis, which is critical for tumor progression and metastasis (Ruiz et al., 2021; Ge et al., 2022). Overloaded copper can also lead to cell death. Since copper is important for the occurrence and progression of cancer, it is of great biological significance to study genes related to copperoptosis (Shanbhag et al., 2021).

Our findings suggest that FDX1 is widely expressed in different normal tissues and is relatively high in the adrenal gland. When tumors were compared with corresponding normal tissues, FDX1 expression was reduced in various cancers, and this high expression was associated with better OS and death-specific survival in some cancer types, such as KIPC.

10.3389/fgene.2022.962028

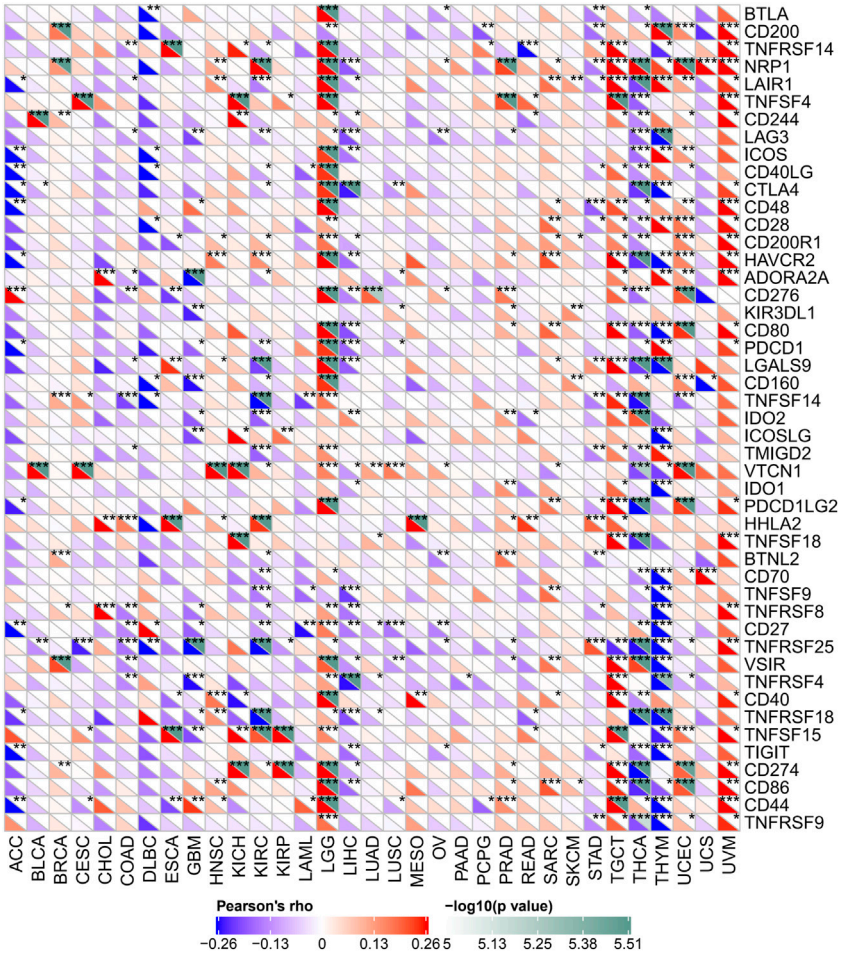


FIGURE 8
Relation between FDX1 mRNA expression levels and mRNA expression at recognized immune checkpoints in multiple tumors from the Cancer Genome Atlas database. The lower triangle refers to the coefficients calculated by Pearson's correlation test, and the upper triangle represents the *p*-value converted by log10. **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

Tumor cells can change the nature of the microenvironment, which in turn can influence tumor growth and spread. Immune cells and stromal cells in the tumor microenvironment can affect cancer prognosis and patient survival outcomes (Ren et al., 2018). TME is strongly associated with tumor occurrence and metastasis (Spill et al., 2016; Liu et al., 2022b; Ye et al., 2022). Previous studies have shown that cytokines in the tumor microenvironment regulate immune function and ultimately suppress immune responses, leading to tumor progression (Hinshaw and Shevde, 2019). Tumor-infiltrating lymphocytes (TILs) in TME have been shown to be independent predictors of prognosis and immunotherapy efficacy in cancer patients (Ohtani, 2007; Azimi et al., 2012). Both immune cells and stromal cells are contained in the tumor environment, and they can determine the role of TME to some extent. Besides, it is reported that immune cells are evidently related to tumorigenesis and development in many researches. Thus, components analyzed in TME contribute to the development of targeted drugs for tumor immunotherapy. We found that FDX1 expression was apparently positively associated with immune cell infiltration in most tumors, whereas in STAD, FDX1 expression was negatively correlated with immune infiltration. In particular, FDX1 expression is also associated with the statistically significant presence of some specific immune checkpoint genes in multiple tumors, such as CTLA4, PDCD1, CD274, NRP1, and VTCN1. Upregulation of this checkpoint gene is associated with escape mechanisms in the immune microenvironment, which suggested that FDX1 plays a role in different immunomodulatory effects in various cancer types.

Our study also found that FDX1 was positively correlated to TME immune, stromal, and ESTIMATE scores in most human cancer types. In addition, the association of FDX1 with TMB and MSI also proves that FDX1 is strongly associated with TME in

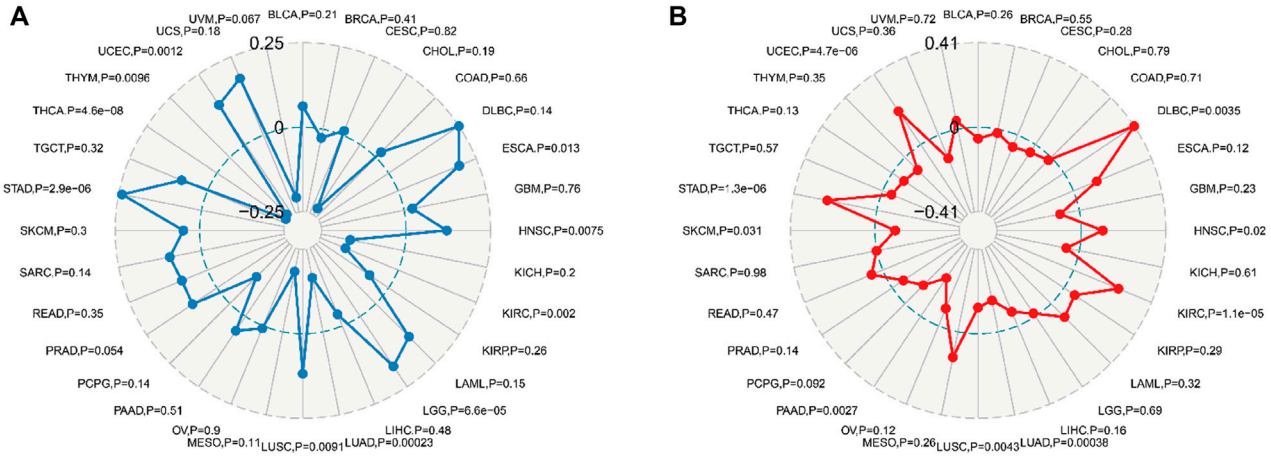


FIGURE 9
Relationship between tumor mutational burden (TMB), microsatellite instability (MSI), and FDX1 mRNA expression levels in different tumors in the Cancer Genome Atlas database. TMB was calculated by the total mutation incidence per million base pairs in each tumor, and MSI was calculated by the total incidence of deletions or insertions per million base pairs of repeats. **(A)** correlation between TMB and FDX1 expression. **(B)** correlation between MSI and FDX1 expression. Spearman correlation test, $p < 0.05$ was evident.

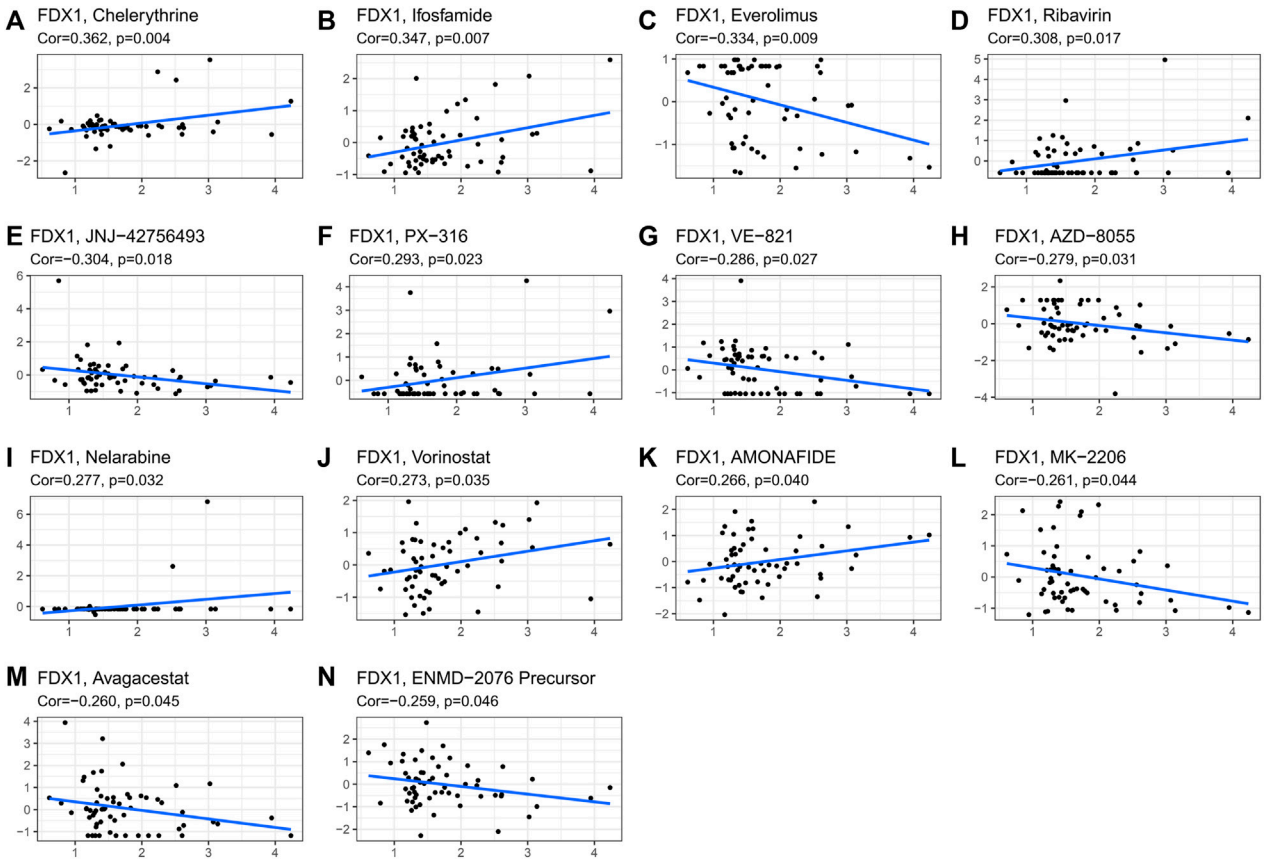


FIGURE 10
Correlation graph with drug IC50. The correlation graph of gene and drug IC50 and the slope of the straight line are the correlation coefficients between gene and drug. [**(A)**: chelerythrine, **(B)**: ifosfamide, **(C)**: everolimus, **(D)**: ribavirin, **(E)**: JNJ-42756493, **(F)**: PX-316, **(G)**: VE-821, **(H)**: AZD-8055, **(I)**: nelarabine, **(J)**: vorinostat, **(K)**: AMONAFIDE, **(L)**: MK-2206, **(M)**: avagacestat, and **(N)**: ENMD-2076 precursor].

human cancer. Previous studies had demonstrated that TMB and MSI were markers of drug response in patients, especially those targeting immune checkpoint inhibitors, such as CTLA4 or PD-1/PD-L1 inhibitors (Overman et al., 2017; Mariathasan et al., 2018; Kwon et al., 2020; Shim et al., 2020). In gastric cancer, an analysis of the MAGIC study showed that MSI-H patients might have worse OS after perioperative treatment. Patients with MSI-H/dMMR (deficiency of MMR, dMMR) had many tumor mutations and a wide range of immunogenicity, so they responded well to PD-1/PD-L1 inhibitors. Here, both TMB and MSI of STAD were positively correlated to FDX1 expression, which would support our claim that, of course, FDX1 might be indicating potential drug response (and MSI) well in STAD.

Using the CellMiner study, the result that high expression of the FDX1 gene was associated with resistance to multiple antitumor drugs was obtained. Among them, FDX1 was negatively correlated with everolimus, JNJ-42756493, VE-821, AZD-8055, FDX1, MK-2206, avagacestat, and ENMD-2076 precursor and positively correlated with chelerythrine, ifosfamide, ribavirin, PX-316, nelarabine, vorinostat, and amondafide. We found that FDX12 could serve as a potential resistance target that could predict tumor cell susceptibility to chemotherapy drugs.

Although our study provides useful indications that FDX1 is involved in tumorigenesis and regulation of the immune environment of tumor cells, it does have some limitations. First, as a pure bioinformatics analysis, it relies entirely on information available in open access databases and has not been confirmed experimentally. Here, the assessment of FDX1 expression was based solely on the mRNA levels reported in the aforementioned database, although this cannot show functional protein levels. For instance, protein activity in normal or cancer cells may be affected by posttranscriptional modifications and/or regulatory proteolysis. Future studies will focus on experimentally the data validation and exploration of possible mechanisms of FDX1 in tumorigenesis. Second, we have shown that in the link between FDX1 expression and TMB, MSI lacks any mechanistic explanation from supporting experimental data. More experimental evidence is needed to prove this.

Conclusion

FDX1 is highly expressed in a variety of tumors, and this high expression is associated with better survival and disease progression, especially for KIRC. FDX1 expression is also associated with immune cell infiltration of tumors, immune checkpoint gene expression, and immunotherapy markers (e.g., TMB and MSI). Taken together, the data suggest that FDX1 provides a valuable new biomarker for several cancers for assessing prognosis and immunotherapy response.

Data availability statement

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

Author contributions

ZH and SH conceived and designed the study. CX and LJ analyzed the data. WL, LY, and FZ analyzed the data. CX wrote the manuscript. All authors have read and approved this manuscript.

Acknowledgments

The authors thank the reviewers for their helpful comments on the manuscript.

Conflict of interest

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

Publisher's note

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

References

Arneth, B. (2019). Tumor microenvironment. *Med. Kaunas. Lith.* 56, E15. doi:10.3390/medicina56010015

Azimi, F., Scolyer, R. A., Rumcheva, P., Moncrieff, M., Murali, R., McCarthy, S. W., et al. (2012). Tumor-infiltrating lymphocyte grade is an independent predictor of sentinel lymph node status and survival in patients with cutaneous melanoma. *J. Clin. Oncol.* 30, 2678–2683. doi:10.1200/JCO.2011.37.8539

Baltaci, A. K., Dundar, T. K., Aksoy, F., and Mogulkoc, R. (2017). Changes in the serum levels of trace elements before and after the operation in thyroid cancer patients. *Biol. Trace Elem. Res.* 175, 57–64. doi:10.1007/s12011-016-0768-2

Blum, A., Wang, P., and Zenklusen, J. C. (2018). SnapShot: TCGA-analyzed tumors. *Cell* 173, 530. doi:10.1016/j.cell.2018.03.059

Ge, E. J., Bush, A. I., Casini, A., Cobine, P. A., Cross, J. R., DeNicola, G. M., et al. (2022). Connecting copper and cancer: From transition metal signalling to metalloplasia. *Nat. Rev. Cancer* 22, 102–113. doi:10.1038/s41568-021-00417-2

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., et al. (2020). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta. Mol. Basis Dis.* 1866, 165916. doi:10.1016/j.bbadis.2020.165916

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: A novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Hinshaw, D. C., and Shevde, L. A. (2019). The tumor microenvironment innately modulates cancer progression. *Cancer Res.* 79, 4557–4566. doi:10.1158/0008-5472.CAN-18-3962

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi:10.1038/nature08987

Hwu, W.-L., Muramatsu, S.-I., and Gidoni-Ben-Zeev, B. (2021). Reduced immunogenicity of intraparenchymal delivery of adeno-associated virus serotype 2 vectors: Brief overview. *Curr. gene Ther.* 22 (3), 185–190. doi:10.2174/1566523221666210922155413

Kwon, M., Hong, J. Y., Kim, S. T., Kim, K. M., and Lee, J. (2020). Association of serine/threonine kinase 11 mutations and response to programmed cell death 1 inhibitors in metastatic gastric cancer. *Pathol. Res. Pract.* 216, 152947. doi:10.1016/j.prp.2020.152947

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. doi:10.1186/s13059-016-1028-7

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, J., Lan, Y., Tian, G., and Yang, J. (2022). A systematic framework for identifying prognostic genes in the tumor microenvironment of colon cancer. *Front. Oncol.* 2, 899156. doi:10.3389/fonc.2022.899156

Liu, X., Yuan, P., Li, R., Zhang, D., An, J., Ju, J., et al. (2022). Predicting breast cancer recurrence and metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. *Comput. Biol. Med.* 146, 105569. doi:10.1016/j.compbiomed.2022.105569

Lv, Z., Qi, L., Hu, X., Mo, M., Jiang, H., Fan, B., et al. (2021). Zic family member 2 (ZIC2): A potential diagnostic and prognostic biomarker for pan-cancer. *Front. Mol. Biosci.* 8, 631067. doi:10.3389/fmolb.2021.631067

Mariathasan, S., Turley, S. J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., et al. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* 554, 544–548. doi:10.1038/nature25501

Ohtani, H. (2007). Focus on TILs: Prognostic significance of tumor infiltrating lymphocytes in human colorectal cancer. *Cancer Immun.* 7, 4.

Overman, M. J., McDermott, R., Leach, J. L., Lonardi, S., Lenz, H. J., Morse, M. A., et al. (2017). Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): An open-label, multicentre, phase 2 study. *Lancet. Oncol.* 18, 1182–1191. doi:10.1016/S1470-2045(17)30422-9

Ren, B., Cui, M., Yang, G., Wang, H., Feng, M., You, L., et al. (2018). Tumor microenvironment participates in metastasis of pancreatic cancer. *Mol. Cancer* 17, 108. doi:10.1186/s12943-018-0858-1

Ruiz, L. M., Libedinsky, A., and Elorza, A. A. (2021). Role of copper on mitochondrial function and metabolism. *Front. Mol. Biosci.* 8, 711227. doi:10.3389/fmolb.2021.711227

Shanbhag, V. C., Gudekar, N., Jasmer, K., Papageorgiou, C., Singh, K., Petris, M. J., et al. (2021). Copper metabolism as a unique vulnerability in cancer. *Biochimica Biophysica Acta - Mol. Cell Res.* 1868, 118893. doi:10.1016/j.bbamcr.2020.118893

Sheftel, A. D., Stehling, O., Pierik, A. J., Elsässer, H. P., Mühlenhoff, U., Webert, H., et al. (2010). Humans possess two mitochondrial ferredoxins, Fdx1 and Fdx2, with distinct roles in steroidogenesis, heme, and Fe/S cluster biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11775–11780. doi:10.1073/pnas.1004250107

Shim, J. H., Kim, H. S., Cha, H., Kim, S., Kim, T. M., Anagnostou, V., et al. (2020). HLA-corrected tumor mutation burden and homologous recombination deficiency for the prediction of response to PD-(L)1 blockade in advanced non-small-cell lung cancer patients. *Ann. Oncol.* 31, 902–911. doi:10.1016/j.annonc.2020.04.004

Spill, F., Reynolds, D. S., Kamm, R. D., and Zaman, M. H. (2016). Impact of the physical microenvironment on tumor progression and metastasis. *Curr. Opin. Biotechnol.* 40, 41–48. doi:10.1016/j.copbio.2016.02.007

Stepien, M., Jenab, M., Freisling, H., Becker, N. P., Czuban, M., Tjønneland, A., et al. (2017). Pre-diagnostic copper and zinc biomarkers and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition cohort. *Carcinogenesis* 38, 699–707. doi:10.1093/carcin/bgx051

Strushkevich, N., MacKenzie, F., Cherkesova, T., Grabovec, I., Usanov, S., Park, H. W., et al. (2011). Structural basis for pregnenolone biosynthesis by the mitochondrial monooxygenase system. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10139–10143. doi:10.1073/pnas.1019441108

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. A Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tang, D., Chen, X., and Kroemer, G. (2022). Cuproptosis: A copper-triggered modality of mitochondrial cell death. *Cell Res.* 32, 417–418. doi:10.1038/s41422-022-00653-7

Tavasolian, F., Hosseini, A. Z., Soudi, S., and Naderi, M. (2020). miRNA-146a improves immunomodulatory effects of MSC-derived exosomes in rheumatoid arthritis. *Curr. Gene Ther.* 20, 297–312. doi:10.2174/1566523220666200916120708

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi:10.5114/wo.2014.47136

Tsvetkov, P., Coy, S., Petrova, B., Dreishpoon, M., Verma, A., Abdusamad, M., et al. (2022). Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science* 375, 1254–1261. doi:10.1126/science.abf0529

Tsvetkov, P., Detappe, A., Cai, K., Keys, H. R., Brune, Z., Ying, W., et al. (2019). Mitochondrial metabolism promotes adaptation to proteotoxic stress. *Nat. Chem. Biol.* 15, 681–689. doi:10.1038/s41589-019-0291-9

Wang, Y., Zhang, L., and Zhou, F. (2022). Cuproptosis: A new form of programmed cell death. *Cell Mol. Immunol.* doi:10.1038/s41423-022-00866-1

Yang, J., Hagen, J., Guntur, K. V., Allette, K., Schuyler, S., Ranjan, J., et al. (2017). A next generation sequencing based approach to identify extracellular vesicle mediated mRNA transfers between cells. *BMC Genomics* 18, 987. doi:10.1186/s12864-017-4359-1

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

Ye, Z., Zhang, Y., Liang, Y., Lang, J., Zhang, X., Zang, G., et al. (2022). Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Curr. Bioinform.* 17, 164–173. doi:10.2174/1574893616666210708143556

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zhang, Z., Ma, Y., Guo, X., Du, Y., Zhu, Q., Wang, X., et al. (2021). FDX1 can impact the prognosis and mediate the metabolism of lung adenocarcinoma. *Front. Pharmacol.* 12, 749134. doi:10.3389/fphar.2021.749134

Zhao, T., Hu, Y., and Cheng, L. (2021). Deep-DRM: A computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 22, bbaa212. doi:10.1093/bib/bbaa212

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 31 August 2022
doi: 10.3389/fgene.2022.934196

A Prognostic Ferroptosis-Related lncRNA Model Associated With Immune Infiltration in Colon Cancer

*Jianzhong Lu, Jinhua Tan and Xiaoqing Yu**

School of Science, Shanghai Institute of Technology, Shanghai, China

Colon cancer (CC) is a common malignant tumor worldwide, and ferroptosis plays a vital role in the pathology and progression of CC. Effective prognostic tools are required to guide clinical decision-making in CC. In our study, gene expression and clinical data of CC were downloaded from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. We identified the differentially expressed ferroptosis-related lncRNAs using the differential expression and gene co-expression analysis. Then, univariate and multivariate Cox regression analyses were used to identify the effective ferroptosis-related lncRNAs for constructing the prognostic model for CC. Gene set enrichment analysis (GSEA) was conducted to explore the functional enrichment analysis. CIBERSORT and single-sample GSEA were performed to investigate the association between our model and the immune microenvironment. Finally, three ferroptosis-related lncRNAs (XXbac-B476C20.9, TP73-AS1, and SNHG15) were identified to construct the prognostic model. The results of the validation showed that our model was effective in predicting the prognosis of CC patients, which also was an independent prognostic factor for CC. The GSEA analysis showed that several ferroptosis-related pathways were significantly enriched in the low-risk group. Immune infiltration analysis suggested that the level of immune cell infiltration was significantly higher in the high-risk group than that in the low-risk group. In summary, we established a prognostic model based on the ferroptosis-related lncRNAs, which could provide clinical guidance for future laboratory and clinical research on CC.

Keywords: colon cancer, ferroptosis, long non-coding RNA, prognostic model, immune microenvironment

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd., China

Reviewed by:
Weiwei Zhang,
East China University of Technology, China
Yan Yang,
Beijing Genomics Institute (BGI), China

****Correspondence:***
Xiaoqing Yu
xqyu@sit.edu.cn

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

Received: *02 May 2022*
Accepted: *13 June 2022*
Published: *31 August 2022*

Citation:
Lu J, Tan J and Yu X (2022) A Prognostic Ferroptosis-Related lncRNA Model Associated With Immune Infiltration in Colon Cancer. Front. Genet. 13:934196. doi: 10.3389/fgene.2022.934196

INTRODUCTION

Colon cancer (CC) has the third most incidences among malignancies, and it is the second most common cause of cancer death in men and women combined (Siegel et al., 2022). The malignant transformation of CC is a multistep process that takes approximately ten years from small clumps to CC (Jemal et al., 2011). Therefore, early diagnosis is essential for improving the prognosis of CC patients. However, the survival of CC patients is poor because of the complexity of the disease, late disease detection, and lack of reliable risk-assessment biomarkers (Lin et al., 2020; Yang C. et al., 2021). Even after treatment, the risk of recurrence and metastasis in CC patients is still high (Chang et al., 2020; Jin et al., 2020). In recent years, more studies have suggested that it is promising to solve the problem by integrating computational techniques with big biomedical data involving multiple types of biomarkers including epigenetic, genetic, and gene expression profiles (Yang Y. et al., 2021; Liu et al., 2021). Therefore, identifying effective

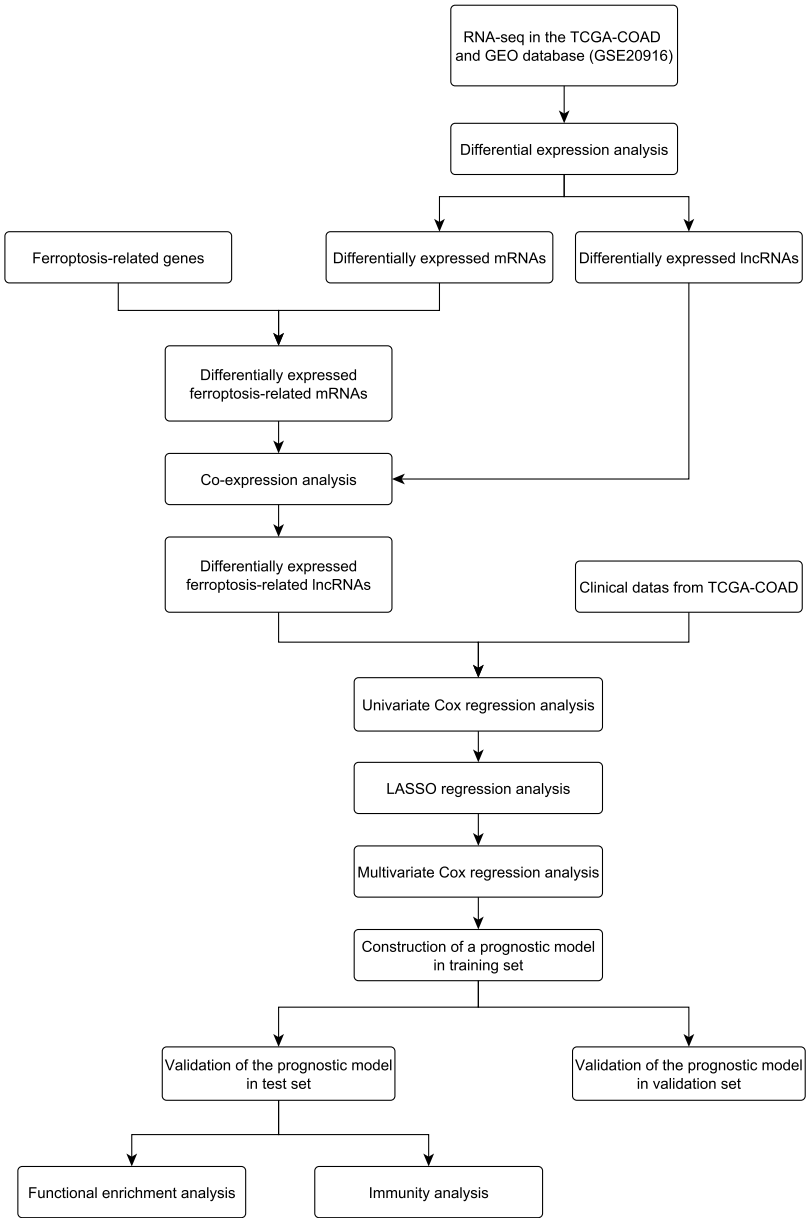


FIGURE 1 | Flowchart of this study.

biomarkers to establish a prognostic model for survival prediction is gaining increasing attention.

lncRNAs are non-protein coding transcripts over 200 nucleotides in length (Mercer et al., 2009). There are more than 50,000 lncRNA genes annotated in the human genome (Borkiewicz et al., 2021). Studies have shown that lncRNAs are often dysregulated during tumorigenesis, which might cause tumor development (Prensner and Chinnaiyan, 2011; Schmitt and Chang, 2016). Therefore, they are used as molecular biomarkers to diagnose and treat many diseases, including CC. For example, Zhou et al. (2019) revealed that lncRNA XIRP2-AS1 has a favorable impact on the overall survival of patients with colon cancer. Tsai et al. (2018) found that lncRNA Linc00659 expression knockdown could accelerate cell apoptosis in CC cells treated with chemotherapy drugs.

Ferroptosis is a newly discovered form of programmed cell death characterized by iron-dependent accumulation of lethal lipid peroxidation (Tang et al., 2018; Mou et al., 2019). Cancer cells are vulnerable to ferroptosis because of their high iron uptake to support fast proliferation (Hassannia et al., 2019). Recently, studies have demonstrated that ferroptosis plays a crucial role in tumorigenesis and cancer therapeutics. Wang et al. (2021) constructed a ferroptosis-related prognostic signature for LUAD and suggested that ferroptosis is a functional and therapeutic target in LUAD. He et al. (2021)

TABLE 1 | Characteristics of CC patients in our study.

Characteristic	Training set (*n* = 185)	Test set (*n* = 185)	GSE72970 (*n* = 124)	GSE17536 (*n* = 177)
Age (years)				
<70	96	107	90	104
≥70	89	78	34	73
Gender				
Female	85	86	50	81
Male	100	99	74	96
T stage				
T1	5	4	1	—
T2	32	33	7	—
T3	134	121	50	—
T4	14	27	37	—
TX	—	—	29	—
N stage (pN)				
N0	118	101	14	—
N1	36	51	28	—
N2	31	33	53	—
NX	—	—	29	—
M stage				
M0	160	150	22	—
M1	25	35	102	—
TNM stage				
I	33	32	0	24
II	82	63	6	57
III	45	55	15	57
IV	25	35	102	39
X	—	—	1	—

have constructed a prognostic risk model based on 10 genes related to ferroptosis and identified potential novel therapeutic targets which improve the individualized treatment of patients with HNSCC. Moreover, considering the critical role of ferroptosis in cancer, many studies proposed ferroptosis-based strategies to identify potential lncRNA biomarkers associated with various cancers. For example, Guo et al. (2021) revealed that ferroptosis-related lncRNAs have the potential to inform immunological research and treatment. Wei et al. (2021) identified that ferroptosis-related lncRNAs have an important prognostic value in gastric cancer. Feng et al. (2022) suggested that ferroptosis and iron metabolism–related lncRNAs can independently predict the overall survival and therapeutic effect in patients with ovarian cancer. Currently, many prognostic models have been proposed based on the ferroptosis-related lncRNAs for colon cancer (Cai et al., 2021; Zhang et al., 2021). However, the functional mechanisms of the ferroptosis-related lncRNAs and the relationship between the prognostic model and the tumor immune microenvironment require further investigation for CC patients.

In this study, three ferroptosis-related lncRNAs were identified as the prognostic biomarkers for CC. The prognostic model based on the ferroptosis-related lncRNAs was constructed for predicting the overall survival of CC patients, which would provide prognostic insights into anticancer therapies and a novel source for immune therapies. The workflow of this study is shown in **Figure 1**.

MATERIALS AND METHODS

Data Collection

In this study, we selected four independent datasets from two different high-throughput platforms, including 458 colon adenocarcinoma (COAD) samples and 41 normal samples from TCGA (https://portal.gdc.cancer.gov/); 111 CC samples, 34 normal samples (GSE20916), 124 colorectal cancer samples (GSE72970), and 177 CC samples (GSE17536) from the GEO (https://www.ncbi.nlm.nih.gov/geo/). The gene expression profiling of the three datasets (GSE20916, GSE72970, and GSE17536) was based on the GPL570 platform. Patients with a survival time of more than 30 days were used for the survival analysis. The detailed clinical characteristics of the patients are shown in **Table 1**. We downloaded 259 ferroptosis-related genes from the FerrDb database (Zhou and Bao, 2020), including 108 driver genes, 69 suppressor genes, and 111 marker genes (**Supplementary Table S1**).

Identification of Differentially Expressed Ferroptosis-Related lncRNAs

In this study, we identified mRNAs and lncRNAs using the Ensembl database (http://ensemblgenomes.org). The expression profile of mRNAs and lncRNAs was extracted from RNA-seq count data, which was normalized using the edgeR package (version 3.32.1). Differentially expressed mRNAs and lncRNAs shared by TCGA-COAD and GSE20916 were identified using the

edgeR and limma R packages [$|\log_2(FoldChange)| > 1$ and $p < 0.05$]. The intersection between the differentially expressed mRNAs (DEmRNAs) and the 259 ferroptosis-related genes was defined as differentially expressed ferroptosis-related mRNAs (DEFR-mRNAs). We constructed the co-expression network with the DEFR-mRNAs and the differentially expressed lncRNAs (DElncRNAs) based on the Pearson correlation analysis to identify the differentially expressed ferroptosis-related lncRNAs (DEFR-lncRNAs). In the co-expression network, the DElncRNAs with $|R^2| > 0.4$ and $p < 0.001$ remained as the DEFR-lncRNAs.

Construction of a DEFR-lncRNA Prognostic Model

Univariate Cox regression analysis was first performed by integrating the gene expression matrix of the DEFR-lncRNAs and the survival data in TCGA-COAD to identify the DEFR-lncRNAs with prognostic relevance for the overall survival (OS). Statistically significant value was set at $p < 0.05$. Moreover, the least absolute shrinkage and selection operator (LASSO) regression analysis was used to avoid overfitting and build a reliable and robust model. Next, the screened DEFR-lncRNAs were validated using the multivariate Cox regression analysis, and the DEFR-lncRNAs associated with the prognosis of CC were obtained. Finally, the prognostic risk score (RS) model was constructed for each patient, which was calculated as follows:

$$\boldsymbol{RS} = \sum_{i=1}^{n} \left[expr(lncRNA_i) \times coef(lncRNA_i) \right],$$

where $expr(lncRNA_i)$ is the gene expression value of $lncRNA_i$, and $coef(lncRNA_i)$ is the corresponding estimated regression coefficient in the multivariate Cox regression analysis.

Enrichment Analysis

Gene set enrichment analysis (GSEA) (http://www.broad.mit.edu/gsea/) is a computational method used to identify whether a pre-defined set of genes shows significant differences between two biological states (Subramanian et al., 2005). GSEA was performed by GSEA software (version 4.2.3). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Hallmark pathways were used to explore the potential pathways and gene sets associated with the model. They were visualized using the ggplot2 R package.

Immunity Analysis

CIBERSORT (https://cibersort.stanford.edu/) is an established computational resource to estimate the abundance of member cell types in a mixed cell population (Newman et al., 2015). In our study, we applied the CIBERSORT algorithm to assess the tumor infiltration levels of 22 immune cell types from the CC patients in TCGA-COAD. It was run using the LM22 signature with 1,000 permutations to estimate the relative fractions of the 22 immune cell types. Moreover, the single-sample gene set enrichment analysis (ssGSEA) was also performed, and 28 immune cell types that are over-represented in the tumor microenvironment were analyzed to understand the association between the prognostic model and immune infiltration (Charoentong et al., 2017).

Statistical Analysis

All statistical analyses were conducted by R software (Version 4.0.2). Univariate Cox regression analysis, LASSO regression analysis, and multivariate Cox regression analysis were performed to identify the DEFR-lncRNAs associated with the prognosis of CC patients. The Kaplan–Meier survival analysis and log-rank test were used to conduct survival analysis. The timeROC R package was used to draw receiver operating characteristic (ROC) curves and quantify the area under the curve (AUC) values. The GSVA R package was used for the ssGSEA.

RESULTS

Identification of Differentially Expressed Ferroptosis-Related lncRNAs

In our study, using the gene type data reported for the genome GRCh38.p13, 19,674 mRNAs and 14,826 lncRNAs were downloaded from TCGA-COAD, and 12,001 mRNAs and 370 lncRNAs were downloaded from GSE20916. The differential expression analysis showed that 4,876 mRNAs and 1,671 lncRNAs were differentially expressed in TCGA-COAD, and 1,370 mRNAs and 44 lncRNAs were differentially expressed in GSE20916. The volcano plots of DEmRNAs and DElncRNAs of TCGA-COAD and GSE20916 are shown in **Figures 2A,B**, respectively. Moreover, 1,157 DEmRNAs and 34 DElncRNAs shared by the two databases were obtained (**Figure 2C**). Then, 30 DEFR-mRNAs were obtained after intersecting 1,157 DEmRNAs and 259 ferroptosis-related genes (**Figure 2D**). Finally, 29 DEFR-lncRNAs were identified using the co-expression analysis, which was shown in the co-expression network (**Figure 2E**).

Construction of a Prognostic Model Based on DEFR-lncRNAs

Based on the 29 DEFR-lncRNAs, we identified five DEFR-lncRNAs (SNHG17, XXbac-B476C20.9, TP73-AS1, SNHG15, and PVT1) that were statistically related to the OS of CC patients using the univariate Cox regression analysis ($p < 0.05$, **Figure 3A**). Then, the five DEFR-lncRNAs were subjected to the LASSO regression analysis. As the values of λ increased, the LASSO coefficients of these five lncRNAs decreased to zero (**Figure 3B**). Moreover, the partial likelihood deviances of different numbers of lncRNAs were revealed by the LASSO regression model, which showed that the model had an optimal performance with the least parameters when $\log(\lambda) = -4.035622$ (**Figure 3C**). Subsequently, the multivariate Cox regression analysis was performed, and three DEFR-lncRNAs (XXbac-B476C20.9, TP73-AS1, and SNHG15) were selected as the prognostic DEFR-lncRNAs for constructing the prognostic model ($p < 0.05$, **Figure 3D**).

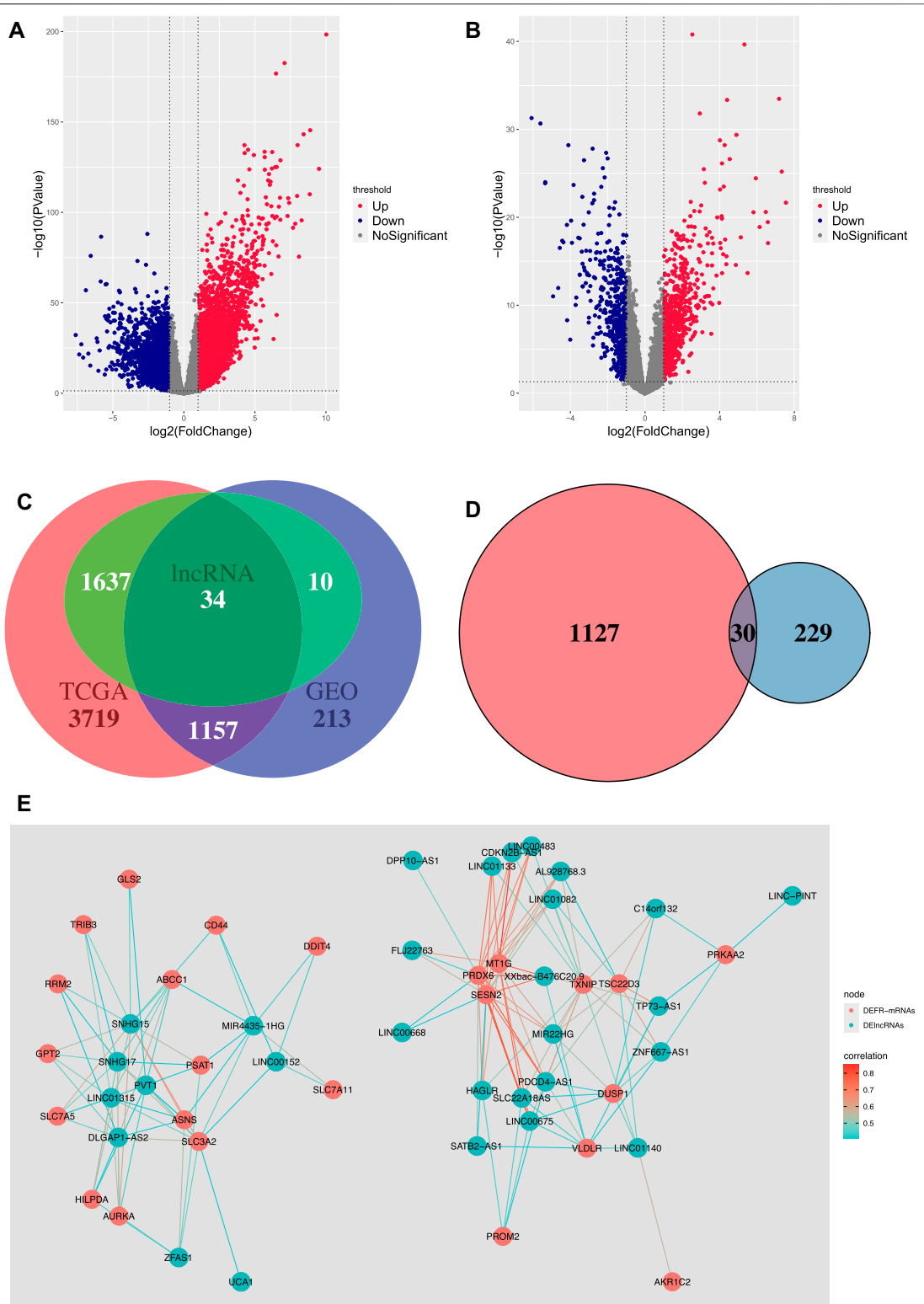


FIGURE 2 | Identification of DEFR-lncRNAs. **(A)** Volcano plot of DEmRNAs and DElncRNAs in TCGA-COAD. **(B)** Volcano plot of DEmRNAs and DElncRNAs in GSE20916. **(C)** Venn diagram of DEmRNAs and DElncRNAs in TCGA-COAD and GSE20916. **(D)** Venn diagram of the shared DEmRNAs and ferroptosis-related genes. Red represents the shared DEmRNAs between TCGA-COAD and GSE20916, and blue represents the ferroptosis-related genes. **(E)** Co-expression network between DEFR-mRNAs and DElncRNAs.

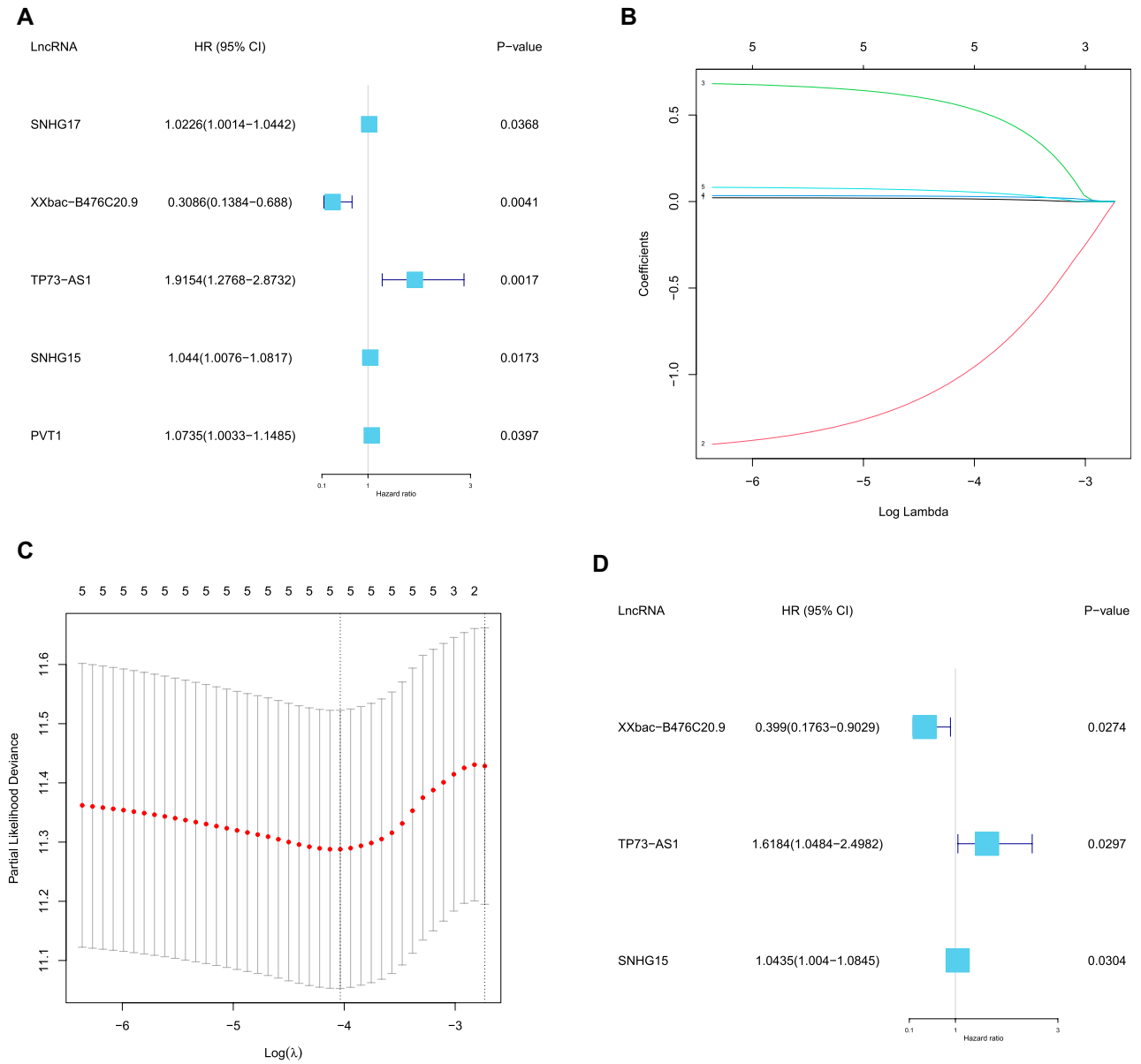


FIGURE 3 | Identification of the prognostic DEFR-lncRNAs. **(A)** Forest map of five DEFR-lncRNAs identified by univariate Cox regression analysis. **(B)** LASSO coefficient profile of the five DEFR-lncRNAs. **(C)** Plots of the cross-validation error rates. **(D)** Forest map of three DEFR-lncRNAs identified by the multivariate Cox regression analysis.

After filtering patients with incomplete gene expression data and clinical information, 370 patients in TCGA-COAD remained in our study, who were divided randomly into the training set and the test set in a 1:1 ratio. The prognostic model was constructed based on the three prognostic DEFR-lncRNAs in the training set. The RS was calculated for each patient using the following equation: $RS = -2.1053 \times expr(\text{XXbac} - \text{B476C20.9}) + 0.6008 \times expr(\text{TP73} - \text{AS1}) + 0.0873 \times expr(\text{SNHG15})$. Patients were classified into high-risk and low-risk groups in the training, test, and whole sets. The cutoff values for the three datasets were the median RS in the training set ($RS = -0.291257$). We observed that the proportion of patients with CC in the high-risk group was significantly higher than that of the low-risk group in the training, test, and whole sets, respectively (**Figures 4A–C**). We also investigated the expression of the three prognostic DEFR-lncRNAs in the high-risk and low-risk groups (**Figures 4D,E**). In the whole set, we can find that the lncRNA XXbac-B476C20.9 was higher expressed in the low-risk group, while the lncRNAs TP73-AS1 and SNHG15 were higher expressed in the high-risk group (**Figure 4F**).

Kaplan–Meier survival curves were plotted to compare the difference in the OS between the high-risk and low-risk groups, which indicated that the patients in the low-risk group had better OS than those in the high-risk group in the training, test, and whole sets (**Figures 5A–C**). Moreover, time-dependent ROC curves were

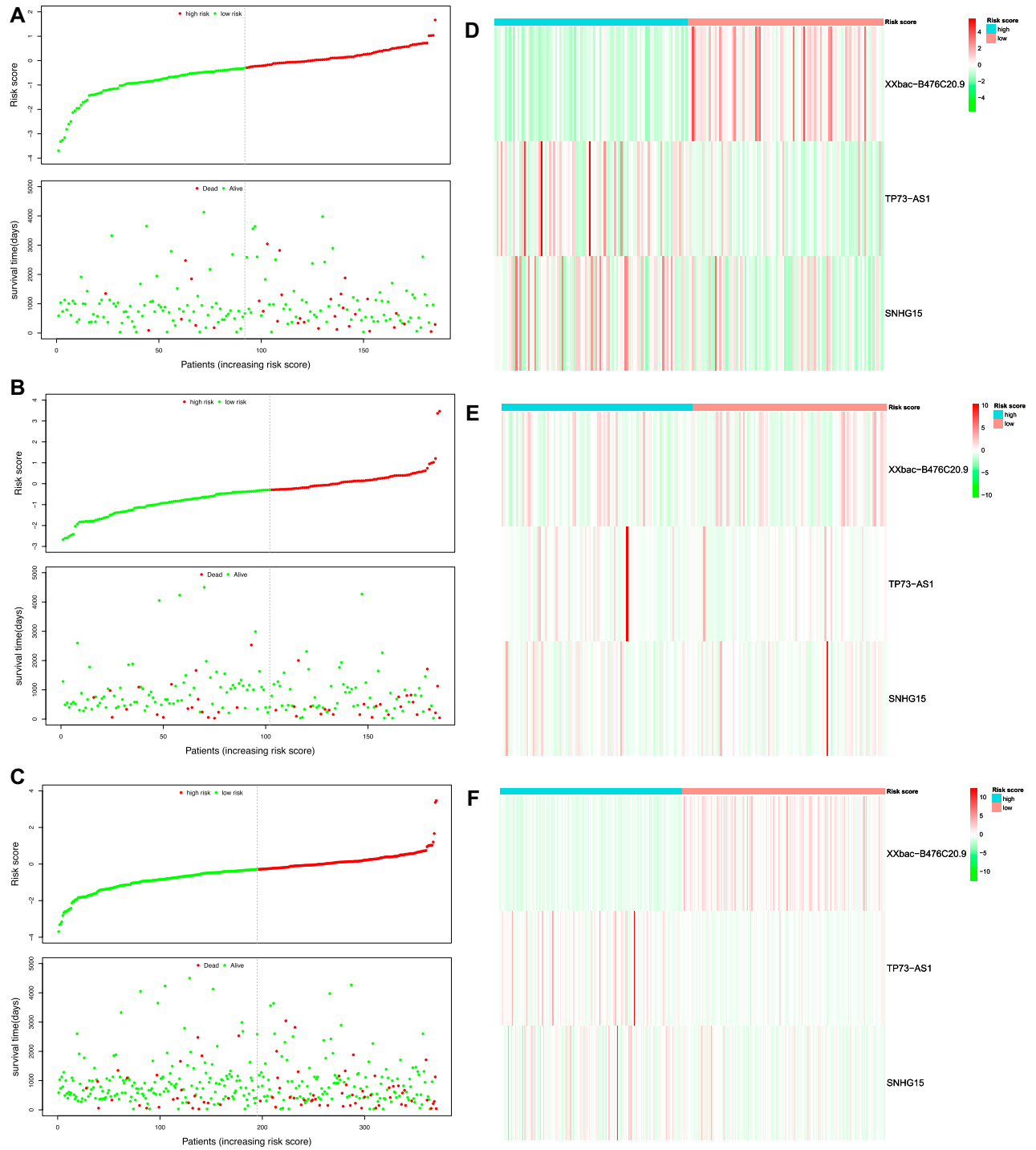


FIGURE 4 | Risk score analysis of the prognostic model in TCGA-COAD. Risk score distribution and survival status of the patients in the training set **(A)**, test set **(B)**, and whole set **(C)**. Expression heatmap of three DEFR-lncRNAs in the training set **(D)**, test set **(E)**, and whole set **(F)**.

plotted to assess the sensitivity and specificity of the 1-, 3-, and 5-year survival predictions of CC patients using the timeROC R package. In the training set, the AUCs used for 1-, 3-, and 5-year OS predictions were 0.72, 0.69, and 0.73, respectively (**Figure 5D**). In the test set, the AUCs used for 1-, 3-, and 5-year OS predictions were 0.63, 0.6, and 0.66, respectively (**Figure 5E**). In the whole set, the AUCs used for 1-, 3-, and 5-year OS predictions were 0.64, 0.63, and 0.66, respectively (**Figure 5F**).

Furthermore, the univariate and multivariate Cox regression analyses were performed to validate the independent predictive

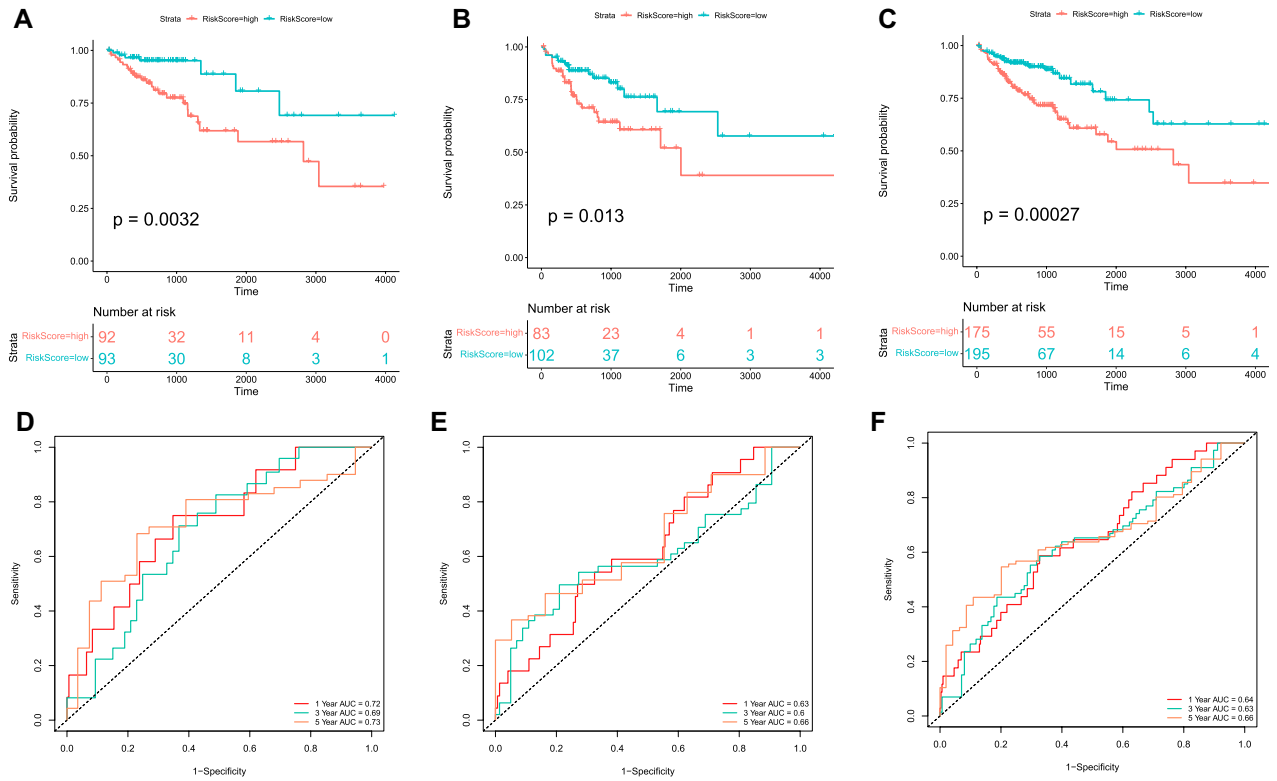


FIGURE 5 | Kaplan–Meier curve and ROC curve of the model in TCGA-COAD. Kaplan–Meier curves of the OS of patients between high-risk and low-risk groups in the training set **(A)**, test set **(B)**, and whole set **(C)**. Time-dependent ROC curve analysis of the RS in the training set **(D)**, test set **(E)**, and whole set **(F)**.

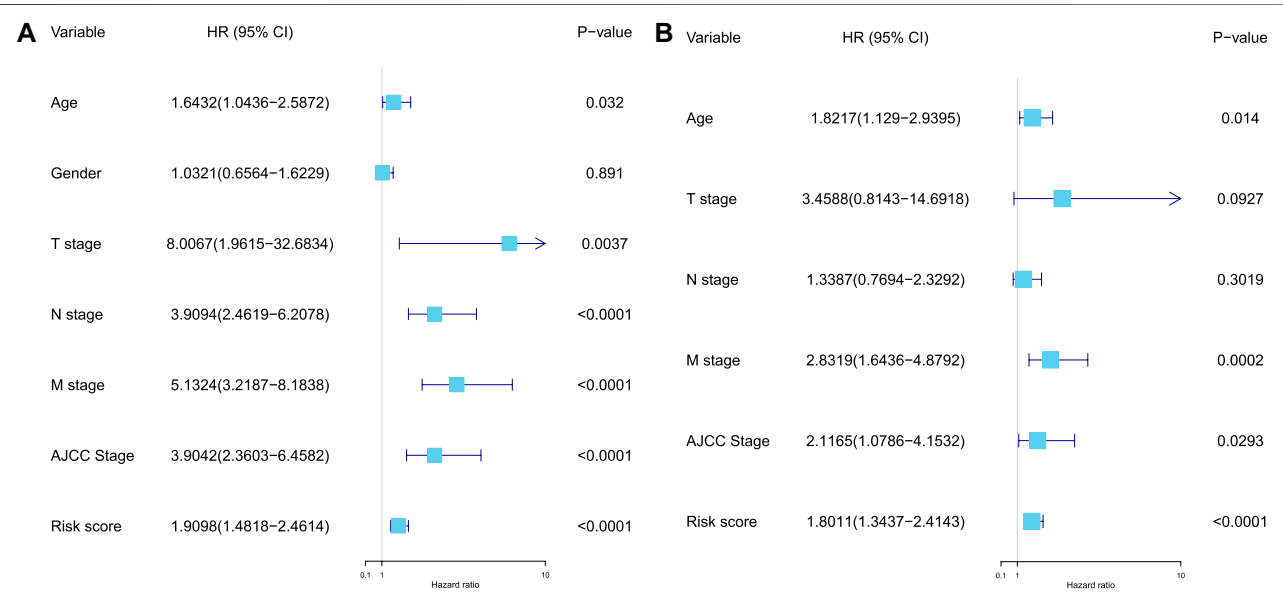


FIGURE 6 | Validation of the independent predictive power of the model. **(A)** Univariate Cox regression analysis of the prognostic variables in the whole set. **(B)** Multivariate Cox regression analysis of the prognostic variables (age, T stage, N stage, M stage, AJCC stage, and RS) in the whole set.

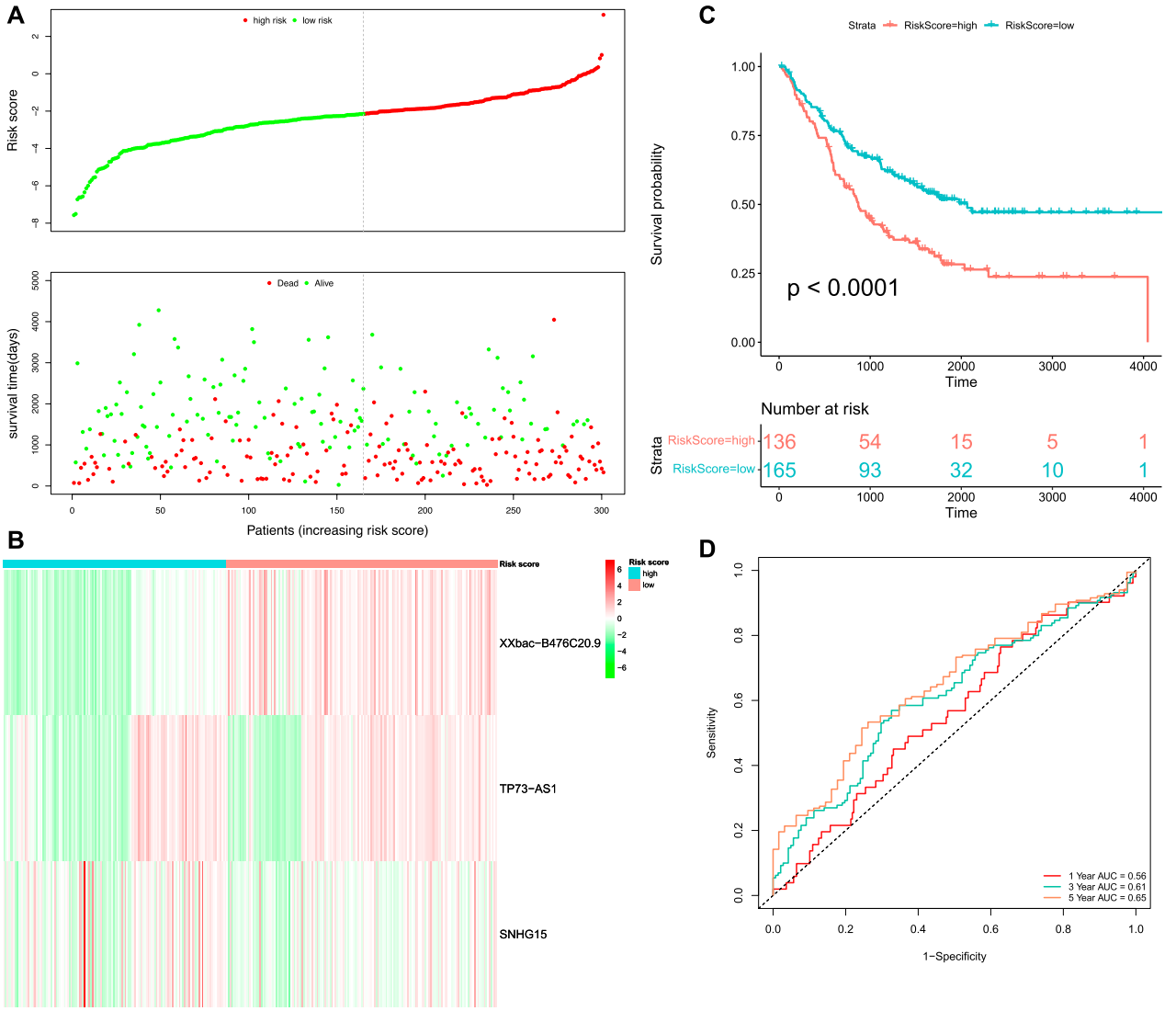


FIGURE 7 | Survival and time-dependent ROC curve analysis of the prognostic model in the validation set. **(A)** Risk score distribution and survival status of the patients. **(B)** Heatmap of three DEFR-lncRNAs. **(C)** Kaplan–Meier curve analysis of the OS between high-risk and low-risk groups. **(D)** Time-dependent ROC curve analysis of the RS.

power of the prognostic RS model for CC patients in the training, test, and whole sets, and the variables (age, gender, T stage, N stage, M stage, AJCC stage, and RS) were used as the possible risk factors. These results revealed that the prognostic model proposed in our study can be used as an independent prognostic factor for CC patients (**Supplementary Table S2**). In the whole set, we found that age, M stage, AJCC stage, and RS were the independent risk factors for CC patients ($p < 0.05$, **Figures 6A,B**).

Verification of the Prognostic Model in the Validation Set

We merged GSE72970 and GSE17536 to form the validation set, which contained 301 tumor samples. We calculated the RS of each patient in the validation set based on the formula of the prognostic RS model. The patients in the validation set were classified into high-risk (n = 136) and low-risk groups (n = 165) according to the optimal cut-off value (RS = −2.150814). The distribution of the RS for each patient and their survival status in the validation set are shown in **Figure 7A**. The death status of the patients increased with the increasing risk score. The expression pattern of the three prognostic DEFR-lncRNAs between the high-risk and low-risk groups is shown as a heatmap in **Figure 7B**. The Kaplan–Meier survival analysis demonstrated that the patients in the high-risk group had a significantly shorter OS than those in the low-risk group ($p < 0.0001$, **Figure 7C**). The AUC values for the 1-, 3-, and 5-year OS in the validation set were 0.56, 0.61, and 0.65, respectively (**Figure 7D**).

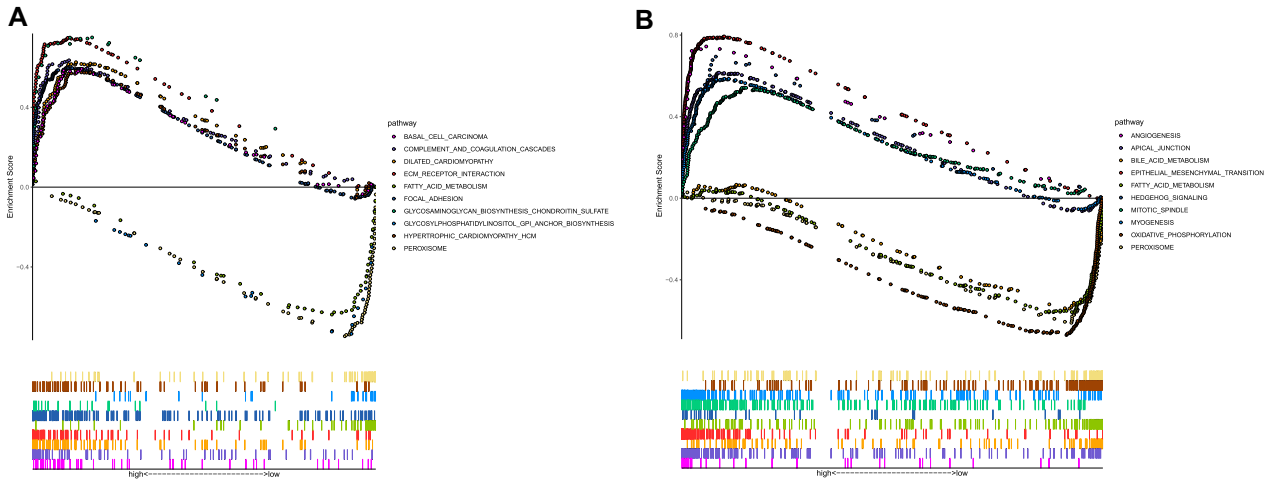


FIGURE 8 | Functional enrichment analysis. **(A)** KEGG pathways with the top 10 NES. **(B)** Hallmark pathways with the top 10 NES.

Functional Enrichment Analysis

The GSEA was performed to investigate the potential pathways and functions connected with high-risk and low-risk groups, and the terms $p < 0.05$ and $FDR < 0.25$ were considered statistically significant. The KEGG pathway analysis showed that peroxisome, glycosylphosphatidylinositol (GPI) anchor biosynthesis, and fatty acid metabolism were enriched in the low-risk group, whereas the extracellular matrix (ECM) receptor interaction, dilated cardiomyopathy, focal adhesion, complement and coagulation cascades, hypertrophic cardiomyopathy (HCM), glycosaminoglycan biosynthesis chondroitin sulfate, and basal cell carcinoma were enriched in the high-risk group (**Figure 8A**). Moreover, the Hallmark pathway analysis also revealed that the high-risk group was mainly enriched for epithelial-mesenchymal transition, apical junction, angiogenesis, hedgehog signaling, myogenesis, and mitotic spindle, whereas the low-risk group was mainly enriched for peroxisome, bile acid metabolism, fatty acid metabolism, and oxidative phosphorylation (**Figure 8B**). Of note, peroxisomes, fatty acid metabolism, and oxidative phosphorylation enriched in the low-risk group were associated with ferroptosis, which have been reported to be closely linked to ferroptosis (Stockwell et al., 2017; Tang and Kroemer, 2020; Ma et al., 2021).

Immune Infiltration Analysis

After the filtration of samples with $p < 0.05$ *via* CIBERSORT, we obtained fractions of 22 immune cell types in 156 CC patients, including 74 patients in the high-risk group and 82 patients in the low-risk group. The relative fractions of 22 immune cell types are shown in **Figure 9A**. From **Figure 9A**, we can find that the highest proportion of patients in the high-risk group was macrophages M0 (24.3%), followed by macrophages M2 (12.9%) and mast cells activated (12.5%). Meanwhile, the highest proportion of patients in the low-risk group was macrophages M0 (17.2%), followed by mast cells activated (12.7%) and macrophages M2 (12.4%). As shown in **Figure 9B**, the distribution of six immune cell types had a significant difference between the high-risk and low-risk groups, which also exhibited higher infiltration of macrophages M0 and T cells regulatory, and lower infiltration of dendritic cells activated, NK cells activated, plasma cells, T cells CD4 memory activated, and T cells CD4 memory resting in the high-risk group. In addition, we also used the ssGSEA method to estimate the infiltration level of the 28 kinds of immune cells that were over-represented in the tumor microenvironment for the 156 CC patients. The results indicated that 12 kinds of immune cells had significant differences between the high-risk and low-risk groups (**Figure 10**). We also found that in addition to type 17 T helper cells, the other 11 kinds of immune cells (central memory CD4 T cells, central memory CD8 T cells, effector memory CD4 T cells, effector memory CD8 T cells, immature dendritic cells, macrophages, MDSC, natural killer cells, natural killer T cells, regulatory T cells, and T follicular helper cells) had a higher infiltration level in the high-risk group than in the low-risk group.

DISCUSSION

With the rapid development of next-generation sequencing technologies, computational tools are used to identify biomarkers and study cancer disease, which is an emergent field in cancer systems biology (Yang J. et al., 2020; Xu et al., 2020). CC is a high-incidence malignant tumor with a poor prognosis. Although targeted drugs can improve the prognosis of patients with CC, the mortality rate among patients remains high (Zhou and Ma, 2019). Therefore, reliable biomarkers must be identified for constructing a prognostic model to assess the prognosis and survival of CC patients.

Ferroptosis is morphologically, biochemically, and genetically distinct from other forms of cell death (Dixon et al., 2012). Previous studies have demonstrated that ferroptosis is involved in tumor immunization and cancer immunotherapy (Wang W. et al., 2019; Xu et al., 2021). Ferroptosis and iron metabolism play

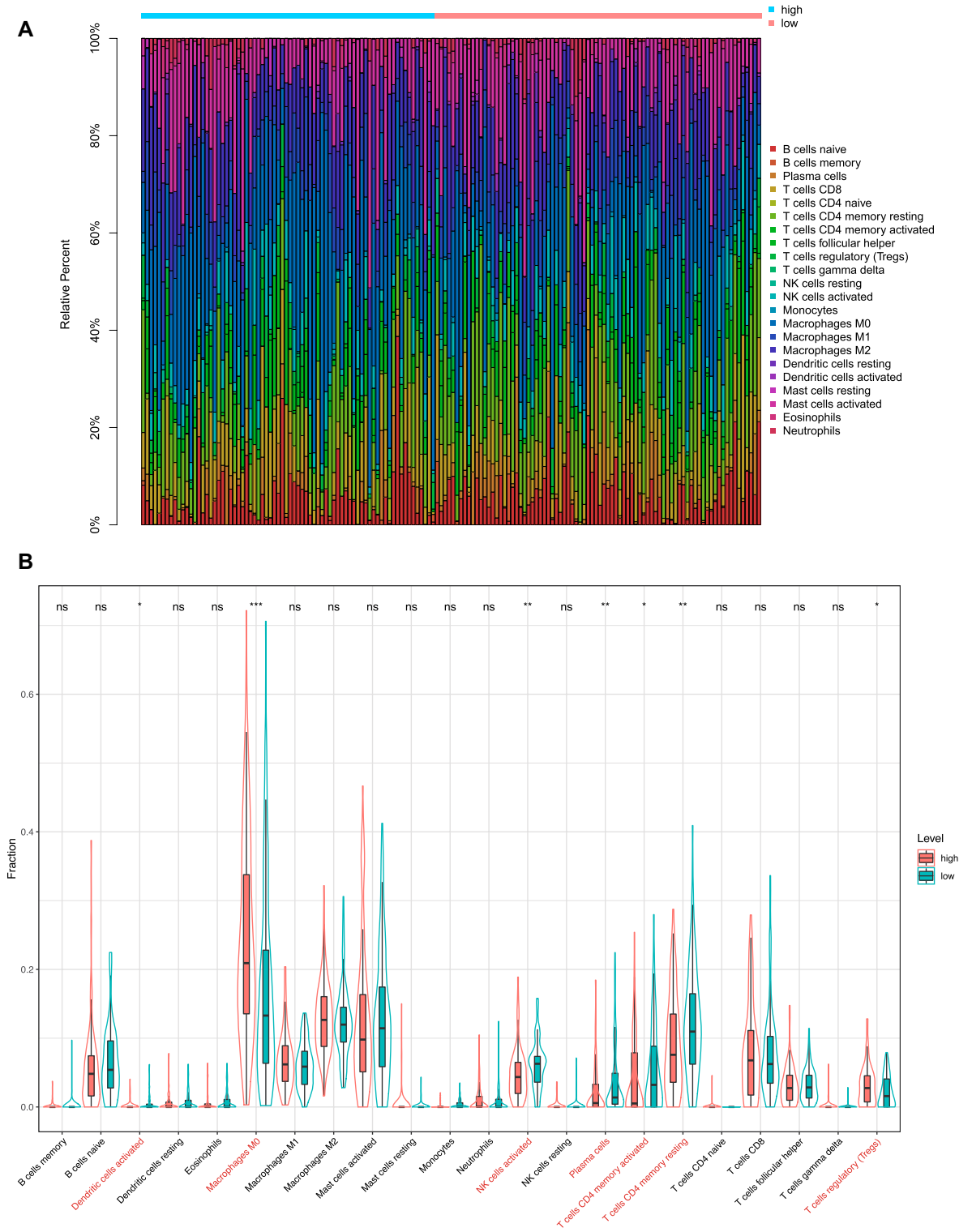


FIGURE 9 | Immunity analysis *via* CIBERSORT. **(A)** Bar graph showing the proportion of 22 immune cell types in CC patients of TCGA-COAD. Column names of the plot are the sample ID. **(B)** Difference in the proportions of 22 immune cell types between patients in the high-risk and low-risk groups. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; and ****$p < 0.0001$; ns, not significant.

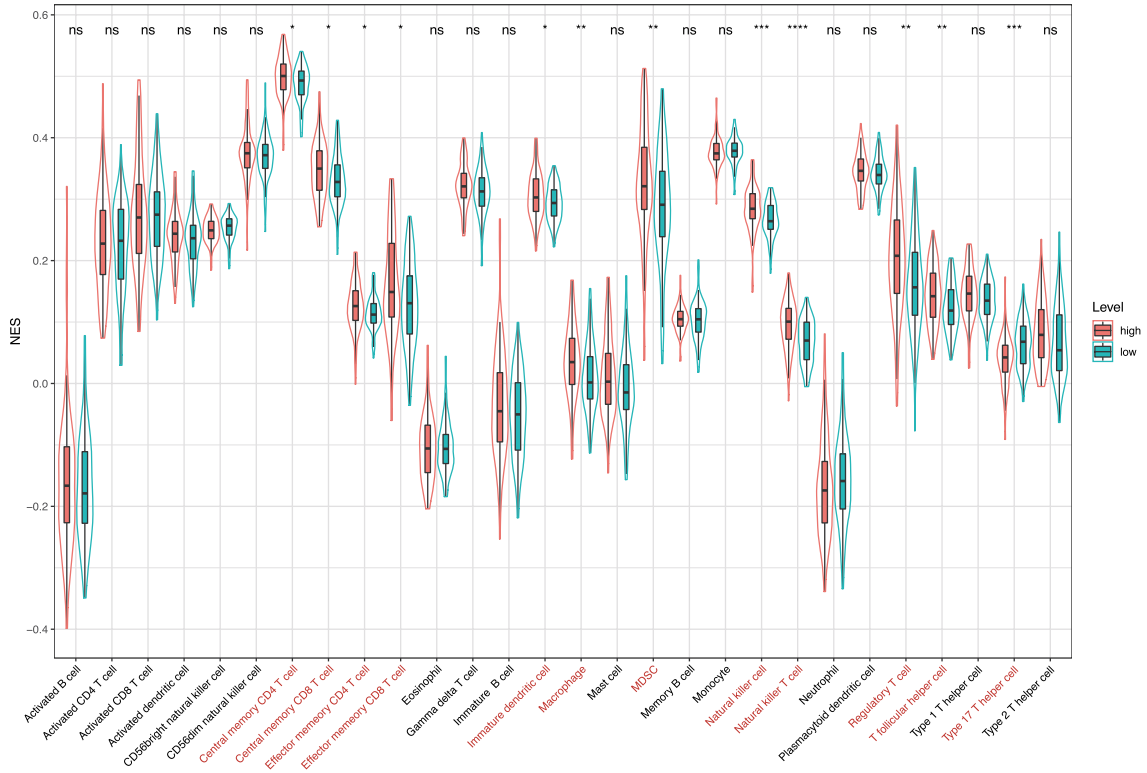


FIGURE 10 | Normalized enrichment scores of 28 kinds of immune cells in the high-risk and low-risk groups. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; and ****$p < 0.0001$; ns, not significant.

important roles in the pathogenesis of cancer. Moreover, ferroptosis-related lncRNA has also attracted attention (Mao et al., 2018; Wang M. et al., 2019; Yang Y. et al., 2020).

In this study, we constructed a prognostic model of three ferroptosis-related lncRNAs (XXbac-B476C20.9, TP73-AS1, and SNHG15) and showed that it had a good predictive ability for the overall survival of CC patients. Interestingly, literature mining revealed that three lncRNAs (XXbac-B476C20.9, TP73-AS1, and SNHG15) had been confirmed to be significantly associated with cancer. For example, the lncRNA XXbac-B476C20.9 was identified as a potential biomarker closely related to the prognosis of CC patients (Huang et al., 2019), which was consistent with our results. The overexpression of lncRNA TP73-AS1 was not only associated with metastasis and advanced clinical stages in colorectal cancer patients (Cai et al., 2018) but also promoted colorectal cancer cell migration and invasion (Li et al., 2019). Patients with high expression of lncRNA SNHG15 displayed a significantly shorter overall survival in COAD (Jiang et al., 2018). Moreover, the deregulation of the lncRNA SNHG15 strongly affected the proliferation, invasion, and tumor formation abilities of colorectal cancer cells (Saeinasab et al., 2019). The aforementioned previous studies further corroborated the results of our study.

We also investigated the underlying molecular mechanism by which the prognostic model is involved in the occurrence and development of CC through the GSEA analysis. Previous studies have also shown that GPI anchor biosynthesis, complement and coagulation cascades, and focal adhesion could play an important role in the progression of colorectal cancer (Cubiella et al., 2018; Gu et al., 2018; Xing et al., 2020). ECM receptor interaction, focal adhesion, and glycosaminoglycan biosynthesis chondroitin sulfate enriched in the high-risk group were related to cell motility, cell proliferation, and cell differentiation, which play a crucial role in the invasion of cancer cells (Han et al., 2021). Moreover, the Hallmark pathway analysis showed that epithelial-mesenchymal transition, apical junction, angiogenesis, and hedgehog signaling were enriched in the high-risk group, which was consistent with a previous study on CC (Yang et al., 2022). It was revealed that the mitotic spindle might lead to tumor formation in multiple tissues including colon cancer (Pussila et al., 2018). Bile acid metabolism was found to impact the microbial composition in colon cancer (Kennedy and Chang, 2020). Therefore, it is plausible that the prognostic model based on the three ferroptosis-related lncRNAs is highly correlated with CC.

Notably, our study found that the infiltration levels of macrophages M0, macrophages M2, and mast cells activated were significantly higher in the high-risk group. It has been shown that macrophages M0 were associated with the survival risk of CC, and the relative fraction of macrophages M0 was significantly increased in CC tissues compared with healthy bowel tissues (Wu et al., 2020). In addition, macrophages M2 induce the epithelial-mesenchymal transition phenotype in CC cells (Lee et al., 2020). The mast cells activated were C3-associated immune cells, where the C3 gene can predict the prognosis of colorectal

adenocarcinoma (Liu and Wang, 2021). After analyzing the 28 kinds of immune cells that are over-represented in the tumor microenvironment, we also found that 12 kinds of immune infiltration cells are significantly different between the high-risk and low-risk groups, especially natural killer cells and natural killer T cells. El-Deeb et al. (2022) have found that the natural killer cells activated by the alginate/κ-carrageenan oral microcapsules lead to apoptosis in the colon cancer Caco-2 cells. Yoshioka et al. (2012) showed that the number of colon tumors and natural killer T cells significantly decreased in the mice in the treated group. In summary, the results indicated that the prognostic model was associated with immune infiltration of CC and might provide a reference for the immunotherapy of CC.

CONCLUSION

In conclusion, we analyzed the lncRNA expression and clinical profiles in TCGA-COAD and GEO databases. Three differentially expressed ferroptosis-related lncRNAs (XXbac-B476C20.9, TP73-AS1, and SNHG15) were identified as biomarkers to establish a prognostic model for CC patients. The limitation to our study is that the prognostic model was constructed and validated on the database publicly available online. Future prospective clinical trials are required to further consolidate the effectiveness of the prognostic model.

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

JL performed data analysis and drafted the manuscript; JT assisted in manuscript writing; and XY designed the study and revised the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 11701379).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.934196/full#supplementary-material

REFERENCES

Borkiewicz, L., Kalafut, J., Dudziak, K., Przybyszewska-Podstawka, A., and Telejko, I. (2021). Decoding LncRNAs. *Cancers* 13 (11), 2643. doi:10.3390/cancers13112643

Cai, H.-j., Zhuang, Z.-c., Wu, Y., Zhang, Y.-y., Liu, X., Zhuang, J.-f., et al. (2021). Development and Validation of a Ferroptosis-Related lncRNAs Prognosis Signature in Colon Cancer. *Bosn J Basic Med Sci* 21 (5), 569–576. doi:10.17305/bjbms.2020.5617

Cai, Y., Yan, P., Zhang, G., Yang, W., Wang, H., and Cheng, X. (2018). Long Non-coding RNA TP73-AS1 Sponges miR-194 to Promote Colorectal Cancer Cell Proliferation, Migration and Invasion via Up-Regulating TGFα. *Cbm* 23 (1), 145–156. doi:10.3233/CBM-181503

Chang, Z., Huang, R., Fu, W., Li, J., Ji, G., Huang, J., et al. (2020). The Construction and Analysis of ceRNA Network and Patterns of Immune Infiltration in Colon Adenocarcinoma Metastasis. *Front. Cell. Dev. Biol.* 8, 688. doi:10.3389/fcell.2020.00688

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 (1), 248–262. doi:10.1016/j.celrep.2016.12.019

Cubiella, J., Clos-Garcia, M., Alonso, C., Martinez-Arranz, I., Perez-Cormenzana, M., Barrenetxea, Z., et al. (2018). Targeted UPLC-MS Metabolic Analysis of Human Faeces Reveals Novel Low-Invasive Candidate Markers for Colorectal Cancer. *Cancers* 10 (9), 300. doi:10.3390/cancers10090300

Dixon, S. J., Lemberg, K. M., Lamprecht, M. R., Skouta, R., Zaitsev, E. M., Gleason, C. E., et al. (2012). Ferroptosis: an Iron-dependent Form of Nonapoptotic Cell Death. *Cell.* 149 (5), 1060–1072. doi:10.1016/j.cell.2012.03.042

El-Deeb, N. M., Ibrahim, O. M., Mohamed, M. A., Farag, M. M. S., Farrag, A. A., and El-Aassar, M. R. (2022). Alginate/κ-carrageenan Oral Microcapsules Loaded with Agaricus Bisporus Polysaccharides MH751906 for Natural Killer Cells Mediated Colon Cancer Immunotherapy. *Int. J. Biol. Macromol.* 205, 385–395. doi:10.1016/j.ijbiomac.2022.02.058

Feng, S., Yin, H., Zhang, K., Shan, M., Ji, X., Luo, S., et al. (2022). Integrated Clinical Characteristics and Omics Analysis Identifies a Ferroptosis and Iron-Metabolism-Related lncRNA Signature for Predicting Prognosis and Therapeutic Responses in Ovarian Cancer. *J. Ovarian Res.* 15 (1), 10. doi:10.1186/s13048-022-00944-y

Gu, C., Wang, X., Long, T., Wang, X., Zhong, Y., Ma, Y., et al. (2018). FSTL1 Interacts with VIM and Promotes Colorectal Cancer Metastasis via Activating the Focal Adhesion Signalling Pathway. *Cell. Death Dis.* 9 (6), 654. doi:10.1038/s41419-018-0695-6

Guo, Y., Qu, Z., Li, D., Bai, F., Xing, J., Ding, Q., et al. (2021). Identification of a Prognostic Ferroptosis-Related lncRNA Signature in the Tumor Microenvironment of Lung Adenocarcinoma. *Cell. Death Discov.* 7 (1), 190. doi:10.1038/s41420-021-00576-z

Han, N., Zhang, Y.-Y., Zhang, Z.-M., Zhang, F., Zeng, T.-Y., Zhang, Y.-B., et al. (2021). High Expression of PDGFA Predicts Poor Prognosis of Esophageal Squamous Cell Carcinoma. *Med. Baltim.* 100 (20), e25932. doi:10.1097/MD.0000000000025932

Hassannia, B., Vandenabeele, P., and Vanden Berghe, T. (2019). Targeting Ferroptosis to Iron Out Cancer. *Cancer Cell.* 35 (6), 830–849. doi:10.1016/j.ccell.2019.04.002

He, D., Liao, S., Xiao, L., Cai, L., You, M., He, L., et al. (2021). Prognostic Value of a Ferroptosis-Related Gene Signature in Patients with Head and Neck Squamous Cell Carcinoma. *Front. Cell. Dev. Biol.* 9, 739011. doi:10.3389/fcell.2021.739011

Huang, W., Liu, Z., Li, Y., Liu, L., and Mai, G. (2019). Identification of Long Noncoding RNAs Biomarkers for Diagnosis and Prognosis in Patients with Colon Adenocarcinoma. *J Cell. Biochem.* 120 (3), 4121–4131. doi:10.1002/jcb.27697

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global Cancer Statistics. *CA A Cancer J. Clin.* 61 (2), 69–90. doi:10.3322/caac.20107

Jiang, H., Li, T., Qu, Y., Wang, X., Li, B., Song, J., et al. (2018). Long Non-coding RNA SNHG15 Interacts with and Stabilizes Transcription Factor Slug and Promotes Colon Cancer Progression. *Cancer Lett.* 425, 78–87. doi:10.1016/j.canlet.2018.03.038

Jin, L., Li, C., Liu, T., and Wang, L. (2020). A Potential Prognostic Prediction Model of Colon Adenocarcinoma with Recurrence Based on Prognostic lncRNA Signatures. *Hum. Genomics* 14 (1), 24. doi:10.1186/s40246-020-00270-8

Kennedy, M. S., and Chang, E. B. (2020). The Microbiome: Composition and Locations. *Prog. Mol. Biol. Transl. Sci.* 176, 1–42. doi:10.1016/bs.pmbts.2020.08.013

Lee, Y. S., Song, S. J., Hong, H. K., Oh, B. Y., Lee, W. Y., and Cho, Y. B. (2020). The FBW7-MCL-1 axis Is Key in M1 and M2 Macrophage-Related Colon Cancer Cell Progression: Validating the Immunotherapeutic Value of Targeting PI3Kγ. *Exp. Mol. Med.* 52 (5), 815–831. doi:10.1038/s12276-020-0436-7

Li, M., Jin, Y., and Li, Y. (2019). LncRNA TP73-AS1 Activates TGF-B1 to Promote the Migration and Invasion of Colorectal Cancer Cell. *Cmar* Vol. 11, 10523–10529. doi:10.2147/CMAR.S228490

Lin, A., Zhang, J., and Luo, P. (2020). Crosstalk between the MSI Status and Tumor Microenvironment in Colorectal Cancer. *Front. Immunol.* 11, 2039. doi:10.3389/fimmu.2020.02039

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-Of-Origin. *Front. Cell. Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, Y., and Wang, X. (2021). Tumor Microenvironment-Associated Gene C3 Can Predict the Prognosis of Colorectal Adenocarcinoma: a Study Based on TCGA. *Clin. Transl. Oncol.* 23 (9), 1923–1933. doi:10.1007/s12094-021-02602-z

Ma, T.-L., Zhou, Y., Wang, C., Wang, L., Chen, J.-X., Yang, H.-H., et al. (2021). Targeting Ferroptosis for Lung Diseases: Exploring Novel Strategies in Ferroptosis-Associated Mechanisms. *Oxidative Med. Cell. Longev.* 2021, 1–21. doi:10.1155/2021/1098970

Mao, C., Wang, X., Liu, Y., Wang, M., Yan, B., Jiang, Y., et al. (2018). A G3BP1-Interacting lncRNA Promotes Ferroptosis and Apoptosis in Cancer via Nuclear Sequestration of P53. *Cancer Res.* 78 (13), canres.3454.2017–3496. doi:10.1158/0008-5472.CAN-17-3454

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10 (3), 155–159. doi:10.1038/nrg2521

Mou, Y., Wang, J., Wu, J., He, D., Zhang, C., Duan, C., et al. (2019). Ferroptosis, a New Form of Cell Death: Opportunities and Challenges in Cancer. *J. Hematol. Oncol.* 12 (1), 34. doi:10.1186/s13045-019-0720-y

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Prensner, J. R., and Chinnaiyan, A. M. (2011). The Emergence of lncRNAs in Cancer Biology. *Cancer Discov.* 1 (5), 391–407. doi:10.1158/2159-8290.CD-11-0209

Pussila, M., Törönen, P., Einarsdottir, E., Katayama, S., Krjutškov, K., Holm, L., et al. (2018). Mlh1 Deficiency in Normal Mouse Colon Mucosa Associates with Chromosomally Unstable Colon Cancer. *Carcinogenesis* 39 (6), 788–797. doi:10.1093/carcin/bgy056

Saeinasab, M., Bahrami, A. R., González, J., Marchese, F. P., Martinez, D., Mowla, S. J., et al. (2019). SNHG15 Is a Bifunctional MYC-Regulated Noncoding Locus Encoding a lncRNA that Promotes Cell Proliferation, Invasion and Drug Resistance in Colorectal Cancer by Interacting with AIF. *J. Exp. Clin. Cancer Res.* 38 (1), 172. doi:10.1186/s13046-019-1169-0

Schmitt, A. M., and Chang, H. Y. (2016). Long Noncoding RNAs in Cancer Pathways. *Cancer Cell.* 29 (4), 452–463. doi:10.1016/j.ccell.2016.03.010

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2022). Cancer Statistics, 2022. *CA A Cancer J. Clin.* 72 (1), 7–33. doi:10.3322/caac.21708

Stockwell, B. R., Friedmann Angeli, J. P., Bayir, H., Bush, A. I., Conrad, M., Dixon, S. J., et al. (2017). Ferroptosis: A Regulated Cell Death Nexus Linking Metabolism, Redox Biology, and Disease. *Cell.* 171 (2), 273–285. doi:10.1016/j.cell.2017.09.021

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Tang, D., and Kroemer, G. (2020). Peroxisome: the New Player in Ferroptosis. *Sig Transduct. Target Ther.* 5 (1), 273. doi:10.1038/s41392-020-00404-3

Tang, M., Chen, Z., Wu, D., and Chen, L. (2018). Ferritinophagy/ferroptosis: Iron-related Newcomers in Human Diseases. *J. Cell. Physiology* 233 (12), 9179–9190. doi:10.1002/jcp.26954

Tsai, K.-W., Lo, Y.-H., Liu, H., Yeh, C.-Y., Chen, Y.-Z., Hsu, C.-W., et al. (2018). Linc00659, a Long Noncoding RNA, Acts as Novel Oncogene in Regulating Cancer Cell Growth in Colorectal Cancer. *Mol. Cancer* 17 (1), 72. doi:10.1186/s12943-018-0821-1

Wang, M., Mao, C., Ouyang, L., Liu, Y., Lai, W., Liu, N., et al. (2019a). Long Noncoding RNA LINC00336 Inhibits Ferroptosis in Lung Cancer by Functioning as a Competing Endogenous RNA. *Cell. Death Differ.* 26 (11), 2329–2343. doi:10.1038/s41418-019-0304-y

Wang, W., Green, M., Choi, J. E., Gijón, M., Kennedy, P. D., Johnson, J. K., et al. (2019b). CD8+ T Cells Regulate Tumour Ferroptosis during Cancer Immunotherapy. *Nature* 569 (7755), 270–274. doi:10.1038/s41586-019-1170-y

Wang, Z., Diao, J., Zhao, X., Xu, Z., and Zhang, X. (2021). Clinical and Functional Significance of a Novel Ferroptosis-related Prognosis Signature in Lung Adenocarcinoma. *Clin. Transl. Med.* 11 (3), e364. doi:10.1002/ctm2.364

Wei, J., Zeng, Y., Gao, X., and Liu, T. (2021). A Novel Ferroptosis-Related lncRNA Signature for Prognosis Prediction in Gastric Cancer. *BMC Cancer* 21 (1), 1221. doi:10.1186/s12885-021-08975-2

Wu, D., Ding, Y., Wang, T., Cui, P., Huang, L., Min, Z., et al. (2020). Significance of Tumor-Infiltrating Immune Cells in the Prognosis of Colon Cancer. *Ott* Vol. 13, 4581–4589. doi:10.2147/OTT.S250416

Xing, S., Wang, Y., Hu, K., Wang, F., Sun, T., and Li, Q. (2020). WGCNA Reveals Key Gene Modules Regulated by the Combined Treatment of Colon Cancer with PHY906 and CPT11. *Biosci. Rep.* 40 (9), BSR20200935. doi:10.1042/BSR20200935

Xu, H., Ye, D., Ren, M., Zhang, H., and Bi, F. (2021). Ferroptosis in the Tumor Microenvironment: Perspectives for Immunotherapy. *Trends Mol. Med.* 27 (9), 856–867. doi:10.1016/j.molmed.2021.06.014

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: an Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinformatics* 36 (10), 3139–3147. doi:10.1093/bioinformatics/btaa109

Yang, C., Huang, S., Cao, F., and Zheng, Y. (2021a). A Lipid Metabolism-Related Genes Prognosis Biomarker Associated with the Tumor Immune Microenvironment in Colorectal Carcinoma. *BMC Cancer* 21 (1), 1182. doi:10.1186/s12885-021-08902-5

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-Positive Breast Cancer Recurrence and Metastasis Risk from Histopathological Images and Clinical Information via Multimodal Deep Learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020a). Human Geroprotector Discovery by Targeting the Converging Subnetworks of Aging and Age-Related Diseases. *GeroScience* 42 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yang, Y., Tai, W., Lu, N., Li, T., Liu, Y., Wu, W., et al. (2020b). lncRNA ZFAS1 Promotes Lung Fibroblast-To-Myofibroblast Transition and Ferroptosis via Functioning as a ceRNA through miR-150-5p/SLC38A1 axis. *Aging* 12 (10), 9085–9102. doi:10.18632/aging.103176

Yang, Y., Yan, X., Li, X., Ma, Y., and Goel, A. (2021b). Long Non-coding RNAs in Colorectal Cancer: Novel Oncogenic Mechanisms and Promising Clinical Applications. *Cancer Lett.* 504, 67–80. doi:10.1016/j.canlet.2021.01.009

Yoshioka, K., Ueno, Y., Tanaka, S., Nagai, K., Onitake, T., Hanaoka, R., et al. (2012). Role of Natural Killer T Cells in the Mouse Colitis-Associated Colon Cancer Model. *Scand. J. Immunol.* 75 (1), 16–26. doi:10.1111/j.1365-3083.2011.02607.x

Zhang, W., Fang, D., Li, S., Bao, X., Jiang, L., and Sun, X. (2021). Construction and Validation of a Novel Ferroptosis-Related lncRNA Signature to Predict Prognosis in Colorectal Cancer Patients. *Front. Genet.* 12, 709329. doi:10.3389/fgene.2021.709329

Zhou, F., Shen, F., Zheng, Z., and Ruan, J. (2019). The LncRNA XIRP2-AS1 Predicts Favorable Prognosis in Colon Cancer. *Ott* Vol. 12, 5767–5778. doi:10.2147/OTT.S215419

Zhou, N., and Bao, J. (2020). FerrDb: a Manually Curated Resource for Regulators and Markers of Ferroptosis and Ferroptosis-Disease Associations. *Database (Oxford)* 2020, baaa021. doi:10.1093/database/baaa021

Zhou, Z., and Ma, J. (2019). Gambogic Acid Suppresses Colon Cancer Cell Activity *In Vitro*. *Exp. Ther. Med.* 18 (4), 2917–2923. doi:10.3892/etm.2019.7912

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

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

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fgene.2022.1008792

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Xiaoxu Yang,
University of California, San Diego, United States
Haiyan Liu,
Changsha Medical University, China

*CORRESPONDENCE
Jidong Lang,
langjidong@hotmail.com

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 01 August 2022
ACCEPTED 22 August 2022
PUBLISHED 15 September 2022

CITATION
Lang J (2022), NanoCoV19: An analytical pipeline for rapid detection of severe acute respiratory syndrome coronavirus 2.
Front. Genet. 13:1008792.
doi: 10.3389/fgene.2022.1008792

NanoCoV19: An analytical pipeline for rapid detection of severe acute respiratory syndrome coronavirus 2

Jidong Lang*

Department of Bioinformatics, Qitan Technology (Beijing) Co., Ltd., Beijing, China

Nanopore sequencing technology (NST) has become a rapid and cost-effective method for the diagnosis and epidemiological surveillance of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) during the coronavirus disease 2019 (COVID-19) pandemic. Compared with short-read sequencing platforms (e.g., Illumina's), nanopore long-read sequencing platforms effectively shorten the time required to complete the detection process. However, due to the principles and data characteristics of NST, the accuracy of sequencing data has been reduced, thereby limiting monitoring and lineage analysis of SARS-CoV-2. In this study, we developed an analytical pipeline for SARS-CoV-2 rapid detection and lineage identification that integrates phylogenetic-tree and hotspot mutation analysis, which we have named NanoCoV19. This method not only can distinguish and trace the lineages contained in the alpha, beta, delta, gamma, lambda, and omicron variants of SARS-CoV-2 but is also rapid and efficient, completing overall analysis within 1 h. We hope that NanoCoV19 can be used as an auxiliary tool for rapid subtyping and lineage analysis of SARS-CoV-2 and, more importantly, that it can promote further applications of NST in public-health and -safety plans similar to those formulated to address the COVID-19 outbreak.

KEYWORDS

nanopore sequencing technology, SARS-CoV-2, hotspot mutation, phylogenetic tree, coronavirus disease 2019 (COVID-19)

1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a causative agent of coronavirus disease 2019 (COVID-19), was identified in late 2019 (Zhu et al., 2020). Shortly thereafter, SARS-CoV-2 spread around the world, causing significant social problems, medical-system stress, and economic stagnation in all countries. It is a positive-sense single-stranded RNA virus with a 29,903 bp genome size, which was published in full in January 2020 (Lu et al., 2020a; Wu et al., 2020). Such publication led to the development of assays for SARS-CoV-2 detection based on real-time polymerase chain reaction (RT-PCR), which has been commonly used as a gold standard for monitoring the COVID-19 pandemic (van Kasteren et al., 2020). Sequencing the

genomes of SARS-CoV-2 at different times and locations and in different populations yields information related to the viral-mutation rate, transmission dynamics, and origin of the disease (Boni et al., 2020). It is also a key technique for understanding the viral lineages that circulate in individual countries and understanding how frequently new variant sources from other geographic regions are introduced. Genome sequencing of SARS-CoV-2 therefore serves to indicate the success of control measures, allow an understanding of how the virus evolves in response to interventions, and inform public response by defining the phylogenetic structure of the disease's outbreaks (Rambaut et al., 2020). Integration of the complete viral genomes and detailed epidemiological data provides a valuable reference for epidemiological investigations into transmission networks and inferences of where cases of unknown origin might have arisen (Lu et al., 2020b; Fauver et al., 2020; Gonzalez-Reiche et al., 2020; Gudbjartsson et al., 2020; Rockett et al., 2020). In addition, several studies have shown that different lineages of SARS-CoV-2 can infect the same person (Fonseca et al., 2021; Tillett et al., 2021; To et al., 2021). Sequencing and analysis of the SARS-CoV-2 genome are essential to confirm reinfections and to rule out disease recurrence. Rapid and reliable sample sequencing in environments such as hospitals is essential to such epidemiological surveillance. Furthermore, large-scale longitudinal monitoring of SARS-CoV-2 genomes also provides important information on the virus's evolution, with important implications for COVID-19 vaccine development (Korber et al., 2020; Li et al., 2020; Uddin et al., 2020; Young et al., 2020).

Excitingly, nanopore sequencing technology (NST) has demonstrated its feasibility and effectiveness in epidemiological surveillance during outbreaks of viral diseases such as Ebola and Zika (Quick et al., 2015; Quick et al., 2016; Quick et al., 2017). Some studies have developed several methods of rapidly sequencing SARS-CoV-2 genomes based on nanopore sequencing platform of companies represented by Oxford Nanopore Technologies (ONT), which is critical for rapid diagnosis and monitoring of the spread of the new coronavirus (Bull et al., 2020; Wang et al., 2021a; Jia et al., 2021). However, the principles and data characteristics of NST, such as non-random systemic errors and many unexpected indels, have a certain effect on analytical results (Magi et al., 2017; Bull et al., 2020). In addition, due to the timeliness requirements of the turnaround time, the sequencing platforms used for SARS-CoV-2 are still primarily based on next-generation sequencing (NGS), with analytical methods mainly focused on the presence of targeted gene regions on the genome. Therefore, we developed an analytical pipeline for rapid detection and lineage identification of SARS-CoV-2, named NanoCoV19, based on NST combined with phylogenetic-tree and hotspot mutation analysis, to distinguish the new coronaviral lineages. We hope that NanoCov19 can further the application of NST in monitoring the direction of COVID-19 outbreaks.

2 Materials and methods

2.1 NanoCoV19 analytical principle

NanoCoV19 consists of two parts: the construction of a reference database, and the data analysis pipeline.

2.1.1 Construction of reference genome sequence and mutation hotspot database for analysis

We downloaded the lineage information of the alpha, beta, gamma, delta, lambda, and omicron variants from RCoV19 (version 4.0) and the corresponding complete genome sequence of SARS-CoV-2 from the National Center for Biotechnology Information (NCBI; Bethesda, MD, United States) virus database (The date of data release used for this paper was 1 June 2022). One genome sequence was randomly selected from the lineage of each variant as a representative reference sequence database (Supplementary Table S1) for phylogenetic-tree analysis. We used MAFFT (v7.487) (Katoh et al., 2002) to perform multiple-sequence alignment on these sequences, and iqtree2 (v2.1.4-beta) (Nguyen et al., 2015) to perform phylogenetic-tree analysis. FigTree (v1.4.4) (https://github.com/rambaut/figtree) was used for visualization to determine whether the selected reference sequences discriminated between viral lineages (Figure 1A).

We also randomly selected 10 complete genome sequences from each lineage. For lineages with < 10 complete genome sequences, all sequences were included in the group. Then, we used NanoSim-H (v1.1.0.4) (Yang et al., 2017) to simulate the error-free nanopore sequencing data of $n \times 1000$ sequencing reads, where n represents the number of complete genomes contained in each variant (Supplementary Table S2). The reference genome (MN908947.3) of SARS-CoV-2 was downloaded from the NCBI database. We used Minimap2 (v2.21-r1071) (Li, 2018) to do the read alignment, and followed by Sambamba (v0.8.0) (Tarasov et al., 2015) for alignment file processing. Longshot (v0.4.1) (Edge and Bansal, 2019) was used to detect mutations. Finally, we used mutation results that were unique to each variant and also present in the *lineages.csv* information published on RCoV19 as a database of hotspot mutations for distinguishing lineages (Figure 1A; Supplementary Table S3).

2.1.2 Data analysis pipeline

As shown in Figure 1B, raw nanopore sequencing data was pre-processed using Porechop (v0.2.4; https://github.com/rrwick/Porechop). Next, we performed statistical analysis on the preprocessed clean data using NanoPlot (v1.38.0)

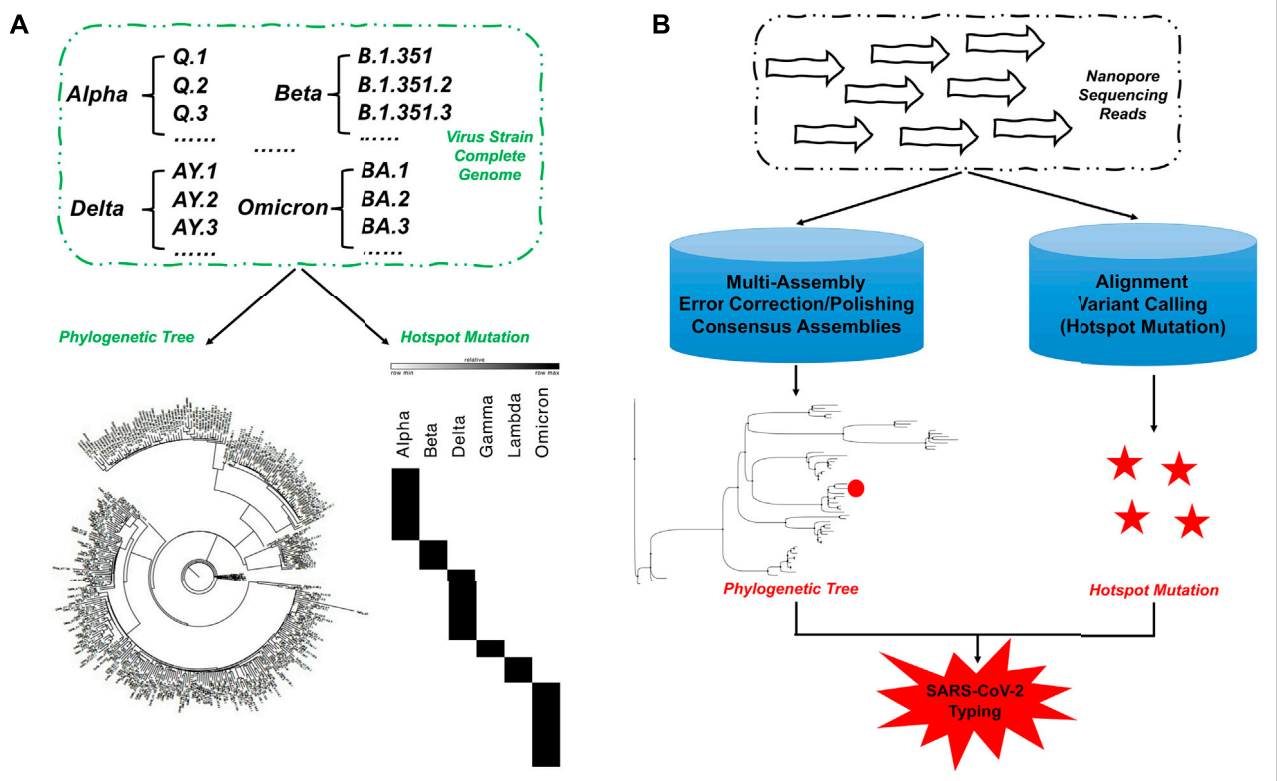


FIGURE 1
Schematic diagram showing the analytical principle of NanoCoV19. **(A)** Construction of a reference sequences and hotspot mutations database. **(B)** Pipeline for lineage analysis of SARS-CoV-2 based on nanopore sequencing data.

(De Coster et al., 2018), after which we employed FlyE (v2.8.3-b1695) (Kolmogorov et al., 2019), Raven (v1.8.1) (Vaser and Šikić, 2021), Canu (Koren et al., 2017), Wtdbg2 [v0.0 (19830203)] (Ruan and Li, 2020), and Trycycler (v0.5.3) (Wick et al., 2021) for data assembly and generation of consensus sequences. Racon (v1.4.20) (Vaser et al., 2017) was used for correction and self-correction after each assembly. In the presence of NGS sequencing data, we polished each error-corrected assembly sequence using Pilon (v1.24) (Walker et al., 2014). We used Samtools (v1.12) (Li et al., 2009) to process the alignment files, and soap.coverage (v2.7.7; https://github.com/gigascience/bgi-soap2/tree/master/tools/soap.coverage) was used for statistical analysis of sequencing depth and genome coverage. The software and parameters used for establishing phylogenetic-tree and hotspot mutation detection were consistent with those described in part (Zhu et al., 2020).

2.2 Testing data set

Ten complete genome sequences that differed from the constructed reference database were randomly selected from the complete genomes of the alpha, beta, gamma, delta, lambda, and omicron variants as data for testing the analytical pipeline. We used NanoSim-H (v1.1.0.4) to simulated nanopore sequencing reads with and without errors. The number of simulated reads was 1000 (Supplementary Table S4). We used nucmer (v3.1; *–mum*) (Marcais et al., 2018) to compare and analyze the assembled draft genome and the corresponding complete genome.

To evaluate the real-world performance of NanoCoV19, the nanopore sequencing data published by Afrad et al. (2021) were also downloaded.

3 Results

3.1 NanoCoV19 performed well on the testing data set

We directly analyzed the phylogenetic tree and detected hotspot mutations of 10 randomly selected complete genomes of the six SARS-CoV-2 variants. The results of phylogenetic-tree (Figure 2A) and hotspot mutation (Figure 2B) analysis were consistent with our expectations: i.e., the concordance

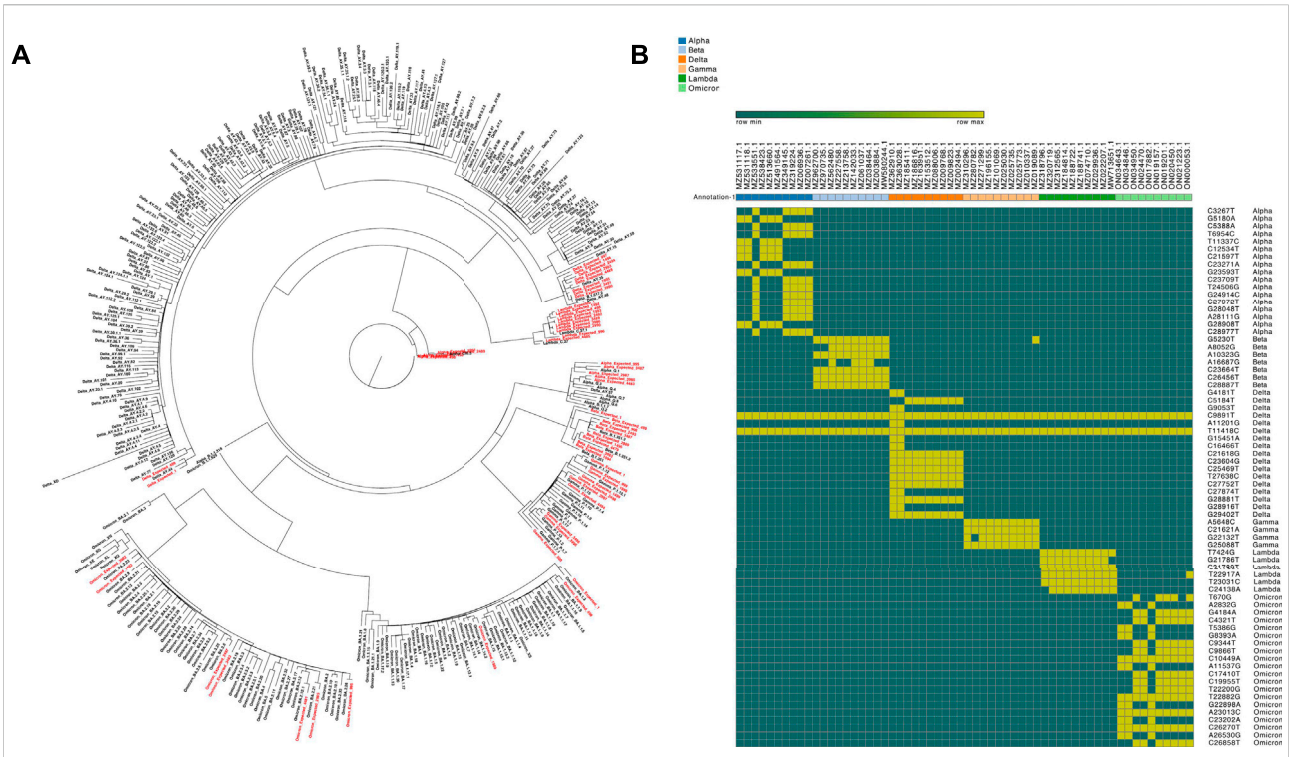


FIGURE 2
Analytical results of simulated sequence data for 60 lineages. **(A)** The result of phylogenetic tree analysis (the red text represents simulated data). **(B)** The heatmap analysis of hotspot mutations.

rate was 100%. Further analysis of the 15 SARS-CoV-2 sub-lineage B.1.617.2 strains published by Afrad et al. (2021) showed that the predicted hotspot mutations were all delta variants (Supplementary Table S5), which was consistent with the classification of pangolin lineage B.1.617.2. However, because the read lengths of the sequencing data were all < 1,000 bp, which was the minimum overlap required, FlyE did not generate effective assembly results, making it impossible to carry out more-detailed lineage analysis.

3.2 The accuracy and integrity of assembly affected the phylogenetic-tree analysis

We used only FlyE assembly results to analyze simulated read data with and without errors. Our results showed that our hotspot mutation analysis results were accurately and effectively for lineage subtyping (Supplementary Tables S6, S7). However, 28 (Figure 3A) and 21 (Figure 3B) simulated samples with and without errors, respectively, were not effectively distinguished after assembly but formed a unique branch and were defined as outlier samples. The remaining assembly results were accurately and effectively performed lineage subtyping. By comparing the assembly results of the outlier samples with their corresponding complete genomes, we found that the outlier results might have been due to the structural problems of the assembled genomes (Figure 3E), indicating that the requirements for completeness and accuracy of the assembly results would be very high when performing cluster analysis on phylogenetic trees. Maybe too many indels or sequence structure problems would lead to serious errors and even failure of lineage analysis, which also reflecting the necessity of comprehensive analysis combined with hotspot mutation analysis.

For the simulated data with errors, we used the assembly results of Raven, FlyE, and Wtdbg2 to combine 10 high-quality assembly results (i.e., the complete genome sequences published by the corresponding lineages). Trycycler was also used to generate consensus sequences. This significantly improved the results: the number of outlier samples dropped to 18 (Figure 3C). Subsequently, after we added 23 high-quality assembly results (the maximum number of sequences that could be input into Trycycler is 26), the number of outlier samples was only 8 (Figure 3D). The lineage analysis results of the remaining simulated data were basically correct.

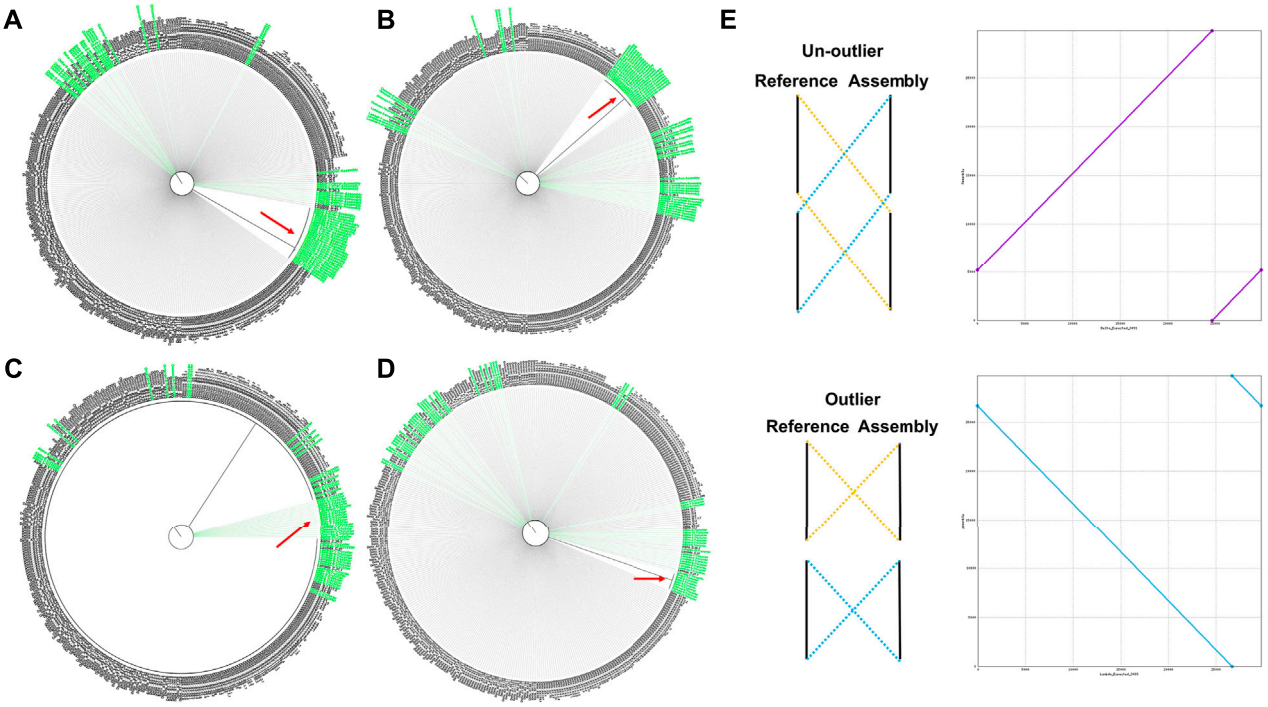


FIGURE 3
Assembly accuracy affects phylogenetic tree analysis. **(A)** The assembly results of FlyE to analyze simulated data with errors. **(B)** The assembly results of FlyE to analyze simulated data without errors. **(C)** The assemblies and consensus results of Trycycler to analyze simulated data with errors, which combination with 10 high-quality assembly results. **(D)** The assemblies and consensus results of Trycycler to analyze simulated data with errors, which combination with 23 high-quality assembly results. **(E)** The structural problems of the assembled draft genomes resulted in the outlier samples, which could not effectively distinguish the lineage.

TABLE 1 Running time during each step of the five tests.

Testing sample		Alpha	Beta	Gamma	Lambda	Omicron
Compute resource	**AMD EPYC 7542 32-core processor, 2T memory, 128 processor (16 processor/task)**					
Data size	Read number	1,000	1,000	1,000	1,000	1,000
	Base number	7,759,122	7,869,879	7,784,216	7,485,683	7,638,683
	Read length N50	9,496	9,553	9,469	9,168	9,134
Data analysis	Data preprocessing	0:05:40	0:06:20	0:05:01	0:07:12	0:05:07
	Assembly-FlyE	0:01:57	0:02:01	0:02:01	0:01:57	0:01:56
	Assembly-Canu	0:02:08	0:02:07	0:02:06	0:01:58	0:02:01
	Assembly-Wtdbg2	0:00:06	0:00:13	0:00:07	0:00:05	0:00:11
	Assembly-raven	0:00:03	0:00:02	0:00:03	0:00:02	0:00:03
	Racon	0:00:15	0:00:21	0:00:21	0:00:18	0:00:18
	Pilon	0:11:16	0:10:44	0:10:28	0:09:52	0:09:08
	Trycycler	0:00:38	0:00:38	0:00:43	0:00:41	0:00:40
	Phylogenetic tree	0:33:58	0:35:27	0:23:48	0:22:28	0:23:02
	Variation calling	0:00:07	0:00:06	0:00:11	0:00:06	0:00:07
	Total time	0:56:08	0:57:59	0:44:49	0:44:39	0:42:33

3.3 Overall analysis time could be controlled within 1 h

Analysis of the 1000-read data from five testing samples showed that on an AMD EPYC 7542 32-core processor with 2 T of memory and 128 processors, when we used 16 processors for each task, the overall analysis time of NanoCoV19 analytical pipeline was controlled within 1 h (Table 1).

4 Discussion

The development of NST has been very rapid (Magi et al., 2018; Wang et al., 2021b), and exciting results have been achieved in many fields, especially metagenomics for pathogen detection (Charalampous et al., 2019; Gu et al., 2021) and animal and/or plant genome assembly (Loman et al., 2015; Vaser et al., 2017; Lang et al., 2022a). Importantly, the advantages of NST in real-time sequencing analysis are self-evident (Payne et al., 2021; Goenka et al., 2022). NST has played a critical role in the tracing and rapid detection of outbreaks of infectious diseases such as COVID-19 (Quick et al., 2016; Quick et al., 2017). Theoretically, with the advantage of long-read lengths in nanopore sequencing, excessive sequencing reads for bacterial- or viral-haplotype assembly might not be required. Our results also showed that the analysis time of NanoCoV19 was controlled within 1 h from input of the 1,000 sequencing reads to end of analysis. Some studies showed that the whole processing time based on nanopore sequencing platforms such as ONT or Qitan Technology (QT) to detect SARS-CoV-2 and other respiratory viruses simultaneously within 6–10 h (Wang et al., 2021a). And the main time consumption was in the wet experimental and libraries sequencing steps. Thereby, we are trying and foresee that the combination of real-time analysis in NST with more-advanced computing resources could control overall analysis time from sample collection to analysis report issuance to within 30 min or even less, yielding significant social and economic benefits. Although ONT's sequencing solutions for SARS-CoV-2 have been established and applied in public-health scenarios (Meredith et al., 2020; Paden et al., 2020), the adoption of this technology has been somewhat limited due to concerns over sequencing accuracy. Given the technical principles and data characteristics of NST (Magi et al., 2017), such as non-random systematic errors and many unexpected indels, the accuracy of SARS-CoV-2 analysis results might be seriously affected. For example, we know that viruses are characterized by low mutation rates (Rambaut, 2020), so sequencing errors might lead to false-positive or false-negative assay results. Therefore, multi-dimensional or multi-aspect consideration, combination, and optimizing iteration may be required for analysis, especially for the infectious virus like SARS-CoV-2.

Although NanoCoV19 benefits in effectiveness from the combination of phylogenetic-tree and hotspot mutation analysis, it still has some shortcomings: 1) The accuracy and sufficiency of the constructed reference sequences and hotspot mutations database in viral-lineage discrimination still need further validation. 2) Continued optimization of the assembly method is still necessary due to the varying performances of different assembly algorithms for assembly results with the same data. For example, we also tried to conduct an assembly analysis on the simulated data using Raven and obtained results that were basically similar to those of FlyE, while the compositions of the outlier samples were different. This confirmed the necessity and high requirements for the quality and integrity of the assembly results before phylogenetic-tree analysis. Therefore, we used Trycycler to integrate multiple assemblies and generate consensus sequence, which is also a more important and worthy of attention in the NanoCoV19 analytical pipeline. However, the intermediate steps required manual selection of the better assembly results so that automation was insufficient. For example, the length of the assembly draft genome and/or the number of scaffolds were very different, so it was necessary to select or even delete some assemblies. Therefore, a method similar to MAECI (Lang, 2022) might also be required to balance accuracy and automation in the assembly results. 3) More tools and/or algorithms are needed for hotspot mutation detection [e.g., PEPPER-Margin-DeepVariant (Shafin et al., 2021) and Nano2NGS-Muta (Lang et al., 2022b)]. 4) NanoCoV19 should be further optimized for analysis time. Some steps could be run in parallel to shorten overall analysis time, although excessive memory consumption might happen, which would require a trade-off between resource consumption and analysis time. 5) As we known, SARS-CoV-2 virus strains are constantly evolving, resulting in the possible generation of many new strain genomes, so the relevant database will be continuously updated. However, NanoCoV19 only analyzes viral lineages with constructed reference database. Knowledge of determination criteria and processing methods for novel (unclassified) lineages is lacking. Therefore, a timely update of the reference database for the complete genome sequences is also required. 6) More actual data validation of NanoCoV19 performance is needed because the published raw sequencing data of SARS-CoV-2 genomes based on nanopore sequencing data are limited.

In summary, we hope that NanoCoV19 can be used as an auxiliary tool for rapid detection and lineage analysis of SARS-CoV-2, and that nanopore sequencers' outstanding advantages of long-read length and real-time sequencing can provide faster and more-accurate solutions for genomic epidemiological surveillance. This would promote the application of NST in the fields of public-health planning and safety, and even offline applications in the international space stations (Castro-Wallace et al., 2017; Carr et al., 2020; Stahl-Rommel et al., 2021).

5 Conclusion

NanoCoV19 is a potential auxiliary tool for rapid detection and lineage analysis of SARS-CoV-2 based on nanopore sequencing technology. It completes all analysis

within 1 h. We hope that it not only can assist in current-day lineage analysis and monitoring of SARS-CoV-2 but also promote the application of NST in related scientific research and clinical settings.

Data availability statement

The link to the RCoV19 database is https://ngdc.cncb.ac.cn/ncov/?lang=en. The SARS-CoV-2's information is also downloaded from the SARS-CoV-2 Data Hub in the National Center for Biotechnology Information (NCBI; Bethesda, MD, USA) virus database. The codes are available at https://github.com/langjidong/NanoCoV19.

Author contributions

JL designed the study, collected, simulated, analyzed and interpreted the data, and wrote the manuscript. All authors approved the final version of the manuscript.

Conflict of interest

Author JL is employed by Qitan Technology (Beijing) Co., Ltd, Beijing, China.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1008792/full#supplementary-material

References

Afrad, M. H., Khan, M. H., Rahman, S. I. A., Bin Manjur, O. H., Hossain, M., Alam, A. N., et al. (2021). Genome sequences of 15 SARS-CoV-2 sublineage B.1.617.2 strains in Bangladesh. *Microbiol. Resour. Announc.* 10, e0056021. doi:10.1128/MRA.00560-21

Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., et al. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* 5, 1408–1417. doi:10.1038/s41564-020-0771-4

Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., et al. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat. Commun.* 11, 6272. doi:10.1038/s41467-020-20075-6

Carr, C. E., Bryan, N. C., Saboda, K. N., Bhattaru, S. A., Ruvkun, G., and Zuber, M. T. (2020). Nanopore sequencing at Mars, Europa, and microgravity conditions. *NPJ Microgravity* 6, 24. doi:10.1038/s41526-020-00113-9

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., et al. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* 7, 18022. doi:10.1038/s41598-017-18364-0

Charalampous, T., Kay, G. L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., et al. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* 37, 783–792. doi:10.1038/s41587-019-0156-5

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi:10.1093/bioinformatics/bty149

Edge, P., and Bansal, V. (2019). Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* 10, 4660. doi:10.1038/s41467-019-12493-y

Fauver, J. R., Petrone, M. E., Hodcroft, E. B., Shioda, K., Ehrlich, H. Y., Watts, A. G., et al. (2020). Coast-to-Coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell.* 181, 990–996. doi:10.1016/j.cell.2020.04.021

Fonseca, V., de Jesus, R., Adelino, T., Reis, A. B., de Souza, B. B., Ribeiro, A. A., et al. (2021). Genomic evidence of SARS-CoV-2 reinfection case with the emerging B.1.2 variant in Brazil. *J. Infect.* 83, 237–279. doi:10.1016/j.jinf.2021.05.014

Goenka, S. D., Gorzynski, J. E., Shafin, K., Fisk, D. G., Pesout, T., Jensen, T. D., et al. (2022). Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nat. Biotechnol.* 40, 1035–1041. doi:10.1038/s41587-022-01221-5

Gonzalez-Reiche, A. S., Hernandez, M. M., Sullivan, M. J., Ciferri, B., Alshammary, H., Obla, A., et al. (2020). Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 369, 297–301. doi:10.1126/science.abc1917

Gu, W., Deng, X., Lee, M., Sucu, Y. D., Arevalo, S., Stryke, D., et al. (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* 27, 115–124. doi:10.1038/s41591-020-1105-z

Gudbjartsson, D. F., Helgason, A., Jonsson, H., Magnusson, O. T., Melsted, P., Norddahl, G. L., et al. (2020). Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.* 382, 2302–2315. doi:10.1056/NEJMoa2006100

Jia, X., Zhang, X., Ling, Y., Zhang, X., Tian, D., Liao, Y., et al. (2021). Application of nanopore sequencing in diagnosis of secondary infections in patients with severe COVID-19. *Zhejiang Da Xue Xue Bao Yi Xue Ban.* 50, 748–754. doi:10.3724/zdxbyxb-2021-0158

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi:10.1038/s41587-019-0072-8

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell.* 182, 812–827. doi:10.1016/j.cell.2020.06.043

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi:10.1101/gr.215087.116

Lang, J., Li, Y., Yang, W., Dong, R., Liang, Y., Liu, J., et al. (2022). Genomic and resistome analysis of alcaligenes faecalis strain PGB1 by nanopore MinION and Illumina Technologies. *BMC Genomics* 23, 316. doi:10.1186/s12864-022-08507-7

Lang, J. (2022). Maeci: A pipeline for generating consensus sequence with nanopore sequencing long-read assembly and error correction. *PLoS One* 17, e0267066. doi:10.1371/journal.pone.0267066

Lang, J., Sun, J., Yang, Z., He, L., He, Y., Chen, Y., et al. (2022). Nano2NGS-Muta: A framework for converting nanopore sequencing data to NGS-liked sequencing data for hotspot mutation detection. *Nar. Genom. Bioinform.* 4, lqac033. doi:10.1093/nargab/lqac033

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., et al. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell.* 182, 1284–1294. doi:10.1016/j.cell.2020.07.012

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi:10.1038/nmeth.3444

Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., et al. (2020). Genomic epidemiology of SARS-CoV-2 in guangdong province, China. *Cell.* 181, 997–1003. doi:10.1016/j.cell.2020.04.023

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* 395, 565–574. doi:10.1016/S0140-6736(20)30251-8

Magi, A., Giusti, B., and Tattini, L. (2017). Characterization of MinION nanopore data for resequencing analyses. *Brief. Bioinform.* 18, 940–953. doi:10.1093/bib/bbw077

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., and D'Aurizio, R. (2018). Nanopore sequencing data analysis: State of the art, applications and challenges. *Brief. Bioinform.* 19, 1256–1272. doi:10.1093/bib/bbx062

Marcais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* 14, e1005944. doi:10.1371/journal.pcbi.1005944

Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., et al. (2020). Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: A prospective genomic surveillance study. *Lancet. Infect. Dis.* 20, 1263–1272. doi:10.1016/S1473-3099(20)30562-4

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300

Paden, C. R., Tao, Y., Queen, K., Zhang, J., Li, Y., Uehara, A., et al. (2020). Rapid, sensitive, full-genome sequencing of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* 26, 2401–2405. doi:10.3201/eid2610.201800

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., and Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* 39, 442–450. doi:10.1038/s41587-020-00746-x

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* 16, 114. doi:10.1186/s13059-015-0677-2

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* 12, 1261–1276. doi:10.1038/nprot.2017.066

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi:10.1038/nature16996

Rambaut, A., Holmes, E. C., O'Toole, A., Hill, V., McCrone, J. T., Ruis, C., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407. doi:10.1038/s41564-020-0770-5

Rambaut, A. (2020). Phylodynamic analysis | 176 genomes | 6 mar 2020. Available at virological.org. (

Rockett, R. J., Arnott, A., Lam, C., Sadsad, R., Timms, V., Gray, K. A., et al. (2020). Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat. Med.* 26, 1398–1404. doi:10.1038/s41591-020-1000-7

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* 17, 155–158. doi:10.1038/s41592-019-0669-3

Shafin, K., Pesout, T., Chang, P. C., Nattestad, M., Kolesnikov, A., Goel, S., et al. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* 18, 1322–1332. doi:10.1038/s41592-021-01299-w

Stahl-Rommel, S., Jain, M., Nguyen, H. N., Arnold, R. R., Aunon-Chancellor, S. M., Sharp, G. M., et al. (2021). Real-time culture-independent microbial profiling onboard the international space station using nanopore sequencing. *Genes. (Basel)* 12, 106. doi:10.3390/genes12010106

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034. doi:10.1093/bioinformatics/btv098

Tillett, R. L., Sevinsky, J. R., Hartley, P. D., Kerwin, H., Crawford, N., Gorzalski, A., et al. (2021). Genomic evidence for reinfection with SARS-CoV-2: A case study. *Lancet. Infect. Dis.* 21, 52–58. doi:10.1016/S1473-3099(20)30764-7

To, K. K., Hung, I. F., Ip, J. D., Chu, A. W., Chan, W. M., Tam, A. R., et al. (2021). Coronavirus disease 2019 (COVID-19) Re-infection by a phylogenetically distinct severe acute respiratory syndrome coronavirus 2 strain confirmed by whole genome sequencing. *Clin. Infect. Dis.* 73, e2946–e2951. doi:10.1093/cid/ciaa1275

Uddin, M., Mustafa, F., Rizvi, T. A., Loney, T., Suwaidi, H. A., Al-Marzouqi, A. H. H., et al. (2020). SARS-CoV-2/COVID-19: Viral genomics, epidemiology, vaccines, and therapeutic interventions. *Viruses* 12, E526. doi:10.3390/v12050526

van Kasteren, P. B., van der Veer, B., van den Brink, S., Wijsman, L., de Jonge, J., van den Brandt, A., et al. (2020). Comparison of seven commercial RT-PCR diagnostic kits for COVID-19. *J. Clin. Virol.* 128, 104412. doi:10.1016/j.jcv.2020.104412

Vaser, R., and Šikić, M. (2021). Time- and memory-efficient genome assembly with Raven. *Nat. Comput. Sci.* 1, 332–336. doi:10.1038/s43588-021-00073-4

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi:10.1101/gr.214270.116

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963. doi:10.1371/journal.pone.0112963

Wang, M., Fu, A., Hu, B., Tong, Y., Liu, R., Liu, Z., et al. (2021). Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. *Small* 17, e2002169. doi:10.1002/smll.202002169

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. doi:10.1038/s41587-021-01108-x

Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Meric, G., Vezina, B., et al. (2021). Trycycler: Consensus long-read assemblies for bacterial genomes. *Genome Biol.* 22, 266. doi:10.1186/s13059-021-02483-z

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi:10.1038/s41586-020-2008-3

Yang, C., Chu, J., Warren, R. L., and Birol, I. (2017). NanoSim: Nanopore sequence read simulator based on statistical characterization. *Gigascience* 6, 1–6. doi:10.1093/gigascience/gix010

Young, B. E., Fong, S. W., Chan, Y. H., Mak, T. M., Ang, L. W., Anderson, D. E., et al. (2020). Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: An observational cohort study. *Lancet* 396, 603–611. doi:10.1016/S0140-6736(20)31757-8

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi:10.1056/NEJMoa2001017

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fgene.2022.1001608

OPEN ACCESS

EDITED BY
Liqian Zhou,
Hunan University of Technology, China

REVIEWED BY
Min Chen,
Hunan Institute of Technology, China
Guanghui Li,
East China Jiaotong University, China

*CORRESPONDENCE
Ling Wu,
wl_nancy08@163.com

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 23 July 2022
ACCEPTED 15 August 2022
PUBLISHED 15 September 2022

CITATION
Su Q, Tan Q, Liu X and Wu L (2022), Prioritizing potential circRNA biomarkers for bladder cancer and bladder urothelial cancer based on an ensemble model.
Front. Genet. 13:1001608.
doi: 10.3389/fgene.2022.1001608

Prioritizing potential circRNA biomarkers for bladder cancer and bladder urothelial cancer based on an ensemble model

Qiongli Su, Qiuhong Tan, Xin Liu and Ling Wu*

Department of Pharmacy, The Affiliated Zhuzhou Hospital Xiangya Medical College CSU, Zhuzhou, Hunan, China

Bladder cancer is the most common cancer of the urinary system. Bladder urothelial cancer accounts for 90% of bladder cancer. These two cancers have high morbidity and mortality rates worldwide. The identification of biomarkers for bladder cancer and bladder urothelial cancer helps in their diagnosis and treatment. circRNAs are considered oncogenes or tumor suppressors in cancers, and they play important roles in the occurrence and development of cancers. In this manuscript, we developed an Ensemble model, CDA-EnRWLRLS, to predict circRNA-Disease Associations (CDA) combining Random Walk with restart and Laplacian Regularized Least Squares, and further screen potential biomarkers for bladder cancer and bladder urothelial cancer. First, we compute disease similarity by combining the semantic similarity and association profile similarity of diseases and circRNA similarity by combining the functional similarity and association profile similarity of circRNAs. Second, we score each circRNA-disease pair by random walk with restart and Laplacian regularized least squares, respectively. Third, circRNA-disease association scores from these models are integrated to obtain the final CDAs by the soft voting approach. Finally, we use CDA-EnRWLRLS to screen potential circRNA biomarkers for bladder cancer and bladder urothelial cancer. CDA-EnRWLRLS is compared to three classical CDA prediction methods (CD-LNLP, DWNN-RLS, and KATZHCDA) and two individual models (CDA-RWR and CDA-LRLS), and obtains better AUC of 0.8654. We predict that circHIPK3 has the highest association with bladder cancer and may be its potential biomarker. In addition, circSMARCA5 has the highest association with bladder urothelial cancer and may be its possible biomarker.

KEYWORDS

bladder cancer, bladder urothelial cancer, circRNA, biomarker, circRNA-disease association, ensemble learning

1 Introduction

Bladder cancer is considered to be the most common cancer in the urinary system (Kamat et al., 2016). It is the fourth most common malignant tumor in men and the eighth most common in women in the Western world. In the United States and Europe, it accounts for 5%–10% among all malignancies in men. The risk with the bladder cancer infection at less than 75 years is 2%–4% in men and 0.5%–1% for women (Kirkali et al., 2005). The incidence of bladder cancer has been increasing (Kamat et al., 2016). The majority of patients with bladder cancer suffer from the less aggressive non-muscle-invasive disease, while 30% of patients suffer from muscle-invasive disease (Lopez-Beltran et al., 2021; Tran et al., 2021; Yang et al., 2021).

Bladder cancer has a metastatic potential, and thus presents a worse prognosis. It is usually grouped into three pathological categories: bladder urothelial carcinoma, bladder squamous cell carcinoma, and bladder adenocarcinoma (Black and Black, 2020; Lopez-Beltran et al., 2021). Bladder urothelial carcinoma accounts for over 90% among all cases of bladder cancer. Furthermore, bladder urothelial carcinoma can be categorized into muscle-invasive bladder cancer, which accounts for about 75% of all cases, and non-muscle-invasive bladder cancer (Kirkali et al., 2005). The all-stage five-year survival rate of bladder urothelial cancer remains approximately 80% (Lopez-Beltran et al., 2021).

Recently, the treatment of bladder cancer has obtained great progresses worldwide. Besides traditional surgical resection, radiotherapy, and chemotherapy, immunotherapy is also a promising avenue for bladder cancer treatment (Gao et al., 2021; Mancini et al., 2021). However, postoperative recurrence and distant metastasis cause five-year survival rates to still be very low for advanced bladder cancer (Fabiano et al., 2021; Roviello et al., 2021). Advanced disease or relapse of radical cystectomy is closely associated with the poor outcomes (Nouhaud et al., 2021). The first-line therapy of metastatic bladder urothelial cancer usually adopts cisplatin-based combinations, and has been unaltered over the last decades (Powles et al., 2021; Renner et al., 2021; Walia et al., 2021). Unfortunately, almost all patients with bladder urothelial cancer will finally progress and die from bladder cancer, despite their initial response to cisplatin-based combinations (Bin Riaz et al., 2021; Lopez-Beltran et al., 2021). Consequently, inferring potential biomarkers for bladder cancer is a good way to diagnose and treat it (Peng et al., 2017; Peng et al., 2018).

With the advance of sequencing technology, there are now massive amounts of RNA data (Ozsolak and Milos, 2011; Peng et al., 2020; Yang et al., 2020; Peng et al., 2022a), which help the prognosis and treatment of various diseases (Xu et al., 2020; Li et al., 2021). Circular RNAs (circRNAs) are a class of single-stranded noncoding RNA molecules that are lack of terminal 5′ caps and 3′ poly(A) tails (He et al., 2017). circRNAs are widely distributed in various organisms. They have circular features, and thus demonstrate more resistance to degradation by exonucleases and stronger stability than linear RNAs (Xia et al., 2018; Li G. et al., 2020). The estimated total number of circRNAs is approximately 1% of one of poly (A) molecules. In addition, the expression levels of the majority of circRNAs are estimated to be 5%–10% of the corresponding linear RNAs (Jeck and Sharpless, 2014; Zhang J. et al., 2021).

Although circRNAs were found in 1976, they were originally considered to be functionless by-products from aberrant RNA splicing and thus did not obtain enough attention over the past 3 decades. However, with the rapid advance of high-throughput sequencing technologies, massive differentially expressed circRNAs have been increasing discovered in human normal and malignant cells (Zhang et al., 2018; Li G. et al., 2020; Yang et al., 2021). circRNAs exist widely in various tissues, serum, and urine. The expression profiles of circRNAs demonstrate strong specificity in cell types, tissues, and developmental stages (Yang et al., 2021). Furthermore, circRNAs can regulate transcription or splicing, translate proteins, interact with RNA-binding proteins, and act as miRNA sponges (Sheng et al., 2018). A large body of evidence shows that circRNAs have dense associations with various diseases, including neurological dysfunction, cardiovascular diseases, and cancer. Here, circRNAs, as miRNA sponges, can inhibit the regulation from downstream cancer target genes. For instance, circCDR1as and circMTO1 can control gene regulation and further indirectly stimulate or inhibit tumors by binding to miR-7 and miR-9 (Vromman et al., 2021).

circRNAs have abundant associations with cancers and thus can be used as candidate cancer biomarkers (Zhang et al., 2018). An increasing amount of evidence has reported that circRNAs present in human biofluids and exosomes, and are a class of potential biomarkers of noninvasive liquid biopsies. For instance, circ-ZEB1.33 is overexpressed in hepatocellular cancer and has close links with the survival of hepatocellular cancer patients (Gong et al., 2018). In particular, substantial studies have demonstrated that circRNAs play key roles in the carcinogenesis and progression of bladder cancer. For example, circRNAs Cdr1as performs anti-oncogenic functions in bladder cancer through microRNA 135a (Li et al., 2018), BCRC-3 suppresses bladder cancer proliferation via sponging miR-182-5p/p27 (Xie et al., 2018), MYLK and circPDSS1 promote bladder cancer progression separately by modulating VEGFA/VEGFR2 signaling pathway and down-regulating miR-16 (Zhong et al., 2017; Yu et al., 2020), PRMT5 supports metastasis of bladder urothelial cancer through Sponging miR-30c (Chen et al., 2018), circSLC8A1 suppresses bladder cancer progression through regulating PTEN (Lu et al., 2019), and circMTO1 inhibits bladder cancer metastasis through sponging miR-221 (Li G. et al., 2019).

Many computational methods have been proposed to identify possible CDAs and further discovered possible circRNA biomarkers for various complex diseases including cancers by case studies (Wang CC. et al., 2021). For example, Lei et al. (Lei et al., 2018) designed a path weighted-based CDA prediction approach (PWCDA). Li et al. (Li Y. et al., 2019; Li

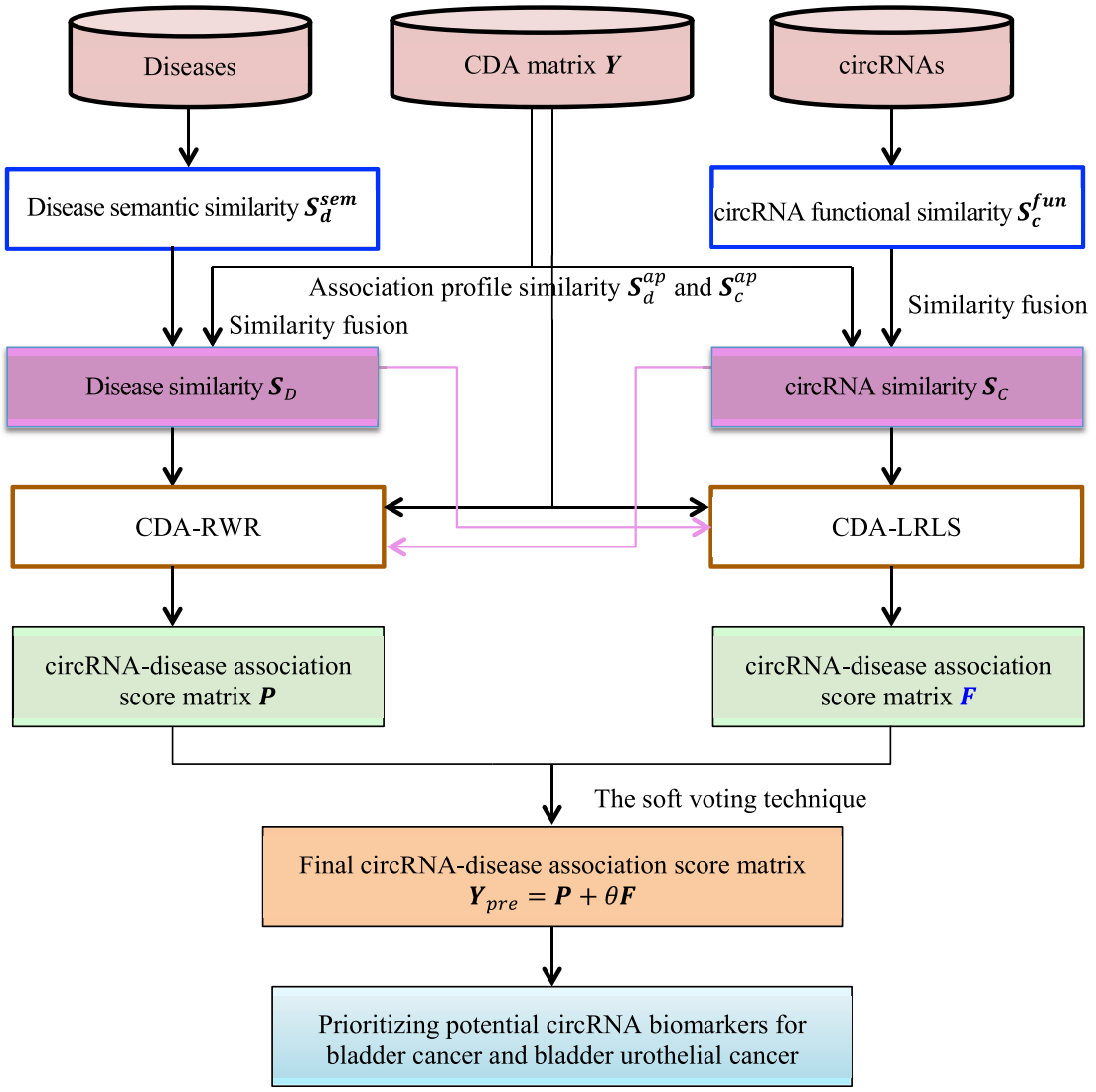


FIGURE 1
Flowchart of the proposed CDA-EnRWLRLS model.

J. et al., 2020) explored two CDA identification models (NCPCDA and DWNCPCDA) based on network consistency projection. Zhang et al. (Zhang et al., 2019) developed a linear neighborhood label propagation algorithm for CDA identification. Deepthi et al. (Deepthi and Jereesh (2020) used autoencoder and deep neural network and explored an ensemble model to predict CDAs. Lu et al. (Lu et al. (2021) improved CDA prediction using convolutional and recurrent neural networks. Wang et al. (Wang et al., 2020; Wang et al. 2021b; Wang et al., 2021c) proposed three CDA identification methods (GCNCDA, MGRCDA, and SGANRDA) based on graph convolutional network, metagraph recommendation, and semi-supervised generative adversarial network, respectively. These methods efficiently predicted possible CDAs.

In this study, inspired by computational CDA prediction methods, we develop an ensemble model, CDA-RWLRLS, to find potential circRNA biomarkers for bladder cancer and bladder urothelial cancer based on known CDAs. CDA-EnRWLRLS first computes circRNA similarity by integrating their functional similarity and association profile similarity, and it computes disease similarity by integrating their semantic similarity and association profile similarity. Second, CDA-EnRWLRLS computes the association probability for each circRNA-disease pair based on random walk with restart and Laplacian regularized least squares. Third, the prediction results obtained by these two models are integrated by the soft voting method. We finally use the proposed CDA-EnRWLRLS model to identify possible

circRNAs associated with bladder cancer and bladder urothelial cancer.

2 Materials and methods

2.1 Materials

2.1.1 Human circRNA-disease associations

circRNA-disease association data can be downloaded from the circR2Disease database (Fan et al., 2018a). This database provides 739 experimentally confirmed CDAs from 661 circRNAs and 100 diseases. We remove redundant elements related to mice and rats and achieve a human circRNA-disease association dataset containing 650 associations between 585 circRNAs and 88 diseases. In particular, suppose that $C = \{c_1, c_2, \ldots, c_m\}$ and $D = \{d_1, d_2, \ldots, d_n\}$ separately denote the sets of m circRNAs and n diseases, then we construct a binary matrix $\boldsymbol{Y} \epsilon R^{m \times n}$ to depict circRNA-disease associations by Eq. 1:

$$Y_{ij} = \begin{cases} 1 & \text{If circRNA } c_i \text{ associates with } d_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2.1.2 Disease semantic similarity

Many studies have computed disease semantic similarity to screen credible noncoding RNAs for a query disease. Inspired by these methods, we investigate disease similarity to improve the prediction performance. Disease semantic similarity can be computed based on corresponding disease ontology. The disease ontology is often represented using a directed acyclic graph and can be downloaded from http://disease-ontology.org/. For two query diseases and corresponding ontology term sets from the two diseases d_i and d_j, their semantic similarity can be scored by the "doSim" function in the DOSE software package, which can be downloaded from http://www.bioconductor.org/packages/release/bioc/html/DOSE.html (Yu et al., 2015). Finally, we compute the semantic similarity matrix S_d^{sem} among n diseases.

2.1.3 circRNA functional similarity

To measure the functional similarity between two circRNAs, we utilize the semantic similarity of two diseases linking to the two circRNAs. In particular, suppose that D_i and D_j denote the disease groups linking to circRNAs c_i and c_j, the functional similarity between c_i and c_j can be computed by Eq. 2:

$$\boldsymbol{S}_c^{fun} = \frac{\sum_{1 \le p \le |D_i|} S\left(d_p, D_j\right) + \sum_{1 \le p \le |D_j|} S\left(d_q, D_i\right)}{|D_i| + |D_j|} \quad (2)$$

and

$$S\left(d_p, D_j\right) = \max_{1 \le t \le |D_j|}\left(S_d^{sem}\left(d_p, d_t\right)\right) \quad (3)$$

where $S(d_p, D_j)$ denotes the similarity between disease d_p linking to circRNA c_i and disease set D_j linking to circRNA c_j.

2.2 Methods

In this manuscript, we develop circRNA-Disease Association prioritization method (CDA-EnRWLRLS) by an Ensemble of Random Walk with restart and Laplacian Regularization Least Squares. First, CDA-EnRWLRLS measures circRNA functional similarity and disease semantic similarity. Second, it computes association profile similarity of circRNAs and diseases, respectively. Third, functional similarity and association profile similarity of circRNAs are combined to obtain the final circRNA similarity. Similarly, disease similarity is fused. Fourth, random walk with restart and Laplacian regularization least squares are used to score each circRNA-disease pair. Fifth, the final association score matrix is obtained by integrating the results from random walk with restart and Laplacian regularization least squares based on the soft voting strategy. Finally, CDA-EnRWLRLS is applied to find possible circRNA biomarkers for bladder cancer and bladder urothelial cancer. The flowchart of CDA-EnRWLRLS is shown in Figure 1.

2.2.1 Association profile similarity of circRNAs and diseases

For two diseases with known ontology terms, we can compute their semantic similarity based on their ontology terms. However, semantic similarity computation may fail for two diseases without ontology terms. Thus, we introduce association profile similarity to further complement similarity measurement of circRNAs and diseases.

Suppose that the association profile $\boldsymbol{Y}(i, :)$ of a circRNA c_i is represented as the ith row of a CDA matrix $\boldsymbol{Y}$. $\boldsymbol{Y}(i, :)$ describes information from all diseases associated with c_i. Association profile similarity between two circRNAs (i.e., (c_i, c_j)) can be computed by Eq. 4:

$$\boldsymbol{S}_c^{ap}\left(c_i, c_j\right) = \exp\left(-\gamma_c \left\|\boldsymbol{Y}(i, :) - \boldsymbol{Y}(j, :)\right\|^2\right)$$
$$\gamma_c = \gamma_c^{'} \Big/ \left(\frac{1}{m}\sum_{k=1}^{m} \left\|\boldsymbol{Y}(k, :)\right\|^2\right) \quad (4)$$

where $\gamma_c^{'}$ is bandwidth parameter and set as the default value of 1.

Similarly, association profile similarity between two diseases (i.e., (d_i, d_j)) can be computed by Eq. 5:

$$\boldsymbol{S}_d^{ap}\left(d_i, d_j\right) = \exp\left(-\gamma_d \left\|\boldsymbol{Y}(:, i) - \boldsymbol{Y}(:, j)\right\|^2\right)$$
$$\gamma_d = \gamma_d^{'} \Big/ \left(\frac{1}{n}\sum_{k=1}^{n} \left\|\boldsymbol{Y}(:, k)\right\|^2\right) \quad (5)$$

where $\gamma_d^{'}$ indicates bandwidth parameter and set as the default value of 1.

2.2.2 Similarity fusion

circRNA functional similarity S_c^{fun}, disease semantic similarity S_d^{sem}, and association profile similarity of circRNAs and diseases (S_c^{ap} and S_d^{ap}) are fused to obtain the final circRNA similarity matrix S_C and disease similarity S_D by Eqs 6, 7:

$$S_C = \alpha_c S_c^{fun} + (1 - \alpha_c) S_c^{ap} \quad (6)$$

$$S_D = \alpha_d S_d^{sem} + (1 - \alpha_d) S_d^{ap} \quad (7)$$

The parameter α_c is used to balance the importance between functional similarity and association profile similarity of circRNAs in Eq. 6 and α_d is used to balance the important between semantic similarity and association profile similarity of diseases in Eq. 7.

2.2.3 Random walk with restart for CDA prediction

Random walk algorithm has been widely used and obtained better performance in various association prediction fields (Peng et al., 2021a). In this study, we utilize Random Walk with Restart for CDA prediction on the heterogeneous circRNA-disease network (CDA-RWR). We first train the random walk with restart model on the CDA dataset and screen possible CDAs with the highest association probability from unknown circRNA-disease pairs on the dataset.

First, circRNA similarity network N_c, disease similarity network N_d, and CDA network N_a are used to build a heterogeneous circRNA-disease network.S_c, S_d, and Y correspond to adjacency matrices of the three networks, respectively. Consequently, the heterogeneous circRNA-disease network can be represented as: $W = \begin{bmatrix} S_C & Y \\ Y^T & S_D \end{bmatrix}$, where Y^T is the transpose of Y.

Second, we compute the transition probability of random walk on the heterogeneous circRNA-disease network. Suppose that $W = \begin{bmatrix} W_{cc} & W_{cd} \\ W_{dc} & W_{dd} \end{bmatrix}$ denote the transition matrix, where W_{cc} and W_{dd} separately indicate the walk within the circRNA network and the disease network, W_{cd} and W_{dc} separately represent the jump from the circRNA network to the disease network and the disease network to the circRNA network. For a known jumping probability μ from the circRNA network to the disease network or from the disease network to the circRNA network, the transition probability from circRNAs c_i to c_j can be calculated by Eq. 8:

$$W_{cc}(i,j) = \begin{cases} \dfrac{S_C(i,j)}{\sum_{k=1}^{m} S_C(i,k)} & if \sum_{k=1}^{n} Y(i,k) = 0 \\ \dfrac{(1-\mu) S_C(i,j)}{\sum_{k=1}^{m} S_C(i,k)} & otherwise \end{cases}, \quad (8)$$

The transition probability from circRNA c_i to disease d_j can be calculated by Eq. 9:

$$W_{cd}(i,j) = \begin{cases} \dfrac{\mu Y(i,j)}{\sum_{k=1}^{n} Y(i,k)} & if \sum_{k=1}^{n} Y(i,k) \neq 0 \\ 0 & otherwise \end{cases}, \quad (9)$$

The transition probability from diseases d_i to d_j can be calculated by Eq. 10:

$$W_{dd}(i,j) = \begin{cases} \dfrac{S_d(i,j)}{\sum_{k=1}^{n} S_d(i,k)} & if \sum_{k=1}^{m} Y(k,i) = 0 \\ \dfrac{(1-\mu) S_d(i,j)}{\sum_{k=1}^{n} S_d(i,k)} & otherwise \end{cases}, \quad (10)$$

The transition probability from disease d_i to circRNA c_j can be calculated by Eq. 11:

$$W_{dv}(i,j) = \begin{cases} \dfrac{\mu Y(j,i)}{\sum_{k=1}^{n} Y(k,i)} & if \sum_{k=1}^{n} Y(k,i) \neq 0 \\ 0 & otherwise \end{cases}, \quad (11)$$

For a query circRNA/disease, it can either stay in the current network with a restart probability $\beta \in (0,1)$ or jump to another network graph. Consequently, we can compute association probability for each circRNA-disease pair at the $(t+1)$-th step by Eq. 12:

$$p_{t+1} = \beta W p_t + (1 - \beta) p_0, \quad (12)$$

where p_t denotes the association probability matrix at the t-th step, p_0 denotes the initial probability and $p_0 = \begin{bmatrix} \lambda u_0 \\ (1-\lambda) v_0 \end{bmatrix}$, where u_0 and v_0 indicate the initial probability on the circRNA and disease network, respectively. When we want to discover possible circRNAs associated with a query disease d_i, it is regarded as a seed in the disease network. Consequently, d_i is assigned as 1 and other disease nodes are 0, thereby building the initial probability matrix of the disease network v_0. All nodes in the circRNA network u_0 are assigned as an equal probability whose sum is 1. The parameter β is used to balance the importance of the circRNA network and the disease network.

2.3 Laplacian regularized least squares for CDA prediction

We can calculate association probability for each circRNA-disease pair based on random walk with restart. However, for random walk with restart, the jump probability is measured by known CDAs and the circRNA and disease similarity matrices.

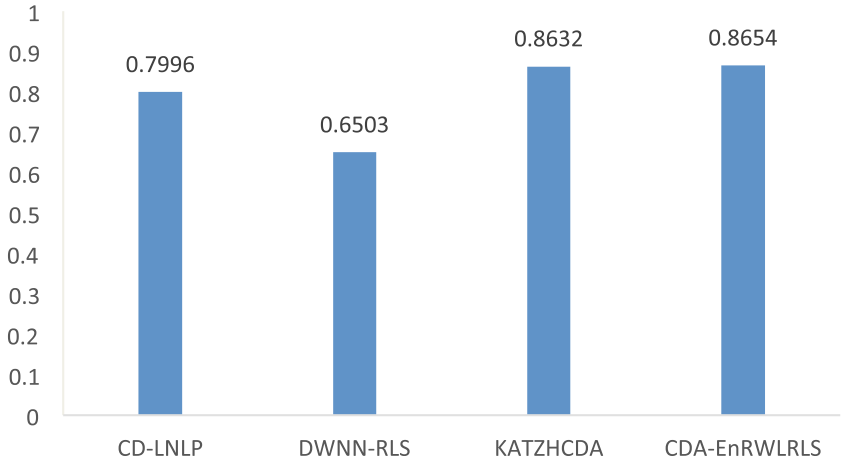


FIGURE 2
The AUC values of CDA-EnRWLRLS and other three method.

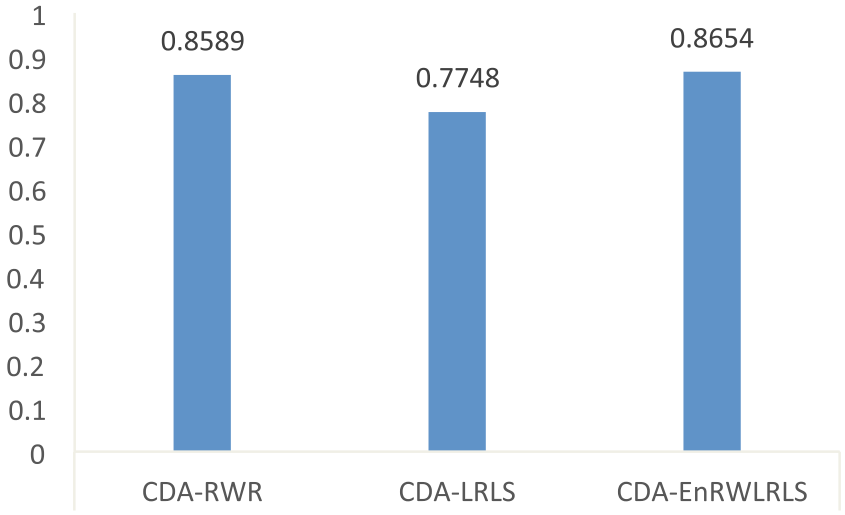


FIGURE 3
The AUC values of CDA-EnRWLRLS and CDA-RWR and CDA-LRLS.

For a circRNA c_i in a CDA network, if two other circRNAs c_j and c_k have the equal similarity with c_i, c_j and c_k will contribute to the jump between nodes at an equal probability. However, the circRNA that exhibits lower similarities with other circRNAs should have more contribution to the jump. Thus, we further use Laplacian regularized least squares (Shen et al., 2022) to compute association probability for each circRNA-disease pair.

First, we compute the circRNA Laplacian matrix $\boldsymbol{L}_c$ and the disease Laplacian matrix $\boldsymbol{L}_d$ by Eqs 13, 14:

$$\boldsymbol{L}_c = (\boldsymbol{A}_c)^{-1/2}(\boldsymbol{A}_c - \boldsymbol{A}_c)(\boldsymbol{A}_c)^{-1/2} \tag{13}$$

$$\boldsymbol{L}_d = (\boldsymbol{A}_d)^{-1/2}(\boldsymbol{A}_d - \boldsymbol{A}_d)(\boldsymbol{A}_d)^{-1/2} \tag{14}$$

where $\boldsymbol{A}_c/\boldsymbol{A}_d$ indicates the diagonal matrix of circRNA/disease similarity matrix and $A_c(i,i)/A_d(j,j)$ is the summation of the i-th/j-th row of $\boldsymbol{S}_C/\boldsymbol{S}_D$.

Second, we define the loss functions of Laplacian regularization least squares in the circRNA and disease spaces based on the Laplacian matrices $\boldsymbol{L}_c$ and $\boldsymbol{L}_d$ by Eqs 15, 16, respectively:

$$\min_{\boldsymbol{F}_c}\left[\left\|\boldsymbol{Y}^T - \boldsymbol{F}_c\right\|_F^2 + \gamma_c\left\|\boldsymbol{F}_c \cdot \boldsymbol{L}_c \cdot (\boldsymbol{F}_c)^T\right\|_F^2\right] \tag{15}$$

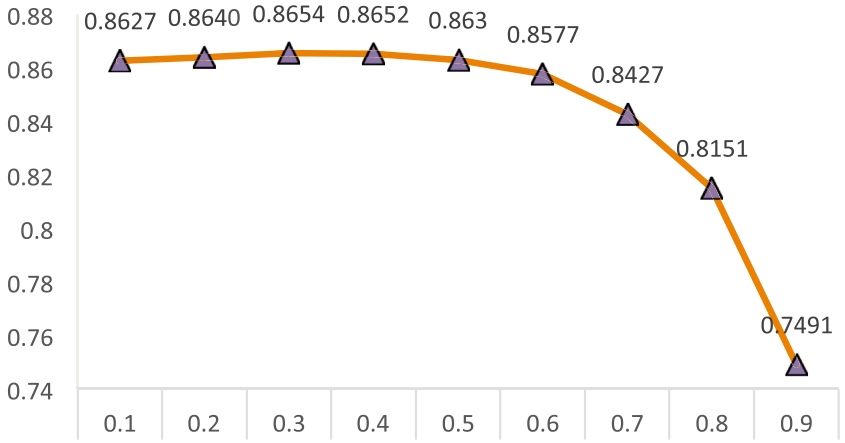


FIGURE 4
The effect of θ on the prediction performance for CDA-EnRWLRLS.

$$\min_{F_d}\left[\|Y - F_d\|_F^2 + \gamma_d\left\|F_d \cdot L_d \cdot (F_d)^T\right\|_F^2\right] \tag{16}$$

where Y^T, $(F_c)^T$, and $(F_d)^T$ separately indicate the transposes of Y, F_c, and F_d, $\|\cdot\|_F$ indicates the Frobenius norm, and γ_c and γ_d indicate trade-off parameters. The Laplacian regularized least square models (15) and (16) can be solved by Eqs 17, 18:

$$F_c = S_C\left(S_C + \gamma_c \cdot L_c \cdot S_c\right)^{-1} Y^T \tag{17}$$

$$F_d = S_d\left(S_d + \gamma_d \cdot L_d \cdot S_d\right)^{-1} Y \tag{18}$$

Finally, the association probability for each circRNA-disease pair by Laplacian regularized least squares can be computed by Eq. 19:

$$F = \frac{1}{2}\left(F_c + F_d\right) \tag{19}$$

2.4 Ensemble learning for CDA prediction

Ensemble learning integrates multiple results from individual models and demonstrates better performance compared to individual models (Zhou et al., 2021a; Peng et al., 2022b). Therefore, in this study, we develop an ensemble learning model by combining random walk with restart and Laplacian regularized least squares to improve the CDA's prediction performance by Eq. 20:

$$Y_{pre} = P + \theta F \tag{20}$$

where Y_{pre} denotes the predicted final CDA score matrix, P and F denote the computed CDA probability matrices based on random walk with restart and Laplacian regularized least squares, respectively. θ is used to weigh the importance of results computed by the above two models.

3 Experiments

3.1 Experimental settings

For similarity computation, the weights between biological feature similarity and association profile similarity α_c and α_d are set as 0.5. For random walk with restart, the restart probability β is set as 0.2, and λ and μ are set as 0.1 and 0.6, respectively. For Laplacian regularized least squares, both γ_c and γ_d are set as 0.95 and 0.2, respectively. For ensemble learning model, θ is set as 0.3. The parameters in other three comparative methods are set as defaults provided by the corresponding methods. We conduct 5-fold cross validation for 10 times. The final prediction performance is from the average value of the 10 experiments. AUC (area under the receiver operating characteristic curve) has been widely used to evaluate the performance of CDA prediction methods. Larger AUC denotes better performance. Thus, we use AUC to measure the performance of our proposed method.

3.2 Performance comparison with five CDA prediction methods

Several comparative experiments are conducted to measure the performance of our proposed CDA-EnRWLRLS model. CD-LNLP (Zhang et al., 2019), DWNN-RLS (Yan et al., 2018), KATZHCDA (Fan et al., 2018b), and CDA-EnRWLRLS are conducted on the preprocessed CDA dataset. CD-LNLP (Zhang et al., 2019) is a linear neighborhood label propagation-based algorithm for CDA prediction. DWNN-RLS (Yan et al., 2018) used regularized least squares to predict

TABLE 1 The inferred top 30 circRNAs associated with bladder cancer.

Rank	circRNAs	Evidence
1	hsa_circ_0000172	circRNADisease
2	hsa_circ_0002495	circRNADisease
3	Chr22: 28943661	circRNADisease
4	Chr5: 158368701	circRNADisease
5	Chr9: 74522734	circRNADisease
6	circRNA BCRC4/hsa_circ_001598/hsa_circ_0001577	circRNADisease
7	hsa_circ_0003221/circPTK2	circRNADisease
8	hsa_circ_0091017	circRNADisease
9	hsa_circ_0002024	circRNADisease
10	circMylk/circRNA-MYLK/hsa_circ_0002768	circRNADisease
11	circTCF25/hsa_circ_0041103	circRNADisease
12	circFAM169A/hsa_circ_0007158	circRNADisease
13	circTRIM24/hsa_circ_0082582	circRNADisease
14	circBC048201/hsa_circ_0061265	circRNADisease
15	hsa_circRNA_100782/circHIPK3/hsa_circ_0000284	Unconfirmed
16	circZFR/hsa_circRNA_103809/hsa_circ_0072088	Unconfirmed
17	Cir-ITCH/hsa_circ_0001141/hsa_circ_001763	Unconfirmed
18	circSMARCA5/hsa_circ_0001445	PMID: 35712125, 35116915, 34482767
19	hsa_circ_0001649	PMID: 35200157
20	CDR1as/ciRS-7/hsa_circ_0001946	PMID: 29694981, 31131537, 33335899

possible CDAs. KATZHCDA (Fan et al., 2018b) discovered CDA candidates based on the KATZ measurement (Zhou et al., 2020). Figure 2 shows the AUC values computed by these four CDA prediction methods.

From Figure 2, we can find that CDA-EnRWLRLS is significantly better than CD-LNLP (Zhang et al., 2019), DWNN-RLS (Yan et al., 2018), and KATZHCDA (Fan et al., 2018b) based on the AUC value. Compared to the three models, CDA-EnRWLRLS obtains the highest AUC of 0.8654, outperforming 7.60%, 24.86%, and 0.25%, respectively. In particular, DWNN-RLS used regularized least squares with Kronecker product kernel for CDA prediction. Disease similarity was computed by their semantic similarity and Gaussian association profile similarity. Meanwhile, circRNA similarity was computed by their Gaussian association profiles. CDA-EnRWLRLS uses an ensemble model to identify possible CDAs. Similar to DWNN-RLS, CDA-EnRWLRLS computes disease similarity. However, CDA-EnRWLRLS computes circRNA similarity by their functional similarity and Gaussian association profile similarity. Furthermore, CDA-EnRWLRLS still computes association score between each circRNA-disease pair using random walk with restart except Laplacian regularized least squares and integrates the results from the two models by the soft voting technique. Therefore, CDA-EnRWLRLS outperforms DWNN-RLS, which demonstrates its powerful CDA prediction ability.

3.3 Performance evaluation of ensemble learning model with individual models

Our proposed CDA-EnRWLRLS model is an ensemble of two state-of-the-art models (i.e., random walk with restart and Laplacian regularized least squares). To evaluate the performance of ensemble learning model and individual models, we conducted 5-fold cross validation experiment for CDA-EnRWLRLS and random walk with restart (CDA-RWR) and Laplacian regularized least squares (CDA-LRLS) on the CDA dataset. Figure 3 shows the AUC values computed by CDA-EnRWLRLS, CDA-RWR, and CDA-LRLS. From Figure 3, we can find that CDA-EnRWLRLS obtains better AUC than two individual models, CDA-RWR and CDA-LRLS, which shows that the proposed ensemble learning-based model can outperforms individual models.

3.4 Evaluation of parameter sensitivity

In this study, we ensemble two individual models, random walk with restart and Laplacian regularized least squares. However, the two models may have different effects on the CDA prediction performance. To evaluate their effect on the performance, we consider θ in the range of [0.1, 0.9] with stride of 0.1. The results are shown in Figure 4.

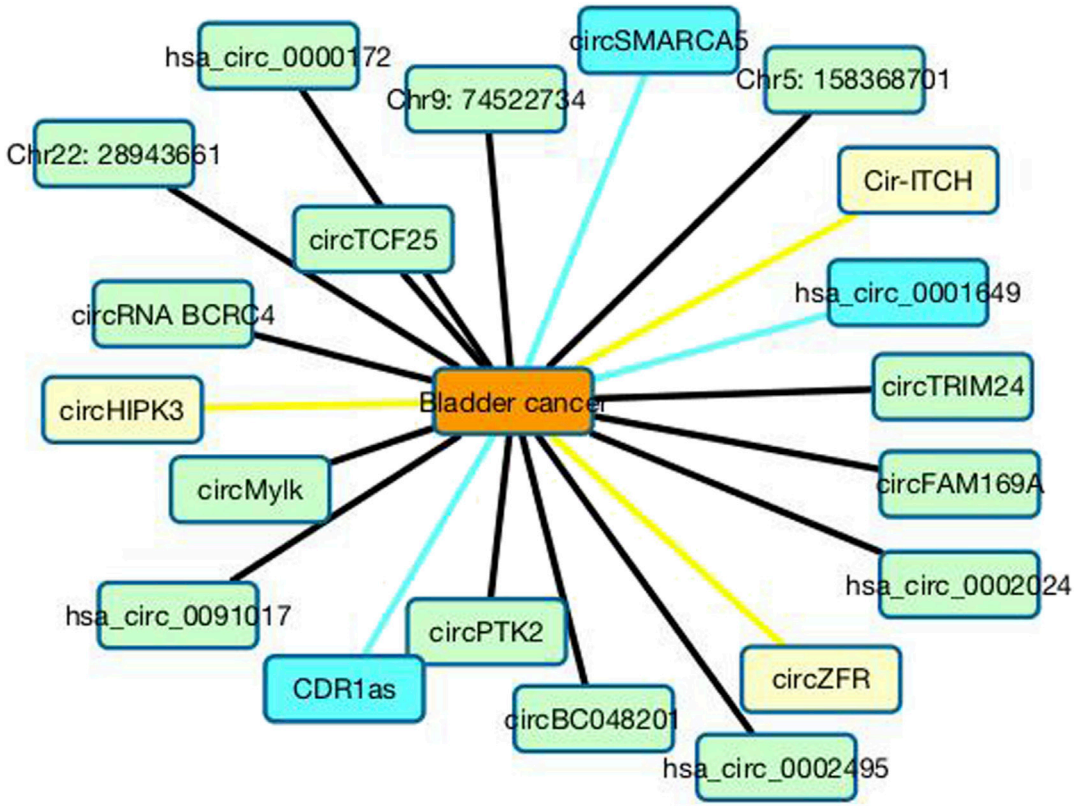


FIGURE 5
Associations between the top 20 circRNAs with bladder cancer. Black lines represent associations that have validated in the CDA dataset. Sky blue lines represent associations that are unknown in the CDA dataset but can be validated by related literatures. Yellow lines represent association that are unknown in the CDA dataset and need validation.

From Figure 4, we can find that AUC computed by CDA-EnRWLRLS gradually increases when the parameter θ is from 0.1 to 0.3. Its computed AUCs gradually decrease when the parameter θ is from 0.3 to 0.9. In other words, CDA-EnRWLRLS obtains the best AUC when the parameter θ is 0.3. Thus, the parameter θ is finally set as 0.3.

3.5 Case study

We consequently compute the association score for each circRNA-disease pair. In particular, we compute association abilities between all circRNAs and bladder cancer and bladder urothelial cancer to analyze any possible associations between these circRNAs and the two cancers, and to further screen for potential circRNA biomarkers for them.

3.5.1 circRNA biomarker analysis for bladder cancer

Bladder cancer is a heterogeneous disease with high morbidity and mortality rates (Kamat et al., 2016). It has been estimated that about 73,510 new cases of bladder cancer were diagnosed in the United States in 2012. During the same period, about 14,880 patients died from bladder cancer (Clark et al., 2013). To analyze circRNA biomarkers for bladder cancer, we compute association between all circRNAs and bladder cancer after training CDA-EnRWLRLS. Table 1 gives the top 20 circRNAs that are predicted to have the highest association scores with bladder cancer.

In the CDA dataset, 15 circRNAs are known to associate with bladder cancer among 585 circRNAs. From Table 1, we can find that the 15 circRNAs are predicted to have the highest association scores with bladder cancer and are ranked as the top 15. Furthermore, we predict that circHIPK3 may associate with bladder cancer with the ranking of 16. Furthermore, circHIPK3 is a promising cancer-related circRNA (Zhang et al., 2020). It can regulate cell growth through sponging multiple miRNAs (Zheng et al., 2016). For instance, circHIPK3 can regulate cell proliferation and migration in hepatocellular cancer by sponging miR-124 (Chen X. et al., 2018), modulate

TABLE 2 The inferred top 30 circRNAs associated with bladder urothelial cancer.

Rank	circRNAs	Evidence
1	hsa_circRNA_100782/circHIPK3/hsa_circ_0000284	circRNADisease
2	circSMARCA5/hsa_circ_0001445	Unconfirmed
3	hsa_circ_0001649	Unconfirmed
4	Cir-ITCH/hsa_circ_0001141/hsa_circ_001763	Unconfirmed
5	CDR1as/ciRS-7/hsa_circ_0001946	PMID: 32658427
6	circZFR/hsa_circRNA_103809/hsa_circ_0072088	Unconfirmed
7	CircDOCK1/hsa_circ_100721	Unconfirmed
8	circRNA_100290/hsa_circ_0013339/hsa_circ_100290	Unconfirmed
9	circPVT1/hsa_circ_0001821	PMID: 34902986
10	hsa_circ_0001313/circCCDC66	Unconfirmed
11	circGFRA1/hsa_circ_005239	Unconfirmed
12	circZNF609/hsa_circ_0000615	Unconfirmed
13	circWDR77/hsa_circ_0013509	Unconfirmed
14	hsa_circ_0000096/circHIAT1/hsa_circ_001013	Unconfirmed
15	circRNA_000167/hsa_circRNA_000167/hsa_circ_0000518	Unconfirmed
16	hsa_circ_0007534	Unconfirmed
17	circPRKCI/hsa_circ_0067934	Unconfirmed
18	hsa_circRNA_103110/hsa_circ_103110/hsa_circ_0004771	Unconfirmed
19	circ-Foxo3/hsa_circ_0006404	PMID: 31903146
20	circFUT8/hsa_circRNA_101368/hsa_circ_0003028	Unconfirmed

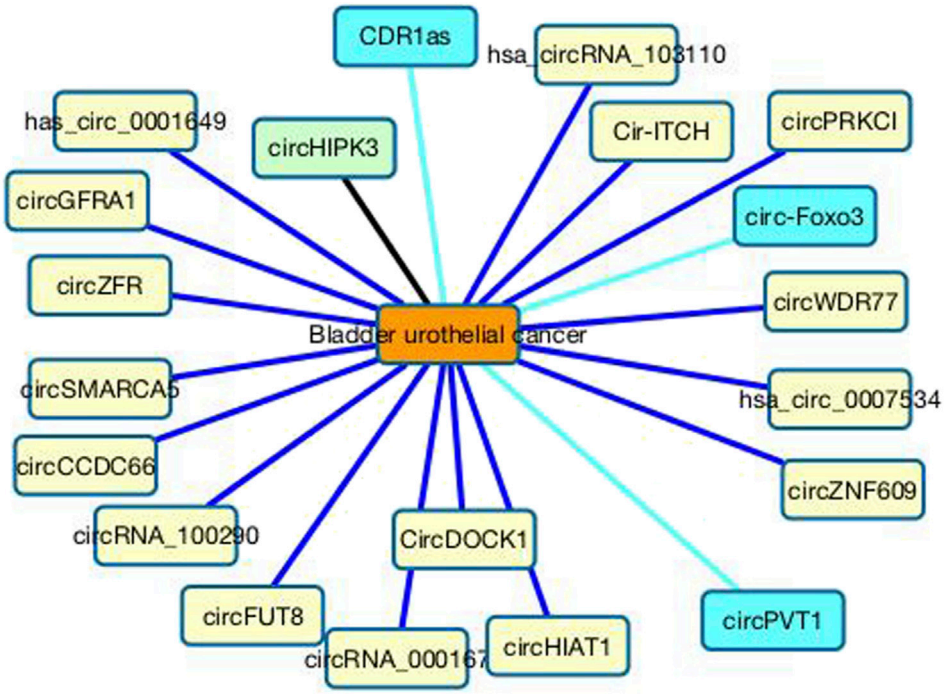


FIGURE 6
Associations between the top 20 circRNAs with bladder urothelial cancer. Black lines represent associations that have validated in the CDA dataset. Sky blue lines represent associations that are unknown in the CDA dataset but can be validated by related literatures. Blue lines represent association that are unknown in the CDA dataset and need validation.

autophagy in STK11 mutant lung cancer (Chen et al., 2020), and promote glioma progression as a prognostic marker (Jin et al., 2018). The overexpression of circHIPK3 can accelerate the proliferation and invasion of prostate cancer cells (Cai et al., 2019). Its inhibition can block angiotensin II-induced cardiac fibrosis (Ni et al., 2019). In this study, we infer that circHIPK3 may be a biomarker of bladder cancer and need experimental validation. Figure 5 shows the association information between the top 20 circRNAs with bladder cancer.

3.5.2 circRNA biomarker analysis for bladder urothelial cancer

Over 90% bladder cancer is bladder urothelial cancer. Bladder urothelial cancer is a common malignancy with high morbidity and mortality worldwide (Cancer Genome Atlas Research Network, 2014). In the United Sates, bladder urothelial cancer is one of the main histologic subtypes (Clark et al., 2013). However, no molecularly targeted agent has been applied to the treatment, until now. To infer potential circRNA biomarkers for bladder urothelial cancer, we compute association scores between all circRNAs and bladder urothelial cancer using CDA-EnRWLRLS. Table 2 gives the top 20 circRNAs that are predicted to have the highest association scores with bladder urothelial cancer.

In the CDA dataset, only one circRNA, circHIPK3, associates with bladder urothelial cancer among all potential 585 circRNAs. We predict that SMARCA5 may associate with bladder urothelial cancer with the ranking of 2. SMARCA5 is a member of the ISWI family that is involved in chromatin remodeling. It can regulate chromosome remodeling through diverse mechanisms, hinder cell proliferation, and assist apoptosis by sponging miRNAs. Its expression may boost the susceptibility of cells to chemotherapy, boost the sensitivity of cancer detection, promote early diagnosis, and help the treatment of chemotherapy-resistant cancers (Qin and Wan, 2022). Its expression level has a certain association with clinical features of many cancers. For instance, SMARCA5 can promote cell proliferation in bladder cancer and prostate cancer (Tan et al., 2019), suppress colorectal cancer progression (Miao et al., 2020), inhibit tumor metastasis in cervical cancer (Zhang X. et al., 2021) and inhibit cell proliferation, migration, and invasion in non-small cell lung cancer (Wang et al., 2019), and boost cell migration and invasion as well as inhibit cell apoptosis in bladder cancer (Kong et al., 2017; Tan et al., 2019). Many studies have reported that circSMARCA5 plays a key role in the occurrence and development of cancer. Moreover, it also serves as a reliable indicator of tumor screening or cancer prognosis evaluation (Qin and Wan, 2022). Therefore, SMARCA5 is a diagnostic and prognostic biomarker of cancer and has obtained wide attention. In this study, we predict that SMARCA5 may be potential biomarker of bladder urothelial cancer; however, this needs validation. Figure 6 shows the association information between the top 20 circRNAs with bladder urothelial cancer.

4 Discussion and conclusion

Bladder cancer, including bladder urothelial cancer, is a common and complex disease. These cancers have caused high morbidity and mortality. The identification of biomarkers for bladder cancer and bladder urothelial cancer can help in their prognosis and treatment. In this manuscript, we developed an ensemble learning model, CDA-EnRWLRLS, to discover potential circRNA biomarkers for the two cancers based on CDA association prediction.

CDA-EnRWLRLS first computes circRNA similarity and disease similarity by fusing semantic similarity and association profile similarity of diseases and functional similarity and association profile similarity of circRNAs. Second, it scores each circRNA-disease pair by random walk with restart and Laplacian regularized least squares, respectively. Third, the results computed by random walk with restart and Laplacian regularized least squares are integrated by the soft voting approach based on ensemble learning. Finally, it is applied to discover potential circRNA biomarkers for bladder cancer and bladder urothelial cancer.

CDA-EnRWLRLS is compared to three classical CDA prediction methods (CD-LNLP, DWNN-RLS, and KATZHCDA) and two individual models (CDA-RWR and CDA-LRLS). The results show that CDA-EnRWLRLS computes relatively better AUC, which demonstrates its relatively powerful CDA prediction ability. We predict that circHIPK3 and SMARCA5 may be potential biomarkers of bladder cancer and bladder urothelial cancer, respectively.

CDA-EnRWLRLS has two advantages: on the one hand, it better fuses biological features and association features of diseases and circRNAs; while on the other hand, it combines two individual classical association prediction models to obtain the powerful association prediction performance from different bioinformatics tools. Although CDA-EnRWLRLS computed better CDA inference ability, the circRNA functional similarity was calculated indirectly by disease semantic similarity. Moreover, its prediction performance needs further improvement. In the future, we will consider biological features of circRNAs and develop more efficient machine learning,

especially ensemble learning models (Zhou et al., 2021a; Peng et al., 2022a) and deep learning models (Peng et al., 2021b; Zhou et al., 2021b; Sun et al., 2022; Yang et al., 2022) to discover potential biomarkers for bladder cancer and bladder urothelial cancer.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, and further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization: QS and LW; Methodology: QS, QT, and LW; Project administration: QS, XL, and LW; Software: QS, QT, XL, and LW; Writing—original draft: QS; Writing—review and editing: QS and LW.

Funding

This research was funded by the Natural Science Foundation of Hunan province (Grant 2020JJ5996).

Conflict of interest

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

Publisher's note

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

References

Bin Riaz, I., Khan, A. M., Catto, J. W. F., and Hussain, S. A. (2021). Bladder cancer: Shedding light on the most promising investigational drugs in clinical trials.. *Expert Opin. Investig. Drugs* 30 (8), 837–855. doi:10.1080/13543784.2021.1948999

Black, A. J., and Black, P. C. (2020). Variant histology in bladder cancer: Diagnostic and clinical implications. *Transl. Cancer Res.* 9 (10), 6565–6575. doi:10.21037/tcr-20-2169

Cai, C., Zhi, Y., Wang, K., Zhang, P., Ji, Z., Xie, C., et al. (2019). CircHIPK3 overexpression accelerates the proliferation and invasion of prostate cancer cells through regulating miRNA-338-3p.. *Onco. Targets. Ther.* 12, 3363–3372. doi:10.2147/OTT.S196931

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma[J]. *Nature* 507 (7492), 315. doi:10.1038/nature12965

Chen, G., Shi, Y., Liu, M., and Sun, J. (2018). circHIPK3 regulates cell proliferation and migration by sponging miR-124 and regulating AQP3 expression in hepatocellular carcinoma.. *Cell. Death Dis.* 9 (2), 1–13. doi:10.1038/s41419-017-0204-3

Chen, X., Chen, R. X., Wei, W. S., Li, Y. H., Feng, Z. H., Tan, L., et al. (2018). PRMT5 circular RNA promotes metastasis of urothelial carcinoma of the bladder through sponging miR-30c to induce epithelial-mesenchymal transition.. *Clin. Cancer Res.* 24 (24), 6319–6330. doi:10.1158/1078-0432.CCR-18-1270

Chen, X., Mao, R., Su, W., Yang, X., Geng, Q., Guo, C., et al. (2020). Circular RNA circHIPK3 modulates autophagy via MIR124-3p-STAT3-PRKAA/AMPKα signaling in STK11 mutant lung cancer.. *Autophagy* 16 (4), 659–671. doi:10.1080/15548627.2019.1634945

Clark, P. E., Agarwal, N., Biagioli, M. C., Eisenberger, M. A., Greenberg, R. E., Herr, H. W., et al. (2013). Bladder cancer.. *J. Natl. Compr. Canc. Netw.* 11 (4), 446–475. doi:10.6004/jnccn.2013.0059

Deepthi, K., and Jereesh, A. S. (2020). An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network.. *Gene* 762, 145040. doi:10.1016/j.gene.2020.145040

Fabiano, E., Durdux, C., Dufour, B., Mejean, A., Thiounn, N., Chretien, Y., et al. (2021). Long-term outcomes after bladder-preserving tri-modality therapy for patients with muscle-invasive bladder cancer.. *Acta Oncol.* 60 (6), 794–802. doi:10.1080/0284186X.2021.1915498

Fan, C., Lei, X., Fang, Z., Jiang, Q., and Wu, F-X. (2018a). CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases[J]. *Database* 2018, bay044. doi:10.1093/database/bay044

Fan, C., Lei, X., and Wu, F. X. (2018b). Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks.. *Int. J. Biol. Sci.* 14 (14), 1950–1959. doi:10.7150/ijbs.28260

Gao, S., Yang, X., Xu, J., Qiu, N., and Zhai, G. (2021). Nanotechnology for boosting cancer immunotherapy and remodeling tumor microenvironment: The horizons in cancer treatment. *ACS Nano* 15 (8), 12567–12603. doi:10.1021/acsnano.1c02103

Gong, Y., Mao, J., Wu, D. I., Wang, X., Li, L., Zhu, L., et al. (2018). Circ-ZEB1.33 promotes the proliferation of human HCC by sponging miR-200a-3p and upregulating CDK6.. *Cancer Cell. Int.* 18 (1), 116–119. doi:10.1186/s12935-018-0602-3

He, J., Xie, Q., Xu, H., Li, J., and Li, Y. (2017). Circular RNAs and cancer.. *Cancer Lett.* 396, 138–144. doi:10.1016/j.canlet.2017.03.027

Jeck, W. R., and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs.. *Nat. Biotechnol.* 32 (5), 453–461. doi:10.1038/nbt.2890

Jin, P., Huang, Y., Zhu, P., Zou, Y., Shao, T., and Wang, O. (2018). CircRNA circHIPK3 serves as a prognostic marker to promote glioma progression by regulating miR-654/IGF2BP3 signaling.. *Biochem. Biophys. Res. Commun.* 503 (3), 1570–1574. doi:10.1016/j.bbrc.2018.07.081

Kamat, A. M., Hahn, N. M., Efstathiou, J. A., Lerner, S. P., Malmstrom, P. U., Choi, W., et al. (2016). Bladder cancer.. *Lancet* 388 (10061), 2796–2810. doi:10.1016/S0140-6736(16)30512-8

Kirkali, Z., Chan, T., Manoharan, M., Algaba, F., Busch, C., Cheng, L., et al. (2005). Bladder cancer: Epidemiology, staging and grading, and diagnosis.. *Urology* 66 (6), 4–34. doi:10.1016/j.urology.2005.07.062

Kong, Z., Wan, X., Zhang, Y., Zhang, P., Zhang, Y., Zhang, X., et al. (2017). Androgen-responsive circular RNA circSMARCA5 is up-regulated and promotes cell proliferation in prostate cancer.. *Biochem. Biophys. Res. Commun.* 493 (3), 1217–1223. doi:10.1016/j.bbrc.2017.07.162

Li, G., Luo, J., Wang, D., Liang, C., Xiao, Q., Ding, P., et al. (2020). Potential circRNA-disease association prediction using DeepWalk and network consistency projection.. *J. Biomed. Inf.* 112, 103624. doi:10.1016/j.jbi.2020.103624

Li, G., Yue, Y., Liang, C., Xiao, Q., Ding, P., and Luo, J. (2019a). Ncpcda: Network consistency projection for circRNA-disease association prediction.. *RSC Adv.* 9 (57), 33222–33228. doi:10.1039/c9ra06133a

Li, J., Sun, D., Pu, W., Wang, J., and Peng, Y. (2020). Circular RNAs in cancer: Biogenesis, function, and clinical significance.. *Trends Cancer* 6 (4), 319–336. doi:10.1016/j.trecan.2020.01.012

Li, P., Yang, X., Yuan, W., Yang, C., Zhang, X., Han, J., et al. (2018). CircRNA-Cdr1as exerts anti-oncogenic functions in bladder cancer by sponging MicroRNA-135a.. *Cell.. Physiol. biochem.* 46 (4), 1606–1616. doi:10.1159/000489208

Li, T., Huang, T., Guo, C., Wang, A., Shi, X., Mo, X., et al. (2021). Genomic variation, origin tracing, and vaccine development of SARS-CoV-2: A systematic review.. *Innovation.* 2 (2), 100116. doi:10.1016/j.xinn.2021.100116

Li, Y., Wan, B., Liu, L., Zhou, L., and Zeng, Q. (2019b). Circular RNA circMTO1 suppresses bladder cancer metastasis by sponging miR-221 and inhibiting epithelial-to-mesenchymal transition.. *Biochem. Biophys. Res. Commun.* 508 (4), 991–996. doi:10.1016/j.bbrc.2018.12.046

Lopez-Beltran, A., Cimadamore, A., Blanca, A., Massari, F., Vau, N., Scarpelli, M., et al. (2021). Immune checkpoint inhibitors for the treatment of bladder cancer. *Cancers* 13 (1), 131. doi:10.3390/cancers13010131

Lu, C., Zeng, M., Wu, F. X., Li, M., and Wang, J. (2021). Improving circRNA–disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks. *Bioinformatics* 36 (24), 5656–5664. doi:10.1093/bioinformatics/btaa1077

Lu, Q., Liu, T., Feng, H., Yang, R., Zhao, X., Chen, W., et al. (2019). Circular RNA circSLC8A1 acts as a sponge of miR-130b/miR-494 in suppressing bladder cancer progression via regulating PTEN.. *Mol. Cancer* 18 (1), 111–113. doi:10.1186/s12943-019-1040-0

Mancini, M., Righetto, M., and Noessner, E. (2021). Checkpoint inhibition in bladder cancer: Clinical expectations, current evidence, and proposal of future strategies based on a tumor-specific immunobiological approach.. *Cancers* 13 (23), 6016. doi:10.3390/cancers13236016

Miao, X., Xi, Z., Zhang, Y., Li, Z., Huang, L., Xin, T., et al. (2020). Circ-SMARCA5 suppresses colorectal cancer progression via downregulating miR-39-3p and upregulating ARID4B.. *Dig. Liver Dis.* 52 (12), 1494–1502. doi:10.1016/j.dld.2020.07.019

Ni, H., Li, W., Zhuge, Y., Xu, S., Wang, Y., Chen, Y., et al. (2019). Inhibition of circHIPK3 prevents angiotensin II-induced cardiac fibrosis by sponging miR-29b-3p.. *Int. J. Cardiol.* 292, 188–196. doi:10.1016/j.ijcard.2019.04.006

Nouhaud, F. X., Chakroun, M., Lenormand, C., ouzaid, I., Peyronnet, B., Gryn, A., et al. (2021). Comparison of the prognosis of primary vs. progressive muscle invasive bladder cancer after radical cystectomy: Results from a large multicenter study. *Urologic Oncol. Seminars Orig. Investigations* 39 (3), 195.e1–195.e6. doi:10.1016/j.urolonc.2020.09.006

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities.. *Nat. Rev. Genet.* 12 (2), 87–98. doi:10.1038/nrg2934

Peng, L., Shen, L., Xu, J., Tian, X., Liu, F., Wang, J., et al. (2021a). Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures[J]. *Sci. Rep.* 11 (1), 1–11. doi:10.1038/s41598-021-83737-5

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022b). EnANNDeep: An ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models[J]. *Interdiscip. Sci. Comput. Life Sci.*, 1–24. doi:10.1007/s12539-021-00483-y

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020). Single-cell RNA-seq clustering: Datasets, models, and algorithms.. *RNA Biol.* 17 (6), 765–783. doi:10.1080/15476286.2020.1728961

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022a). Cell-cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: Data resources and computational strategies.. *Brief. Bioinform.* 23 (4), bbac234. doi:10.1093/bib/bbac234

Peng, L. H., Chen, Y. Q., Ma, N., and Chen, X. (2017). Narrmda: Negative-aware and rating-based recommendation algorithm for miRNA-disease association prediction.. *Mol. Biosyst.* 13 (12), 2650–2659. doi:10.1039/c7mb00499k

Peng, L. H., Sun, C. N., Guan, N. N., Qiang, J., and Chen, X. (2018). Hnmda: Heterogeneous network-based miRNA-disease association prediction.. *Mol. Genet. Genomics* 293 (4), 983–995. doi:10.1007/s00438-018-1438-1

Peng, L., Wang, C., Tian, X., Zhou, L., and Li, K. (2021b). Finding lncRNA-protein interactions based on deep learning with dual-net neural architecture[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 2021, 29. doi:10.1109/TCBB.2021.3116232

Powles, T., Csőszi, T., Özgüroğlu, M., Matsubara, N., Geczi, L., Cheng, S. Y. S., et al. (2021). Pembrolizumab alone or combined with chemotherapy versus chemotherapy as first-line therapy for advanced urothelial carcinoma (KEYNOTE-361): A randomised, open-label, phase 3 trial.. *Lancet. Oncol.* 22 (7), 931–945. doi:10.1016/S1470-2045(21)00152-2

Qin, H., and Wan, R. (2022). The emerging roles of circSMARCA5 in cancer[J]. *J. Oncol.*, 2022.

Renner, A., Burotto, M., Valdes, J. M., Roman, J. C., and Walton-Diaz, A. (2021). Neoadjuvant immunotherapy for muscle invasive urothelial bladder carcinoma: Will it change current standards? *Ther. Adv. Urol.* 13, 17562872211029779. doi:10.1177/17562872211029779

Roviello, G., Catalano, M., Santi, R., Palmieri, V. E., Vannini, G., Galli, I. C., et al. (2021). Immune checkpoint inhibitors in urothelial bladder cancer: State of the art and future perspectives.. *Cancers* 13 (17), 4411. doi:10.3390/cancers13174411

Shen, L., Liu, F. X., Huang, L., Liu, G. Y., Zhou, L. Q., and Peng, L. H. (2022). VDA-RWLRLS: An anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Sheng, J. Q., Liu, L., Wang, M. R., and Li, P. Y. (2018). Circular RNAs in digestive system cancer: Potential biomarkers and therapeutic targets.. *Am. J. Cancer Res.* 8 (7), 1142–1156.

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations via graph neural network.. *Brief. Bioinform.* 23 (4), bbac266. doi:10.1093/bib/bbac266

Tan, Y., Zhang, T., and Liang, C. (2019). Circular RNA SMARCA5 is overexpressed and promotes cell proliferation, migration as well as invasion while inhibits cell apoptosis in bladder cancer.. *Transl. Cancer Res.* 8 (5), 1663–1671. doi:10.21037/tcr.2019.08.08

Tran, L., Xiao, J. F., Agarwal, N., Duex, J. E., and Theodorescu, D. (2021). Advances in bladder cancer biology and therapy.. *Nat. Rev. Cancer* 21 (2), 104–121. doi:10.1038/s41568-020-00313-1

Vromman, M., Vandesompele, J., and Volders, P. J. (2021). Closing the circle: Current state and perspectives of circular RNA databases.. *Brief. Bioinform.* 22 (1), 288–297. doi:10.1093/bib/bbz175

Walia, A. S., Sweis, R. F., Agarwal, P. K., Kader, A. K., and Modi, P. K. (2021). Cost-effectiveness of immune checkpoint inhibitors in urothelial carcinoma-A review.. *Cancers* 14 (1), 73. doi:10.3390/cancers14010073

Wang, C. C., Han, C. D., Zhao, Q., and Chen, X. (2021a). Circular RNAs and complex diseases: From experimental results to computational models.. *Brief. Bioinform.* 22 (6), bbab286. doi:10.1093/bib/bbab286

Wang, L., Yan, X., You, Z. H., Zhou, X., Li, H. Y., and Huang, Y. A. (2021c). Sganrda: Semi-supervised generative adversarial networks for predicting circRNA-disease associations.. *Brief. Bioinform.* 22 (5), bbab028. doi:10.1093/bib/bbab028

Wang, L., You, Z. H., Huang, D. S., and Li, J-Q. (2021b). Mgrcda: Metagraph recommendation method for predicting CircRNA-disease association[J]. *IEEE Trans. Cybern.* 2021. doi:10.1109/TCYB.2021.3090756

Wang, L., You, Z. H., Li, Y. M., Zheng, K., and Huang, Y. A. (2020). Gcncda: A new method for predicting circRNA-disease associations based on graph convolutional network algorithm.. *PLoS Comput. Biol.* 16 (5), e1007568. doi:10.1371/journal.pcbi.1007568

Wang, Y., Li, H., Lu, H., and Qin, Y. (2019). Circular RNA SMARCA5 inhibits the proliferation, migration, and invasion of non-small cell lung cancer by miR-19b-3p/HOXA9 axis.. *Onco. Targets. Ther.* 12, 7055–7065. doi:10.2147/OTT.S216320

Xia, L., Song, M., Sun, M., Wang, F., and Yang, C. (2018).. Circular RNAs, 171–187. doi:10.1007/978-981-13-1426-1_14Circular RNAs as biomarkers for cancer[J]

Xie, F., Li, Y., Wang, M., Huang, C., Tao, D., Zheng, F., et al. (2018). Circular RNA BCRC-3 suppresses bladder cancer proliferation through miR-182-5p/p27 axis.. *Mol. Cancer* 144, 1–12. doi:10.1186/s12943-018-0892-z

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: An accurate imputation tool for single-cell RNA-seq data.. *Bioinformatics* 36 (10), 3139–3147. doi:10.1093/bioinformatics/btaa109

Yan, C., Wang, J., and Wu, F. X. D. W. N. N-R. L. S. (2018). Regularized least squares method for predicting circRNA-disease associations[J]. *BMC Bioinforma.* 19 (19), 73–81.

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning.. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases.. *Geroscience* 42 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yang, X., Ye, T., Liu, H., Lv, P., Duan, C., Wu, X., et al. (2021). Expression profiles, biological functions and clinical significance of circRNAs in bladder cancer[J]. *Mol. cancer* 20 (1), 1–25. doi:10.1186/s12943-020-01300-8

Yu, G., Wang, L. G., Yan, G. R., and He, Q. Y. (2015). Dose: An R/bioconductor package for disease ontology semantic and enrichment analysis.. *Bioinformatics* 31 (4), 608–609. doi:10.1093/bioinformatics/btu684

Yu, Q., Liu, P., Han, G., Xue, X., and Ma, D. (2020). CircRNA circPDSS1 promotes bladder cancer by down-regulating miR-16.. *Biosci. Rep.* 40 (1), BSR20191961. doi:10.1042/BSR20191961

Zhang, J., Hou, L., Zuo, Z., Ji, P., Zhang, X., Xue, Y., et al. (2021a). Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long.. *Nat. Biotechnol.* 39 (7), 836–845. doi:10.1038/s41587-021-00842-6

Zhang, W., Yu, C., Wang, X., and Liu, F. (2019). Predicting CircRNA-disease associations through linear neighborhood label propagation method. *Ieee Access* 7, 83474–83483. doi:10.1109/access.2019.2920942

Zhang, X., Zhang, Q., Zhang, K., Wang, F., Qiao, X., and Cui, J. (2021b). Circ SMARCA5 inhibited tumor metastasis by interacting with SND1 and downregulating the YWHAB gene in cervical cancer. *Cell. Transpl.* 30, 096368972098378. doi:10.1177/0963689720983786

Zhang, Y., Liu, Q., and Liao, Q. (2020). CircHIPK3: A promising cancer-related circular RNA.. *Am. J. Transl. Res.* 12 (10), 6694–6704.

Zhang, Z., Yang, T., and Xiao, J. (2018). Circular RNAs: Promising biomarkers for human diseases.. *EBioMedicine* 34, 267–274. doi:10.1016/j.ebiom.2018.07.036

Zhao, Q., Yang, Y., Ren, G., Ge, E., and Fan, C. (2019). Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations.. *IEEE Trans. Nanobioscience* 18 (4), 578–584. doi:10.1109/TNB.2019.2922214

Zheng, Q., Bao, C., Guo, W., Li, S., Chen, J., Chen, B., et al. (2016). Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs.. *Nat. Commun.* 7 (1), 1–13. doi:10.1038/ncomms11215

Zhong, Z., Huang, M., Lv, M., He, Y., Duan, C., Zhang, L., et al. (2017). Circular RNA MYLK as a competing endogenous RNA promotes bladder cancer progression through modulating VEGFA/VEGFR2 signaling pathway.. *Cancer Lett.* 403, 305–317. doi:10.1016/j.canlet.2017.06.027

Zhou, L., Wang, J., Liu, G., Lu, Q., Dong, R., Tian, G., et al. (2020). Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method.. *Genomics* 112 (6), 4427–4434. doi:10.1016/j.ygeno.2020.07.044

Zhou, L., Duan, Q., Tian, X., Tang, J., and Peng, L. H. (2021a). LPI-HyADBS: A hybrid framework for lncRNA-protein interaction prediction integrating feature selection and classification[J]. *BMC Bioinforma.* 22 (1), 1–31.

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021b). LPI-deepGBDT: A multiple-layer deep framework based on gradient boosting decision trees for lncRNA-protein interaction identification.. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 September 2022
DOI 10.3389/fgene.2022.980437

OPEN ACCESS

EDITED BY
Liqian Zhou,
Hunan University of Technology, China

REVIEWED BY
Guohua Huang,
Shaoyang University, China
Ying Liang,
Jiangxi Agricultural University, China

*CORRESPONDENCE
Chun Qiu,
13976242127@139.com

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

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 28 June 2022
ACCEPTED 01 August 2022
PUBLISHED 22 September 2022

CITATION
Zhai S, Li X, Wu Y, Shi X, Ji B and Qiu C (2022), Identifying potential microRNA biomarkers for colon cancer and colorectal cancer through bound nuclear norm regularization. *Front. Genet.* 13:980437. doi: 10.3389/fgene.2022.980437

Identifying potential microRNA biomarkers for colon cancer and colorectal cancer through bound nuclear norm regularization

Shengyong Zhai[1†], Xiaoling Li[2,3†], Yan Wu[4], Xiaoli Shi[4], Binbin Ji[4] and Chun Qiu[5*]

[1]Department of General Surgery, Weifang People's Hospital, Shandong, China, [2]The Second Department of Oncology, Beidahuang Industry Group General Hospital, Harbin, China, [3]Heilongjiang Second Cancer Hospital, Harbin, China, [4]Geneis Beijing Co., Ltd., Beijing, China, [5]Department of Oncology, Hainan General Hospital, Haikou, China

Colon cancer and colorectal cancer are two common cancer-related deaths worldwide. Identification of potential biomarkers for the two cancers can help us to evaluate their initiation, progression and therapeutic response. In this study, we propose a new microRNA-disease association identification method, BNNRMDA, to discover potential microRNA biomarkers for the two cancers. BNNRMDA better combines disease semantic similarity and Gaussian Association Profile Kernel (GAPK) similarity, microRNA function similarity and GAPK similarity, and the bound nuclear norm regularization model. Compared to other five classical microRNA-disease association identification methods (MIDPE, MIDP, RLSMDA, GRNMF, AND LPLNS), BNNRMDA obtains the highest AUC of 0.9071, demonstrating its strong microRNA-disease association identification performance. BNNRMDA is applied to discover possible microRNA biomarkers for colon cancer and colorectal cancer. The results show that all 73 known microRNAs associated with colon cancer in the HMDD database have the highest association scores with colon cancer and are ranked as top 73. Among 137 known microRNAs associated with colorectal cancer in the HMDD database, 129 microRNAs have the highest association scores with colorectal cancer and are ranked as top 129. In addition, we predict that hsa-miR-103a could be a potential biomarker of colon cancer and hsa-mir-193b and hsa-mir-7days could be potential biomarkers of colorectal cancer.

KEYWORDS

colon cancer, colorectal cancer, microRNA, biomarker, microRNA-disease association, bound nuclear norm regularization

1 Introduction

Cancers are seriously threatening and endangering human health (Yang et al., 2013; Liu et al., 2021; Yang et al., 2022). Colon cancer and colorectal cancer are two of leading causes of cancer-related deaths worldwide (Lee et al., 2018; Piawah and Venook, 2019). Patients with colon cancer only have a survival rate of 10% when diagnosed at late stage. More importantly, colon cancer shows a higher incidence rate in elder populations. The survival rate of patients with colon cancer is densely associated with the size, location, and stage of the tumor. Metastasis may be the leading cause of deaths for patients suffered from late-stage colon cancer. Thus, understanding the mechanisms of colon cancer could contribute to designing more strong therapeutic options (Ma et al., 2021).

Nowadays, patients with colorectal cancer show a younger trend. In the last decade, incidence rates and death rates of colorectal cancers separately increased by 22 and 13% among adults under 50 years in the United State. However, their precise aetiologic factors still remain unknown. Many evidence demonstrate that early screening of colorectal cancer can reduce their incidence and mortality. Thus, the identification of diagnosis or prognosis biomarkers can contribute to assessment of tumour initiation, progression and therapeutic response for colorectal cancer (Sampath et al., 2021).

Many researches show that numerous RNA data play important roles in the development and metastasis of various diseases including cancers and COVID-19 (Huang et al., 2017; Peng L. et al., 2020; Xu et al., 2020; Yang et al., 2020; Zhang et al., 2021; Peng L. et al., 2022; Shen et al., 2022; Tian et al., 2022). In particular, noncoding RNAs could be biomarkers to boost drug design (Liu et al., 2020; Meng et al., 2022). For example, lncRNAs and circRNAs have been used as biomarkers of cancers (Peng et al., 2021a; Peng et al., 2021b; Chen et al., 2021; Li et al., 2021; Verduci et al., 2021; Wang et al., 2021; Peng L. H. et al., 2022). MicroRNAs (miRNAs) are a class of small non-coding RNAs with 22–24 nucleotides in length (Li et al., 2018; Chen et al., 2020). MicroRNAs can bind to mRNAs of target genes to inhibit expression of these genes. In addition, a few microRNAs may suppress tumors while other microRNAs may affect the progression and metastasis of tumors.

The dysfunction of microRNAs is densely linked to the inflammation of colon cancer. For example, Ma et al. (Ma et al., 2021) found that M2 macrophage-derived exosomal miR-155-5p may have an association with the immune escape of cells in colon cancer. Pagotto et al. (Pagotto et al., 2022) observed that the miR-483 gene could have a responsive to glucose availability for colon cancer. Miao et al. (Miao et al., 2021) identified that miR-4284 could be a therapeutic target in colon cancer. Dougherty et al. (Dougherty et al., 2021) inferred that the upregulations of microRNA-143 and microRNA-145 have close linkages with colonocytes suppresses colitis and inflammation-related colon cancer. Zhang et al. (Zhang et al., 2021) suggested that microRNA-24-3p could heighten the resistance of colon cancer cell to MTX. Yue et al. (Yue et al., 2021) reported that NEDD4 could trigger colon cancer progression through microRNA-340-5p suppression. In summary, the identification of microRNAs in the blood, tissues, and faecal matter will help us use these microRNA as biomarkers in early detection of colon cancer and thus design strong targeted therapeutic strategies for inflammation-mediated colon cancer (Peng et al., 2018; Sampath et al., 2021).

More importantly, microRNAs densely link to the carcinogenic process of colorectal cancer. For example, microRNA-143-3p can limit colorectal cancer metastases (Guo et al., 2019), microRNA-375-3p can boost chemosensitivity to 5-fluorouracil through targeting thymidylate synthase in colorectal cancer (Xu et al., 2020), microRNA-451a influences colorectal cancer proliferation (Ruhl et al., 2018), and microRNA-146a can inhibit tumorigenic inflammation of colorectal cancer (Garo et al., 2021). Biomarkers are an important strategy in early screening, prognostication, survival, and treatment response prediction for cancers. Therefore, microRNAs have been explored as biomarkers in colorectal cancer (Peng LH. et al., 2020; Ogunwobi et al., 2020).

Recently, many researchers have been devoted to microRNA biomarker identification for cancer including colon cancer and colorectal cancer by computational microRNA-disease association prediction (Peng et al., 2017; Li et al., 2021). Huang et al. (Huang et al., 2021) innovatively represented microRNA-disease-type triples as a tensor and further designed a tensor decomposition model to detect new microRNA-disease associations. Li et al. (Li et al., 2021) considered that the abnormal expression of microRNAs is densely associated with the evolution and progression of human diseases and inferred disease-related microRNAs as new biomarkers through a graph auto-encoder model. Chen et al. (Chen et al., 2021) designed a deep learning model for microRNA-disease association identification based on deep belief network. Wang et al. (2022)) pretrained a stacked autoencoder to predict potential microRNA-disease associations in an unsupervised manner. These methods effectively improved microRNA biomarker identification of human complex diseases.

In this study, we design a MicroRNA-Disease Association prediction algorithm (BNNRMDA) to find potential microRNA biomarkers for colon cancer and colorectal cancer based on disease semantic similarity, microRNA functional similarity, Gaussian association profile kernel (GAPK) similarity, and the Bound Nuclear Norm Regularization model.

2 Materials and methods

2.1 Data

2.1.1 Dataset

Experimentally confirmed microRNA-disease association data can be downloaded from the HMDD database provided by Li et al.

(Li et al., 2014). The hierarchical structures between diseases can be downloaded from the MeSH database (https://www.nlm.nih.gov/mesh/). Experimentally supported microRNA-gene interactions can be downloaded from TarBase (Vergoulis et al., 2012), miRTarBase (Hsu et al., 2014), and miRecords (Xiao et al., 2009). We acquired microRNA-disease associations between 495 microRNAs and 378 diseases, hierarchical structures for 4,663 diseases, and 38,089 microRNA-gene interactions between 477 microRNAs and 12,422 genes. Finally, we obtained 4,791 associations between 353 microRNAs and 327 diseases after removing microRNAs without target genes and diseases without hierarchical structures.

2.1.2 Disease semantic similarity

For a known disease d, it can be described as a directed acyclic graph (DAG) based on the MeSH descriptor: $DAG_d = (d, T_d, E_d)$ where T_d denotes the set of nodes that contains d and all its ancestors, and E_d represents corresponding direct edges. Given a disease $t \in T_d$, its semantic contribution to d can be defined as Eq. 1:

$$D_d(t) = \begin{cases} 1 & if\ t \neq d \\ max\{\Delta * D_d(t') | t' \in children\ of\ t\} & if\ t \neq d \end{cases} \tag{1}$$

where Δ denotes the semantic contribution decay factor ($\Delta = 0.5$) (Wang et al., 2010). In general, two diseases d_i and d_j are more similar when they share more common ancestors. Thus, pairwise semantic similarity between d_i and d_j can be defined as Eq. 2:

$$S^d(d_i, d_j) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}(t) + D_{d_j}(t)\right)}{\sum_{t \in T_{d_i}} D_{d_i}(t) + \sum_{t \in T_{d_j}} D_{d_j}(t)} \tag{2}$$

2.1.3 MicroRNA functional similarity

MicroRNA similarity can be computed based on microRNA-gene associations and gene functional network. First, the associated log-likelihood scores $LLS(g_i, g_j)$ between two genes g_i and g_j can be calculated using HumanNet (Lee et al., 2011).

Second, $LLS(g_i, g_j)$ is normalized by Eq. 3:

$$LLS_N(g_i, g_j) = \frac{LLS(g_i, g_j) - LLS_{min}}{LLS_{max} - LLS_{min}} \tag{3}$$

where LLS_{min} and LLS_{max} represent the minimum and maximum associated log-likelihood scores computed by HumanNet, respectively.

Third, similarity between g_i and g_j can be calculated by Eq. 4:

$$S^g(g_i, g_j) = \begin{cases} 1 & g_i = g_j \\ 0 & e(g_i, g_j) \notin HumanNet \\ LLS_N(g_i, g_j) & e(g_i, g_j) \in HumanNet \end{cases} \tag{4}$$

where $e(g_i, g_j)$ indicates interaction between g_i and g_j.

Finally, the functional similarity between two microRNAs m_i and m_j can be computed by Eq. 5 based on their associated genes:

$$S^m(m_i, m_j) = \frac{\sum_{g \in G_i} S(g, G_j) + \sum_{g \in G_j} S(g, G_i)}{|G_i| + |G_j|} \tag{5}$$

where G_i and G_j denotes the gene sets associated with m_i and m_j, respectively, $|G_i|$ and $|G_j|$ denote corresponding cardinalities, respectively, and $S(g, G) = max_{g_i \in G}\{S^g(g, g_i)\}$.

2.1.4 GAPK similarity

For a known disease d_i in a microRNA-disease association matrix $X_{a \times b}$, let the i th row of X denotes its Gaussian association profile $GAP(d_i)$ to represent its association features with all diseases. GAPK similarity between diseases d_i and d_j can be measured by Eq. 6.

$$\begin{aligned} G_D(d_i, d_j) &= \exp\left(-\gamma_d \left\| GAP(d_i) - GAP(d_j) \right\|^2\right) \\ \gamma_d &= \gamma'_d \Big/ \left(\frac{1}{a}\sum_{k=1}^{a} \|GAP(d_k)\|^2\right) \end{aligned} \tag{6}$$

where γ_d indicates normalized kernel bandwidth according to parameter γ'_d, and a indicates the number of diseases.

Similarly, for a known microRNA m_i, let the i th column of X denotes its Gaussian association profile $GAP(m_i)$ to describe its association features with all microRNAs. GAPK similarity between microRNAs m_i and m_j can be measured by Eq. 7:

$$\begin{aligned} G_M(m_i, m_j) &= \exp\left(-\gamma_m \|GAP(m_i) - GAP(m_j)\|^2\right) \\ \gamma_m &= \gamma'_m \Big/ \left(\frac{1}{b}\sum_{k=1}^{b} \|GAP(m_k)\|^2\right) \end{aligned} \tag{7}$$

where γ_m indicates normalized kernel bandwidth according to parameter γ'_m, and b indicates the number of microRNAs.

2.1.5 Similarity fusion

Disease semantic similarity S^d and GAPK similarity G_d are fused to calculate the final disease similarity matrix S_D by Eq. 8:

$$S_D = wG_D + (1 - w)S^d \tag{8}$$

where the parameter w is applied to measure the weight between disease semantic similarity and GAPK similarity.

MicroRNA functional similarity S^m and GAPK similarity G_m are fused to calculate the final microRNA similarity matrix by Eq. 9:

$$S_M = wG_M + (1 - w)S^m \tag{9}$$

where the parameter w is applied to measure the weight between microRNA functional similarity and GAPK similarity.

2.2 Heterogeneous microRNA-disease network construction

A heterogeneous microRNA-disease network is created by fusing microRNA similarity network, disease similarity network,

and microRNA-disease association network. Each edge in similarity network is weighted based on the computed similarity. The heterogeneous microRNA-disease network can be described using a bipartite graph $G(M, D, E)$, where M and D separately represent microRNA set and disease set, $E(G) = \{e_{ij}\} \subseteq M \times D$ represents the microRNA-disease edge set. The adjacency matrix of $G(M, D, E)$ is described as Eq. 10.

$$W = \begin{bmatrix} W_{mm} & W_{md} \\ W_{md}^T & W_{dd} \end{bmatrix} \tag{10}$$

where W_{md} denotes known microRNA-disease association matrix, W_{mm} and W_{dd} denotes the adjacency matrices about microRNA similarity network and disease similarity network, respectively. Hence, the adjacency matrix can be rewritten as Eq. 11.

$$W = \begin{bmatrix} S_M & X_{md} \\ X_{md}^T & S_D \end{bmatrix} \tag{11}$$

2.3 BNNRMDA model

In known microRNA-disease association dataset, majority of microRNA-disease pairs are unknown-associated. Inspired by the bound nuclear norm regularization model provided by Yang et al. (Yang et al., 2019), in this study, we design the bounded nuclear norm regularization-based MDA prediction method to score each unknown microRNA-disease pair. We describe microRNA-disease association inference as a matrix completion problem and construct model (12) to predict new microRNA-disease associations in microRNA-disease association matrix:

$$\begin{aligned} &\min_Y rank(Y) \\ &\text{subject to } \mathrm{P}_\Omega(Y) = \mathrm{P}_\Omega(W) \end{aligned} \tag{12}$$

where Y denotes a matrix need to complete, $rank(Y)$ denotes the rank of Y, $W \in \mathcal{R}^{(m+n)\times(m+n)}$ denotes a known microRNA-disease association matrix, Ω denotes a set containing all index pairs (i, j) that correspond to known microRNA-disease associations in W, and P_Ω represents a projection operator on Ω by Eq. 13:

$$(\mathrm{P}_\Omega(Y))_{ij} = \begin{cases} Y_{ij}, & (i,j) \in \Omega \\ 0, & (i,j) \notin \Omega \end{cases} \tag{13}$$

Model (12) is a non-convex model and difficult to solve. Thus, we transform it to a nuclear norm model through the nuclear norm optimization method proposed by Candes et al. (2013) by Eq. 14:

$$\begin{aligned} &\min_A \|Y\|_* \\ &\text{subject to } \mathrm{P}_\Omega(Y) = \mathrm{P}_\Omega(W) \end{aligned} \tag{14}$$

where Y_* represents the nuclear norm of Y.

Because the value of each element in microRNA and disease similarity matrices S_m and S_d is in the range of [0,1] and the value of each element in microRNA-disease association matrix X_{md} is 1 or 0, the computed microRNA-disease association scores are restricted to [0,1]. Higher score indicates bigger association probability for one microRNA-disease pair. But the elements in Y are in the range of $(-\infty, +\infty)$. Therefore, we add a bounded constraint to Eq. 14 to make the computed scores in [0, 1]. Considering the affect of data noise on the prediction performance, in addition, we develop a rank minimization-based matrix completion model by Eq. 15:

$$\begin{aligned} &\min_A \|Y\|_* \\ &\text{subject to } \|\mathrm{P}_\Omega(Y) - \mathrm{P}_\Omega(W)\|_F \le \epsilon \end{aligned} \tag{15}$$

where $\|.\|_F$ indicates Frobenius norm and ϵ represents the noise level.

We introduce a soft regularization term to tolerate data noise considering the difficulty in selecting an appropriate parameter in Eq. 15. Consequently, a bound nuclear norm regularization model is built to infer potential microRNA-disease associations by Eq. 16:

$$\begin{aligned} &\min_Y \|Y\|_* + \frac{\alpha}{2}\|\mathrm{P}_\Omega(Y) - \mathrm{P}_\Omega(W)\|_F^2 \\ &\text{subject to } 0 \le Y \le 1 \end{aligned} \tag{16}$$

where the parameter α is applied to weigh the importance between the nuclear norm and the error term.

Consequently, we introduce an auxiliary matrix Z and define model 17) to optimize model (16):

$$\begin{aligned} &\min_Y \|Y\|_* + \frac{\alpha}{2}\|\mathrm{P}_\Omega(Z) - \mathrm{P}_\Omega(W)\|_F^2 \\ &\text{subject to } Y = Z \\ &0 \le W \le 1 \end{aligned} \tag{17}$$

where $Y_1 = \mathrm{P}_\Omega(W)$.

Thus, the corresponding augmented Lagrange function is written as Eq. 18:

$$\begin{aligned} L(Z, Y, L, \alpha, \beta) = \|Y\|_* + \frac{\alpha}{2}\|\mathrm{P}_\Omega(Z) - \mathrm{P}_\Omega(W)\|_F^2 \\ + T_r\left(L^T(Y - Z)\right) + \frac{\beta}{2}\|Y - Z\|_F^2 \end{aligned} \tag{18}$$

where L and β represent the Lagrange multiplier and penalty parameter, respectively.

At the t -th iteration, we alternatively compute one of $\boldsymbol{Y}_{k+1}$, $\boldsymbol{Z}_{k+1}$ and $\boldsymbol{L}_{k+1}$ by fixing other two values according to the solution from Yang et al. (Yang et al., 2019). Finally, microRNA-disease association matrix $\boldsymbol{Z}_{md}^*$ is updated through completing the unlabeled elements in $\boldsymbol{Z}_{md}$.

3 Experiments

3.1 Experimental settings and evaluation

In this study, we perform five-fold cross validation for 10 times to investigate the microRNA-disease association

TABLE 1 AUCs of microRNA-disease association prediction methods under cross validation.

Method	MIDPE	MIDP	RLSMDA	GRNMF	LPLNS	BNNRMDA
AUC	0.7820	0.8256	0.8555	0.8963	0.9034	0.9071

TABLE 2 The inferred top 30 microRNAs associated with colon cancer except for 73 known microRNAs.

Rank	MicroRNA	Evidence	Rank	MicroRNA	Evidence
1	hsa-mir-200a	25371200	16	hsa-mir-99a	Unconfirmed
2	hsa-mir-375	29930763	17	hsa-mir-195	26064276
3	hsa-mir-222	27855613	18	hsa-mir-96	Unconfirmed
4	hsa-mir-30d	28651493	19	hsa-mir-148a	Unconfirmed
5	hsa-mir-103a	Unconfirmed	20	hsa-mir-98	28025745
6	hsa-mir-100	28032929	21	hsa-mir-34c	https://doi.org/10.1166/jbt. 2018.1859
7	hsa-mir-181a	25977338	22	hsa-mir-182	Unconfirmed
8	hsa-mir-133a	29930763	23	hsa-mir-20b	33044899
9	hsa-mir-429	Unconfirmed	24	hsa-mir-124	30980700
10	hsa-mir-224	Unconfirmed	25	hsa-mir-7	26648422
11	hsa-mir-93	22180714	26	hsa-mir-193b	31007734
12	hsa-mir-25	23435373	27	hsa-mir-210	27611932
13	hsa-mir-181b	18172508	28	hsa-mir-10a	Unconfirmed
14	hsa-mir-183	Unconfirmed	29	hsa-mir-138	Unconfirmed
15	hsa-mir-153	Unconfirmed	30	hsa-mir-196a	Unconfirmed

inference ability of BNNRMDA. During five-fold cross validation, 80% of elements in microRNA-disease association matrix X are randomly chosen as the training set and the remaining are taken as the test set. Parameters α, β, w , and γ' are set by grid search. We find that BNNRMDA obtain the best AUC when the four parameters are set as $\alpha = 1, \beta = 10, w = 0.3$, and $\gamma' = 0.5$, respectively. Therefore, we set the four parameters as corresponding values. In addition, AUC is widely used to measure the performance of association prediction methods, and thus we use it to measure the performance of BNNRMDA.

3.2 Performance measurement

To measure the microRNA-disease association prediction performance of BNNRMDA, we compare it with MIDPE (Xuan et al., 2015), MIDP (Xuan et al., 2015), RLSMDA (Chen and Yan, 2014), GRNMF (Xiao et al., 2018), and LPLNS (Li et al., 2018). MIDP (Xuan et al., 2015) and MIDPE (Xuan et al., 2015) are two random walk-based microRNA-disease association prediction methods. MIDP is used to detect association information for microRNAs related to diseases. MIDPE is used to detect association information through the bilayer network. RLSMDA (Chen and Yan, 2014) is a semi-supervised learning-based microRNA-disease association inference framework. GRNMF (Xiao et al., 2018) is a graph regularized non-negative matrix factorization-based microRNA-disease association prediction model. In addition, GRNMF built an association probability profile for each disease or miRNA based on a weighted nearest K neighbor profiles. LPLNS (Li et al., 2018) combined label propagation and linear neighborhood similarity for microRNA-disease association prediction. MIDP, MIDPE, RLSMDA, GRNMF, and LPLNS obtained better AUCs for microRNA-disease association prediction. Table 1 shows the AUC values of six microRNA-disease association prediction methods under cross validation.

From Table 1, we can find that BNNRMDA obtains better AUC of 0.9071 than MIDPE, MIDP, RLSMDA, GRNMF, and LPLNS. Compared to MIDPE, MIDP, RLSMDA, GRNMF, and LPLNS, BNNRMDA increases the performance of 13.79, 8.98, 5.69, 1.19, and 0.41% based on the AUC value, respectively. The results show that our proposed BNNRMDA

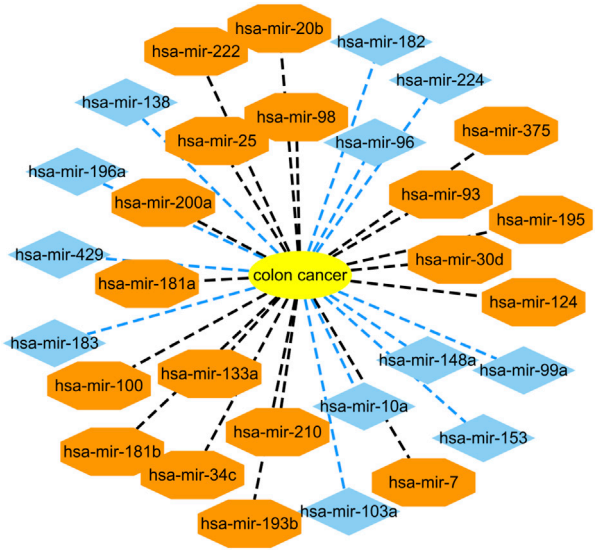


FIGURE 1
Associations between the predicted top 30 microRNAs and colon cancer except for known 73 microRNA-colon cancer associations in the HMDD database that are predicted to have the highest association scores with colon cancer. Black dot lines denote associations between microRNAs and colon cancer and these associations have been reported by publications. Blue dot lines denote associations between microRNAs and colon cancer and these associations are unknown and need to experimental validation.

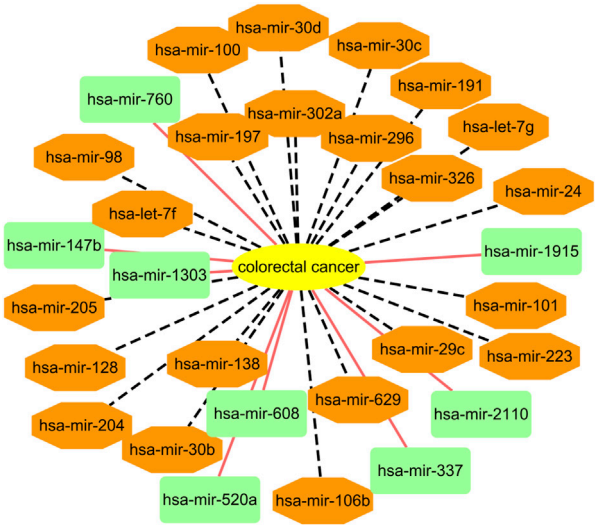


FIGURE 2
Associations between the predicted top 30 microRNAs and colorectal cancer except for known 129 microRNA-colorectal cancer associations in the HMDD database. Black dot lines denote associations between microRNAs and colorectal cancer and these associations have been reported by publications. Orange solid lines denote associations between microRNAs and colorectal cancer and these associations are unknown and need to experimental validation.

method can effectively predict new microRNA-disease associations.

3.3 Case study

In the above section, we have computed the performance of BNNRMDA. The results show that BNNRMDA obtains better AUC and outperforms other five microRNA-disease association prediction methods. We continue to implement case analyses to identify possible microRNA biomarkers for colon cancer and colorectal cancer.

3.3.1 Inferring possible microRNA biomarkers for colon cancer

Colon cancer is a common malignant tumor and has a very high incidence rate in adult with age of 40–50 (Zhu et al., 2020; Liu et al., 2021). More importantly, it has no any symptoms in the early stage. Therefore, it is important to infer possible biomarkers to boost the diagnosis and treatment for colon cancer (Liu et al., 2021). Among the HMDD dataset, there are 73 known microRNAs associated with colon cancer among 353 microRNAs. Based on the proposed BNNRMDA method, we compute the association score for each microRNA-disease pair. The results show that all 73 known microRNAs associated with colon cancer in the HMDD database have the highest association scores with colon cancer and are ranked as top 73. We continue to investigate the following 30 miRNAs that have higher association scores with colon cancer and are ranked as 74–103. The results are shown in Table 2 and Figure 1. From Table 2 and Figure 1, we can find that 18 microRNAs are confirmed to associate with colon cancer by literature retrieval. In addition, 12 microRNAs are inferred to associate with colon cancer and are potential biomarkers of colon cancer.

In addition, we infer that microRNA hsa-mir-103a may associate with colon cancer. Wnt signaling pathway is hyper-activated in many human cancers. Therefore, Wnt pathway demonstrates promising diagnostic and therapeutic effect in cancer medicine. Fasihi et al. (2018) found that hsa-miR-103a may be a possible regulator of Wnt signaling pathway by detecting its effect on Wnt pathway components in colorectal cancer-originated cell lines and its expression in colorectal cancer tissues. They also found that hsa-miR-103a has an upregulation function in colorectal cancer tissues through RT-qPCR and its overexpression could cause elevated Wnt activity. Therefore, we infer that hsa-miR-103a could be a potential biomarker of colon cancer (Fasihi et al., 2017).

TABLE 3 The inferred top 30 microRNAs associated with colorectal cancer except for 129 known microRNAs.

Rank	MicroRNA	Evidence	Rank	MicroRNA	Evidence
1	hsa-mir-191	18079988	16	hsa-mir-223	27759076
2	hsa-mir-760	the HMDD database	17	hsa-mir-100	25973296
3	hsa-mir-337	the HMDD database	18	hsa-mir-204	25209181
4	hsa-mir-1915	the HMDD database	19	hsa-let-7g	18172508
5	hsa-mir-24	30375302	20	hsa-mir-106b	34070923
6	hsa-mir-520a	the HMDD database	21	hsa-mir-296	28209128
7	hsa-mir-101	30797148	22	hsa-let-7f	29805607
8	hsa-mir-138	27248318	23	hsa-mir-29c	29262657
9	hsa-mir-608	the HMDD database	24	hsa-mir-30c	25799050
10	hsa-mir-1303	the HMDD database	25	hsa-mir-30b	32112903
11	hsa-mir-629	30042169	26	hsa-mir-302a	31754405
12	hsa-mir-2110	the HMDD database	27	hsa-mir-326	25760058
13	hsa-mir-147b	the HMDD database	28	hsa-mir-98	34370878
14	hsa-mir-205	29488611	29	hsa-mir-128	30257253
15	hsa-mir-197	30106114	30	hsa-mir-30d	28651493

3.3.2 Inferring possible microRNA biomarkers for colorectal cancer

Colorectal cancer is the third leading cause of cancer-related deaths in the United States. In the United State, there are about 1.85 million cases and 850 thousand deaths annually. In 2020, there are 53,200 colorectal cancer deaths in the United State. Among new colorectal cancer diagnoses, approximately 20% of patients suffered from metastatic disease and approximately 25% of patients suffered from localized disease that may later develop metastases. Of patients who are diagnosed as metastatic colorectal cancer, about 70–75% of patients survive more than 1 year, about 30–35% patients survive more than 3 years, and less than 20% patients survive more than 5 years (Xie et al., 2020; Biller and Schrag, 2021).

Among the HMDD dataset, there are 137 known microRNAs associated with colorectal cancer among 353 microRNAs. Based on the proposed BNNRMDA method, we compute the association score for each microRNA-colorectal cancer pair. The results show that 129 known microRNAs associated with colorectal cancer in the HMDD database have the highest association scores with colorectal cancer and are ranked as top 129. We continue to investigate the following 30 miRNAs that have higher association scores with colorectal cancer and are ranked as 130–159. The results are shown in Table 3 and Figure 2. From Table 3 and Figure 2, we can find that 8 microRNAs are known to associate with colorectal cancer in the HMDD database. In addition, the remaining 22 microRNAs are inferred to associate with colorectal cancer and are reported by publications. The results confirm the strong microRNA identification performance of BNNRMDA for colorectal cancer. In addition, we predict that hsa-mir-193b and hsa-mir-7 days may associate with colorectal cancer and need validation.

4 Conclusion

Colon cancer and colorectal cancer are two of leading causes of cancer-related deaths worldwide and are seriously threatening human health. Inference of diagnosis or prognosis biomarkers for colon cancer and colorectal cancer can help to evaluate their initiation, progression and therapeutic response. In this study, we developed a new microRNA-disease association prediction method, BNNRMDA, to find possible microRNA biomarkers for colon cancer and colorectal cancer. BNNRMDA effectively integrated disease semantic similarity and GAPK similarity, microRNA function similarity and GAPK similarity, and bound nuclear norm regularization.

Compared to other five classical microRNA-disease association prediction methods, BNNRMDA obtains the best AUC of 0.9071, demonstrating its powerful microRNA-disease association prediction performance. We continue to use the proposed BNNRMDA method for finding possible microRNA biomarkers for colon cancer and colorectal cancer. The results show that hsa-miR-103a could be a potential biomarker of colon cancer and hsa-mir-193b and hsa-mir-7 days could be potential biomarkers of colorectal cancer.

Our proposed BNNRMDA method fully considers the affect of Gaussian association profile similarity on the prediction performance. In addition, the bound nuclear norm regularization approach can effectively learn the intrinsic distribution of data. Therefore, BNNRMDA significantly outperform other MDA prediction methods. Although BNNRMDA obtains better AUC, its performance including AUC, precision, recall, and accuracy need to further improve. In the future, we will improve the bound nuclear norm regularization model to discover possible biomarkers for colon cancer and colorectal cancer.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization: S-YZ and CQ; Methodology: S-YZ, X-LL and CQ; Project administration: CQ, YW, X-LS and B-BJ; Software: S-YZ, X-LL and CQ; Writing-original draft: S-YZ, X-LL and CQ; Writing-review and editing: S-YZ, and CQ.

Funding

This work was supported by the Medical and Health Science Technology Development Program in Shandong Province (202104080159) and Science Technology Development Program in Weifang City (2021YX007).

Conflict of interest

Authors YW, XS, and BJ were employed by the company Geneis (Beijing) Co. Ltd. In addition, this manuscript was conducted by a multicenter study initiated by the corresponding author (CQ). Geneis Beijing Co., Ltd. Contacted the doctors in hospitals around the country.

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

Publisher's note

References

Biller, L. H., and Schrag, D. (2021). Diagnosis and treatment of metastatic colorectal cancer: A review. *Jama* 325 (7), 669–685. doi:10.1001/jama.2021.0106

Candes, E., and Recht, B. (2013). Simple bounds for recovering low-complexity models[J]. *Mathematical Programming* 141 (1), 577–589.

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference[J]. *Scientific reports* 4 (1), 1–10.

Chen, B., Xia, Z., Deng, Y. N., Yang, Y., Zhang, P., Zhu, H., et al.(2019). Emerging microRNA biomarkers for colorectal cancer diagnosis and prognosis. *Open Biol.* 9 (1), 180212. doi:10.1098/rsob.180212

Chen, H., Guo, R., Li, G., Zhang, W., and Zhang, Z. (2020). Comparative analysis of similarity measurements in miRNAs with applications to miRNA-disease association predictions. *BMC Bioinforma.* 21 (1), 176–214. doi:10.1186/s12859-020-3515-9

Chen, S., Yang, X., Yu, C., Zhou, W., Xia, Q., Liu, Y., et al.(2021). The potential of circRNA as a novel diagnostic biomarker in cervical cancer. *J. Oncol.* 2021, 2021–2026. doi:10.1155/2021/5529486

Chen, X., Li, T. H., Zhao, Y., Wang, C. C., and Zhu, C. C. (2021). Deep-belief network for predicting potential miRNA-disease associations. *Brief. Bioinform.* 22 (3), bbaa186. doi:10.1093/bib/bbaa186

Dougherty, U., Mustafi, R., Zhu, H., Zhu, X., Deb, D., Meredith, S. C., et al. (2021). Upregulation of polycistronic microRNA-143 and microRNA-145 in colonocytes suppresses colitis and inflammation-associated colon cancer. *Epigenetics* 16 (12), 1317–1334. doi:10.1080/15592294.2020.1863117

Fasihi, A. M., Soltani, B., and Atashi, A. (2018). Introduction of hsa–miR–103a and hsa–miR–1827 and hsa–miR–137 as new regulators of Wnt signaling pathway and their relation to colorectal carcinoma[J]. *J. cellular biochemistry* 119 (7), 5104–5117.

Garo, L. P., Ajay, A. K., Fujiwara, M., Gabriely, G., Raheja, R., Kuhn, C., et al. (2021). MicroRNA-146a limits tumorigenic inflammation in colorectal cancer. *Nat. Commun.* 12 (1), 2419–2516. doi:10.1038/s41467-021-22641-y

Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., et al. (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* 42 (D1), D78–D85. doi:10.1093/nar/gkt1266

Huang, F., Yue, X., Xiong, Z., Yu, Z., Liu, S., and Zhang, W. (2021). Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief. Bioinform.* 22 (3), bbaa140. doi:10.1093/bib/bbaa140

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses. *Bioinformatics* 33 (20), 3195–3201. doi:10.1093/bioinformatics/btx390

Lee, C. H., Im, E. J., Moon, P. G., and Baek, M. C. (2018). Discovery of a diagnostic biomarker for colon cancer through proteomic profiling of small extracellular vesicles. *BMC cancer* 18 (1), 1058–1111. doi:10.1186/s12885-018-4952-y

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 21 (7), 1109–1121. doi:10.1101/gr.118992.110

Li, G., Luo, J., Xiao, Q., Liang, C., and Ding, P. (2018). Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity. *J. Biomed. Inf.* 82, 169–177. doi:10.1016/j.jbi.2018.05.005

Li, T., Huang, T., Guo, C., Wang, A., Shi, X., Mo, X., et al. (2021). Genomic variation, origin tracing, and vaccine development of SARS-CoV-2: A systematic review. *Innovation.* 2 (2), 100116. doi:10.1016/j.xinn.2021.100116

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42 (D1), D1070–D1074. doi:10.1093/nar/gkt1023

Li, Z., Li, J., Nie, R., et al. (2021). A graph auto-encoder model for miRNA-disease associations prediction[J]. *Briefings Bioinforma.* (4), 22, bbaa240. doi:10.1093/bib/bbaa240

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, D., Huang, Y., Nie, W., Zhang, J., and Deng, L. (2021). Smalf: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. *BMC Bioinforma.* 22 (1), 219–318. doi:10.1186/s12859-021-04135-2

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9, 886. doi:10.3389/fcell.2021.619330

Ma, Y. S., Wu, T. M., Ling, C. C., Yu, F., Zhang, J., Cao, P. S., et al. (2021). M2 macrophage-derived exosomal microRNA-155-5p promotes the immune escape of colon cancer by downregulating ZC3H12B. *Mol. Ther. Oncolytics* 20, 484–498. doi:10.1016/j.omto.2021.02.005

Meng, Y., Lu, C., Jin, M., Xu, J., Zeng, X., and Yang, J. (2022). A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief. Bioinform.* 23, bbab581. doi:10.1093/bib/bbab581

Miao, X., Li, Z., Zhang, Y., and Wang, T. (2021). MicroRNA-4284 inhibits colon cancer epithelial-mesenchymal transition by down-regulating Perilipin 5. *STEMedicine* 2 (6), e85. doi:10.37175/stemedicine.v2i6.85

Ogunwobi, O. O., Mahmood, F., and Akingboye, A. (2020). Biomarkers in colorectal cancer: Current research and future prospects. *Int. J. Mol. Sci.* 21 (15), 5311. doi:10.3390/ijms21155311

Pagotto, S., Colorito, M. L., Nicotra, A., Apuzzo, T., Tinari, N., Protasi, F., et al. (2022). A perspective analysis: microRNAs, glucose metabolism, and drug resistance in colon cancer stem cells. *Cancer Gene Ther.* 29 (1), 4–9. doi:10.1038/s41417-021-00298-5

Peng, L., Chen, Y., Ma, N., and Chen, X. (2017). Narrmda: Negative-aware and rating-based recommendation algorithm for miRNA-disease association prediction. *Mol. Biosyst.* 13 (12), 2650–2659. doi:10.1039/c7mb00499k

Peng, L. H., Sun, C. N., Guan, N. N., Li, J. Q., and Chen, X. (2018). Hnmda: Heterogeneous network-based miRNA-disease association prediction. *Mol. Genet. Genomics* 293 (4), 983–995. doi:10.1007/s00438-018-1438-1

Peng, L. H., Zhou, L. Q., Chen, X., and Piao, X. (2020b). A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front. Bioeng. Biotechnol.* 8, 40. doi:10.3389/fbioe.2020.00040

Peng, L., Tian, X., Shen, L., et al. (2020a). Drug designing and repositioning against severe acute respiratory coronavirus 2 (SARS-Cov-2) through computational simulation: Current progress and hopes. *Front. Genet.* 5, 1–5. doi:10.23880/oajmb-16000168

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022a). Cell-cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: Data resources and computational strategies. *Brief. Bioinform.* 23 (4), bbac234. doi:10.1093/bib/bbac234

Peng, L. H., Tan, J. W., Tian, X. F., et al. (2022b). EnANNDeep: An ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models[J]. *Interdiscip. Sci. Comput. Life Sci.* 14, 1–24.

Peng, L. H., Wang, C., Tian, X. F., Zhou, L. Q., and Li, K. Q. (2021b). Finding lncRNA-protein interactions based on deep learning with dual-net neural architecture. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 29, 1. doi:10.1109/TCBB.2021.3116232

Peng, L. H., Yuan, R. Y., Shen, L., Gao, P. F., and Zhou, L. Q. (2021a). LPI-EnEDT: An ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification[J]. *BioData Min.* 14 (1), 1–22.

Piawah, S., and Venook, A. P. (2019). Targeted therapy for colorectal cancer metastases: A review of current methods of molecularly targeted therapy and the use of tumor biomarkers in the treatment of metastatic colorectal cancer. *Cancer* 125 (23), 4139–4147. doi:10.1002/cncr.32163

Ruhl, R., Rana, S., Kelley, K., Espinosa-Diez, C., Hudson, C., Lanciault, C., et al. (2018). microRNA-451a regulates colorectal cancer proliferation in response to radiation. *BMC cancer* 18 (1), 517–519. doi:10.1186/s12885-018-4370-1

Sampath, S. S., Venkatabalasubramanian, S., and Ramalingam, S. (2021). Role of MicroRNAs in the progression and metastasis of colon cancer. *Endocr. Metab. Immune Disord. Drug Targets* 21 (1), 35–46. doi:10.2174/1871530320666200825184924

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: An anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Tian, X., Shen, L., Gao, P., Huang, L., Liu, G., Zhou, L., et al. (2022). Discovery of potential therapeutic drugs for COVID-19 through logistic matrix factorization with kernel diffusion. *Front. Microbiol.*, 13, 740382. doi:10.3389/fmicb.2022.740382

Verduci, L., Tarcitano, E., Strano, S., Yarden, Y., and Blandino, G. (2021). CircRNAs: Role in human diseases and potential use as biomarkers. *Cell Death Dis.* 12 (5), 468–512. doi:10.1038/s41419-021-03743-3

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., et al. (2012). TarBase 6.0: Capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.* 40 (D1), D222–D229. doi:10.1093/nar/gkr1161

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26 (13), 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, S., Dong, Y., Gong, A., Kong, H., Gao, J., Hao, X., et al. (2021). Exosomal circRNAs as novel cancer biomarkers: Challenges and opportunities. *Int. J. Biol. Sci.* 17 (2), 562–573. doi:10.7150/ijbs.48782

Wang, C. C., Li, T. H., and Huang, L. (2022). Prediction of potential miRNA–disease associations based on stacked autoencoder[J]. *Briefings in Bioinformatics* 23 (2), 1–11.

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 (1), D105–D110. doi:10.1093/nar/gkn851

Xiao, Q., Luo, J., and Liang, C. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations[J]. *Bioinformatics* 34 (2), 239–248.

Xie, Y. H., Chen, Y. X., and Fang, J. Y. (2020). Comprehensive review of targeted therapy for colorectal cancer[J]. *Signal Transduct. Target. Ther.* 5 (1), 1–30. doi:10.1038/s41392-020-0116

Xu, F., Ye, M. L., Zhang, Y. P., Li, W. J., Li, M. T., Wang, H. Z., et al. (2020). MicroRNA-375-3p enhances chemosensitivity to 5-fluorouracil by targeting thymidylate synthase in colorectal cancer. *Cancer Sci.* 111 (5), 1528–1541. doi:10.1111/cas.14356

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: An accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36 (10), 3139–3147. doi:10.1093/bioinformatics/btaa109

Xuan, P., Han, K., and Guo, Y. (2015). Prediction of potential disease-associated microRNAs based on random walk[J]. *Bioinformatics* 31 (11), 1805–1815.

Yang, J., Grünewald, S., and Wan, X. F. (2013). Quartet-net: A quartet-based method to reconstruct phylogenetic networks. *Mol. Biol. Evol.* 30 (5), 1206–1217. doi:10.1093/molbev/mst040

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35 (14), i455–i463. doi:10.1093/bioinformatics/btz331

Yuan, Y., Liu, W., Zhang, Y., and Sun, S. (2018). CircRNA circ_0026344 as a prognostic biomarker suppresses colorectal cancer progression via microRNA-21 and microRNA-31. *Biochem. Biophys. Res. Commun.* 503 (2), 870–875. doi:10.1016/j.bbrc.2018.06.089

Yue, M., Yun, Z., Li, S., Yan, G., and Kang, Z. (2021). NEDD4 triggers FOXA1 ubiquitination and promotes colon cancer progression under microRNA-340-5p suppression and ATF1 upregulation. *RNA Biol.* 18 (11), 1981–1995. doi:10.1080/15476286.2021.1885232

Zhang, H. W., Shi, Y., Liu, J. B., Wang, H. M., Wang, P. Y., Wu, Z. J., et al. (2021). Cancer-associated fibroblast-derived exosomal microRNA-24-3p enhances colon cancer cell resistance to MTX by down-regulating CDX2/HEPH axis. *J. Cell. Mol. Med.* 25 (8), 3699–3713. doi:10.1111/jcmm.15765

Zhu, J., Xu, Y., Liu, S., Qiao, L., Sun, J., and Zhao, Q. (2020). MicroRNAs associated with colon cancer: New potential prognostic markers and targets for therapy. *Front. Bioeng. Biotechnol.* 8, 176. doi:10.3389/fbioe.2020.00176

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 27 September 2022
DOI 10.3389/fgene.2022.1017539

OPEN ACCESS

EDITED BY
Jialiang Yang,
Geneis (Beijing) Co. Ltd., China

REVIEWED BY
Junlin Xu,
Hunan University, China
Lijuan Zhu,
Shijiazhuang Tiedao University, China

*CORRESPONDENCE
Sheng Zheng,
zs19860909@163.com

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 12 August 2022
ACCEPTED 08 September 2022
PUBLISHED 27 September 2022

CITATION
Leng X, Yang J, Liu T, Zhao C, Cao Z, Li C, Sun J and Zheng S (2022), A bioinformatics framework to identify the biomarkers and potential drugs for the treatment of colorectal cancer. *Front. Genet.* 13:1017539. doi: 10.3389/fgene.2022.1017539

A bioinformatics framework to identify the biomarkers and potential drugs for the treatment of colorectal cancer

Xiaogang Leng, Jianxiu Yang, Tie Liu, Chunbo Zhao, Zhongzheng Cao, Chengren Li, Junxi Sun and Sheng Zheng*

Department of Colorectal and Anal Surgery, Weifang People's Hospital, Weifang, China

Colorectal cancer (CRC), a common malignant tumor, is one of the main causes of death in cancer patients in the world. Therefore, it is critical to understand the molecular mechanism of CRC and identify its diagnostic and prognostic biomarkers. The purpose of this study is to reveal the genes involved in the development of CRC and to predict drug candidates that may help treat CRC through bioinformatics analyses. Two independent CRC gene expression datasets including The Cancer Genome Atlas (TCGA) database and GSE104836 were used in this study. Differentially expressed genes (DEGs) were analyzed separately on the two datasets, and intersected for further analyses. 249 drug candidates for CRC were identified according to the intersected DEGs and the Crowd Extracted Expression of Differential Signatures (CREEDS) database. In addition, hub genes were analyzed using Cytoscape according to the DEGs, and survival analysis results showed that one of the hub genes, *TIMP1* was related to the prognosis of CRC patients. Thus, we further focused on drugs that could reverse the expression level of *TIMP1*. Eight potential drugs with documentary evidence and two new drugs that could reverse the expression of *TIMP1* were found among the 249 drugs. In conclusion, we successfully identified potential biomarkers for CRC and achieved drug repurposing using bioinformatics methods. Further exploration is needed to understand the molecular mechanisms of these identified genes and drugs/small molecules in the occurrence, development and treatment of CRC.

KEYWORDS

colorectal cancer, differential expressed gene, hub gene, survival analysis, *TIMP1*, drug repurposing

Introduction

Colorectal cancer (CRC) is the most common subtype in gastrointestinal cancers, and its early symptoms are unobvious, which results in a high mortality rate. The continuous rise of new cases and deaths of CRC will lead to a significant increase in the economic burden globally (Rogler, 2014; Arnold et al., 2017; Hong et al., 2021; Liu et al., 2021). As

the second leading cause of cancer death worldwide (Zhao et al., 2020; Sung et al., 2021), CRC has become a major global public health concern. Studies have shown that the clinical tumor stage at diagnosis affects the prognosis of patients. The 5-years relative survival rate of patients with stage I was 90%, while that of patients with stage IV was only 10% (Siegel et al., 2012; O'Connell et al., 2004; Yang et al., 2022). Currently, various diagnostic strategies for CRC include both invasive and non-invasive methods. Invasive methods rely on endoscopy and imaging. Imaging tests such as nuclear magnetic resonance (NMR) and computed tomography (CT) can be used to diagnose severe focal lesions, but both tests are expensive (Grassetto et al., 2012; Swiderska et al., 2014). Hence, there is an urgent need for alternative, cheap and easy-to-measure screening methods. Despite recent advances in treatment and multidisciplinary care, CRC patients continue to suffer from serious adverse reactions, which can impair prognosis and reduce survival (McQuade et al., 2017; Kong et al., 2020). The developing drugs with low toxicity, especially drug repositioning (Liu et al., 2020; Meng et al., 2022) is of great significance for improving the clinical treatment and reducing adverse reactions.

The improvement of molecular biology technology provides opportunities to develop more curative effect and enhance the outcomes of CRC. With the progress of high-throughput sequencing technology, gene expression profiling methods, such as RNA sequencing (RNA-seq), have been applied to scientific research and become a hot field of gene expression research (Saito et al., 2018; Deshiere et al., 2019; Zhang et al., 2021). The molecular mechanism of CRC holds the key to the prognosis and treatment response of patients, and is of great potential for the clinical practice (De Sousa et al., 2013; Sadanandam et al., 2013; Nguyen and Duong, 2018; Cheng et al., 2020; Cheng et al., 2021; Liu et al., 2022). Therefore, understanding of the molecular mechanism in the occurrence and development of CRC will help to develop novel therapies to optimize the treatment response throughout the disease course. In recent years, a large number of relevant CRC sequencing data have been generated, archived, and stored in public databases (Guo et al., 2017). Researches combining high-throughput sequencing data and bioinformatics analysis has gradually become a hot spot (Alves Martins et al., 2019; Zhao et al., 2019). Here, bioinformatics analysis of RNA-seq data of CRC patients may provide insights for drug repositioning for the treatment of CRC.

In this study, bioinformatics analysis was used to identify biomarkers of CRC and potential drugs that can improve the outcomes of CRC patients. Specifically, based on the TCGA data set and GSE104836 data set, we compared the transcriptome data of tumor samples and normal samples to identify differentially expressed genes (DEGs) on the two independent datasets. The DEGs were intersected for further analysis. Then these DEGs were further explored to detect the enriched GO terms and KEGG pathways. From those DEGs, latent drugs that can improve the prognosis of patients from the Crowd Extracted Expression of Differential Signatures (CREEDS) were also predicted. In addition, the hub genes in the protein-protein interaction (PPI) network were discovered according to the DEGs and survival analysis was carried out on these hub genes. Finally, drug candidates could reverse hub genes were also predicted by CREEDS and validated by literatures.

Materials and methods

Data collection

RNA-seq data of CRC patients were downloaded from the Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/) and the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104836). Meanwhile, the associated clinical information of 478 tumor samples and 41 normal samples from TCGA, and 10 patients and 10 healthy controls from the GSE104836 dataset was obtained.

Differentially expressed gene analysis

DESeq2 is a R package that can identify DEGs from raw count data. It uses the contraction estimation of discrete and the fold change of the gene expression to improve the stability and interpretability of the estimation, which makes the more quantitative analysis focus on intensity (Love et al., 2014). DEGs in CRC tumor samples and normal samples were detected using DESeq2 package with the criteria of p-value < 0.001 and log2 |fold change| $\geqq$ 2.

Functional and pathway enrichment analysis of DEGs

After DEG analysis of the TCGA dataset and GSE104836 dataset, overlapping DEGs were screened, and then enrichment analysis of KEGG pathway and GO (The Gene Ontology, 2019) including biological process (BP), cellular component (CC), and molecular function (MF) were carried out to reveal the altered biological characteristics of CRC. The R packages "clusterProfiler" and "ggplot" were used to visualize the results of the enrichment analysis.

PPI network and hub genes analysis

The online database STRING (http://string-db.org) was used to develop a PPI network of DEGs, and the minimum required interaction score was 0.7. The Cytoscape software was used to visualize the PPI network and to analyze the structural properties

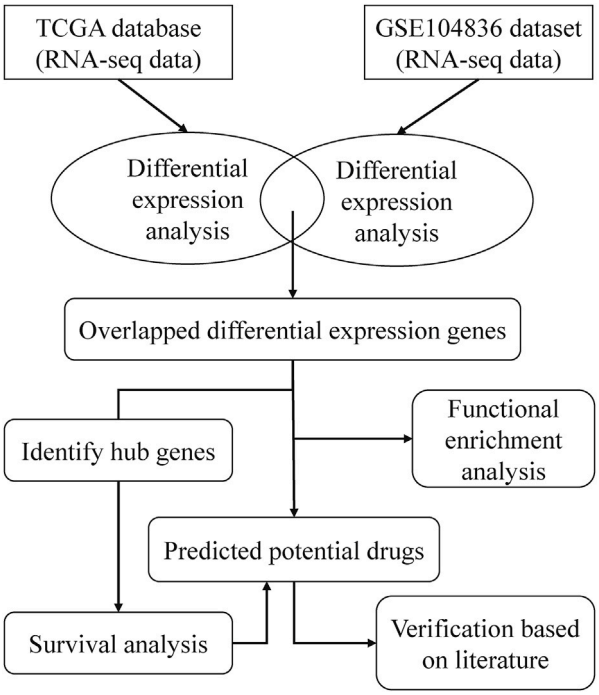


FIGURE 1
A brief workflow for drug repurposing.

of the constructed network. The cytoHubba plug-in was used to identify hub genes in the PPI network.

Potential drug identification

The CREEDS database consists of gene expression characteristics induced by single drug perturbation, which can be used to identify the relationship between genes, diseases, and drugs. To identify potential drugs for the treatment of CRC, we used the CREEDS database to find drugs that can reverse the DEGs. Specifically, for each drug in the CREEDS database, we calculated the *p*-value of the overlapping genes between downregulated genes of the drug and upregulated DEGs in CRC by hypergeometric test, and similarly, calculate the *p*-value of the overlapping genes between upregulated genes of the drug and downregulated DEGs in CRC. The drugs with any of the two *p*-value lower than 0.05 could be taken as candidates that could reverse the DEGs and might treat the CRC.

Survival analysis

We obtained the OS time of all patients in TCGA database, and estimated the survival probability of CRC patients using Kaplan-Meier method. Kaplan-Meier survival curve was used to estimate the 50th percentile (median) of survival time and compare the survival distribution of two or more groups. Log-rank test was also used to compare the survival differences between groups. *p*-value <0.05 was considered to have significant differences between groups. The data were analyzed by R software.

TABLE 1 General clinical information of CRC patients included in this study.

Characteristics	No	
Type		
	Tumor	478
	Normal	41
Average age	67.04	
Gender	Female	247
	Male	272
Tumor stage	I	85
	II	209
	III	140
	IV	73
	Unknown	12

Results

A framework of CRC related drugs repurposing

To find drugs that can be used to treat CRC, we proposed a bioinformatics pipeline of drug repurposing based on transcriptome data. The workflow was shown in Figure 1. After downloading the RNA-seq data from TCGA and GEO databases, we performed DEG analysis and pathway enrichment analysis. Then, the hub genes of DEGs were identified and survival analysis was done on the hub genes. According to the DEGs and CREEDS, drugs that could reverse the DEGs were identified, and 10 drugs can reverse the survival-related hub gene were further investigated. Finally, according to some previous studies, the effectiveness of the newly discovered drugs was verified.

Patient characteristics

The RNA-seq data involved 478 tumor samples and 41 normal samples. There were 247 women and 272 men. 85 cases were at clinical stage I, 209 cases were at stage II, 140 cases were at stage III and 73 cases were at stage IV. Their average age was ~67 years old. The clinical features of patients from the TCGA dataset were shown in Table 1.

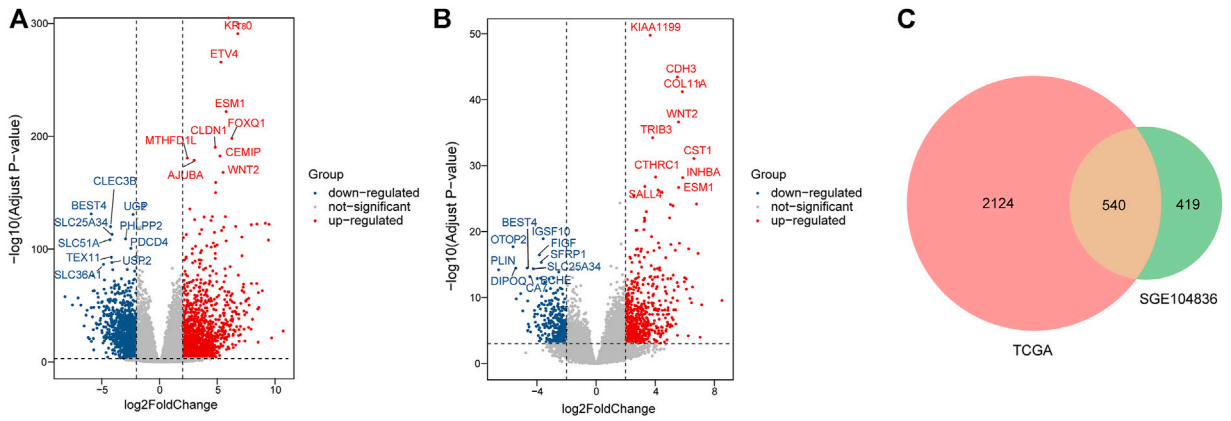


FIGURE 2
Identification of DEGs between tumor tissues and normal tissues in CRC patients. **(A–B)**. Differential expression volcanic map of **(A)**TCGA and **(B)** GEO dataset. Red dots indicate significant up-regulation, blue dots indicate significant down-regulation, and gray dots indicate genes with no significant changes. **(C)**. Venn plot for DEGs detected in two datasets.

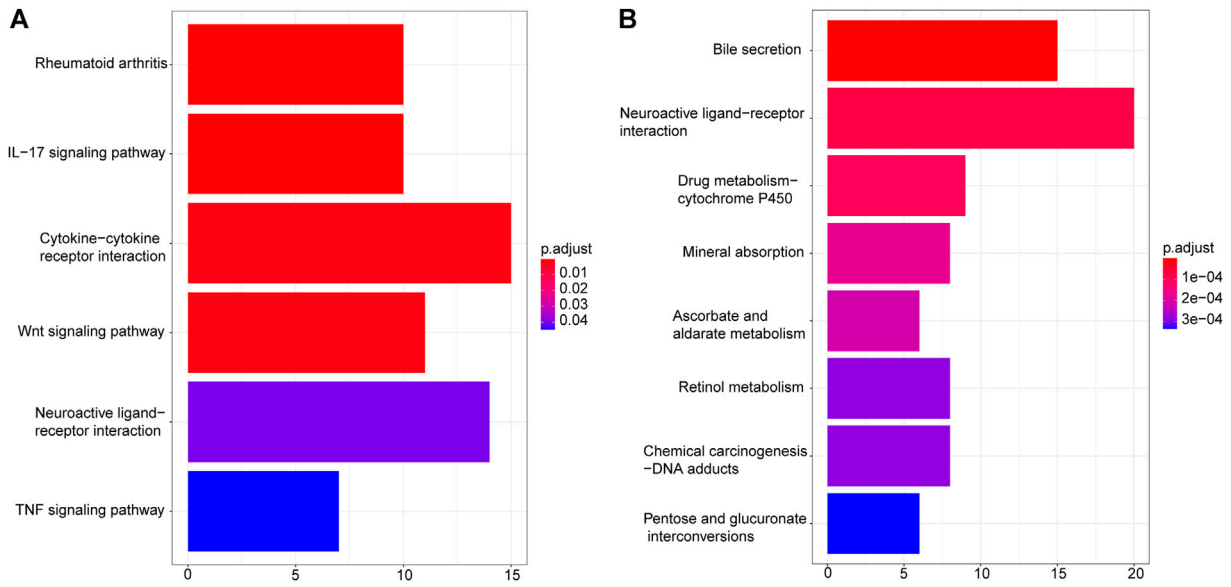


FIGURE 3
KEGG enrichment analysis of DEGs. **(A)**. Upregulated DEGs. **(B)** Downregulated DEGs.

DEGs identification

In total, 2664 DEGs (1537 upregulated genes and 1127 downregulated genes) and 959 DEGs (563 upregulated genes and 396 downregulated genes) were extracted from TCGA (Figure 2A) and GSE104836 (Figure 2B) datasets respectively using p-value < 0.001 and |log2 |fold change| ≧ 2 as the cut-off criteria. A total of 540 DEGs (276 upregulated genes and 264 downregulated genes) were identified in both datasets (Figure 2C).

Enrichment Analysis

To understand the possible biological mechanisms that cause the identified changes in the transcriptome data, we conducted

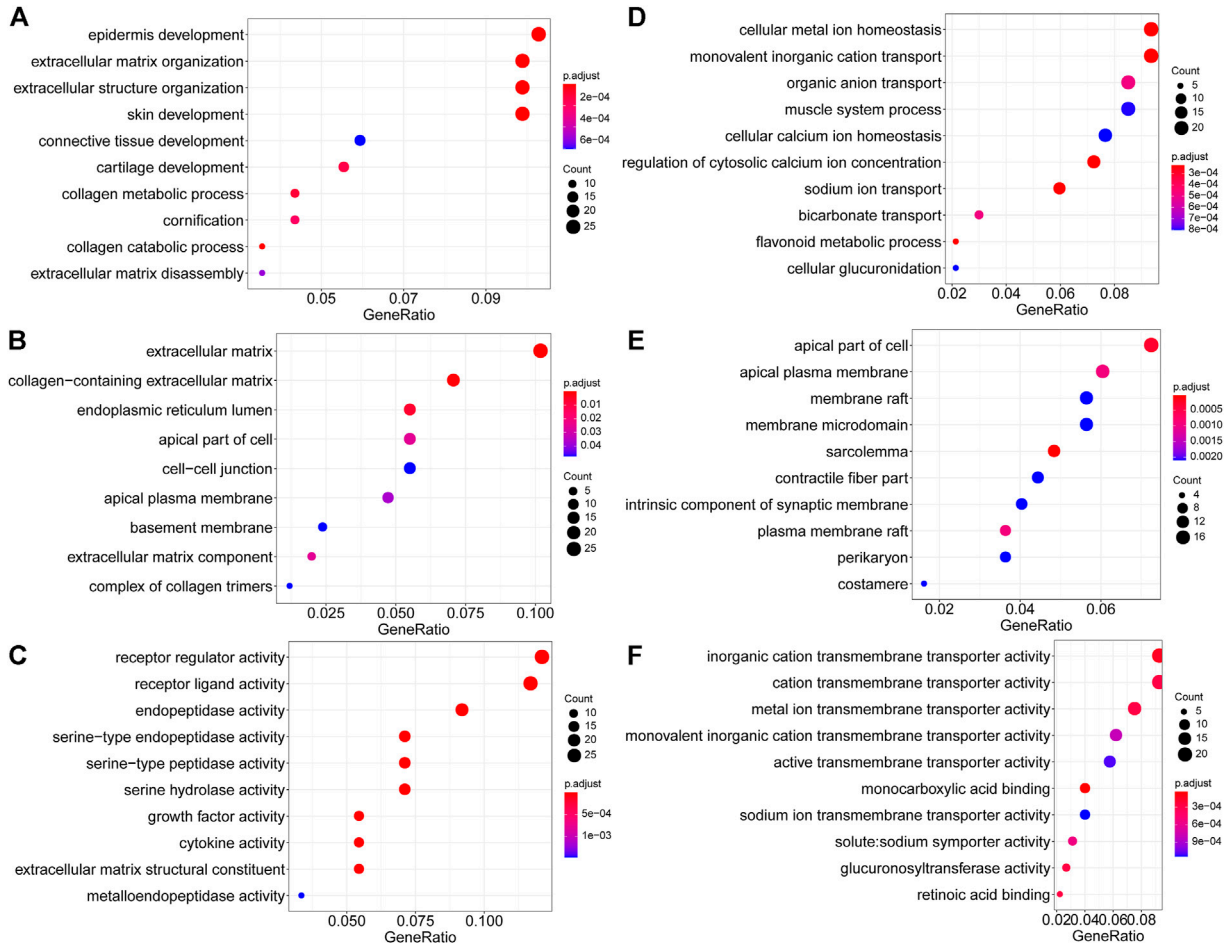


FIGURE 4
GO enrichment analysis of DEGs. **(A–C)** Upregulated DEGs. **(A)**. Biological process. **(B)** Cellular component. **(C)** Molecular function. **(D–F)**. Downregulated DEGs. **(D)**. Biological process. **(E)** Cellular component. **(F)** Molecular function.

the enrichment analysis on the overlapped DEGs using KEGG and GO databases. KEGG pathway enrichment results showed that upregulated DEGs were enriched in "Rheumatoid arthritis", "IL−17 signaling pathway", "Cytokine−cytokine receptor interaction", "Wnt signaling pathway", "Neuroactive ligand−receptor interaction", and "TNF signaling pathway" (Figure 3A), while downregulated DEGs were enriched in "Bile secretion", "Neuroactive ligand−receptor interaction", "Drug metabolism − cytochrome P450", "Mineral absorption", "Ascorbate and aldarate metabolism", "Retinol metabolism", "Chemical carcinogenesis − DNA adducts", and "Pentose and glucuronate interconversions" (Figure 3B).

GO terms cover biological process (BP), cellular component (CC), and molecular function (MF). For upregulated DEGs, the enriched BP terms included "epidermis development", "extracellular matrix organization", "extracellular structure organization", "skin development", "connective tissue development", "cartilage development", "collagen metabolic process", "cornification", "collagen catabolic process" (Figure 4A). In the CC group, upregulated DEGs were primarily enriched in "extracellular matrix", "collagen−containing extracellular matrix", "endoplasmic reticulum lumen", "apical part of cell", "cell−cell junction", "apical plasma membrane", "basement membrane", "extracellular matrix component" and "complex of collagen trimers" (Figure 4B). And enriched MF-related terms of upregulated DEGs were "receptor regulator activity", "receptor ligand activity", "endopeptidase activity", "serine−type endopeptidase activity", "serine−type peptidase activity", "serine hydrolase activity", "growth factor activity", "cytokine activity" and "extracellular matrix structural constituent" (Figure 4C). For downregulated DEGs, the enriched BP terms were "cellular metal ion homeostasis", "monovalent inorganic cation transport", "organic anion transport", "muscle system process", "cellular calcium ion homeostasis", "regulation of cytosolic calcium ion

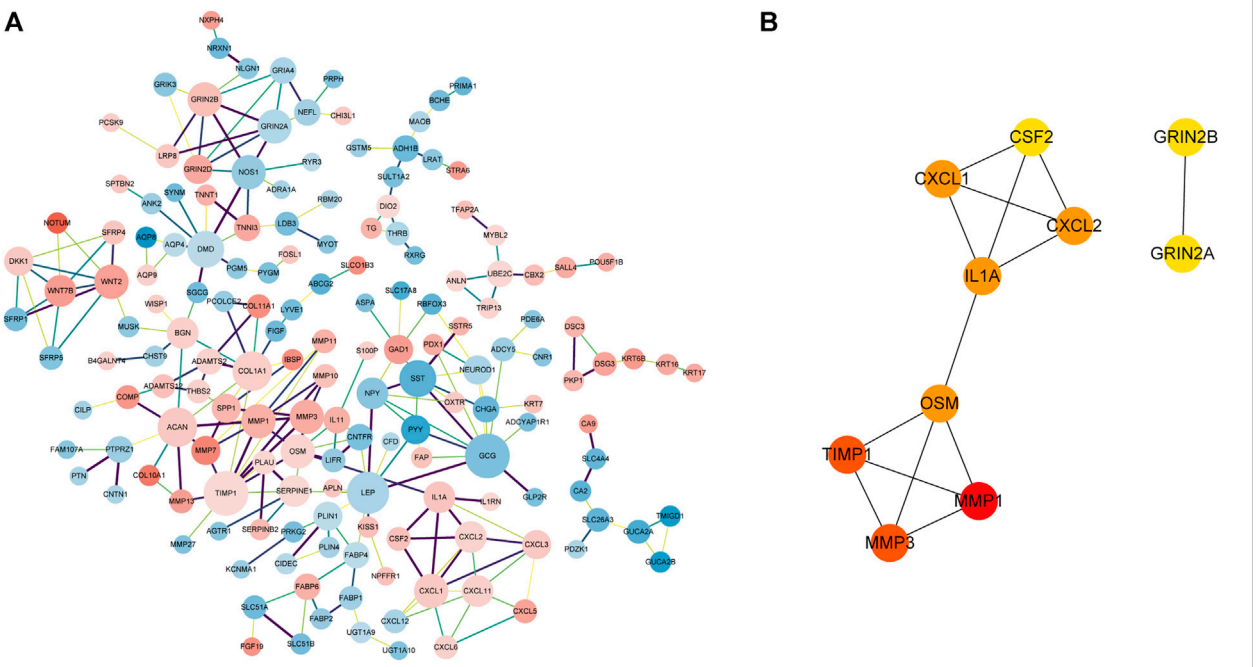


FIGURE 5
PPI network of DEGs and hub genes in the network. **(A)**. PPI network of DEGs with the interaction score>0.7. The pink nodes indicate significantly upregulated genes and the blue nodes indicate significantly downregulated genes. The edge thickness is proportional to the combined score of the connected genes. The size of the node is proportional to the value of log2|FC|. **(B)** Top 10 hub genes with a higher degree of connectivity.

concentration", "sodium ion transport", "bicarbonate transport", "flavonoid metabolic process" and "cellular glucuronidation" (Figure 4D). In the CC group, the downregulated DEGs were enriched in "apical part of cell", "apical plasma membrane", "membrane raft", "membrane microdomain", "sarcolemma", "contractile fiber part", "intrinsic component of synaptic membrane", "plasma membrane raft", "perikaryon" and "costamere" (Figure 4E). The enriched MF-related terms of the downregulated DEGs were "inorganic cation transmembrane transporter activity", "cation transmembrane transporter activity", "metal ion transmembrane transporter activity", "monovalent inorganic cation transmembrane transporter activity", "active transmembrane transporter activity", "monocarboxylic acid binding", "sodium ion transmembrane transporter activity", "solute:sodium symporter activity", "glucuronosyltransferase activity" and "retinoic acid binding" (Figure 4F).

Hub genes in the PPI network of DEGs

Based on the STRING online database (http://string-db.org) and Cytoscape software, a PPI network of 164 DEGs and 241 edges was constructed. The minimum required interaction score of each edge were bigger than 0.7 (Figure 5A) which excludes 376 DEGs. The top 10 hub genes according to the node degree were *MMP1, MMP3, TIMP1, OSM, IL1A, CXCL1, CXCL2, CSF2, GRIN2A,* and *GRIN2B* (Figure 5B).

Correlation between hub genes expression and overall survival

To examine the potential relationship between DEGs and overall survival (OS), a weighted Kaplan Meier survival curves were generated from TCGA data. The survival curves of the top four hub genes were shown in Figure 6, which shown that only *TIMP1* is associated with OS (p-value<0.05), and its high expression led to poor prognosis (Figure 6A). Other hub genes are not significantly associated with OS (Figures 6B–D) and Supplementary Figure S1.

Identification of potential drugs

249 potential drugs were predicted according to the DEGs. For example, we plotted five drugs for upregulated DEGs and

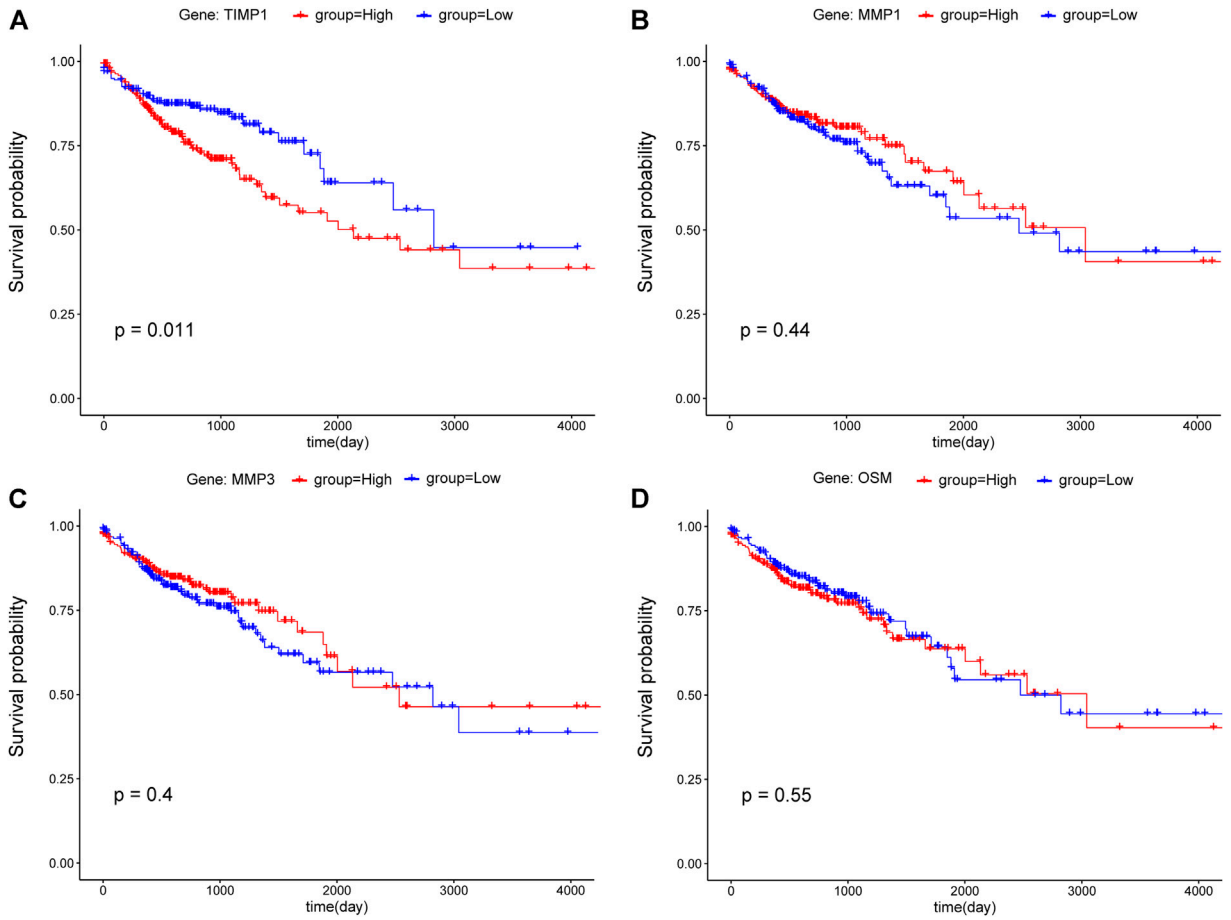


FIGURE 6
Kaplan-Meier survival curves of CRC patients for the top four hub genes including **(A)** *TIMP1*, **(B)** *MMP1*, **(C)** *MMP3*, and **(D)** *OSM*. According to the median value, gene expression was divided into two groups (red: high; blue: low), and the *p*-value<0.05 was considered statistically significant.

five drugs for downregulated DEGs in Figure 7. Figure 7 indicated that formaldehyde, glucocorticoid|dexamethasone, paclitaxel|eribulin, messenger RNA|inhibitor, and eribulin|paclitaxel could reverse upregulated DEGs. fluoxetine|sucrose|antidepressant|imipramine, nevirapine, sucrose|antidepressant|imipramine|L-proline residue|, imipramine|sucrose|antidepressant|, and histone|N-methyl-D-aspartic acid could reverse the downregulated DEGs.

Since *TIMP1* is significantly related to the OS of CRC patients, and the high expression of *TIMP1* is correlated to a poor prognosis, we next looked for drugs/small molecules that can reverse the expression of *TIMP1*, which might improve the prognosis of CRC patients. We provided details of the top 10 drugs that can reverse the hub gene *TIMP1* in Table 2, including formaldehyde, paclitaxel|eribulin, erlotinib|dimethyl sulfoxide, glucocorticoid|dexamethasone, antagonist, trichostatin A, rosiglitazone, inhibitor, retinoic acid, and cisplatin. Among them, eight drugs/small molecules were confirmed to be related to *TIMP1* or CRC. It is reported that exposure to formaldehyde can reduce *TIMP1* expression (Kang et al., 2022).

Discussion

In recent decades, CRC, including colon and rectal cancer, has become one of the main causes of cancer-related death around the world (Fuccio et al., 2018; Røed Skårderud et al., 2018; He et al., 2020a; He et al., 2020b). Therefore, it is urgent to find more effective prevention and treatment to reverse this problem (Teer et al., 2017). With the recent progress in the field of medicine and biotechnology, many preclinical and clinical studies have been carried out to reveal the potential mechanism of CRC liver metastasis. Identifying cancer-related marker genes through gene-targeted therapy is a new and effective potentially powerful treatment for CRC (Okugawa

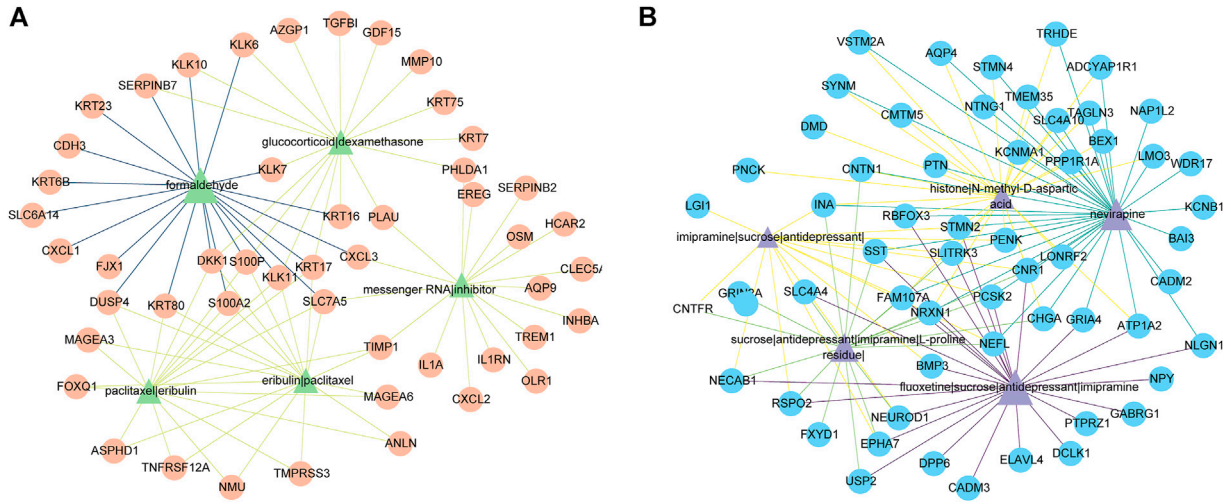


FIGURE 7
The predicted top five drugs and their gene networks. The color and thickness of the edges are inversely proportional to the *p*-value of drugs and DEGs. **(A)** Five drugs (green triangles) and 50 upregulated DEGs (orange circles). **(B)** Five drugs (purple triangles) and 59 downregulated DEGs (blue circles).

TABLE 2 Top 10 drugs for *TIMP1* that were significantly associated with survival rate of CRC patients.

Gene name	Drug/Small molecule	*p*-value	Possible effect	Evidence
TIMP1	formaldehyde	1.04403630163397E-06	Formaldehyde is a colorless, irritant, highly active and toxic environmental pollutant, which is used in various industries and products. Inhaled formaldehyde is a human and animal carcinogen that can cause genotoxicity, such as the formation of reactive oxygen species and DNA damage	PMID:35379891
	paclitaxel|eribulin	8.64715E-06	A well-known anticancer agent with a unique mechanism of action. It is considered to be one of the most successful natural anticancer drugs	Unconfirmed
	erlotinib|dimethyl sulfoxide	1.74273E-05	It can interfere with a variety of cellular processes, such as cell proliferation, differentiation, apoptosis and cycle	PMID: 32911099
	glucocorticoid|dexamethasone	3.5055E-05	It has pharmacological effects of anti-inflammatory, anti-endotoxin, inhibiting immunity, anti-shock and enhancing stress response	PMID: 21789017
	antagonist	3.66683E-05	It can bind to receptors and has strong affinity without intrinsic activity ($\alpha = 0$) drugs	Unconfirmed
	trichostatin A	0.0000703775443677834	trichostatin A (TSA), a histone deacetylase (HDAC) inhibitor	PMID: 21520296
	rosiglitazone	0.000141022683961323	Rosiglitazone is a thiazolidinedione insulin sensitizer. Its mechanism of action is similar to that of specific peroxisome proliferator activator γ Type a receptor	PMID: 29743857
	Inhibitor	0.000282045	Inhibitors of proteinases or antibodies against certain proteolytic enzymes can prevent tumor invasion and metastasis in experimental conditions	PMID: 23202950
	retinoic acid	0.000564894	Retinoic acid (RA) signal transduction is an important and conservative way to regulate cell proliferation and differentiation. In addition, disturbed RA signaling is associated with the occurrence and progression of cancer	PMID: 34877501
	cisplatin	0.000758387	Cisplatin is an inorganic platinum complex, which can be inhibited by the formation of DNA adducts in tumor cells	PMID: 32329836; PMID: 20607860

et al., 2015; Guo et al., 2017). High throughput sequencing technology provides a new perspective on the genome, transcriptome, and epigenome characteristics of cancer. In this study, we aim to reveal the hub gene of CRC through bioinformatics methods and identify potential drugs or small molecules, to improve the predictive power of CRC and provide a valuable theoretical basis for the clinical treatment of CRC patients.

First, RNA-seq data and clinical information of 478 CRC tumor samples and 41 healthy control samples were downloaded from TCGA. In addition, RNA-seq data of 10 tumor samples and 10 normal samples were obtained from the GSE104836 dataset. Using DESeq2 to detect the DEGs from TCGA and GEO respectively, 2664 DEGs were identified from TCGA, 959 DEGs were identified from the GSE104836 data set, and 540 DEGs appeared in both datasets, including 276 upregulated genes and 264 downregulated genes. KEGG pathway enrichment results showed that upregulated DEGs are enriched in "Rheumatoid arthritis", "IL−17 signaling pathway", "Cytokine−cytokine receptor interaction", "Wnt signaling pathway", "Neuroactive ligand−receptor interaction", and "TNF signaling pathway" (Figure 3A). It has been reported that IL − 17 is able to regulate colorectal tumor cells and inhibits their production of cxcl9/10 chemokines, thus prevents the infiltration of CD8 + CTLs and Tregs into CRC tumor, thereby promoting the development of CRC (Chen et al., 2019). Wnt signaling pathway is the key medium of tissue homeostasis and repair. Almost all CRC tumors show overactivation of Wnt pathway (Schatoff et al., 2017; Bian et al., 2020). GO enrichment analysis shows that epidermis development, extracellular matrix, and receptor regulator activity are the most significantly abundant upregulated DEGs in biological processes, cellular components, and molecular function categories. Downregulated DEGs are enriched in "Bile secretion", "Neuroactive ligand−receptor interaction", "Drug metabolism−cytochrome P450", "Mineral absorption", "Ascorbate and aldarate metabolism", "Retinol metabolism", "Chemical carcinogenesis−DNA adducts", and "Pentose and glucuronate interconversions" (Figure 3B). Previous studies have shown that a high-fat diet promotes the secretion of bile acids, thereby inducing the formation of precancerous lesions and/or aggravating the occurrence of colon tumors (Ocvirk and O'Keefe, 2021). Neuroactive ligand-receptor interactions were associated with other gastrointestinal cancers (Yu et al., 2021). The lack and deficiency of minerals may be related to cancer and increase the risk of cancer; For example, effective absorption of vitamin D can prevent colorectal cancer (Takada and Makishima, 2017).

To identify the key regulating genes in CRC development, a PPI network was constructed based on overlapping DEGs. In this network, edges with association scores <0.7 were filtered out. The PPI network obtained based STRING online database has 164 nodes, and the top 10 hub genes, including *MMP1, MMP3, TIMP1, OSM, IL1A, CXCL1, CXCL2, CSF2, GRIN2A, and GRIN2B*, were identified using Cytoscape. Among them, *TIMP1* is a soluble protein that can be released from endometrial cells, fibroblasts, and cancer cells, which are correlated with the prognosis of various cancers (Peng et al., 2011; Wang et al., 2013). The Kaplan–Meier survival analysis of Zheng et al. showed that *TIMP1* expression was upregulated in CRC tissues and was also connected with poor prognosis in GEPIA datasets (p-value = 0.02) (Zheng et al., 2020). Song et al. (2016) reported that *TIMP1* depletion can inhibit the proliferation, migration, and invasion of colon cancer cells, and inhibit the tumorigenesis and metastasis of CRC. Consistent with these studies, our results show that *TIMP1* was up-regulated in CRC samples compared with matched normal tissue samples, and its high expression was associated with poor OS in CRC patients.

Based on DEGs and CREEDS, we made drug predictions for all DEGs (Yang et al., 2020). Previous studies have shown the anti-migration and anti-invasion effects of imipramine, an FDA-approved antidepressant oral drug, on CRC cells (Liu et al., 2016; Alburquerque-González et al., 2020). Fluoxetine has been shown to induce antitumor activity. It was found that fluoxetine could selectively induce concentration-dependent apoptosis in human CRC cells by changing mitochondrial membrane potential and inducing phosphatidylserine translocation to the outer membrane (Marcinkute et al., 2019). In addition, 10 potential drugs were identified to reverse the expression of *TIMP1*. It has been shown that after glucocorticoid treatment, the expression level of *TIMP1* in patients with idiopathic pulmonary fibrosis (IPF) were significantly lower than those before glucocorticoid treatment ($p < 0.05$) (Zhang et al., 2015). Dexamethasone is a synthetic steroid with anti-inflammatory, anti-allergic, and immunosuppressive properties (Sinner, 2019). Trichostatin A is a histone deacetylase (HDAC) inhibitor, which inhibits the growth of CRC cells and induces G1 cell cycle arrest and apoptosis by regulating the downstream target of the JAK2/STAT3 signal (Xiong et al., 2012). A study on the effect of cisplatin on the invasion of ovarian cancer cells showed that the use of cisplatin could reduce the expression of *TIMP1* by 5.0 times ($p < 0.05$) (Karam et al., 2010). It is worth noting that there is no relevant evidence that paclitaxel|eribulin, and Antiagonist are related to the expression of *TIMP1* or the outcome of CRC. Further experiments are needed to verify their effectiveness of action, which may provide a basis for guiding the treatment of CRC patients.

Overall, this study revealed the altered gene expressions and enriched pathways in CRC based on bioinformatics analyses and provides insights for further screening of effective biomolecules for CRC treatment intervention, which is of clinical significance. However, the current research has some limitations. First, because the candidate prognosis-related central DEGs were detected using the data from two independent databases, more datasets were needed to confirm our discoveries. Secondly, experimental methods such as PCR were also needed to verify the DEGs. Third, clinical trials were needed to identify effects of the predicted drugs.

Conclusion

Our study effectively identified several candidate drug targets through differentially gene expression analysis, hub gene analysis and survival analysis for CRC treatment. We revealed compounds that have the potential to reverse the expressions

of the identified DEGs. These findings provide new directions for the diagnosis and treatment of CRC.

Data availability statement

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

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

SZ conceived and designed the study. XL, JY, and TL performed the experiments. CZ and ZC analyzed the data. XL wrote the manuscript. CL and JS contributed to literature review and language editing. All authors have read and approved this manuscript.

Conflicts of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1017539/full#supplementary-material

References

Alburquerque-González, B., Bernabe-Garcia, M., Montoro-Garcia, S., Bernabe-Garcia, A., Rodrigues, P. C., Ruiz Sanz, J., et al. (2020). New role of the antidepressant imipramine as a Fascin1 inhibitor in colorectal cancer cells. *Exp. Mol. Med.* 52 (2), 281–292. doi:10.1038/s12276-020-0389-x

Alves Martins, B. A., de Bulhoes, G. F., Cavalcanti, I. N., Martins, M. M., de Oliveira, P. G., and Martins, A. M. A. (2019). Biomarkers in colorectal cancer: The role of translational proteomics research. *Front. Oncol.* 9, 1284. doi:10.3389/fonc.2019.01284

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66 (4), 683–691. doi:10.1136/gutjnl-2015-310912

Bian, J., Dannappel, M., Wan, C., and Firestein, R. (2020). Transcriptional regulation of wnt/β-catenin pathway in colorectal cancer. *Cells* 9 (9), E2125. doi:10.3390/cells9092125

Chen, J., Ye, X., Pitmon, E., Lu, M., Wan, J., Jellison, E. R., et al. (2019). IL-17 inhibits CXCL9/10-mediated recruitment of CD8(+) cytotoxic T cells and regulatory T cells to colorectal tumors. *J. Immunother. Cancer* 7 (1), 324. doi:10.1186/s40425-019-0757-z

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., et al. (2021). gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res.* 50, D795–D800. doi:10.1093/nar/gkab786

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48 (D1), D554-D560–D560. doi:10.1093/nar/gkz843

De Sousa, E. M. F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P. M. H., et al. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.* 19 (5), 614–618. doi:10.1038/nm.3174

Deshiere, A., Berthet, N., Lecouturier, F., Gaudaire, D., and Hans, A. (2019). Molecular characterization of Equine Infectious Anemia Viruses using targeted sequence enrichment and next generation sequencing. *Virology* 537, 121–129. doi:10.1016/j.virol.2019.08.016

Fuccio, L., Repici, A., Hassan, C., Ponchon, T., Bhandari, P., Jover, R., et al. (2018). Why attempt *en bloc* resection of non-pedunculated colorectal adenomas? A systematic review of the prevalence of superficial submucosal invasive cancer after endoscopic submucosal dissection. *Gut* 67 (8), 1464–1474. doi:10.1136/gutjnl-2017-315103

Grassetto, G., Capirci, C., Marzola, M. C., Rampin, L., Chondrogiannis, S., Musto, A., et al. (2012). Colorectal cancer: Prognostic role of 18F-FDG-PET/CT. *Abdom. Imaging* 37 (4), 575–579. doi:10.1007/s00261-011-9789-7

Guo, Y., Bao, Y., Ma, M., and Yang, W. (2017). Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis. *Int. J. Mol. Sci.* 18 (4), E722. doi:10.3390/ijms18040722

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., et al. (2020). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta. Mol. Basis Dis.* 1866 (11), 165916. doi:10.1016/j.bbadis.2020.165916

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: A novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Hong, J., Lin, X., Hu, X., Wu, X., and Fang, W. (2021). A five-gene signature for predicting the prognosis of colorectal cancer. *Curr. Gene Ther.* 21 (4), 280–289. doi:10.2174/1566523220666201012151803

Kang, D. S., Lee, N., Shin, D. Y., Jang, Y. J., Lee, S. H., Lim, K. M., et al. (2022). Network-based integrated analysis for toxic effects of high-concentration formaldehyde inhalation exposure through the toxicogenomic approach. *Sci. Rep.* 12 (1), 5645. doi:10.1038/s41598-022-09673-0

Karam, A. K., Santiskulvong, C., Fekete, M., Zabih, S., Eng, C., and Dorigo, O. (2010). Cisplatin and PI3kinase inhibition decrease invasion and migration of human ovarian carcinoma cells and regulate matrix-metalloproteinase expression. *Cytoskelet. Hob.* 67 (8), 535–544. doi:10.1002/cm.20465

Kong, F., Zou, H., Liu, X., He, J., Zheng, Y., Xiong, L., et al. (2020). miR-7112-3p targets PERK to regulate the endoplasmic reticulum stress pathway and apoptosis

induced by photodynamic therapy in colorectal cancer CX-1 cells. *Photodiagnosis Photodyn. Ther.* 29, 101663. doi:10.1016/j.pdpdt.2020.101663

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, J., Lan, Y., Tian, G., and Yang, J. (2022). A systematic framework for identifying prognostic genes in the tumor microenvironment of colon cancer. *Front. Oncol.* 12, 899156. doi:10.3389/fonc.2022.899156

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia. *Sci. Rep.* 6, 22811. doi:10.1038/srep22811

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Marcinkute, M., Afshinjavid, S., Fatokun, A. A., and Javid, F. A. (2019). Fluoxetine selectively induces p53-independent apoptosis in human colorectal cancer cells. *Eur. J. Pharmacol.* 857, 172441. doi:10.1016/j.ejphar.2019.172441

McQuade, R. M., Stojanovska, V., Bornstein, J. C., and Nurgali, K. (2017). Colorectal cancer chemotherapy: The evolution of treatment and new approaches. *Curr. Med. Chem.* 24 (15), 1537–1557. doi:10.2174/0929867324666170111152436

Meng, Y., Lu, C., Jin, M., Xu, J., Zeng, X., and Yang, J. (2022). A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief. Bioinform.* 23 (2), bbab581. doi:10.1093/bib/bbab581

Nguyen, H. T., and Duong, H. Q. (2018). The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy. *Oncol. Lett.* 16 (1), 9–18. doi:10.3892/ol.2018.8679

O'Connell, J. B., Maggard, M. A., and Ko, C. Y. (2004). Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *J. Natl. Cancer Inst.* 96 (19), 1420–1425. doi:10.1093/jnci/djh275

Ocvirk, S., and O'Keefe, S. J. D. (2021). Dietary fat, bile acid metabolism and colorectal cancer. *Semin. Cancer Biol.* 73, 347–355. doi:10.1016/j.semcancer.2020.10.003

Okugawa, Y., Grady, W. M., and Goel, A. (2015). Epigenetic alterations in colorectal cancer: Emerging biomarkers. *Gastroenterology* 149 (5), 1204–1225. doi:10.1053/j.gastro.2015.07.011.e12

Peng, L., Yanjiao, M., Ai-guo, W., Pengtao, G., Jianhua, L., Ju, Y., et al. (2011). A fine balance between CCNL1 and TIMP1 contributes to the development of breast cancer cells. *Biochem. Biophys. Res. Commun.* 409 (2), 344–349. doi:10.1016/j.bbrc.2011.05.021

Røed Skårderud, M., Polk, A., Kjeldgaard Vistisen, K., Larsen, F. O., and Nielsen, D. L. (2018). Efficacy and safety of regorafenib in the treatment of metastatic colorectal cancer: A systematic review. *Cancer Treat. Rev.* 62, 61–73. doi:10.1016/j.ctrv.2017.10.011

Rogler, G. (2014). Chronic ulcerative colitis and colorectal cancer. *Cancer Lett.* 345 (2), 235–241. doi:10.1016/j.canlet.2013.07.032

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* 19 (5), 619–625. doi:10.1038/nm.3175

Saito, M., Momma, T., and Kono, K. (2018). Targeted therapy according to next generation sequencing-based panel sequencing. *Fukushima J. Med. Sci.* 64 (1), 9–14. doi:10.5387/fms.2018-02

Schatoff, E. M., Leach, B. I., and Dow, L. E. (2017). Wnt signaling and colorectal cancer. *Curr. Colorectal Cancer Rep.* 13 (2), 101–110. doi:10.1007/s11888-017-0354-9

Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., et al. (2012). Cancer treatment and survivorship statistics, 2012. *Ca. Cancer J. Clin.* 62 (4), 220–241. doi:10.3322/caac.21149

Sinner, B. (2019). [Perioperative dexamethasone]. *Anaesthesist* 68 (10), 676–682. doi:10.1007/s00101-019-00672-x

Song, G., Xu, S., Zhang, H., Wang, Y., Xiao, C., Jiang, T., et al. (2016). TIMP1 is a prognostic marker for the progression and metastasis of colon cancer through FAK-PI3K/AKT and MAPK pathway. *J. Exp. Clin. Cancer Res.* 35 (1), 148. doi:10.1186/s13046-016-0427-7

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Swiderska, M., Choromanska, B., Dabrowska, E., Konarzewska-Duchnowska, E., Choromanska, K., Szczurko, G., et al. (2014). The diagnostics of colorectal cancer. *Contemp. Oncol.* 18 (1), 1–6. doi:10.5114/wo.2013.39995

Takada, I., and Makishima, M. (2017). Control of inflammatory bowel disease and colorectal cancer by synthetic vitamin D receptor ligands. *Curr. Med. Chem.* 24 (9), 868–875. doi:10.2174/0929867323666161202145509

Teer, J. K., Zhang, Y., Chen, L., Welsh, E. A., Cress, W. D., Eschrich, S. A., et al. (2017). Evaluating somatic tumor mutation detection without matched normal samples. *Hum. Genomics* 11 (1), 22. doi:10.1186/s40246-017-0118-2

The Gene Ontology, C. (2019). The gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47 (D1), D330-D338–D338. doi:10.1093/nar/gky1055

Wang, Y. Y., Li, L., Zhao, Z. S., and Wang, H. J. (2013). Clinical utility of measuring expression levels of KAP1, TIMP1 and STC2 in peripheral blood of patients with gastric cancer. *World J. Surg. Oncol.* 11, 81. doi:10.1186/1477-7819-11-81

Xiong, H., Du, W., Zhang, Y. J., Hong, J., Su, W. Y., Tang, J. T., et al. (2012). Trichostatin A, a histone deacetylase inhibitor, suppresses JAK2/STAT3 signaling via inducing the promoter-associated histone acetylation of SOCS1 and SOCS3 in human colorectal cancer cells. *Mol. Carcinog.* 51 (2), 174–184. doi:10.1002/mc.20777

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., et al. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. *Comput. Biol. Med.* 146, 105516. doi:10.1016/j.compbiomed.2022.105516

Yu, J., Zhang, Q., Wang, M., Liang, S., Huang, H., Xie, L., et al. (2021). Comprehensive analysis of tumor mutation burden and immune microenvironment in gastric cancer. *Biosci. Rep.* 41 (2), BSR20203336. doi:10.1042/BSR20203336

Zhang, H. T., Fang, S. C., Wang, C. Y., Wang, W., Wu, J., Wang, C., et al. (2015). MMP-9 1562C>T gene polymorphism and efficacy of glucocorticoid therapy in idiopathic pulmonary fibrosis patients. *Genet. Test. Mol. Biomarkers* 19 (11), 591–597. doi:10.1089/gtmb.2015.0057

Zhang, Y., Xiang, J., Tang, L., Li, J., Lu, Q., Tian, G., et al. (2021). Identifying breast cancer-related genes based on a novel computational framework involving KEGG pathways and PPI network modularity. *Front. Genet.* 12, 596794. doi:10.3389/fgene.2021.596794

Zhao, B., Baloch, Z., Ma, Y., Wan, Z., Huo, Y., Li, F., et al. (2019). Identification of potential key genes and pathways in early-onset colorectal cancer through bioinformatics analysis. *Cancer control.* 26 (1), 1073274819831260. doi:10.1177/1073274819831260

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci. U. S. A.* 117 (14), 7568–7569. doi:10.1073/pnas.2000499117

Zheng, Z., Xie, J., Xiong, L., Gao, M., Qin, L., Dai, C., et al. (2020). Identification of candidate biomarkers and therapeutic drugs of colorectal cancer by integrated bioinformatics analysis. *Med. Oncol.* 37 (11), 104. doi:10.1007/s12032-020-01425-2

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 28 October 2022
DOI 10.3389/fgene.2022.1037120

OPEN ACCESS

EDITED BY
Jialiang Yang,
Geneis (Beijing) Co., Ltd., China

REVIEWED BY
Huihui Wang,
China Medical University, China
Xiang Hu,
Hunan Normal University, China
Xiao Liang,
Shanghai JiaoTong University, China

*CORRESPONDENCE
Feng Zhang,
felix.f.zhang@outlook.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 05 September 2022
ACCEPTED 10 October 2022
PUBLISHED 28 October 2022

CITATION
Xu C-Y, Zeng X-X, Xu L-F, Liu M and Zhang F (2022), Circular RNAs as diagnostic biomarkers for gastric cancer: A comprehensive update from emerging functions to clinical significances.
Front. Genet. 13:1037120.
doi: 10.3389/fgene.2022.1037120

Circular RNAs as diagnostic biomarkers for gastric cancer: A comprehensive update from emerging functions to clinical significances

Chun-Yi Xu[1,2†], Xi-Xi Zeng[2,3†], Li-Feng Xu[2,3], Ming Liu[3,4,5] and Feng Zhang[2,3*]

[1]Zhejiang Chinese Medical University, Hangzhou, China, [2]Core Facility, Quzhou People's Hospital, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou, China, [3]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [4]The Joint Innovation Center for Engineering in Medicine, Quzhou, China, [5]University of Electronic Science and Technology of China, Chengdu, China

The incidence and mortality of gastric cancer ranks as a fouth leading cause of cancer death worldwide, especially in East Asia. Due to the lack of specific early-stage symptoms, the majority of patients in most developing nations are diagnosed at an advanced stage. Therefore, it is urgent to find more sensitive and reliable biomarkers for gastric cancer screening and diagnosis. Circular RNAs (circRNAs), a novel type of RNAs with covalently closed loops, are becoming a latest hot spot in the field of. In recent years, a great deal of research has demonstrated that abnormal expression of circRNAs was associated with the development of gastric cancer, and suggested that circRNA might serve as a potential biomarker for gastric cancer diagnosis. In this review, we summarize the structural characteristics, formation mechanism and biological function of circRNAs, and elucidate research progress and existing problems in early screening of gastric cancer.

KEYWORDS

circular RNA, gastric cancer, biomarker, bodily fluids, diagnosis

Abbreviations: AUC, area under curve; CEA, carcinoembryonic antigen; CDDP, cisplatin; ciRNAs, intronic circRNAs; circRNA, circular RNA; ecircRNAs, exonic circRNAs; EIciRNAs, exon–intron circular RNAs; GC, gastric cancer; HuR, human antigen R; IRES, internal ribosome entry site; MBL, Muscleblind protein; miRNA, microRNA; RNase, ribonuclease; RBPs, RNA binding proteins; rRNA, ribosomal RNA; TEM, transmission electron microscopy.

1 Introduction

Gastric cancer (GC) is one of the most prevalent forms of cancer. It ranks fifth and fourth in morbidity and mortality respectively among all tumors. There are geographical and populational distribution differences in different regions, among which East Asia, South America, Central America and Eastern Europe have higher incidence rates than those of other regions (2020). Particularly in Japan, South Korea, and China, gastric cancer is one of the most commonly diagnosed cancers (Bray et al., 2018). Early-stage gastric cancer lacks specific symptoms, making it difficult to detect. Approximately two-thirds of gastric cancer patients in China are diagnosed at an advanced stage, which lacks effective treatments (Shen et al., 2013). Even given the neoadjuvant therapy combined with surgery, the 5-year progression-free rate in patients with advanced gastric cancer is only 20%–30% (Sitarz et al., 2018). At present, endoscopic biopsy and histopathological examination are the gold standards for gastric cancer diagnosis. However, due to the discomfort caused during gastroscopy, general acceptance of endoscopy by the population as a screening approach remains low. In particular, endoscopy is restrictive in elderly patients and patients with cardiopulmonary insufficiency (Yao 2013). In addition, traditional laboratory tumor markers such as CEA, CA19-9, CA12-5, and CA72-4 have poor sensitivity and specificity in the detection of gastric cancer in the early stage (Sekiguchi and Matsuda 2020). At present, there is an urgent clinical demand for more reliable biomarkers to strengthen the detection of gastric cancer especially in the early stage.

Circular RNAs (circRNAs) are a type of closed-loop non-coding RNA without the 3'end poly-A structure and the 5'end cap structure (Kristensen et al., 2019). In recent years, the rapid development of genome microarray and whole-genome sequencing technology promotes the discovery of novel circRNAs. Research has demonstrated that abnormal expressions of circRNAs are associated with cancer development, and have proposed them as potential biomarkers for cancer diagnosis, including gastric cancer. In this review, we summarize and discuss findings in this field thus far, providing a comprehensive update on the application of circRNAs in the screening and diagnosis of gastric cancer.

2 Overview of circular RNAs

2.1 The biogenesis and classification of circular RNAs

CircRNAs are molecules of single-stranded RNA that have been covalently closed into a circular structure. Unlike linear RNA, circRNAs lack 5′ to 3′ polarity and polyadenylation [poly(A)] tail. Alternative exon splicing generates linear RNA, whereas circRNAs are typically generated by back splicing the 3′ end of the exon to the upstream exon or the 5′ end of itself. CircRNAs usually contains one to five exons. Therefore, circRNAs are resistant to ribonuclease (RNase) and exonuclease degradation, with a half-life of up to 48 h (Kristensen et al., 2019). The long half-life and tissue-specific expression pattern of circRNAs make them more appealing as diagnostic markers compared to other forms of RNAs.

CircRNAs can be divided into exonic circRNAs (ecircRNAs) formed only by exon sequences, intronic circRNAs (ciRNAs) formed by intron sequences, and exon-intron circular RNAs (EIciRNAs) composed of both exon and intron sequences, depending on their source (Zhu et al., 2019). The circularization process of circRNAs has been intensely studied, and several models have been investigated and validated. 1) Lasso-driven circularization model: the splice donor and splice acceptor form a lasso containing exons connected through covalent bonding, thereby forming ecircRNAs. 2) Intron pair-driven circularization model: complementary bases flanking introns bind together and bring two adjacent exons together. Introns are then removed by the spliceosome. Subsequently the splicing sites are joined to form EIciRNAs or ecircRNAs. 3) Intron circularization model: The remaining lasso introns in the pre-mRNA are circularized by the GU-rich sequence near the 5′ splice site and the C-rich sequence near the branch point. The circularized introns are further cut to form stable ciRNA. This ciRNA, which forms a lasso structure by connecting its two ends, can resist exonuclease degradation and has high stability. These structural characteristics are of great significance in the screening of cancers (Kristensen et al., 2019; Su et al., 2019).

The dynamics of the circularization of circRNAs are influenced by numerous factors, Zhang et al. (2014) showed that exon circularization depended on the complementary sequences of flanking introns, and the efficiency of circularization is controlled by the rivalry between RNA pairing across flanking introns and within individual introns. In addition, it was reported that proteins such as MBL (Muscleblind protein) were involved in the formation of circRNAs. MBL has binding sites on the flanking introns of its pre-mRNA that can promote the circularization of circRNAs (Ashwal-Fluss et al., 2014).

2.2 The biological functions of circular RNAs

In recent years, extensive research has been conducted on the biological functions of circRNAs, and several major functions have been elucidated. Firstly, circRNAs can act as competitive inhibitors of miRNA by binding to miRNAs, also known as "miRNA sponges," or as target mimics to inhibit the activity of a specific miRNA (Hansen et al., 2013). For example, ciRS-7 indirectly up-regulates the expressions of miR-7 target genes by binding to miR-7 and therefore participating in processes such

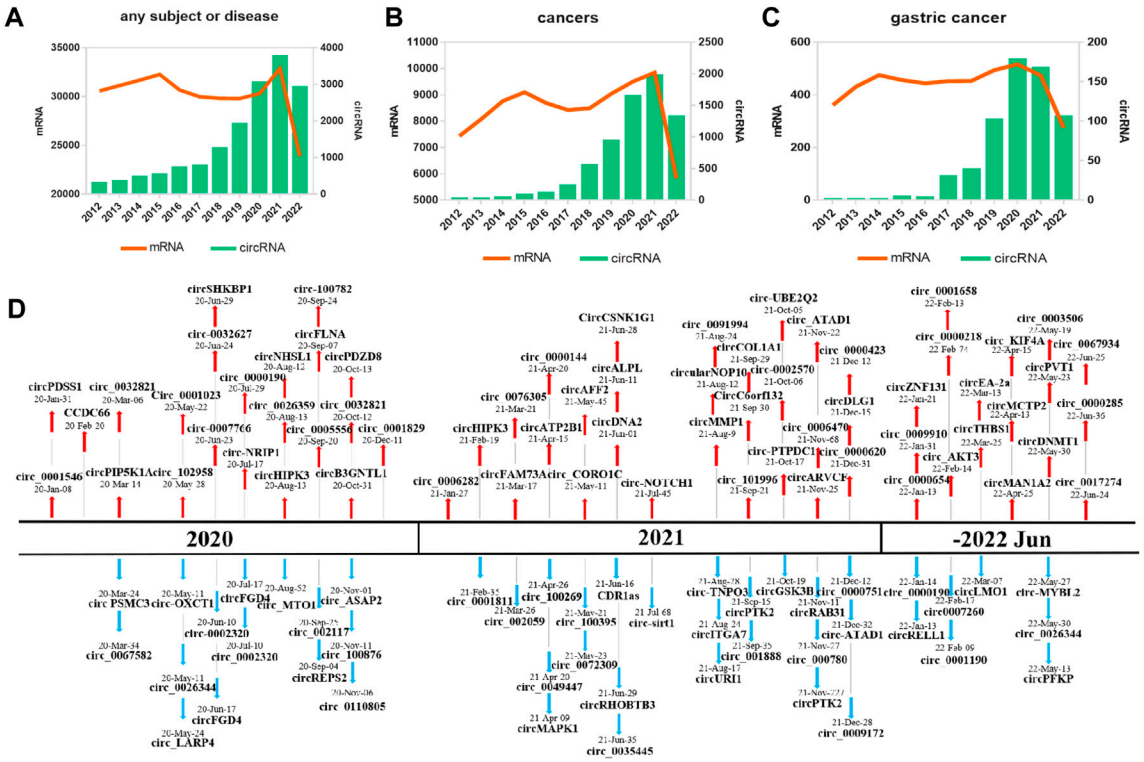


FIGURE 1
Research on and discovery of circRNAs in GC. The amount of research, as quantified by the annual number of peer-reviewed publications, has been relatively stable for mRNAs (orange line) but not for circRNAs (blue bars) in the following categories: **(A)** an overall, for any subject or disease; **(B)** cancers; **(C)** GC. **(D)** Increasing numbers of novel circRNAs were identified from 2020 to July 2022.

as insulin secretion, myocardial infarction and gastric cancer (GC) progression (Zheng et al., 2017; Pan et al., 2018a). Secondly, circRNAs interact with RNA binding proteins (RBPs) and thus indirectly affect the signaling pathways downstream of RBPs (Du et al., 2017). Thirdly, circRNAs work with U1 snRNP to stimulate the transcription of their parental genes (Li et al., 2015b). A few circRNAs can also function as templates for protein translation (Pan et al., 2018b).

3 CircRNAs in gastric cancer

3.1 Abnormal expression of circular RNAs in gastric cancer

A comprehensive review was conducted by searching PubMed for articles with the keywords ("circular RNA" and "gastric cancer") published over the past 10 years (January 2012–August 2022). Multiple studies have explored that the discovery and characterization of circRNAs in GC has increased annually, while protein-coding gene (mRNA) discovery research has remained stable (Figures 1A–C). These results show a rising fascination with circRNAs and their involvement in GC. Overall, related studies have validated 115 circRNAs (67 upregulated and 48 downregulated) in the past 3 years (Figure 1D).

Thousands of circRNAs have been identified by circRNA-specific microarrays and RNA-seq in GC tissues, cells, blood, and exosomes from patients with GC (Figure 2). Most of the gastric cancer-associated circRNAs are expressed in cancer tissues, with only a few circRNAs in body fluids. CircRNAs in plasma is easier to use for disease prediction and therapeutic efficacy judgment due to differences in ease of access in tissues.

3.1.1 Dysregulated circular RNAs in gastric cancer cells

Using high-throughput RNA-seq, Guo et al. (2022a) analyzed circRNA expression profiles in PBS-treated and Helicobacter pylori-infected AGS cells. As compared to the control, among 18,308 different circRNA candidates, the experiment yielded 101 significantly differentially expressed circRNAs, including 84 upregulated and 17 downregulated circRNAs. In addition, numerous studies have reported that circRNAs in gastric cell lines are dysregulated. CircAKT3 was

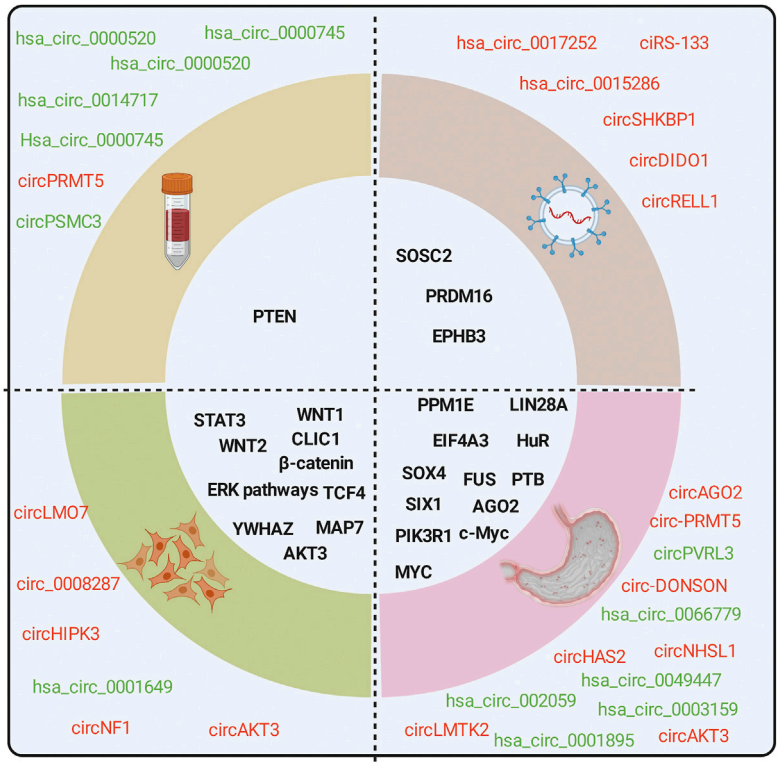


FIGURE 2
CircRNAs are associated with the hallmarks of GC. CircRNAs are differentially expressed in GC tissues, cells, exosomes, and blood from patients with GC compared with normal controls. Red for upregulation, green for downregulation.

identified as being overexpressed in MKN-7 and HGC-27 cells compared to GES-1 cells (Huang et al., 2019). Consistent with these findings, Yang et al. (2021a) determined that the expression level of circHIPK3 was elevated in gastric cancer cell lines compared with normal gastric cell lines. In addition, the expression of circLMO7 was significantly higher in gastric cancer cells than in GES-1 cells (Cao et al., 2021). These studies suggest that circRNA promotes the progression of gastric cancer.

3.1.2 Dysregulated circular RNAs in gastric cancer tissues

To identify the circRNAs involved in GC tumorigenesis, a recent study detected differential circRNA expression between GC tissues and adjacent noncancerous tissues. In a study by Shao et al. (2017), among the 308 significantly differentially expressed circRNAs, there were 107 (34.74%) upregulated ones. However, the majority (65.26%) of circRNAs were found to be down-regulated in cancer tissues. In addition, Zhang et al. (2017b) detected 3,071 expressed circRNA indicators in the six pairs of tumors and adjacent normal mucosal specimens, among these circRNAs, 46 indicators revealed different expression levels. Another study performed large-scale gene screening in three pairs of GC tissues using high-throughput sequencing, 25,303 circRNAs were detected in the screening. Of these circRNAs, 2,007 DECs were identified based on the filter criteria of $|FC| \geq 2$, $p < 0.05$ (Kong et al., 2019). Based on RNA-seq, Jie et al. (2020) found most of these circRNAs originated from exons, And thirteen candidate circRNAs were significantly downregulated and 9 were upregulated which were analyzed by ggplot2 between 30 pairs of gastric cancer and adjacent normal cancer tissues. Wang et al. (2021b) applied ribosomal RNA (rRNA)-depleted RNA-seq analysis of five-paired GC and normal tissues to systematically characterize the genome-wide landscape of circRNAs in GC. The result displayed 4485 circRNAs in GC and 5008 circRNAs in normal tissue. Among the dysregulated circRNAs, 245 candidates were significantly dysregulated (152 downregulated and 93 upregulated) in GC.

These sequencing and bioinformatics analysis illustrate the dysregulation of circRNA profiles in GC, However, the precise role and internal mechanisms of circRNAs in GC remain elusive.

3.1.3 Dysregulated circular RNAs in blood from patients with gastric cancer

Liquid biopsy is a noninvasive technique that utilizes body fluids such as blood, urine, and gastric juice to determine the disease state (Reimers et al., 2019). Identifying circulating tumor

markers in blood and other bodily fluids has been one of the research focuses in this area (Batth et al., 2017). Recently, circular RNAs (circRNAs) have attracted considerable attention in tumor biopsies as detection and quantitative biomarker (de Fraipont et al., 2019). Although research on circRNAs is in its infancy, numerous studies have indicated their potential as useful biomarkers for the diagnosis and prognosis of cancer (Arnaiz et al., 2019).

For instance, increased expression of serum circSHKBP1 (hsa_circ_0000936) level was significantly associated with poor survival and advanced TNM stage (Xie et al., 2020). CircPSMC3 was downregulated in plasmas in GC patients. Lower circPSMC3 expression was associated with a higher TNM stage and shorter overall survival in GC patients (Rong et al., 2019). It was discovered that hsa_circ_0000520 was significantly down-regulated in gastric cancer plasm compared to normal control. The hsa_circ_0000520 plasma concentration was linked to CEA expression based on clinicopathological characteristics. (Sun et al., 2018). Also, hsa_circ_0000745 was downregulated in GC plasma samples compared with healthy controls ($p < 0.001$). The plasma hsa_circ_0000745 levels were correlated with the stage of tumor-node metastasis. And the AUC of plasma hsa_circ_0000745 was elevated in conjunction with the level of carcinoembryonic antigen (CEA), which suggests plasma hsa_circ_0000745 is a good diagnostic biomarker (Huang et al., 2017). Furthermore, the group reported that hsa_circ_0000181 levels in plasma from GC patients were significantly lower than those in adjacent non-tumorous tissues and in healthy individuals ($p < 0.001$). In addition, the sensitivity of plasma hsa_circ_0000181 were 85.2% and 99.0% respectively (Zhao et al., 2018). A previous study demonstrated that hsa_circ_0000211, hsa_circ_0000284 and hsa_circ_0004771 exhibited identical expression profiles when analyzed by distinct techniques (RNA-Seq and RT-qPCR) and distinct sample types (tissue and blood) (Reis-das-Mercês et al., 2022).

At present, traditional circulating tumor markers in the setting of clinical laboratories such as CEA and CA19-9 have low specificity and sensitivity (Sekiguchi and Matsuda 2020), which limited their clinical application. On the other hand, studies have confirmed that circRNAs exist not only in tissues but also in human serum, plasma and other bodily fluids, especially enriched in microvesicles and exosomes (Li et al., 2015a). Therefore, circRNAs have the potential to be candidates as non-invasive tumor markers.

3.1.4 Dysregulated circRNAs in gastric cancer exosomes

Exosomes are nano-sized vesicles secreted by various cells that express exosome markers such as TSG101, HSP70, CD9, and CD63 but not albumin or calnexin (Feng et al., 2019; Hon et al., 2019). Transmission electron microscopy (TEM) images of exosomes typically depict translucent cup-shaped or spherical structures with diameters ranging from 30 to 150 nm (Xu et al., 2018; Mathieu et al., 2019). In recent years, it has been discovered that exosomes transport miRNAs, lncRNAs, proteins, and even circRNAs for intercellular signal transduction (Xu et al., 2018; Mathieu et al., 2019).

Systematic administration of circDIDO1 through exosome-mediated gene suppressed the tumorigenicity and aggressiveness of GC *in vitro* and *in vivo*, indicating that RGD-Exo-circDIDO1 could be employed as a nanomedicine for the treatment of GC (Guo et al., 2022b). In addition, GC cells' exosomal hsa_circ_0017252 inhibited GC progression by inhibiting macrophage M2-like polarization. These findings enhance our fundamental comprehension of GC and suggest a novel strategy for developing more effective GC treatments (Song et al., 2022). The expression level of exosomal hsa_circ_0015286 decreased significantly in GC patients following surgery. Patients with low hsa_circ_0015286 expression had a longer overall survival than those with high expression. Exosomal hsa_circ_0015286 may be a promising noninvasive biomarker for the diagnosis and prognostic evaluation of GC (Zheng et al., 2022). CircRELL1 is transmissible *via* exosomal communication, and exosomal circRELL1 inhibited the malignant behavior of GC *in vivo* and *in vitro*. This work reveals a promising novel circulating diagnostic biomarker and treatment target for GC (Sang et al., 2022). These circRNAs may play regulatory functions in the start of GC and may serve as biomarkers for the diagnosis of GC in liquid biopsies.

3.2 Molecular mechanisms of circular RNAs in gastric cancer

In an authoritative review, Kristensen et al. (2019) summarized that circRNAs perform regulatory roles may exert their biological functions by acting as miRNA sponges or decoys, protein sponges or decoys, enhancers of protein function, protein scaffolding, protein recruitment and templates for translation. The majority of circRNAs serve as microRNA (miRNA) sponges or decoys, shielding target mRNAs from miRNA-dependent destruction, thus inhibiting the activities of the corresponding miRNAs. The circRNAs that show different biological functions in GC are summarized in Table 1.

CircNHSL1 acts as a sponge for miR-1306-3p to alleviate its suppression of SIX1 target. Enhanced expression of circNHSL1 promotes invasion and metastasis of gastric cancer (Zhu et al., 2019). Functionally, circPVT1 serves as a sponge for miR-125 family members to stimulate cell proliferation (Chen J et al., 2017). Overexpression of

TABLE 1 CircRNAs associated with human gastric cancer.

CircRNAs	Deregulation	Mechanism (target genes)	Functions	References
circPVT1	Increased	miR-125	Cell growth	Chen et al. (2017)
circLMTK2	Increased	miR-150-5p	Cell growth and metastasis	Wang et al. (2019a)
circAGO2	Increased	miR-224-5p, miR-143-3p	Cell growth, invasion, and metastasis	Chen et al. (2019b)
circ-DONSON	Increased	SOX4	Cell growth and invasion	Ding et al. (2019)
circFNDC3B	Increased	E-cadherin, CD44	Cell migration and invasion	Hong et al. (2019)
circAKT3	Increased	miR-198, PIK3R1	Cell growth and apoptosis	Huang et al. (2019)
circRBMS3	Increased	miR-153, SNAI1	Cell growth and invasion	Li et al. (2019)
circPDSS1	Increased	miR-186-5p, NEK2	Cell cycle and apoptosis	Ouyang et al. (2019)
circNF1	Increased	miR-16	Cell growth	Wang et al. (2019c)
ciRS-133	Increased	miR-133	White adipose tissue browning, cancer-associated cachexia	Zhang et al. (2019a)
circDLST	Increased	miR-502-3p, NRAS/MEK1/ERK1/2	Cell viability, invasion, and metastasis	Zhang et al. (2019b)
circCACTIN	Increased	miR-331-3p, TGFBR1	Cell growth and metastasis	Zhang et al. (2019c)
circNRIP1	Increased	miR-149-5p	Cell growth and metastasis	Zhang et al. (2019d)
circNHSL1	Increased	miR-1306-3p	Cancer invasion and metastasis	Zhu et al. (2019)
circSERPINE2	Increased	miR-375, YWHAZ	Cell growth	Liu et al. (2019)
circHIPK3	Increased	miR-637, AKT1	Cell growth	Yang et al. (2021a)
circPRMT5	Increased	miR-145, miR-1304	Cell growth and metastasis	Du et al. (2019)
circSFMBT2	Increased	miR-182-5p	Cell growth	Li et al. (2020)
hsa_circ_0078607	Increased	miR-188-3p	Cell growth	Bian et al. (2021)
circSMAD4	Increased	miR-1276, CTNNB1	Cell growth	Wang et al. (2021a)
circLMO7	Increased	miR-30a-3p, WNT2/β-Catenin	Cell growth and metastasis	Cao et al. (2021)
circDUSP16	Increased	miR-145-5p	Cell growth and invasion	Zhang et al. (2020)
circOSBPL10	Increased	miR-136-5p, WNT2	Cell growth and metastasis	Wang et al. (2019b)
circSHKBP1	Increased	miR-582-3p, HUR/VEGF	Cell growth and metastasis	Xie et al. (2020)
circHECTD1	Increased	miR-137, PBX3	Cell growth	Lu et al. (2021)
hsa_circ_0081143	Increased	miR-646, CDK6	Cell growth and invasion	Lu et al. (2021)
circHAS2	Increased	miR-944, PPM1E	Cell growth and invasion	Ma et al. (2021a)
hsa_circ_0000993	Decreased	miR-214-5p	Cell growth and metastasis	Zhong et al. (2018)
circHuR	Decreased	HuR	Cell growth and invasion	Yang et al. (2019)
circHIAT1	Decreased	miR-21	Cell growth and migration	Quan et al. (2020)
circLARP4	Decreased	miR-424, LATS1	Cell growth and invasion	Zhang et al. (2017a)
circCUL2	Decreased	mir-142-3p, VAMP3	Cell growth and metastasis	Peng et al. (2020)
circPSMC3	Decreased	miR-296-5p	Cell growth and migration	Rong et al. (2019)
circRNA_100,269	Decreased	miR-630	Cell growth	Zhang et al. (2017c)
circYAP1	Decreased	miR-367-5p	Cell growth and invasion	Liu et al. (2018)
circFAT1(e2)	Decreased	miR-548g, RUNX1	Cell growth and metastasis	Fang et al. (2019)
circMCTP2	Decreased	miR-99a-5p, MTMR3	Cell proliferation and apoptosis	Sun et al. (2020)
circREPS2	Decreased	miR-558, RUNX3/β-catenin	Cell growth and migration	Guo et al. (2020)
circCCT3	Decreased	miR-613, VEGFA/VEGFR2	Cell migration and invasion	Hou et al. (2021)
circCCDC9	Decreased	miR-6792-3p, CAV1	Cell growth	Luo et al. (2020)
circSPECC1	Decreased	miR-526b, KDM4A/YAP1	Cell growth and invasion	Chen et al. (2019a)
circMRPS35	Decreased	KAT7/FOXO1/3a	Cell growth and invasion	Jie et al. (2020)
circRPPH1	Decreased	miR-512-5p, STAT1	Cell growth	Huang et al. (2021)
circRHOBTB3	Decreased	miR-654-3p, p21	Cell growth	Deng et al. (2020)
circMAPK1	Decreased	MAPK1	Cell growth and invasion	Jiang et al. (2021)

circLMTK2 enhances gastric cell proliferation, migration and invasion *in vitro* and *in vivo*. CircLMTK2 absorbs miR-150-5p and then indirectly regulates the expression of c-Myc to promote gastric cancer carcinogenesis (Wang et al., 2019a). *In vitro* and *in vivo* studies indicate that circAGO2 enhances the development, invasion, and dissemination of gastric cancer cells. Mechanistic studies demonstrate that circAGO2 physically interacts with the human antigen R (HuR) protein to assist the HuR-repressed actions of AGO2-miRNA complexes that promote cancer progression (Chen et al., 2019b). The silencing of circDONSON substantially inhibited GC cell proliferation, migration, and invasion, while promoting apoptosis. Functionally, circDONSON recruits the NURF complex to the promoter of SOX4 and initiates its transcription to facilitate gastric cancer growth and metastasis (Ding et al., 2019). CircPDSS1 enhanced GC cell cycle and reduced apoptosis by preventing miR-186-5p from targeting NEK2 to promote apoptosis. Therefore, circPDSS1 may serve as a biomarker and therapeutic target for the treatment of GC (Ouyang et al., 2019).

GSPT1-238aa, a novel protein encoded by circGSPT1, was discovered as a selective translation driven by IRES. GSPT1-238aa modulates autophagy can interact with vimentin/Beclin1/14-3-3 complex *via* the PI3K/AKT/mTOR signaling pathway in GC cells (Hu et al., 2022). What's more, AXIN1-295aa as a novel protein encoded by circAXIN1, it functions as an oncogenic protein to promote GC tumorigenesis and progression by activating the Wnt signaling pathway, suggesting a potential therapeutic target for GC (Peng et al., 2021).

Another study revealed that circST3GAL6 controlled apoptosis and autophagy *via* FOXP2-mediated transcriptional regulation of the MET axis *via* the miR-300/FOXP2 axis, which may represent a viable GC treatment target (Xu et al., 2022b). Ebv-circRPMS1 binds to Sam68 to promote its physical contact with the METTL3 promotor, resulting in transactivation of METTL3 and development of cancer (Zhang et al., 2022b). Ebv-circLMP2A interacted with KHSRP to increase the KHSRP-mediated degradation of VHL mRNA, resulting in an accumulation of HIF1 under hypoxia, which was crucial in controlling tumor angiogenesis in EBVaGC and might be a good therapeutic target for EBVaGC (Du et al., 2022). Circ-TNPO3 can bind competitively with IGF2BP3 and reduce IGF2BP3's capacity to stabilize MYC mRNA, ultimately inhibiting the proliferation and metastasis of GC (Yu et al., 2021).

Most of the circRNAs were located in the cytoplasm, However, circGSK3B was mainly identified in the nucleus. CircGSK3B is able to interact directly with EZH2, inhibiting the binding of EZH2 and H3K27me3 to the RORA promoter (Ma et al., 2021b).

3.3 Biological functions of circular RNAs in gastric cancer

3.3.1 The functions of circular RNAs in gastric cancer: Based on *in vivo* evidence

Due to the large number of circRNAs studied, a large number of circRNAs are reported every year for the role in gastric cancer. Nevertheless, most of these studies are based on the data from *in vitro* cell culture. *In vivo* investigation provides much more In-depth perspectives for these circRNAs. In particular, we summarize here relevant studies of circRNAs with relatively well-established functional studies *in vivo* in gastric cancer to help us understand which circRNAs functions have received focused attention and a more comprehensive understanding.

Numerous factors contribute to the biological makeup of GC. A recent study demonstrated that HOTAIR upregulation was associated with shorter overall survival in patients with gastric cancer, as well as advanced pathological stage, larger tumor size, and extensive metastasis. In addition, HOTAIR overexpression promoted the progression of gastric carcinoma *in vitro* and *in vivo via* regulating HER2 expression as a ceRNA of miR-331-3p (Liu et al., 2014). Furthermore, circAKT3 (hsa_circ_0000096) was significantly downregulated in gastric cancer tissues relative to nearby nontumorous tissues and normal gastric epithelial cells ($p < 0.001$). CircAKT3 might stimulate PIK3R1 expression *via* sponging miR-198, thereby increasing DNA damage repair and preventing apoptosis *in vivo* and *in vitro* in GC cells (Huang et al., 2019). The knockdown of hsa_circ_0000096 markedly decreased cell proliferation and migration *in vivo* (Li et al., 2017a). In GC tissues and cells, the amount of circCUL2, which is stable and restricted to the cytoplasm, was drastically decreased. Overexpression of circCUL2 decreased *in vivo* tumorigenicity (Peng et al., 2020).

3.3.2 The role of circular RNAs in gastric cancer *in vitro*

Silencing circRBMS3 decreased GC cell proliferation and invasion through sponging miR-153 *in vitro* (Li et al., 2019). Loss- and gain-of-function experiments indicate that circNF1 greatly increases GC cell proliferation (Wang et al., 2019c). Moreover, circCACTIN could function as a sponge of miRNA-331-3p and modulate the mRNA expression of TGFBR1. Knockdown of circCACTIN reduced the capability of cells proliferation, migration and invasion in GC cells (Zhang et al., 2019c).

4 The clinical values of circular RNAs in gastric cancer

As alluded to earlier, many circRNAs do not have high sensitivity on their own, however, the combination of these circRNAs with other tumor markers or circRNAs can

dramatically improve the sensitivity and specificity in early gastric cancer screening. For example, the combined detection of hsa_circ_0001017 and hsa_circ_0061276 in gastric cancer tissues and patient plasma has a high diagnostic value, with AUC as high as 0.966, and sensitivity and specificity of 95.5% and 95.7%, respectively (Li et al., 2018). The independent AUCs of hsa_circ_0000096 and hsa_circ_002059, both downregulated in gastric cancer tissue, were 0.82 and 0.73, but the AUC increased to 0.91 when these two circRNAs were used in combination (Li et al., 2017a). In addition, hsa_circ_0000745 was reduced in gastric cancer patients' plasma with an AUC of merely 0.683, but when hsa_circ_0000745 was combined with CEA, the AUC rose dramatically to 0.775 (Huang et al., 2017). These studies revealed that combining multiple circRNAs or with traditional diagnostic markers such as CEA, and CA19-9 can increase the sensitivity, specificity and accuracy of circRNAs based on gastric cancer diagnosis and prognosis. It provides new perspectives for developing circRNAs as diagnostic markers for early gastric cancer screening.

Although early-stage gastric cancer is highly curable through surgery, the majority of patients are diagnosed at an advanced stage, and therefore missed the window of opportunity for surgery. The high incidence and high mortality of gastric cancer urgently call for an early screening program. Detecting biomarkers in bodily fluids is a patient-friendly approach as bodily fluids are easy to obtain with likely higher patient compliance than endoscopy. The aforementioned studies demonstrated the potential of circRNAs as diagnostic markers for gastric cancer. Firstly, circRNAs have advantages compared with linear RNAs such as stable expression and a high degree of conservation. Secondly, an array of circRNAs is closely associated with gastric cancer development and risk factors, and their abnormal expression can signal tumor development and thus be used as screening biomarkers. Thirdly, in addition to their presence in tissues, circRNAs can also be sampled non-invasively in plasma and other bodily fluids. Lastly, with the development of technology, the detection of circRNAs will become more sensitive and cost-effective.

4.1 Diagnostic biomarkers

CA72-4 is currently the standard biomarker for early diagnosis of GC; however, its sensitivity and specificity are not optimal. CircRNAs exhibit distinct expression patterns in the tumor tissues and blood of patients with GC versus those of healthy controls. Thus, they are regarded as promising biomarkers for tissue or liquid biopsies in the diagnosis of GC.

Xie et al. (2018) found that the expression of hsa_circ_0074362 was downregulated in both gastritis and gastric cancer tissues. Due to the association between gastritis and a high risk of gastric cancer, hsa_circ_0074362 was proposed to be an early indication of gastric cancer. Albeit circ_0074362 was not able to function as an independent diagnostic marker of gastric cancer since it has a relatively low sensitivity (0.362), the level of hsa_circ_0074362 was associated with the serum tumor biomarker CA19-9 and lymph node metastasis. Therefore, by combining with other clinical markers, hsa_circ_0074362 may still hold the potential for gastric cancer screening. Another circRNA with a potential diagnostic value is hsa_circ_0001649, which is down-regulated in GC tissues. The ROC curve showed a sensitivity and specificity of 71.1% and 81.6% respectively, and the AUC was 0.834. These findings suggested that hsa_circ_0001649 could be used as a highly accurate, specific, and sensitive biomarker for gastric cancer (Li et al., 2017b).

Perhaps more interesting was the finding that the expression level of exosomal hsa_circ_0015286 decreased significantly in GC patients following surgery, suggesting that exosomal hsa_circ_0015286 may be a promising noninvasive biomarker for the diagnosis and prognostic evaluation of GC (Zheng et al., 2022). Likewise, in another study, a panel of 8 circRNAs as non-invasive, liquid-biopsy biomarkers that could serve as possible diagnostic biomarkers for the early diagnosis of GC were identified (Roy et al., 2022).

4.2 Prognostic biomarkers

Secondary prevention, including early identification, early diagnosis, and early treatment, can improve GC patients' prognosis. CircRNAs have been increasingly recognized as potential biomarkers for prognosis. CircLARP4, for example, was mostly located in the cytoplasm and regulated the biological behaviors of GC cells by sponging miR-424. Meanwhile, the decreased expression of circLARP4 in GC tissues was an independent predictive factor for the overall survival of GC patients (Zhang et al., 2017a) (Figure 3A). Further *in vivo* investigations verified that the combination treatment of circUBE2Q2 knockdown and STAT3 inhibitor had synergistic effects on the inhibition of gastric cancer growth, suggesting that targeting circUBE2Q2 may increase the sensitivity of targeted therapies to gastric cancer (Yang et al., 2021b) (Figure 3B). In addition, the differential expression of serum hsa_circ_0007507 among GC, post-operative GC, gastritis, intestinal metaplasia and relapsed patients, suggests it would be useful as a new diagnostic and dynamic monitoring biomarker for GC (Zhang et al., 2021) (Figure 3C). More specifically, the study of Li et al. (2017a) showed that hsa_circ_0000096 was significantly downregulated in gastric cancer tissues ($p < 0.001$) and the AUC was as high as 0.82, indicating high diagnostic accuracy.

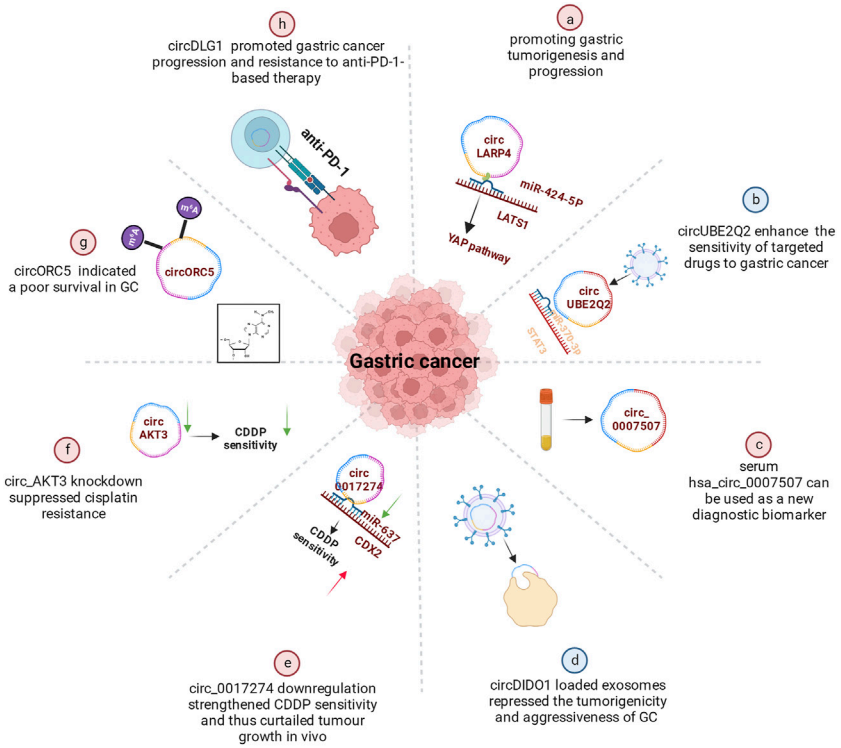


FIGURE 3
Functional roles of circRNAs in GC. **(A)** CircRNAs promoting gastric tumorigenesis and progression (e.g., circLARP4). A few circRNAs also regulated CDDP resistance **(B)**, cisplatin resistance **(C)** or promoted gastric cancer progression and resistance to anti-PD-1-based therapy **(D)**. **(E)** Some circRNAs indicated a poor survival in GC (e.g., circORC5). **(F)** Serum hsa_circ_0007507 can be used as a new diagnostic biomarker. **(G)** Certain circRNAs could enhance the sensitivity of targeted drugs to gastric cancer (e.g., circUBE2Q2). **(H)** circDIDO1 could repressed the tumorigenicity.

4.3 Therapeutic targets

A bunch of researches have revealed the relevance of circRNAs in GC and their link with GC carcinogenesis and development, and indicated that circRNAs have the potential to act as therapeutic targets in GC.

For instance, the overexpression of circAKT3 in GC patients undergoing cisplatin (CDDP) therapy was substantially linked with aggressive features and constituted an independent risk factor for disease-free survival (DFS). circAKT3 was expressed at a higher level in CDDP-resistant GC tissues and cells than in CDDP-sensitive samples. Clinicopathological characteristics demonstrated that the level of hsa_circ_0000520 in GC tissues was adversely correlated with TNM stage and that the amount of CEA expression in GC plasma was correlated with TNM stage (Sun et al., 2018).

Another finding showed that circDIDO1 inhibited the advancement of GC through regulating the miR-1307-3p/SOSC2 axis (Figure 3D). Systemic injection of RGD-modified, circDIDO1-loaded exosomes inhibited the tumorigenicity and aggressiveness of GC *in vitro* and *in vivo*, indicating that RGD-Exo-circDIDO1 could be employed as a nanomedicine for the treatment of GC (Guo et al., 2022b).

Hu et al. (2022) confirmed that GSPT1-238aa, a new protein encoded by circGSPT1, inhibits the development of GC tumors. They also shed light on the function and molecular mechanisms behind GSPT1-238aa in GC and suggest that this protein constitutes a unique therapeutic target for GC. Moreover, another circRNA, Circ-MTO1, correlates with less lymph node metastasis, prolonged DFS, and improved chemotherapy sensitivity in gastric cancer (Chang et al., 2022).

4.4 Drug resistance

Currently, circular RNAs in significant numbers are now linked to the emergence of treatment resistance and the onset of cancers. By regulating the miR-383-5p/FGF7 axis, knockdown of circLRCH3 reduced GC OXA resistance, providing a prospective therapeutic target for GC chemoresistance (Xiang et al., 2022). As shown in a study by Xu et al. (2022a), circ0017274 was upregulated in GC tissues and cells resistant to CDDP, while

miR-637 was lower (Figure 3E). Reducing the abundance of circ_0017274 not only alleviated CDDP resistance but also induced cell cycle arrest in GC cells. The xenograft models further demonstrated that circ0017274 downregulation increased CDDP sensitivity and consequently inhibited *in vivo* tumor growth. By acting on miR-637/CDX2 in CDDP-resistant GC cells, circ0017274 downregulation improved CDDP sensitivity.

It was reported that ICA decreased GC cell survival and induced pyroptosis by modulating the hsa_circ_0003159/miR-223-3p/NLRP3 axis both *in vitro* and *in vivo*. ICA suppresses the proliferation of GC cells *via* modulating the hsa_circ_0003159/miR-223-3p/NLRP3 signalling pathway (Zhang et al., 2022a). Circ_AKT3 knockdown decreased cisplatin resistance in cisplatin-resistant GC cells *via* the miR-206/PTPN14 axis (Shi and Wang 2022) (Figure 3F). Furthermore, the METTL14-mediated m6A alteration of circORC5 inhibits the progression of gastric cancer by modulating the miR-30c-2-3p/AKT1S1 axis (Fan et al., 2022) (Figure 3G). In addition, through targeting PRKAA2, circCPM plays a vital role in regulating GC autophagy and 5-FU resistance. It could serve as a novel theoretical foundation for evaluating the therapeutic efficacy of GC and reversing 5-FU chemoresistance (Fang et al., 2022). CircDLG1 was highly increased in distant metastatic lesions and anti-PD-1-resistant gastric cancer tissues, and was linked with an aggressive tumor phenotype and poor prognosis in gastric cancer patients treated with anti-PD-1 drugs (Chen et al., 2021) (Figure 3H).

On the basis of this mounting evidence, circRNAs play increasingly crucial roles in the regulation of drug development.

5 Perspectives

5.1 Insights and limitations of current research

The clinical application of circRNAs has broad prospects, but it still faces many difficulties. First of all, the cost of circRNA testing is still higher than that of existing gastroscopy testing, which limits its application at the population-scale for early gastric cancer screening. Secondly, the research on circRNAs is still in its infancy, and the diagnostic accuracy and consistency are less than optimal. Current research shows that the sensitivity, specificity and diagnostic accuracy of different circRNAs are highly variable, and gastroscopic pathological biopsy is still the "gold standard" for clinical diagnosis of gastric cancer. Therefore, it still needs more research to screen out the most efficient circRNA candidates and study their values in combination with traditional tumor markers to achieve the best diagnostic results. Furthermore, current research on circRNAs focuses on tissue and blood samples as the source, and the research on circRNAs in other types of bodily fluids is scarce. Since circRNAs are abundant and stably expressed in other types of bodily fluids (Pardini et al., 2019), they should be explored further in the future. Overall, there is still a long way to go until we can establish circRNAs as non-invasive tumor markers in the clinical settings.

5.2 Innovative suggestions for future research

In recent years, a growing number of studies have uncovered fundamental aspects of circRNAs and produced many surprising results indicating that circRNAs are important in biology and pathobiology; consequently, circRNA-related research is advancing at a constant and rapid rate. Nonetheless, a global and exhaustive understanding of circRNAs associated with GC early detection is still lacking.

In the section that follows, we propose a number of innovative and challenging directions in the field of circRNAs in future. First, the majority of current investigation is still carried out in cells and animals, how to facilitate translation toward clinical application would be a hot topic in the future. Second, circRNAs as a stable and can be widely detected in many types of body fluids, it is urged to confirm their potential as novel drugs, therapeutic targets, or biomarkers. In addition, Whether or not a circRNA with a low abundance can achieve measurable effects remains debatable. Unlike most current studies that explore the effect of a single circRNA on a specific physiological process, the investigation in future should focus on a group of circRNAs with similar functions that affect the physiological processes.

6 Summary and outlook

As a newly discovered type of RNA molecule, circRNAs have important biological functions and the potential to become an early biomarker for gastric cancer screening. The volume of research on this topic has been steadily growing in the past several years. Multiple studies have revealed the potential of circRNAs as biomarkers for gastric cancer as they are highly conservative and differentially expressed in gastric cancer patients, while the current research is still limited in scope. The interactions between circRNAs, miRNAs and RBPs and the mechanisms underlying their functions in gastric cancer are not yet fully understood. However, with these mechanistic questions being studied and answered, circRNAs will likely to become a novel marker in the early screening of gastric cancer to improve the survival rate of patients.

Author contributions

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

Funding

This work was supported by National Natural Science Foundation of China (81502304), Zhejiang Provincial Natural Science Foundation (LGF22G010009), Medical and Health Technology Projects of Zhejiang Province, China (2017KY696, 2019PY089), Quzhou technology projects, China (2022K46, 2018K20).

Conflict of interest

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

Publisher's note

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

References

Arnaiz, E., Sole, C., Manterola, L., Iparraguirre, L., Otaegui, D., and Cjsicb, Lawrie (2019). CircRNAs and cancer: Biomarkers and master regulators. *Semin. Cancer Biol.* 58, 90–99. doi:10.1016/j.semcancer.2018.12.002

Ashwal-Fluss, R., Meyer, M., Pamudurti, N., Ivanov, A., Bartok, O., Hanan, M., et al. (2014). circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* 56, 55–66. doi:10.1016/j.molcel.2014.08.019

Batth, I., Mitra, A., Manier, S., Ghobrial, I., Menter, D., Kopetz, S., et al. (2017). Circulating tumor markers: Harmonizing the yin and yang of CTCs and ctDNA for precision medicine. *Ann. Oncol.* 28, 468–477. doi:10.1093/annonc/mdw619

Bian, W., Liu, Z., Chu, Y., and Xing, X. J. A-cd (2021). Silencing of circ_0078607 prevents development of gastric cancer and inactivates the ERK1/2/AKT pathway through the miR-188-3p/RAP1B axis. *Anticancer. Drugs* 32, 909–918. doi:10.1097/CAD.0000000000001083

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L., and Jemal Ajcacjfc (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

Cao, J., Zhang, X., Xu, P., Wang, H., Wang, S., Zhang, L., et al. (2021). Circular RNA circLMO7 acts as a microRNA-30a-3p sponge to promote gastric cancer progression via the WNT2/β-catenin pathway. *J. Exp. Clin. Cancer Res.* 40, 6. doi:10.1186/s13046-020-01791-9

Chang, C., Zheng, A., Wang, P., and Teng, X. J. Jocla (2022). Circular RNA mitochondrial translation optimization 1 correlates with less lymph node metastasis, longer disease-free survival, and higher chemotherapy sensitivity in gastric cancer. *J. Clin. Lab. Anal.* 36, e23918. doi:10.1002/jcla.23918

Chen, D., Sheng, H., Zhang, D., Jin, Y., Zhao, B., Chen, N., et al. (2021). The circular RNA circDLG1 promotes gastric cancer progression and anti-PD-1 resistance through the regulation of CXCL12 by sponging miR-141-3p. *Mol. Cancer* 20, 166. doi:10.1186/s12943-021-01475-8

Chen, J., Li, Y., Zheng, Q., Bao, C., He, J., Chen, B., et al. (2017). Circular RNA profile identifies circPVT1 as a proliferative factor and prognostic marker in gastric cancer. *Cancer Lett.* 388, 208–219. doi:10.1016/j.canlet.2016.12.006

Chen, L., Wang, L., and Ma, X. J. B. (2019a). Communications brCirc_SPECC1 enhances the inhibition of miR-526b on downstream KDM4A/YAP1 pathway to regulate the growth and invasion of gastric cancer cells. *Biochem. Biophys. Res. Commun.* 517, 253–259. doi:10.1016/j.bbrc.2019.07.065

Chen, Y., Yang, F., Fang, E., Xiao, W., Mei, H., Li, H., et al. (2019b). Circular RNA circAGO2 drives cancer progression through facilitating HuR-repressed functions of AGO2-miRNA complexes. *Cell Death Differ.* 26, 1346–1364. doi:10.1038/s41418-018-0220-6

de Fraipont, F., Gazzeri, S., Cho, W., and Bjfig, Eymin (2019). Circular RNAs and RNA splice variants as biomarkers for prognosis and therapeutic response in the liquid biopsies of lung cancer patients. *Front. Genet.* 10, 390. doi:10.3389/fgene.2019.00390

Deng, G., Mou, T., He, J., Chen, D., Lv, D., Liu, H., et al. (2020). Circular RNA circRHOBTB3 acts as a sponge for miR-654-3p inhibiting gastric cancer growth. *J. Exp. Clin. Cancer Res.* 39, 1. doi:10.1186/s13046-019-1487-2

Ding, L., Zhao, Y., Dang, S., Wang, Y., Li, X., Yu, X., et al. (2019). Circular RNA circ-DONSON facilitates gastric cancer growth and invasion via NURF complex dependent activation of transcription factor SOX4. *Mol. Cancer* 18, 45. doi:10.1186/s12943-019-1006-2

Du, W., Li, D., Guo, X., Li, P., Li, X., Tong, S., et al. (2019). Circ-PRMT5 promotes gastric cancer progression by sponging miR-145 and miR-1304 to upregulate MYC. *Artif. Cells Nanomed. Biotechnol.* 47, 4120–4130. doi:10.1080/21691401.2019.1671857

Du, W., Zhang, C., Yang, W., Yong, T., Awan, F., and Yang, B. J. T. (2017). Identifying and characterizing circRNA-protein interaction. *Theranostics* 7, 4183–4191. doi:10.7150/thno.21299

Du, Y., Zhang, J. Y., Gong, L. P., Feng, Z. Y., Wang, D., Pan, Y. H., et al. (2022). Hypoxia-induced ebv-circLMP2A promotes angiogenesis in EBV-associated gastric carcinoma through the KHSRP/VHL/HIF1α/VEGFA pathway. *Cancer Lett.* 526, 259–272. doi:10.1016/j.canlet.2021.11.031

Erratum (2020). Erratum: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 70, 313. doi:10.3322/caac.21609

Fan, H., Chen, Z., Chen, X., Chen, M., Yi, Y., Zhu, J., et al. (2022). METTL14-mediated m^6A modification of circORC5 suppresses gastric cancer progression by regulating miR-30c-2-3p/AKT1S1 axis. *Mol. Cancer* 21, 51. doi:10.1186/s12943-022-01521-z

Fang, J., Hong, H., Xue, X., Zhu, X., Jiang, L., Qin, M., et al. (2019). A novel circular RNA, circFAT1(e2), inhibits gastric cancer progression by targeting miR-548g in the cytoplasm and interacting with YBX1 in the nucleus. *Cancer Lett.* 442, 222–232. doi:10.1016/j.canlet.2018.10.040

Fang, L., Lv, J., Xuan, Z., Li, B., Li, Z., He, Z., et al. (2022). Circular CPM promotes chemoresistance of gastric cancer via activating PRKAA2-mediated autophagy. *Clin. Transl. Med.* 12, e708. doi:10.1002/ctm2.708

Feng, W., Gong, H., Wang, Y., Zhu, G., Xue, T., Wang, Y., et al. (2019). circIFT80 functions as a ceRNA of miR-1236-3p to promote colorectal cancer progression. *Mol. Ther. Nucleic Acids* 18, 375–387. doi:10.1016/j.omtn.2019.08.024

Guo, R., Cui, X., Li, X., Zang, W., Chang, M., Sun, Z., et al. (2022a). CircMAN1A2 is upregulated by *Helicobacter pylori* and promotes development of gastric cancer. *Cell Death Dis.* 13, 409. doi:10.1038/s41419-022-04811-y

Guo, X., Dai, X., Liu, J., Cheng, A., Qin, C., and Wang, Z. J. MtNa (2020). Circular RNA circREPS2 acts as a sponge of miR-558 to suppress gastric cancer progression by regulating RUNX3/β-catenin signaling. *Mol. Ther. Nucleic Acids* 21, 577–591. doi:10.1016/j.omtn.2020.06.026

Guo, Z., Zhang, Y., Xu, W., Zhang, X., and Jiang, J. J. Jotm (2022b). Engineered exosome-mediated delivery of circDIDO1 inhibits gastric cancer progression via regulation of MiR-1307-3p/SOCS2 Axis. *J. Transl. Med.* 20, 326. doi:10.1186/s12967-022-03527-z

Hansen, T., Jensen, T., Clausen, B., Bramsen, J., Finsen, B., Damgaard, C., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi:10.1038/nature11993

Hon, K., Ab-Mutalib, N., Abdullah, N., Jamal, R., and Abu, N. J. Sr (2019). Extracellular Vesicle-derived circular RNAs confers chemoresistance in Colorectal cancer. *Sci. Rep.* 9, 16497. doi:10.1038/s41598-019-53063-y

Hong, Y., Qin, H., Li, Y., Zhang, Y., Zhuang, X., Liu, L., et al. (2019). FNDC3B circular RNA promotes the migration and invasion of gastric cancer cells via the regulation of E-cadherin and CD44 expression. *J. Cell. Physiol.* 234, 19895–19910. doi:10.1002/jcp.28588

Hou, J., Men, X., Yang, L., Han, E., Han, C., and Liu, L. J. Bmj (2021). CircCCT3 acts as a sponge of miR-613 to promote tumor growth of pancreatic cancer through regulating VEGFA/VEGFR2 signaling. *Balk. Med. J.* 38, 229–238. doi:10.5152/balkanmedj.2021.21145

Hu, F., Peng, Y., Chang, S., Luo, X., Yuan, Y., Zhu, X., et al. (2022). Vimentin binds to a novel tumor suppressor protein, GSPT1-238aa, encoded by circGSPT1 with a selective encoding priority to halt autophagy in gastric carcinoma. *Cancer Lett.* 545, 215826. doi:10.1016/j.canlet.2022.215826

Huang, M., He, Y., Liang, L., Huang, Q., and Zhu, Z. J. Wjog (2017). Circular RNA hsa_circ_0000745 may serve as a diagnostic marker for gastric cancer. *World J. Gastroenterol.* 23, 6330–6338. doi:10.3748/wjg.v23.i34.6330

Huang, X., Li, Z., Zhang, Q., Wang, W., Li, B., Wang, L., et al. (2019). Circular RNA AKT3 upregulates PIK3R1 to enhance cisplatin resistance in gastric cancer via miR-198 suppression. *Mol. Cancer* 18, 71. doi:10.1186/s12943-019-0969-3

Huang, Y., Zheng, W., Ji, C., Wang, X., Yu, Y., Deng, X., et al. (2021). Circular RNA circRPPH1 promotes breast cancer progression via circRPPH1-miR-512-5p-STAT1 axis. *Cell Death Discov.* 7, 376. doi:10.1038/s41420-021-00771-y

Jiang, T., Xia, Y., Lv, J., Li, B., Li, Y., Wang, S., et al. (2021). A novel protein encoded by circMAPK1 inhibits progression of gastric cancer by suppressing activation of MAPK signaling. *Mol. Cancer* 20, 66. doi:10.1186/s12943-021-01358-y

Jie, M., Wu, Y., Gao, M., Li, X., Liu, C., Ouyang, Q., et al. (2020). CircMRPS35 suppresses gastric cancer progression via recruiting KAT7 to govern histone modification. *Mol. Cancer* 19, 56. doi:10.1186/s12943-020-01160-2

Kong, S., Yang, Q., Tang, C., Wang, T., Shen, X., and Ju, S. J. Fig (2019). Identification of hsa_circ_0001821 as a novel diagnostic biomarker in gastric cancer via comprehensive circular RNA profiling. *Front. Genet.* 10, 878. doi:10.3389/fgene.2019.00878

Kristensen, L., Andersen, M., Stagsted, L., Ebbesen, K., Hansen, T., and Kjems, J. J. NrG. (2019). The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* 20, 675–691. doi:10.1038/s41576-019-0158-7

Li, G., Xue, M., Yang, F., Jin, Y., Fan, Y., and Li, W. J. Jocp (2019). CircRBMS3 promotes gastric cancer tumorigenesis by regulating miR-153-SNAI1 axis. *J. Cell. Physiol.* 234, 3020–3028. doi:10.1002/jcp.27122

Li, J., Zhang, G., Wu, G. J. C., and biology, m. (2020). Effect of paeonol on proliferation, apoptosis, migration, invasion and glutamine of gastric cancer cells via circSFMBT2/miR-665 axis. *Cell. Mol. Biol.* 66, 33–40. doi:10.14715/cmb/2020.66.8.6

Li, P., Chen, H., Chen, S., Mo, X., Li, T., Xiao, B., et al. (2017a). Circular RNA 0000096 affects cell growth and migration in gastric cancer. *Br. J. Cancer* 116, 626–633. doi:10.1038/bjc.2016.451

Li, T., Shao, Y., Fu, L., Xie, Y., Zhu, L., Sun, W., et al. (2018). Plasma circular RNA profiling of patients with gastric cancer and their droplet digital RT-PCR detection. *J. Mol. Med.* 96, 85–96. doi:10.1007/s00109-017-1600-y

Li, W., Song, Y., Zhang, H., Zhou, Z., Xie, X., Zeng, Q., et al. (2017b). Decreased expression of Hsa_circ_00001649 in gastric cancer and its clinical significance, *Dis Markers*. 2017. 4587698. doi:10.1155/2017/4587698

Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015a). Circular RNA is enriched and stable in exosomes: A promising biomarker for cancer diagnosis. *Cell Res.* 25, 981–984. doi:10.1038/cr.2015.82

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., et al. (2015b). Exon-intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* 22, 256–264. doi:10.1038/nsmb.2959

Liu, H., Liu, Y., Bian, Z., Zhang, J., Zhang, R., Chen, X., et al. (2018). Circular RNA YAP1 inhibits the proliferation and invasion of gastric cancer cells by regulating the miR-367-5p/p27 Kip1 axis. *Mol. Cancer* 17, 151. doi:10.1186/s12943-018-0902-1

Liu, J., Song, S., Lin, S., Zhang, M., Du, Y., Zhang, D., et al. (2019). Circ-SERPINE2 promotes the development of gastric carcinoma by sponging miR-375 and modulating YWHAZ. *Cell Prolif.* 52, e12648. doi:10.1111/cpr.12648

Liu, X., Sun, M., Nie, F. Q., Ge, Y. B., Zhang, E. B., Yin, D. D., et al. (2014). Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Mol. Cancer* 13, 92. doi:10.1186/1476-4598-13-92

Lu, Y., Li, L., Li, L., Wu, G., and Liu, G. J. Cci (2021). Circular RNA circHECTD1 prevents Diosbulbin-B-sensitivity via miR-137/PBX3 axis in gastric cancer. *Cancer Cell Int.* 21, 264. doi:10.1186/s12935-021-01957-1

Luo, Z., Rong, Z., Zhang, J., Zhu, Z., Yu, Z., Li, T., et al. (2020). Circular RNA circCCDC9 acts as a miR-6792-3p sponge to suppress the progression of gastric cancer through regulating CAV1 expression. *Mol. Cancer* 19, 86. doi:10.1186/s12943-020-01203-8

Ma, S., Gu, X., Shen, L., Chen, Y., Qian, C., Shen, X., et al. (2021a). CircHAS2 promotes the proliferation, migration, and invasion of gastric cancer cells by regulating PPM1E mediated by hsa-miR-944. *Cell Death Dis.* 12, 863. doi:10.1038/s41419-021-04158-w

Ma, X., Chen, H., Li, L., Yang, F., Wu, C., Tao, K. J. Joe, et al. (2021b). CircGSK3B promotes RORA expression and suppresses gastric cancer progression through the prevention of EZH2 trans-inhibition. *J. Exp. Clin. Cancer Res.* 40, 330. doi:10.1186/s13046-021-02136-w

Mathieu, M., Martin-Jaular, L., Lavieu, G., and Cjncb, Théry (2019). Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication. *Nat. Cell Biol.* 21, 9–17. doi:10.1038/s41556-018-0250-9

Ouyang, Y., Li, Y., Huang, Y., Li, X., Zhu, Y., Long, Y., et al. (2019). CircRNA circPDSS1 promotes the gastric cancer progression by sponging miR-186-5p and modulating NEK2. *J. Cell. Physiol.* 234, 10458–10469. doi:10.1002/jcp.27714

Pan, H., Li, T., Jiang, Y., Pan, C., Ding, Y., Huang, Z., et al. (2018a). Overexpression of circular RNA ciRS-7 abrogates the tumor suppressive effect of miR-7 on gastric cancer via PTEN/PI3K/AKT signaling pathway. *J. Cell. Biochem.* 119, 440–446. doi:10.1002/jcb.26201

Pan, J., Meng, X., Jiang, N., Jin, X., Zhou, C., Xu, D., et al. (2018b). Insights into the noncoding RNA-encoded peptides. *Protein Pept. Lett.* 25, 720–727. doi:10.2174/0929866525666180809142326

Pardini, B., Sabo, A., Birolo, G., and Calin, G. J. C. (2019). Noncoding RNAs in extracellular fluids as cancer biomarkers: The new frontier of liquid biopsies.*Cancers Basel*, 11. 1170. doi:10.3390/cancers11081170

Peng, L., Sang, H., Wei, S., Li, Y., Jin, D., Zhu, X., et al. (2020). circCUL2 regulates gastric cancer malignant transformation and cisplatin resistance by modulating autophagy activation via miR-142-3p/ROCK2. *Mol. Cancer* 19, 156. doi:10.1186/s12943-020-01270-x

Peng, Y., Xu, Y., Zhang, X., Deng, S., Yuan, Y., Luo, X., et al. (2021). A novel protein AXIN1-295aa encoded by circAXIN1 activates the Wnt/β-catenin signaling pathway to promote gastric cancer progression. *Mol. Cancer* 20, 158. doi:10.1186/s12943-021-01457-w

Quan, J., Dong, D., Lun, Y., Sun, B., Sun, H., Wang, Q., et al. (2020). Circular RNA circHIAT1 inhibits proliferation and epithelial-mesenchymal transition of gastric cancer cell lines through downregulation of miR-21. *J. Biochem. Mol. Toxicol.* 34, e22458. doi:10.1002/jbt.22458

Reimers, N., Kjcc, Pantel, and medicine, l. (2019). Liquid biopsy: Novel technologies and clinical applications. *Clin. Chem. Lab. Med.* 57, 312–316. doi:10.1515/cclm-2018-0610

Reis-das-Mercês, L., Vinasco-Sandoval, T., Pompeu, R., Ramos, A., Anaissi, A., Demachki, S., et al. (2022). CircRNAs as potential blood biomarkers and key elements in regulatory networks in gastric cancer, *Int. J. Mol. Sci.* 23. 650. doi:10.3390/ijms23020650

Rong, D., Lu, C., Zhang, B., Fu, K., Zhao, S., Tang, W., et al. (2019). CircPSMC3 suppresses the proliferation and metastasis of gastric cancer by acting as a competitive endogenous RNA through sponging miR-296-5p. *Mol. Cancer* 18, 25. doi:10.1186/s12943-019-0958-6

Roy, S., Kanda, M., Nomura, S., Zhu, Z., Toiyama, Y., Taketomi, A., et al. (2022). Diagnostic efficacy of circular RNAs as noninvasive, liquid biopsy biomarkers for early detection of gastric cancer. *Mol. Cancer* 21, 42. doi:10.1186/s12943-022-01527-7

Sang, H., Zhang, W., Peng, L., Wei, S., Zhu, X., Huang, K., et al. (2022). Exosomal circRELL1 serves as a miR-637 sponge to modulate gastric cancer progression via regulating autophagy activation. *Cell Death Dis.* 13, 56. doi:10.1038/s41419-021-04364-6

Sekiguchi, M., and Matsuda, T. J. Sr (2020). Limited usefulness of serum carcinoembryonic antigen and carbohydrate antigen 19-9 levels for gastrointestinal and whole-body cancer screening, *Sci. Rep.* 10, 18202. doi:10.1038/s41598-020-75319-8

Shao, Y., Li, J., Lu, R., Li, T., Yang, Y., Xiao, B., et al. (2017). Global circular RNA expression profile of human gastric cancer and its clinical significance. *Cancer Med.* 6, 1173–1180. doi:10.1002/cam4.1055

Shen, L., Shan, Y. S., Hu, H. M., Price, T. J., Sirohi, B., Yeh, K. H., et al. (2013). Management of gastric cancer in Asia: Resource-stratified guidelines. *Lancet. Oncol.* 14, e535–e547. doi:10.1016/S1470-2045(13)70436-4

Shi, W., and Wang, F. J. Om (2022). circ_AKT3 knockdown suppresses cisplatin resistance in gastric cancer. *Open Med.* 17, 280–291. doi:10.1515/med-2021-0355

Sitarz, R., Skierucha, M., Mielko, J., Offerhaus, G., Maciejewski, R., Polkowski, W. J. Cm, et al. (2018). Gastric cancer: Epidemiology, prevention, classification, and treatment. *Cancer Manag. Res.* 10, 239–248. doi:10.2147/CMAR.S149619

Song, J., Xu, X., He, S., Wang, N., Bai, Y., Li, B., et al. (2022). Exosomal hsa_circ_0017252 attenuates the development of gastric cancer via inhibiting macrophage M2 polarization. *Hum Cell.* 35. 1499-1511. doi:10.1007/s13577-022-00739-9

Su, M., Xiao, Y., Ma, J., Tang, Y., Tian, B., Zhang, Y., et al. (2019). Circular RNAs in cancer: Emerging functions in hallmarks, stemness, resistance and roles as potential biomarkers. *Mol. Cancer* 18, 90. doi:10.1186/s12943-019-1002-6

Sun, G., Li, Z., He, Z., Wang, W., Wang, S., Zhang, X., et al. (2020). Circular RNA MCTP2 inhibits cisplatin resistance in gastric cancer by miR-99a-5p-mediated induction of MTMR3 expression. *J. Exp. Clin. Cancer Res.* 39, 246. doi:10.1186/s13046-020-01758-w

Sun, H., Tang, W., Rong, D., Jin, H., Fu, K., Zhang, W., et al. (2018). Hsa_circ_0000520, a potential new circular RNA biomarker, is involved in gastric carcinoma. *Cancer Biomark.* 21, 299–306. doi:10.3233/CBM-170379

Wang, L., Li, B., Yi, X., Xiao, X., Zheng, Q., and Ma, L. J. Cp (2021a). Circ_SMAD4 promotes gastric carcinogenesis by activating wnt/β-catenin pathway. *Cell Prolif.* 54, e12981. doi:10.1111/cpr.12981

Wang, S., Tang, D., Wang, W., Yang, Y., Wu, X., Wang, L., et al. (2019a). circLMTK2 acts as a sponge of miR-150-5p and promotes proliferation and metastasis in gastric cancer. *Mol. Cancer* 18, 162. doi:10.1186/s12943-019-1081-4

Wang, S., Zhang, X., Li, Z., Wang, W., Li, B., Huang, X., et al. (2019b). Circular RNA profile identifies circOSBPL10 as an oncogenic factor and prognostic marker in gastric cancer. *Oncogene* 38, 6985–7001. doi:10.1038/s41388-019-0933-0

Wang, X., Li, J., Bian, X., and Lin, W (2021b). CircURI1 interacts with hnRNPM to inhibit metastasis by modulating alternative splicing in gastric cancer, *Biol. Sci.* 118. e2012881118. doi:10.1073/pnas.2012881118

Wang, Z., Ma, K., Pitts, S., Cheng, Y., Liu, X., Ke, X., et al. (2019c). Novel circular RNA circNF1 acts as a molecular sponge, promoting gastric cancer by absorbing miR-16. *Endocr. Relat. Cancer* 26, 265–277. doi:10.1530/ERC-18-0478

Xiang, C., Li, R., Qiu, H., Zuo, E., Zhang, Y., Shan, L., et al. (2022). Circular RNA circLRCH3 promotes oxaliplatin resistance in gastric cancer through the modulation of the miR-383-5p/FGF7 axis. *Histol Histopathol.* 3, 18506. doi:10.14670/HH-18-506

Xie, M., Yu, T., Jing, X., Ma, L., Fan, Y., Yang, F., et al. (2020). Exosomal circSHKBP1 promotes gastric cancer progression via regulating the miR-582-3p/HUR/VEGF axis and suppressing HSP90 degradation. *Mol. Cancer* 19, 112. doi:10.1186/s12943-020-01208-3

Xie, Y., Shao, Y., Sun, W., Ye, G., Zhang, X., Xiao, B., et al. (2018). Downregulated expression of hsa_circ_0074362 in gastric cancer and its potential diagnostic values. *Biomark. Med.* 12, 11–20. doi:10.2217/bmm-2017-0114

Xu, B., Guo, J., and Chen, M. J. C. (2022a). Circ_0017274 acts on miR-637/CDX2 axis to facilitate cisplatin resistance in gastric cancer. *Pharmacology e, physiology.* 49. 1105-1115. doi:10.1111/1440-1681.13692

Xu, P., Zhang, X., Cao, J., Yang, J., Chen, Z., Wang, W., et al. (2022b). The novel role of circular RNA ST3GAL6 on blocking gastric cancer malignant behaviours through autophagy regulated by the FOXP2/MET/mTOR axis. *Clin. Transl. Med.* 12, e707. doi:10.1002/ctm2.707

Xu, R., Rai, A., Chen, M., Suwakulsiri, W., Greening, D., and Simpson, R. J. NrCo (2018). Extracellular vesicles in cancer - implications for future improvements in cancer care. *Nat. Rev. Clin. Oncol.* 15, 617–638. doi:10.1038/s41571-018-0036-9

Yang, D., Hu, Z., Zhang, Y., Zhang, X., Xu, J., Fu, H., et al. (2021a). CircHIPK3 promotes the tumorigenesis and development of gastric cancer through miR-637/AKT1 pathway. *Front. Oncol.* 11, 637761. doi:10.3389/fonc.2021.637761

Yang, F., Hu, A., Li, D., Wang, J., Guo, Y., Liu, Y., et al. (2019). Circ-HuR suppresses HuR expression and gastric cancer progression by inhibiting CNBP transactivation. *Mol. Cancer* 18, 158. doi:10.1186/s12943-019-1094-z

Yang, J., Zhang, X., Cao, J., Xu, P., Chen, Z., Wang, S., et al. (2021b). Circular RNA UBE2Q2 promotes malignant progression of gastric cancer by regulating signal transducer and activator of transcription 3-mediated autophagy and glycolysis. *Cell Death Dis.* 12, 910. doi:10.1038/s41419-021-04216-3

Yao, K. (2013). The endoscopic diagnosis of early gastric cancer. *Ann. Gastroenterol.* 26, 11–22.

Yu, T., Ran, L., Zhao, H., Yin, P., Li, W., Lin, J., et al. (2021). Circular RNA circ-TNPO3 suppresses metastasis of GC by acting as a protein decoy for IGF2BP3 to regulate the expression of MYC and SNAIL. *Mol. Ther. Nucleic Acids* 26, 649–664. doi:10.1016/j.omtn.2021.08.029

Zhang, F., Yin, Y., Xu, W., Song, Y., Zhou, Z., Sun, X., et al. (2022a). Icariin inhibits gastric cancer cell growth by regulating the hsa_circ_0003159/miR-223-3p/NLRP3 signaling axis. *Hum. Exp. Toxicol.* 41, 9603271221097363. doi:10.1177/09603271221097363

Zhang, H., Zhu, L., Bai, M., Liu, Y., Zhan, Y., Deng, T., et al. (2019a). Exosomal circRNA derived from gastric tumor promotes white adipose browning by targeting the miR-133/PRDM16 pathway. *Int. J. Cancer* 144, 2501–2515. doi:10.1002/ijc.31977

Zhang, J., Du, Y., Gong, L. P., Shao, Y. T., Pan, L. J., Feng, Z. Y., et al. (2022b). ebv-circRPMS1 promotes the progression of EBV-associated gastric carcinoma via Sam68-dependent activation of METTL3. *Cancer Lett.* 535, 215646. doi:10.1016/j.canlet.2022.215646

Zhang, J., Hou, L., Liang, R., Chen, X., Zhang, R., Chen, W., et al. (2019b). CircDLST promotes the tumorigenesis and metastasis of gastric cancer by sponging miR-502-5p and activating the NRAS/MEK1/ERK1/2 signaling. *Mol. Cancer* 18, 80. doi:10.1186/s12943-019-1015-1

Zhang, J., Liu, H., Hou, L., Wang, G., Zhang, R., Huang, Y., et al. (2017a). Circular RNA_LARP4 inhibits cell proliferation and invasion of gastric cancer by sponging miR-424-5p and regulating LATS1 expression. *Mol. Cancer* 16, 151. doi:10.1186/s12943-017-0719-3

Zhang, L., Song, X., Chen, X., Wang, Q., Zheng, X., Wu, C., et al. (2019c). Circular RNA CircCACTIN promotes gastric cancer progression by sponging MiR-331-3p and regulating TGFBR1 expression. *Int. J. Biol. Sci.* 15, 1091–1103. doi:10.7150/ijbs.31533

Zhang, W., Zheng, M., Kong, S., Li, X., Meng, S., Wang, X., et al. (2021). Circular RNA hsa_circ_0007507 may serve as a biomarker for the diagnosis and prognosis of gastric cancer. *Front. Oncol.* 11, 699625. doi:10.3389/fonc.2021.699625

Zhang, X., Wang, H., Zhang, Y., Lu, X., Chen, L., and Yang, L. J. C. (2014). Complementary sequence-mediated exon circularization. *Cell* 159, 134–147. doi:10.1016/j.cell.2014.09.001

Zhang, X., Wang, S., Wang, H., Cao, J., Huang, X., Chen, Z., et al. (2019d). Circular RNA circNRIP1 acts as a microRNA-149-5p sponge to promote gastric cancer progression via the AKT1/mTOR pathway. *Mol. Cancer* 18, 20. doi:10.1186/s12943-018-0935-5

Zhang, Y., Li, J., Yu, J., Liu, H., Shen, Z., Ye, G., et al. (2017b). Circular RNAs signature predicts the early recurrence of stage III gastric cancer after radical surgery. *Oncotarget* 8, 22936–22943. doi:10.18632/oncotarget.15288

Zhang, Y., Liu, H., Li, W., Yu, J., Li, J., Shen, Z., et al. (2017c). CircRNA_100269 is downregulated in gastric cancer and suppresses tumor cell growth by targeting miR-630. *Aging* 9, 1585–1594. doi:10.18632/aging.101254

Zhang, Z., Wang, C., Zhang, Y., Yu, S., Zhao, G., Xu, J. J. GcojotI. G. C. A., et al. (2020). CircDUSP16 promotes the tumorigenesis and invasion of gastric cancer by sponging miR-145-5p. *Gastric Cancer* 23, 437–448. doi:10.1007/s10120-019-01018-7

Zhao, C., Sun, J., Dang, Z., Su, Q., Yang, J. J. P., and research, practice (2022). Circ_0000775 promotes the migration, invasion and EMT of hepatic carcinoma cells by recruiting IGF2BP2 to stabilize CDC27. *Pathol. Res. Pract.* 235, 153908. doi:10.1016/j.prp.2022.153908

Zhao, Q., Chen, S., Li, T., Xiao, B., and Zhang, X. J. Jocla (2018). Clinical values of circular RNA 0000181 in the screening of gastric cancer. *J. Clin. Lab. Anal.* 32, e22333. doi:10.1002/jcla.22333

Zheng, P., Gao, H., Xie, X., and Lu, P. J. PorP. (2022). Plasma exosomal hsa_circ_0015286 as a potential diagnostic and prognostic biomarker for gastric cancer. *Pathol. Oncol. Res.* 28, 1610446. doi:10.3389/pore.2022.1610446

Zheng, X., Zhang, M., and Xu, M. J. N. (2017). Detection and characterization of ciRS-7: A potential promoter of the development of cancer. *Neoplasma*, 64, 321–328. doi:10.4149/neo_2017_301

Zhong, S., Wang, J., Hou, J., Zhang, Q., Xu, H., Hu, J., et al. (2018). Circular RNA hsa_circ_0000993 inhibits metastasis of gastric cancer cells. *Epigenomics* 10, 1301–1313. doi:10.2217/epi-2017-0173

Zhu, Z., Rong, Z., Luo, Z., Yu, Z., Zhang, J., Qiu, Z., et al. (2019). Circular RNA circNHSL1 promotes gastric cancer progression through the miR-1306-3p/SIX1/vimentin axis. *Mol. Cancer* 18, 126. doi:10.1186/s12943-019-1054-7

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 January 2023
DOI 10.3389/fgene.2022.1023615

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology, China

REVIEWED BY
Guanghui Li,
East China Jiaotong University, China
Li Zejun,
Professional Services Review, Australia

*CORRESPONDENCE
Shijun Li,
cflishijun6588@sina.com

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

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 20 August 2022
ACCEPTED 10 October 2022
PUBLISHED 20 January 2023

CITATION
Li S, Chang M, Tong L, Wang Y, Wang M and Wang F (2023), Screening potential lncRNA biomarkers for breast cancer and colorectal cancer combining random walk and logistic matrix factorization.
Front. Genet. 13:1023615.
doi: 10.3389/fgene.2022.1023615

Screening potential lncRNA biomarkers for breast cancer and colorectal cancer combining random walk and logistic matrix factorization

Shijun Li*†, Miaomiao Chang†, Ling Tong, Yuehua Wang, Meng Wang and Fang Wang

Department of Pathology, Chifeng Municipal Hospital, Chifeng, China

Breast cancer and colorectal cancer are two of the most common malignant tumors worldwide. They cause the leading causes of cancer mortality. Many researches have demonstrated that long noncoding RNAs (lncRNAs) have close linkages with the occurrence and development of the two cancers. Therefore, it is essential to design an effective way to identify potential lncRNA biomarkers for them. In this study, we developed a computational method (LDA-RWLMF) by integrating random walk with restart and Logistic Matrix Factorization to investigate the roles of lncRNA biomarkers in the prognosis and diagnosis of the two cancers. We first fuse disease semantic and Gaussian association profile similarities and lncRNA functional and Gaussian association profile similarities. Second, we design a negative selection algorithm to extract negative LncRNA-Disease Associations (LDA) based on random walk. Third, we develop a logistic matrix factorization model to predict possible LDAs. We compare our proposed LDA-RWLMF method with four classical LDA prediction methods, that is, LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM. The results from 5-fold cross validation on the MNDR dataset show that LDA-RWLMF computes the best AUC value of 0.9312, outperforming the above four LDA prediction methods. Finally, we rank all lncRNA biomarkers for the two cancers after determining the performance of LDA-RWLMF, respectively. We find that 48 and 50 lncRNAs have the highest association scores with breast cancer and colorectal cancer among all lncRNAs known to associate with them on the MNDR dataset, respectively. We predict that lncRNAs HULC and HAR1A could be separately potential biomarkers for breast cancer and colorectal cancer and need to biomedical experimental validation.

KEYWORDS

breast cancer, colorectal cancer, lncRNA, biomarker, lncRNA-disease association, random walk, logistic matrix factorization

1 Introduction

Breast cancer is the second leading cause of cancer-related death in women worldwide and the most common malignant tumor among US woman (Sun et al., 2017; DeSantis et al., 2019; Yang et al., 2013; Waks and Winer, 2019). During the past 25 years, breast cancer mortality rate showed a substantial increase in the world (Garrido-Castro et al., 2019). This increasing rate is one threaten to health for women in the world, in particular women from developing and low-income regions. More than 1.5 million women were diagnosed to breast cancer every year, which accounts for 25% among all women with cancers (Sun et al., 2017). In 2018, breast cancer accounts for approximately 24% of new cancer cases and approximately 15% of cancer deaths in women (Heer et al., 2020). In 2019, it is estimated that about 268,600 new patients suffer from invasive breast cancer and 48,100 patients suffer from ductal carcinoma *in situ* among US women. Moreover, 41,760 women may die from breast cancer in the same year (DeSantis et al., 2019). About 13% of women may suffer from invasive breast cancer in lifetime (DeSantis et al., 2019). The incident rate of breast cancer will increase by more than 46% by 2040 (Heer et al., 2020). Consequently, breast cancer has been one essential problem to be solved around the world.

However, the precise mechanisms of breast cancer remain unclear (Barzaman et al., 2020). Systemic treatment of breast cancer patients mainly consists of chemotherapy, endocrine treatment, and targeted therapy (Campos-Parra et al., 2018). In spite of rapid progress in different treatment strategies, accumulating patients show recurrence of the disease and decreased survival because of therapy resistance, which increases metastasis rates (Sledge et al., 2014). Once the metastasis occurs, the 5-year overall survival rate may be below 25% (Siegel et al., 2013).

Colorectal cancer is the third most frequent cancer and the second most death-caused cancer. It is estimated that there are about 1.9 million new cases and 0.9 million death cases worldwide in 2020 (Xi and Xu, 2021). Of new diagnose cases, 20% of patients have metastases and another 25% with localized disease may later develop metastases (Biller and Schrag, 2021). Its incidence is high in developed countries and is increasing in low- and middle-income countries, which poses a challenge to global public health (Biller and Schrag, 2021; Xi and Xu, 2021).

In this situation, it is essential to discover novel molecular biomarkers that can characterize therapy response for breast cancer and colorectal cancer. We can extend the overall survival rates of patients and delay or prevent the two cancers from metastases based on molecular biomarkers (Campos-Parra et al., 2018). Consequently, screening reliable biomarker is a research hotspot on the diagnosis and treatment of cancer including breast cancer and colorectal cancer (Huang et al., 2019; Yang et al., 2020; Peng et al., 2022a).

A substantial number of evidence suggest that over 80% of the human genome can be transcribed into non-coding RNAs, such as microRNAs (Peng et al., 2017; Peng et al., 2018; Chen et al., 2019; Huang et al., 2021), circle RNAs (Zhao et al., 2019; Lan et al., 2022), and long non-coding RNAs (lncRNAs) (Zhang et al., 2021a; Peng et al., 2021a; Peng et al., 2022b; Zhou et al., 2021a; Zhou et al., 2021b). In particular, lncRNAs obtain emerging interest as diagnostic biomarkers and therapeutic targets (Chandra Gupta and Nandan Tripathi, 2017; Guo et al., 2022). Differential expression of lncRNAs forms specific patterns to various complex diseases including cancer (Wahlestedt C, 2013). Once the regulation effects of lncRNAs are detected, they are promising therapeutic targets.

LncRNAs are closely related to breast cancer and colorectal cancer. For example, lncRNA BCRT1, MaTAR25, DSCAM-AS1, and CDC6 can promote breast cancer progression (Niknafs et al., 2016; Kong et al., 2019a; Chang et al., 2020; Liang et al., 2020), BCRT4 can induce signaling transduction in breast cancer (Xing et al., 2015), LINC00673 can promote cell proliferation of breast cancer (Qiao et al., 2019), and BORG can cause breast cancer metastasis and disease recurrence (Gooding et al., 2017). SNHG11, FEZF1-AS1, RP11, and DLEU1 have been reported to novel biomarkers of colorectal cancer (Bian et al., 2018; Liu et al., 2018; Wu et al., 2019; Xu et al., 2020). Thus, many computational models have been developed to discover lncRNA biomarkers for cancers (Peng et al., 2020a; Shen et al., 2022; Sun et al., 2022), for instance, rotation forest (Guo et al., 2019), KATZ measure (Chen, 2015), collaborative deep learning (Lan et al., 2020), matrix factorization (Fu et al., 2018; Wang et al., 2021a), network consistency projection (Li et al., 2019), and graph autoencoder (Shi et al., 2021).

In this manuscript, inspired by the association prediction method provided by Peng et al. (2020b), we develop a computational method, LDA-RWLMF, to predict LncRNA-Disease Associations (LDAs). LDA-RWLMF integrates random walk and Logistic Matrix Factorization to discover the roles of lncRNA biomarkers in the prognosis and diagnosis for breast cancer and colorectal cancer. First, we compute disease similarity and lncRNA similarity. Second, we first use random walk to extract negative LDAs. Third, we explored a logistic matrix factorization model to predict possible LDAs. The results from 5-fold cross validation show that LDA-RWLMF computes the best AUC value of 0.9312 on the MNDR dataset. Finally, we rank all lncRNA biomarkers for breast cancer and colorectal cancer after determining the performance of LDA-RWLMF.

2 Datasets

2.1 LncRNA-disease associations

Human LDA dataset was collected from the MNDR database (Cui et al., 2018; Fan et al., 2020) (http://www.rna-society.org/mndr/index.html). There are 1,529 LDAs between 89 diseases and 190 lncRNAs after preprocessing. For an LDA matrix between n lncRNAs and m diseases, we use $\boldsymbol{Y} \epsilon \mathrm{R}^{\mathrm{n} \times \mathrm{m}}$ to describe the association information by Eq. 1:

$$\boldsymbol{Y}_{ij} = \begin{cases} 1 \text{ If lncRNA } l_i & \text{associates with } d_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2.2 Disease semantic similarity

We use the method provided by Fan et al. (2020) to compute disease semantic similarity based on the MeSH descriptors. Disease semantic similarity method provided by Fan et al. (2020) was based on LNCSIM1 and LNCSIM2 provided by Chen (2015). For a disease A, suppose that T_A represents its ancestor node set, E_A denotes all edge set, its Directed Acyclic Graph (DAG) is represented as $DAG_A = \{T_A, E_A\}$. For a disease term $t \in T_A$ in DAG_A, its semantic contribution to A is calculated by Eq. 2 (Chen, 2015):

$$SV_A^1(t) = \begin{cases} 1 & t = A \\ \max\ (\Delta \times SV_A^1(t') \Big| t' \in C(t) & t \neq A \end{cases} \quad (2)$$

where $C(t)$ indicates the children of t, Δ indicates the sematic contribution factor related to edges that link t' to t, and Δ was usually set as 0.5 (Wang et al., 2010).

The above equation demonstrates that terms at the same layer from DAG_A have the same semantic contribution to A. But if two terms t_1 and t_2 are in the same layer of DAG_A and t_1 appears in less in DAG_A than t_2, the conclusion from t_1 will be more specific than one from t_2, thus, $SV_A^1(t_1)$ is higher than $SV_A^1(t_2)$.

In this case, we compute the second semantic contribution of term $t \in T_A$ to disease A by Eq. 3:

$$SV_A^2(t) = -log\frac{Dags(t)}{D} \quad (3)$$

where D indicates the number of diseases in MeSH, $Dags(t)$ indicates the number of DAGs that contain the disease term t. And the semantic contribution of t in DAG_A can be defined by Eq. 4:

$$SV_A^3(t) = \begin{cases} 1 & t = A \\ \max\ ((\Delta + \nabla) SV_A^3(t') \Big| t' \in C(t) & t \neq A \end{cases} \quad (4)$$

where ∇ indicates the contribution factor related to information content, and is computed by Eq. 5:

$$\nabla = \frac{\max_{k \in K}(Dags(k)) - dags(t)}{D} \quad (5)$$

where K indicates the disease set in MeSH.

Furthermore, the contribution of all terms in DAG_A to the disease A is computed by Eq. 6:

$$SV(A) = \sum_{t \in T_A} SV_A^3(t) \quad (6)$$

Finally, the semantic similarity between two diseases (A and B) can be computed by Eq. 7:

$$S_d^s(A, B) = \frac{\sum_{t \in T_A \cap T_B}(SV_A^3(t) + SV_B^3(t))}{SV(A) + SV(B)} \quad (7)$$

2.3 LncRNA functional similarity

We use the method provided by Fan et al. (Fan et al., 2020) to compute lncRNA functional similarity. Let that $DG(u)$ [or $DG(v)$] indicate diseases linking to lncRNA u (or v) on LDA matrix, the similarity between two lncRNAs u and v is obtained through disease semantic similarity in DG (u) and DG (v). A disease semantic similarity sub-matrix is first constructed. In the constructed matrix, rows and columns are diseases in $DG(u) \cup DG(v)$, and each element indicates the semantic similarity between diseases. Suppose that d_u indicate a disease in DG (u), the similarity between d_u and $DG(v)$ is computed by Eq. 8:

$$S(d_u, DG(v)) = \max_{d \in DG(v)}(S_d(d_u, d)) \quad (8)$$

Similarly, the similarity between d_v and DG (u) is computed by Eq. 9:

$$S(d_v, DG(u)) = \max_{d \in DG(u)}(S_d(d_v, d)) \quad (9)$$

And the similarity of $DG(u) \rightarrow DG(v)$ is computed by Eq. 10:

$$S_{u \rightarrow v} = \sum_{d \in DG(u)} S(d, DG(v)) \quad (10)$$

And similarity of $DG(v) \rightarrow DG(u)$ is computed by Eq. 11:

$$S_{v \rightarrow u} = \sum_{d \in DG(v)} S(d, DG(u)) \quad (11)$$

The similarity between lncRNAs u and v is measured based on the disease semantic similarity by Eq. 12:

$$S_l^f(u, v) = \frac{S_{u \rightarrow v} + S_{v \rightarrow u}}{|DG(u)| + |DG(v)|} \quad (12)$$

where $|DG(u)|$ and $|DG(v)|$ are the number of diseases in $DG(u)$ and $DG(v)$.

3 Methods

We want to compute association probability for each lncRNA-disease pair based on disease semantic similarity and lncRNA functional similarity. The pipeline is shown in Figure 1.

3.1 Gaussian association profile similarity and similarity fusion

In this section, we use Gaussian Association Profile (GAP) to compute the GAP similarity of diseases and lncRNAs. For a lncRNA l_i, its GAP $AP(l_i)$ is denoted using the i th row of $\boldsymbol{Y}$. The GAP similarity of lncRNAs l_i and l_j is defined by Eq. 13:

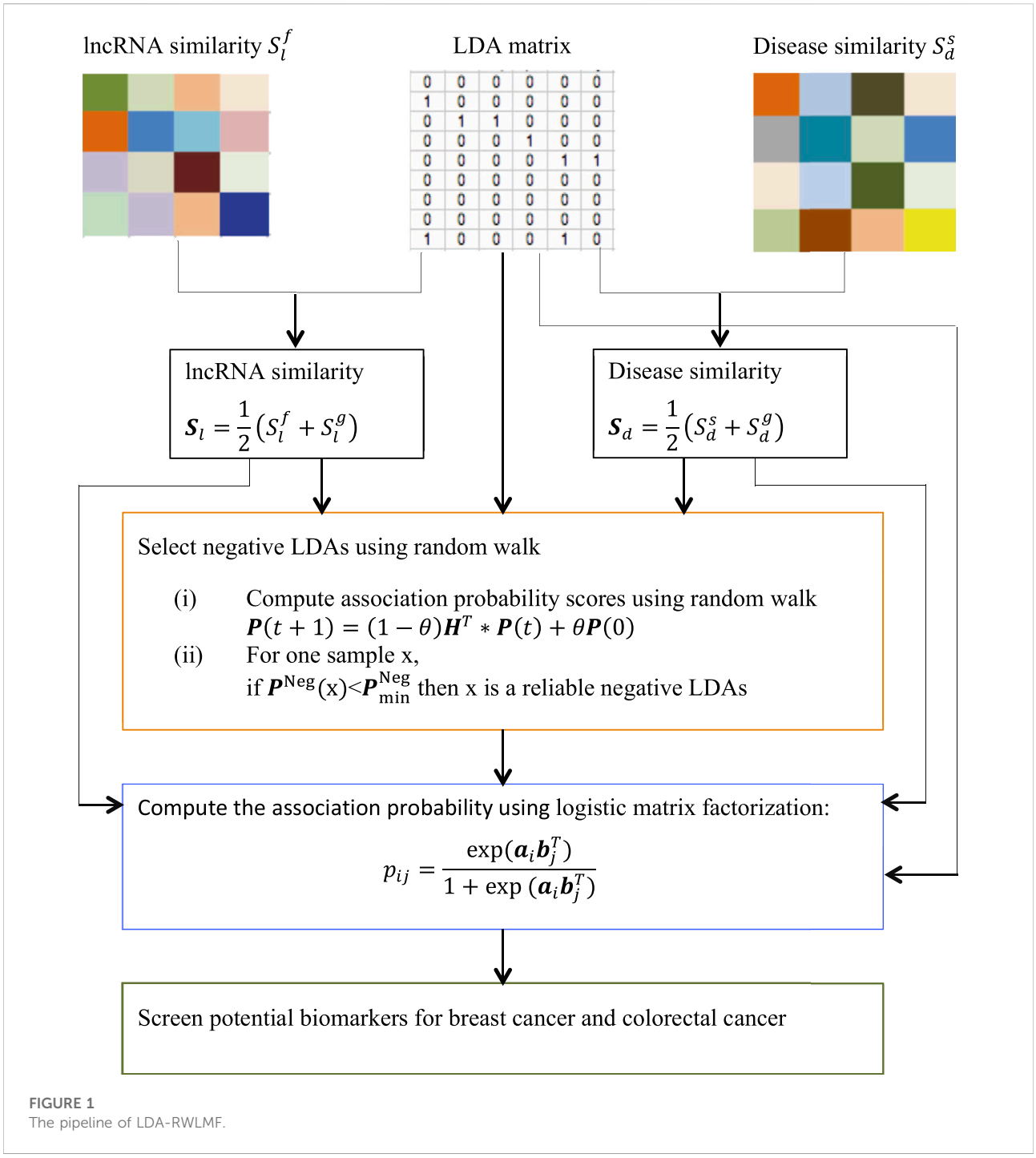


FIGURE 1
The pipeline of LDA-RWLMF.

$$S_l^g(l_i, l_j) = \exp\left(-\gamma_l \left\| AP(l_i) - AP(l_j) \right\|^2\right) \quad (13)$$

where $\gamma_l = \gamma_l' / (\frac{1}{n}\sum_{k=1}^{n} \| AP(l_i) \|^2)$ is the normalized kernel bandwidth with parameter γ_l'. Thus, the lncRNA similarity matrix $\boldsymbol{S}_l$ is computed by Eq. 14:

$$\boldsymbol{S}_l = \frac{1}{2}(S_l^f + S_l^g) \quad (14)$$

Similarly, the disease GAP similarity $\boldsymbol{S}_d$ can be computed.

3.2 Screening negative LDAs

There are not negative LDAs in the MNDR dataset. Credible negative LDAs help improve LDA prediction performance and further more effectively find potential lncRNA biomarkers for

breast cancer and colorectal cancer. Peng et al. (2021b) developed a random walk with restart-based virus-drug association prediction method and obtained better performance. Inspired by the method provided by Peng et al. (2021b), we first compute association probability for each lncRNA-disease pair through random walk with restart and then screen credible negative LDAs.

We first constructed a heterogeneous network composed of lncRNA similarity network, disease similarity network, and LDA network. lncRNA similarity matrix $\boldsymbol{S}_l$, disease similarity matrix $\boldsymbol{S}_d$, and LDA matrix Y are used as the adjacency matrices related to the heterogeneous network. The adjacency matrix related to the heterogeneous network is represented as Eq. 15:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{S}_l & \boldsymbol{Y} \\ \boldsymbol{Y}^T & \boldsymbol{S}_d \end{bmatrix} \tag{15}$$

where $\boldsymbol{Y}^T$ denotes the transpose of $\boldsymbol{Y}$.

We then compute transition probability on the heterogeneous graph. Suppose that $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{ll} & \boldsymbol{H}_{ld} \\ \boldsymbol{H}_{dl} & \boldsymbol{H}_{dd} \end{bmatrix}$ indicate transition probability matrix, where $\boldsymbol{H}_{ll}$ and $\boldsymbol{H}_{dd}$ indicate the walks within lncRNA similarity network and disease similarity network, respectively, $\boldsymbol{H}_{ld}$ and $\boldsymbol{H}_{dl}$ indicate the jumps between networks. For an lncRNA/disease, when there is an association between the lncRNA/disease and diseases/lncRNAs, the node will either continue to walk in the current network based on a transition probability $\lambda \in [0, 1]$ or jump between the above four networks.

The i -th lncRNA will walk to the j -th lncRNA through the transition probability $H_{ll}(i, j)$ by Eq. 16:

$$\boldsymbol{H}_{ll}(i,j) = \begin{cases} \dfrac{S_l(i,j)}{\sum_{k=1}^{n} S_l(i,k)}, & \text{if } \sum_{k=1}^{m} \boldsymbol{Y}(i,k) = 0 \\ \dfrac{(1-\lambda)S_l(i,j)}{\sum_{k=1}^{n} S_l(i,k)}, & otherwise \end{cases} \tag{16}$$

or jump to a disease d_j through the transition probability $H_{ld}(i, j)$ by Eq. 17:

$$\boldsymbol{H}_{ld}(i,j) = \begin{cases} \dfrac{\lambda Y(i,j)}{\sum_{k=1}^{m} Y(i,k)}, & \text{if } \sum_{k=1}^{m} \boldsymbol{Y}(i,k) \neq 0 \\ 0, & otherwise \end{cases} \tag{17}$$

Similarly, the i -th disease d_i will walk to the j -th disease d_j through the transition probability $\boldsymbol{H}_{dd}(i, j)$ by Eq. 18:

$$\boldsymbol{H}_{dd}(i,j) = \begin{cases} \dfrac{S_d(i,j)}{\sum_{k=1}^{m} S_d(i,k)}, & \text{if } \sum_{k=1}^{n} \boldsymbol{Y}(k,i) = 0 \\ \dfrac{(1-\lambda)S_d(i,j)}{\sum_{k=1}^{m} S_d(i,k)}, & otherwise \end{cases} \tag{18}$$

or jump to an lncRNA l_j through the transition probability $H_{dl}(i, j)$ by Eq. 19:

$$\boldsymbol{H}_{dl}(i,j) = \begin{cases} \dfrac{\lambda Y(i,j)}{\sum_{k=1}^{n} Y(k,i)}, & \text{if } \sum_{k=1}^{m} \boldsymbol{Y}(k,i) \neq 0 \\ 0, & otherwise \end{cases} \tag{19}$$

At the $t-$ th step, the association probability matrix between all lncRNA-disease pairs on the heterogeneous network is computed by Eq. 20:

$$\boldsymbol{P}(t+1) = (1-\theta)\boldsymbol{H}^T * \boldsymbol{P}(t) + \theta \boldsymbol{P}(0) \tag{20}$$

where $\boldsymbol{H}^T$ indicates the transpose of $\boldsymbol{H}$, and θ is the restarting probability. $\boldsymbol{P}(0)$ indicates the initial probability with $p_i(0) = \begin{bmatrix} (1-\eta)v_i \\ \eta s_i \end{bmatrix}$, where v_i and s_j indicate the initial probability distributions on disease similarity network and lncRNA similarity network, respectively. And $\eta \in [0, 1]$ is used to control the restarting probability in these two similarity networks. If $\eta < 0.5$, the particle will more tend to restart from one of the seed microbes than from one of the seed diseases.

In the second step, we consider known LDAs as positive sample set P, unknown lncRNA-disease pairs as unlabeled set U and propose a PU learning approach to screen credible negative LDA sample set RN. The method contains the following six steps:

Step 1. Randomly screening positive sample subset D from P

Step 2. Adding D into U;

Step 3. Considering $P - D$ as positive samples, $U + D$ as negative samples;

Step 4. Obtaining LDA score matrix $\boldsymbol{S}^{Neg}$ using random walk with restart;

Step 5. Ranking lncRNA-disease pairs in $\boldsymbol{D}$ based on $\boldsymbol{S}_{min}^{Neg}$ and obtaining the minimum score $\boldsymbol{S}_{min}^{Neg}$ in D;

Step 6. For every lncRNA-disease pair x in U:

If $\boldsymbol{S}^{Neg}(x) < \boldsymbol{S}_{min}^{Neg}$ then $RN = RN \cup x$.

3.3 LDA prediction based on logistic matrix factorization

Logistic matrix factorization has been applied to multiple areas (Liu et al., 2020; Tang et al., 2021; Tian et al., 2022). Inspired by the approaches, we develop a logistic matrix factorization-based LDA prediction method, LDA-RWLMF.

Assume that both lncRNAs and diseases are mapped to r-dimensional shared latent spaces ($r \ll n, m$), thus an lncRNA l_i or disease d_i can be represented as a latent vector $\boldsymbol{a}_i \in \Re^{1\times t}$ or $\boldsymbol{b}_i \in \Re^{1\times t}$. The association probability p_{ij} between l_i and d_i is calculated by Eq. 12:

TABLE 1 AUCs of LDA identification approaches on the MNDR dataset.

Dataset	LNCSIM1	LNCSIM2	ILNCSIM	IDSSIM	LDA-RWLMF
the MNDR dataset	0.9251	0.9280	0.9267	0.9302	0.9312

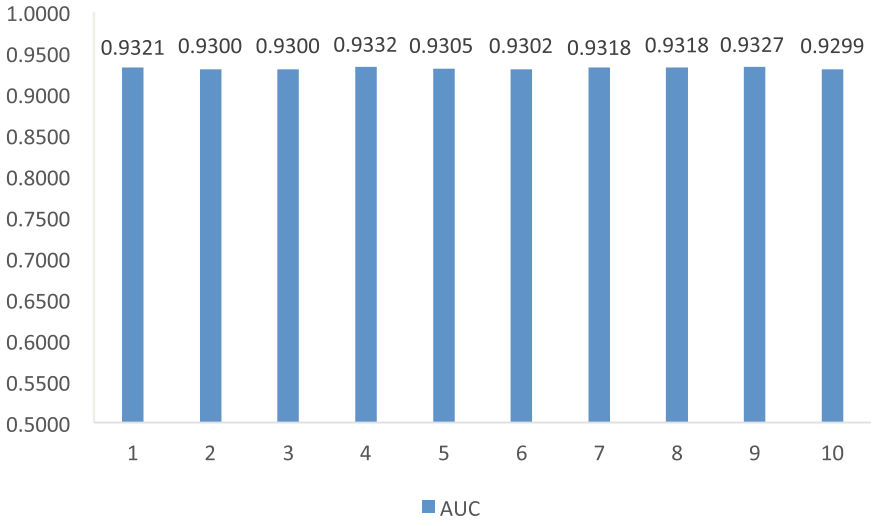


FIGURE 2
The AUC of LDA-RWLMF from 10 time cross validation (t = 1, 2, 3, . . . , 10).

$$p_{ij} = \frac{\exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)}{1 + \exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)} \tag{21}$$

The latent vector matrix of all lncRNAs or diseases can be represented as $\boldsymbol{A} \in \Re^{n \times r}$ or $\boldsymbol{B} \in \Re^{m \times r}$ where $\boldsymbol{a}_i$ or $\boldsymbol{b}_i$ indicates the i th or j th row in $\boldsymbol{A}$ or $\boldsymbol{B}$. In addition, known LDAs are more credible than unknown lncRNA-disease pairs. Thus, we assign higher confidence values to known LDAs than unknown lncRNA-disease pairs. Similar to Peng et al. (2020b), we use a constant c to assess the importance of known LDAs and construct a prediction model by Eq. 22:

$$\begin{aligned} p(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{B}) = & \left(\prod_{1 \le i \le n, 1 \le j \le m, y_{ij}=1} \left[p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{\left(1-y_{ij}\right)} \right]^c \right) \\ & \times \left(\prod_{1 \le i \le n, 1 \le j \le m, y_{ij}=0} \left[p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{\left(1-y_{ij}\right)} \right] \right) \\ = & \prod_{i=1}^{n} \prod_{j=1}^{m} p_{ij}^{cy_{ij}} \left(1 - p_{ij}\right)^{\left(1-y_{ij}\right)} \end{aligned} \tag{22}$$

Model (21) can be optimized based on the Bayesian distribution by Eq. 23:

$$\begin{aligned} \min_{\boldsymbol{A},\boldsymbol{B}} & \sum_{i=1}^{m} \sum_{j=1}^{n} \left(1 + cy_{ij} - y_{ij}\right) \log\left[1 + \exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)\right] - cy_{ij} \boldsymbol{a}_i \boldsymbol{b}_j^T + \frac{\lambda_l}{2} \\ & \| \boldsymbol{A} \|_F^2 + \frac{\lambda_d}{2} \| \boldsymbol{B} \|_F^2 \end{aligned} \tag{23}$$

where λ_l and λ_d are two parameters, $\|\boldsymbol{A}\|_F$ indicates the Frobenius norm of $\boldsymbol{A}$. (Zhang et al. 2019a; Zhang et al. 2019b) integrated linear neighborhood information to model (22) to predict various associations. Similarly, we fuse neighborhood information to Eq. 23 by Eq. 24:

$$\begin{aligned} \min_{\boldsymbol{A},\boldsymbol{B}} & \sum_{i=1}^{m} \sum_{j=1}^{n} \left(1 + cy_{ij} - y_{ij}\right) \ln\left[1 + \exp\left(\boldsymbol{a}_i \boldsymbol{b}_j^T\right)\right] - cy_{ij} \boldsymbol{a}_i \boldsymbol{b}_j^T \\ & + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{A}^T \left(\lambda_l \boldsymbol{I} + \alpha \boldsymbol{L}_l\right) \boldsymbol{A} + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{B}^T \left(\lambda_d \boldsymbol{I} + \alpha \boldsymbol{L}_d\right) \boldsymbol{B}\right.\right. \end{aligned} \tag{24}$$

where $tr\ (\cdot)$ indicates the trace of the matrix. $\boldsymbol{L}_l$ and $\boldsymbol{L}_d$ indicate the corresponding Laplacian matrix of $\boldsymbol{A}$ and $\boldsymbol{B}$. $\boldsymbol{L}_l = (\boldsymbol{D}_l + \tilde{\boldsymbol{D}}_l) - (\boldsymbol{A} + \boldsymbol{A}^T)$ where $\boldsymbol{D}_l$ and $\tilde{\boldsymbol{D}}_l$ are two diagonal matrices and $\boldsymbol{D}_l(i,i) = \sum_{j=1}^{m} a_{ij}$ and $\tilde{\boldsymbol{D}}_l(i,i) = \sum_{i=1}^{m} a_{ij}$. Similarly, $\boldsymbol{L}_d$ can be computed.

We compute $\boldsymbol{A}$ and $\boldsymbol{B}$ by solving Eq. 24 through an alternating gradient ascent approach.

TABLE 2 The rankings of the predicted top 48 lncRNAs according to association with breast cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
1	CASC2	Known	25	PVT1	Known
2	DLEU2	Known	26	RMST	Known
3	MIR17HG	Known	27	TRAF3IP2-AS1	Known
4	DSCAM-AS1	Known	28	HCP5	Known
5	SNHG4	Known	29	LINC00271	Known
6	TCL6	Known	30	GHET1	Known
7	XIST	Known	31	SNHG3	Known
8	CBR3-AS1	Known	32	TDRG1	Known
9	MIAT	Known	33	DAOA-AS1	Known
10	CCAT2	Known	34	BACE1-AS	Known
11	SOX2-OT	Known	35	NAMA	Known
12	GAS5	Known	36	BDNF-AS	Known
13	PCA3	Known	37	SNHG11	Known
14	MALAT1	Known	38	UCA1	Known
15	BANCR	Known	39	SNHG16	Known
16	WT1-AS	Known	40	MIR100HG	Known
17	PANDAR	Known	41	H19	Known
18	HNF1A-AS1	Known	42	TERC	Known
19	HAR1B	Known	43	MEG3	Known
20	CCDC26	Known	44	SPRY4-IT1	Known
21	BCAR4	Known	45	DANCR	Known
22	PDZRN3-AS1	Known	46	KCNQ1OT1	Known
23	HIF1A-AS2	Known	47	IFNG-AS1	Known
24	CRNDE	Known	48	HOTAIR	Known

TABLE 3 The rankings of the remaining 41 lncRNAs according to association with breast cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
49	HULC	PMID: 31824174, 33107484, 33745450	70	ZFAT-AS1	Unconfirmed
50	CCAT1	Known	71	PTENP1	PMID: 28731027, 29085464, 29212574, 31196157
51	NPTN-IT1	Unconfirmed	72	HIF1A-AS1	Unconfirmed
52	PCAT1	PMID: 32853955, 28989584, 33850635, 32220602	73	SRA1	Known
53	HAR1A	PMID: 26942882	74	MINA	Unconfirmed
54	LSINCT5	Known	75	DLEU1	Known
55	TUG1	PMID: 28950664, 27848085, 30098551, 33380806	76	PSORS1C3	Unconfirmed
56	MIR155HG	Unconfirmed	77	LINC00032	Unconfirmed
57	DGCR5	PMID: 32521856	78	WRAP53	Unconfirmed
58	IGF2-AS	PMID: 33175607	79	7SK	Unconfirmed
59	BCYRN1	Known	80	RRP1B	Unconfirmed
60	EPB41L4A-AS1	PMID: 35181612	81	MYCNOS	Unconfirmed
61	PINK1-AS	Unconfirmed	82	PRINS	Unconfirmed
62	DNM3OS	Unconfirmed	83	ATP6V1G2-DDX39B	Unconfirmed
63	ADAMTS9-AS2	PMID: 30840279	84	MKRN3-AS1	Unconfirmed
64	MIR31HG	lncRNADisease	85	NRON	Unconfirmed
65	BOK-AS1	Unconfirmed	86	MESTIT1	Unconfirmed
66	ESRG	Unconfirmed	87	LINC00162	Unconfirmed
67	KCNQ1DN	Unconfirmed	88	DISC2	Unconfirmed
68	ATXN8OS	PMID: 31173245, 33385064, 33477683	89	SCAANT1	Unconfirmed
69	CDKN2B-AS1	Known			

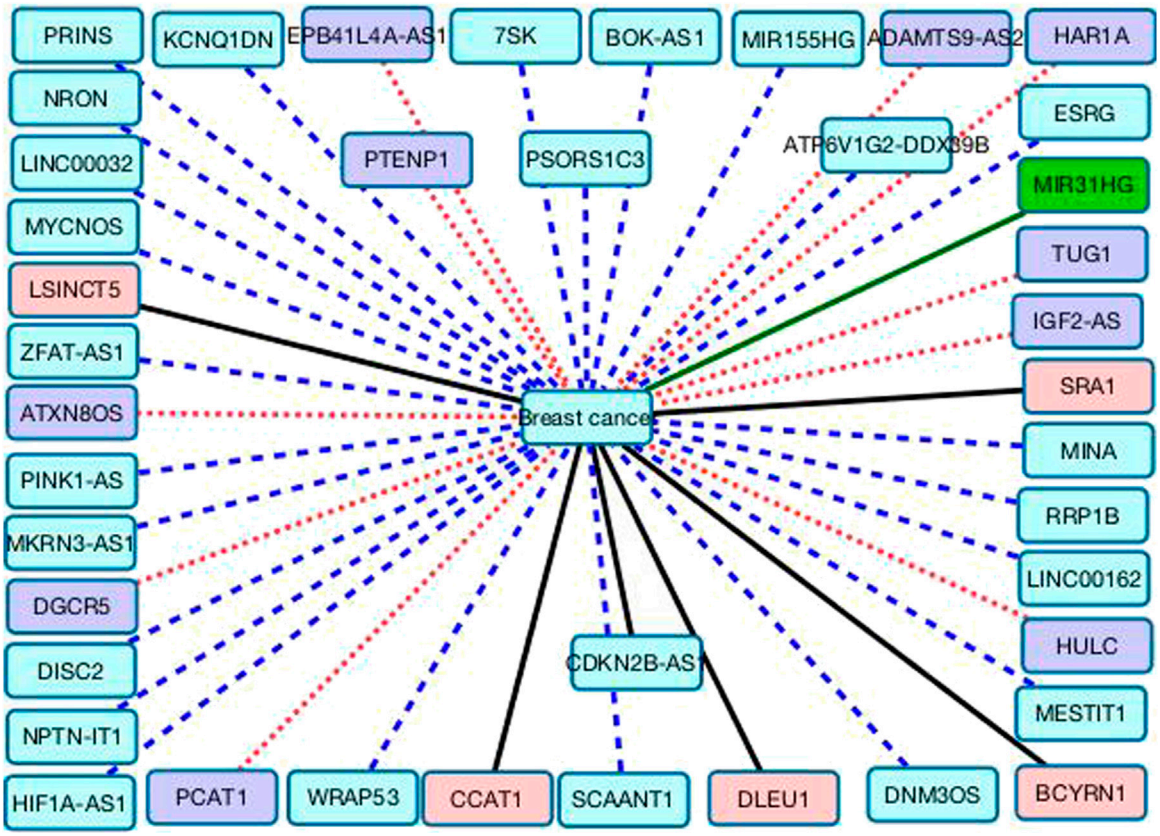


FIGURE 3
The associations between the remaining 41 lncRNAs and breast cancer.

Finally, lncRNA-disease association score $\boldsymbol{Y}_{fin}(i, j)$ for each lncRNA-disease pair can be computed by Eq. 25:

$$\boldsymbol{Y}_{fin} = \boldsymbol{A}\boldsymbol{B}^{T} \tag{25}$$

4 Results

4.1 Experimental settings

We conduct 5-fold cross validation for 10 times to investigate the performance of LDA-RWLMF. AUC is used to evaluate the prediction accuracy of LDA identification models. AUC is the area under the true positive rate (TPR)-false positive rate (FPR) curve, where TPR and FPR are defined by Eqs 26, 27:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{26}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{27}$$

where TP, FP, TN, FN represent the number of true positives, false positives, true negatives, false negatives, respectively. Higher AUC is, better the prediction performance is. In addition, parameters in LDA-RWLMF are set to defaults provided by Peng et al. (2020b). And parameters in the other four comparison LDA prediction methods (LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM) are set to the same values provided by corresponding methods.

4.2 Performance comparison with other methods

To measure the performance of the proposed LDA-RWLMF method, we compare it with four other representative LDA inference approaches on the MNDR dataset. That is, LNCSIM1 (Chen, 2015), LNCSIM2 (Chen, 2015), ILNCSIM (Huang et al., 2016), and IDSSIM (Fan et al., 2020). LNCSIM1 and LNCSIM2 used Laplacian regularized least squares to predict possible LDAs based on disease DAGs and the information content, respectively. ILNCSIM first combined the hierarchical structure of disease DAG and the information content to compute disease similarity and then used Laplacian

TABLE 4 The rankings of the identified top 50 lncRNAs associated with colorectal cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
1	SOX2-OT	Known	26	NAMA	Known
2	DLEU2	Known	27	WT1-AS	Known
3	CASC2	Known	28	TDRG1	Known
4	TCL6	Known	29	GHET1	Known
5	TRAF3IP2-AS1	Known	30	CRNDE	Known
6	DSCAM-AS1	Known	31	XIST	Known
7	GAS5	Known	32	MALAT1	Known
8	MIR17HG	Known	33	RMST	Known
9	HAR1B	Known	34	SNHG3	Known
10	CCDC26	Known	35	BACE1-AS	Known
11	CBR3-AS1	Known	36	MIR100HG	Known
12	PANDAR	Known	37	IFNG-AS1	Known
13	MIAT	Known	38	DANCR	Known
14	SNHG4	Known	39	SNHG16	Known
15	HIF1A-AS2	Known	40	SNHG11	Known
16	HNF1A-AS1	Known	41	TERC	Known
17	PCA3	Known	42	KCNQ1OT1	Known
18	BANCR	Known	43	MEG3	Known
19	LINC00271	Known	44	HULC	Known
20	PDZRN3-AS1	Known	45	UCA1	Known
21	CCAT2	Known	46	SPRY4-IT1	Known
22	BCAR4	Known	47	PCAT1	Known
23	DAOA-AS1	Known	48	HOTAIR	Known
24	BDNF-AS	Known	49	PVT1	Known
25	HCP5	Known	50	CCAT1	Known

regularized least squares to infer new LDAs. IDSSIM designed a weighted K nearest neighbor approach to identify potential associations between lncRNAs and diseases by integrating disease semantic similarity and lncRNA functional similarity. Table 1 gives the AUC values of the four LDA identification methods and our proposed LDA-RWLMF on the MNDR dataset.

The results from Table 1 demonstrate that LDA-RWLMF computes the highest AUC compared to LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM on the MNDR dataset. Figure 2 gives the results of LDA-RWLMF from 10 time cross validation. From Figure 2, we can find that AUC obtain by LDA-RWLMF is relatively steady during 10 time cross validation.

4.3 Case study

4.3.1 lncRNA biomarker identification for breast cancer

Breast cancer is the commonest life-threatening cancer in women (Key et al., 2001; Sharma, et al., 2010). lncRNAs play important roles in epigenetic regulation, transcriptional regulation and post-transcriptional regulation and have been potential biomarkers of many diseases. Substantial publications have reported that lncRNAs affect proliferation and apoptosis, invasion and metastasis, and cancer stemness of breast cancer. For example, LSINCT5 and Zfas one can promote the proliferation of breast cancer, HOTAIR suppresses invasion and migration of breast cancer, SOX2OT induces SOX2 expression in breast cancer, and SRA is the expression activator of breast cancer (Sun et al., 2017). We want to conduct case analyses to find possible lncRNA biomarkers for breast cancer based on the proposed LDA-RWLMF model.

In the MNDR dataset, there are 89 lncRNAs that may associate with breast cancer, where 54 lncRNAs have been experimentally validated to associate with the cancer and 35 lncRNAs have unknown associations with it. We use the proposed LDA-RWLMF method to rank the 89 lncRNAs for breast cancer. The results are shown in Tables 2, 3. Table 2 demonstrates the ranking results of the predicted top 48 lncRNAs according to the computed association score with breast cancer on the MNDR dataset. These 48 lncRNAs are known to link to breast cancer on the MNDR dataset and are ranked as top 48.

Table 3 gives the rankings of the remaining 41 lncRNAs according to the association scores with breast cancer on the

TABLE 5 The rankings of the remaining 41 lncRNAs according to association with breast cancer on the MNDR dataset.

Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
51	HAR1A	Unconfirmed	71	ZFAT-AS1	Unconfirmed
52	NPTN-IT1	known	72	SRA1	Unconfirmed
53	TUG1	known	73	PSORS1C3	Unconfirmed
54	IGF2-AS	PMID: 32853944, 30581274	74	HIF1A-AS1	Unconfirmed
55	LSINCT5	known	75	MINA	Unconfirmed
56	DGCR5	PMID: 31452812	76	LINC00032	Unconfirmed
57	H19	known	77	WRAP53	Unconfirmed
58	EPB41L4A-AS1	PMID: 32557646	78	DLEU1	Unconfirmed
59	MIR155HG	PMID: 34562123,31228357	79	RRP1B	Unconfirmed
60	CDKN2B-AS1	known	80	7SK	Unconfirmed
61	MIR31HG	PMID: 30447009,35733512,34485123	81	PRINS	Unconfirmed
62	ESRG	PMID: 34896077	82	MYCNOS	Unconfirmed
63	BCYRN1	PMID: 30114690,32944001,31773686	83	ATP6V1G2-DDX39B	Unconfirmed
64	BOK-AS1	Unconfirmed	84	MKRN3-AS1	Unconfirmed
65	PINK1-AS	Unconfirmed	85	NRON	Unconfirmed
66	KCNQ1DN	Unconfirmed	86	SCAANT1	Unconfirmed
67	ATXN8OS	Unconfirmed	87	DISC2	Unconfirmed
68	DNM3OS	Unconfirmed	88	MESTIT1	Unconfirmed
69	PTENP1	Unconfirmed	89	LINC00162	Unconfirmed
70	ADAMTS9-AS2	Unconfirmed			

MNDR dataset. Among all lncRNAs unknown to associate with breast cancer on the MNDR dataset, lncRNA HULC is predicted to link to breast cancer with the highest association scores. Shi et al. (2016) observed that HULC can act as an oncogene biomarker in triple-negative breast cancer and as an independent possible poor prognostic factor in patients suffered from triple-negative breast cancer. Wang et al. (2019) found that HULC can promote the development of breast cancer through regulating the expression of LYPD1. Gavgani et al. (2020) investigated that the HULC knockdown can induce apoptosis and suppress cellular migration in breast cancer cells.

PCAT1 may link to breast cancer with the ranking of three among all lncRNAs unknown to associate with breast cancer on the MNDR dataset. Several studies have reported that PCAT1 can associate with breast cancer although its association with the cancer on the MNDR dataset is unobserved. Abdollahzadeh et al. (2020) reported that the altered regulation of PCAT1 may play crucial roles in the development and pathogenesis of breast cancer. Sarrafzadeh et al. (2017) assessed the expression of PCAT-1 through real-time reverse transcription polymerase chain reaction in breast tumor samples from 47 breast cancer patients and found that PCAT-1 may involve in the pathogenesis of breast cancers. Wang et al. (2021a) observed that PCAT-1 can facilitate breast cancer progression by binding to RACK1 and thus boosting oxygen-independent stability of HIF-1α. Tang et al. (2022) detect that PCAT1 can regulate the expression of PITX2 in breast cancer.

In addition, we predict that nephronectin intronic transcript 1 (NPTN-IT1, also known as lncRNA-LET) may have relationship with breast cancer. NPTN-IT1 has been reported to associate with bladder cancer through attenuating the expression of the target of miR-145 and ILF3 in bladder cancer (Zhang et al., 2021b). It was significantly down-regulated in multiple tumor tissues of colorectal cancer. It also has a regulation role in hypoxia signaling of hepatocellular carcinoma (Sun et al., 2013) and was highly expressed in HepG2 cells (Kong et al., 2019b). We hope that association between three lncRNAs (HULC, NPTN-IT1, and PCAT1) and breast cancer can be validated through wet experiments. Figure 3 shows the associations between the 41 lncRNAs that are ranked as the last 41 and breast cancer. Black solid lines represent known LDAs in the MNDR database. Green solid lines represent LDAs that can be observed in the lncRNA disease database. Red dots lines represent LDAs that are predicted to be potential lncRNA biomarkers of breast cancer and can be confirmed by related publications. Blue equal dash lines represent unknown LDAs.

4.3.2 lncRNA biomarker identification for colorectal cancer

Colorectal cancer is a heterogeneous disease. It has high morbidity and mortality. lncRNAs demonstrate dense associations with colorectal cancer. In this study, we conduct case analyses to identify possible lncRNA

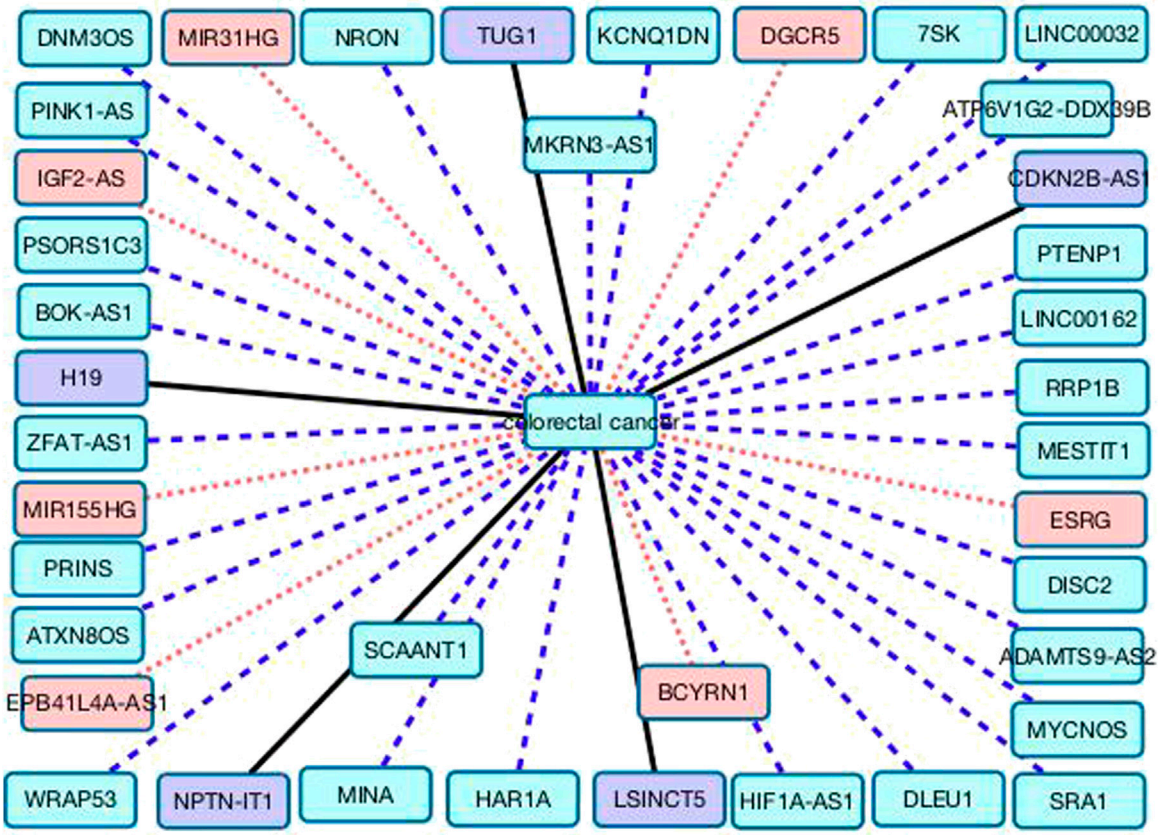


FIGURE 4
The associations between the remaining 39 lncRNAs and colorectal cancer.

biomarkers for colorectal cancer based on LDA-RWLMF. In the MNDR dataset, 89 lncRNAs possibly associate with colorectal cancer, where 55 lncRNAs have been validated to be the biomarkers of the cancer and remaining 34 lncRNAs have not been validated. We use LDA-RWLMF to compute the association scores between all 89 lncRNAs and colorectal cancer and rank the 89 lncRNAs for colorectal cancer. The results are shown in Tables 4, 5. Table 4 shows the rankings of the identified top 50 lncRNAs according to the computed association score with colorectal cancer on the MNDR dataset. The 50 lncRNAs are known to associate with colorectal cancer on the MNDR dataset and are ranked as top 50.

Table 5 gives the rankings of the remaining 39 lncRNAs according to the association scores with colorectal cancer on the MNDR dataset. Among all lncRNAs unknown association with colorectal cancer on the MNDR dataset, lncRNA HAR1A is inferred to link to colorectal cancer with the highest association scores. HAR1A is a favorable prognostic biomarker for patients. Shi et al. (2019) analyzed the expression profiles of HAR1A using RT-qPCR and found its expression level was significantly lower in hepatocullular cancer. Chen et al. (2020) have still reported that the HAR1A expression levels were reduced in hepatocellular carcinoma tissues.

Figure 4 gives the associations between the remaining 39 lncRNAs and colorectal cancer. Black solid lines represent known LDAs in the MNDR database. Red dots lines represent LDAs that are predicted to be potential lncRNA biomarkers of breast cancer and can be confirmed by related publications. Blue equal dash lines represent unknown LDAs.

5 Discussion and conclusion

Breast cancer and colorectal cancer are the most frequent cancers with high mortality rates. They demonstrate very high heterogeneity at molecular and clinical levels. With the fast development of next generation sequencing technologies, we can more accurately characterize the human genome. lncRNAs act mainly as gene expression regulators. The dysregulation of lncRNAs may destroy the normal transcriptional landscape and thus cause malignant transformation. In addition, their highly specific expression

and functional tertiary structure force them to be as promising diagnostic biomarkers and potential targets for various diseases including breast cancer and colorectal cancer.

In this study, we proposed a computational lncRNA-disease association method (LDA-RWLMF) to identify potential biomarkers for breast cancer and colorectal cancer. First, a random walk with restart method was designed to extract negative LDAs. Second, a logistic matrix factorization model was explored to infer possible associations between lncRNAs and diseases. Finally, all lncRNAs are ranked according to association scores with breast cancer and colorectal cancer on the MNDR dataset.

We conduct 5-fold cross validation for 10 times to compare LDA-RWLMF with state-of-the-art LDA prediction models on the MNDR dataset, that is, LNCSIM1, LNCSIM2, ILNCSIM, and IDSSIM. The results show that LDA-RWLMF computes the best AUC values of 0.9312. We predict that lncRNAs (HULC, NPTN-IT1, and PCAT1) may be possible biomarkers of breast cancer and colorectal cancer.

Our proposed LDA-RWLMF method has two disadvantages. First, it extracted credible negative LDA samples. In the area of association prediction, there are no negative association samples because of the limitation of biomedical experiments, which causes relatively poor performance. Thus, we designed a negative LDA extraction method based on PU learning. Second, the logistic matrix factorization model can effectively discover possible associations between two biological entities. Thus, we used the model to identify new LDAs. In addition, diseases and lncRNAs exhibit abundant biological features. In this study, we failed to consider these diverse features. In the future, we will further integrate more biological information to improve LDA prediction.

In the future, we will further design more effective negative sample screening method based on positive-unlabeled learning. In addition, we will also develop deep learning model for LDA prediction. We anticipate that the proposed LDA-RWLMF method can help design therapeutic regimens for personalized treatment of breast cancer and colorectal cancer and thus opportunely inhibit its recurrence.

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

Conceptualization: SL, MC, and LT; Methodology: SL, MC, YW, MW, and FW; Project administration: SL; Software: SL and MC; Writing-original draft: SL; Writing-review and editing: SL, MC.

Conflict of interest

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

Publisher's note

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

References

Abdollahzadeh, R., Mansoori, Y., Azarnezhad, A., Daraei, A., Paknahad, S., Mehrabi, S., et al. (2020). Expression and clinicopathological significance of AOC4P, PRNCR1, and PCAT1 lncRNAs in breast cancer. *Pathol. Res. Pract.* 216 (10), 153131. doi:10.1016/j.prp.2020.153131

Barzaman, K., Karami, J., Zarei, Z., Hosseinzadeh, A., Kazemi, M. H., Moradi-Kalbolandi, S., et al. (2020). Breast cancer: Biology, biomarkers, and treatments. *Int. Immunopharmacol.* 84, 106535. doi:10.1016/j.intimp.2020.106535

Bian, Z., Zhang, J., Min, L., Feng, Y., Xue, W., Jia, Z., et al. (2018). LncRNA-FEZF1-AS1 promotes tumor proliferation and metastasis in colorectal cancer by regulating PKM2 signaling. *Clin. Cancer Res.* 24 (19), 4808–4819. doi:10.1158/1078-0432.CCR-17-2967

Biller, L. H., and Schrag, D. (2021). Diagnosis and treatment of metastatic colorectal cancer: A review. *Jama* 325 (7), 669–685. doi:10.1001/jama.2021.0106

Campos-Parra, A. D., López-Urrutia, E., Orozco Moreno, L. T., Lopez-Camarillo, C., Meza-Menchaca, T., Figueroa Gonzalez, G., et al. (2018). Long non-coding RNAs as new master regulators of resistance to systemic treatments in breast cancer. *Int. J. Mol. Sci.* 19 (9), 2711. doi:10.3390/ijms19092711

Chandra Gupta, S., and Nandan Tripathi, Y. (2017). Potential of long non-coding RNAs in cancer patients: From biomarkers to therapeutic targets. *Int. J. Cancer* 140 (9), 1955–1967. doi:10.1002/ijc.30546

Chang, K. C., Diermeier, S. D., Yu, A. T., Brine, L. D., and Spector, D. L. (2020). MaTAR25 lncRNA regulates the Tensin1 gene to impact breast cancer progression [J]. *Nat. Commun.* 11 (1), 1–19.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41 (D1), D983–D986. doi:10.1093/nar/gks1099

Chen, X. (2015). Katzlda: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5 (1), 16840–16911. doi:10.1038/srep16840

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019). MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 20 (2), 515–539. doi:10.1093/bib/bbx130

Chen, Y., Guo, Y., Chen, H., and Ma, F. (2020). Long non-coding RNA expression profiling identifies a four-long non-coding RNA prognostic signature for isocitrate dehydrogenase mutant glioma. *Front. Neurol.* 11, 573264. doi:10.3389/fneur.2020.573264

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: An updated resource of ncRNA-disease associations in mammals *Nucleic Acids Res.* 46 (D1), D371–D374. doi:10.1093/nar/gkx1025

DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., et al. (2019). Breast cancer statistics. *Ca. Cancer J. Clin.* 69 (6), 438–451. doi:10.3322/caac.21583

Duffy, M. J., Synnott, N. C., and Crown, J. (2018). Mutant p53 in breast cancer: Potential as a therapeutic target and biomarker. *Breast Cancer Res. Treat.* 170 (2), 213–219. doi:10.1007/s10549-018-4753-7

Fan, W., Shang, J., Li, F., Sun,, Y., and Liu, J. X. (2020). Idssim: An lncRNA functional similarity calculation model based on an improved disease semantic similarity method[J]. *BMC Bioinforma.* 21 (1), 1–14.

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34 (9), 1529–1537. doi:10.1093/bioinformatics/btx794

Garrido-Castro, A. C., Lin, N. U., and Polyak, K. (2019). Insights into molecular classifications of triple-negative breast cancer: Improving patient selection for treatment. *Cancer Discov.* 9 (2), 176–198. doi:10.1158/2159-8290.CD-18-1177

Gavgani, R. R., Babaei, E., Hosseinpourfeizi, M. A., Fakhrjou, A., and Montazeri, V. (2020). Study of long non-coding RNA highly upregulated in liver cancer (HULC) in breast cancer: A clinical & *in vitro* investigation. *Indian J. Med. Res.* 152 (3), 244–253. doi:10.4103/ijmr.IJMR_1823_18

Gooding, A. J., Zhang, B., Jahanbani, F. K., Gilmore, H. L., Chang, J. C., Valadkhan, S., et al. (2017). The lncRNA BORG drives breast cancer metastasis and disease recurrence. *Sci. Rep.* 7 (1), 1–18. doi:10.1038/s41598-017-12716-6

Guo, Z. H., You, Z. H., Wang, Y. B., Yi, H. C., and Chen, Z. H. (2019). A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest. *IScience* 19, 786–795. doi:10.1016/j.isci.2019.08.030

Guo, Z., Hui, Y., Kong, F., and Lin, X. (2022). Finding lung-cancer-related lncRNAs based on laplacian regularized least squares with unbalanced Bi-random walk. *Front. Genet.* 13, 933009. doi:10.3389/fgene.2022.933009

Heer, E., Harper, A., Escandor, N., Sung, H., McCormack, V., and Fidler-Benaoudia, M. M. (2020). Global burden and trends in premenopausal and postmenopausal breast cancer: A population-based study. *Lancet. Glob. Health* 8 (8), e1027–e1037. doi:10.1016/S2214-109X(20)30215-1

Huang, F., Yue, X., Xiong, Z., Yu, Z., Liu, S., and Zhang, W. (2021). Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief. Bioinform.* 22 (3), bbaa140. doi:10.1093/bib/bbaa140

Huang, Y. A., Chen, X., You, Z. H., Huang, D. S., and Chan, K. C. C. (2016). Ilncsim: Improved lncRNA functional similarity calculation model. *Oncotarget* 7 (18), 25902–25914. doi:10.18632/oncotarget.8296

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47 (D1), D1013–D1017. doi:10.1093/nar/gky1010

Key, T. J., Verkasalo, P. K., and Banks, E. (2001). Epidemiology of breast cancer. *Lancet. Oncol.* 2 (3), 133–140. doi:10.1016/S1470-2045(00)00254-0

Kong, J., Qiu, Y., Li, Y., Zhang, H., and Wang, W. (2019b). TGF-β1 elevates P-gp and BCRP in hepatocellular carcinoma through HOTAIR/miR-145 axis. *Biopharm. Drug Dispos.* 40 (2), 70–80. doi:10.1002/bdd.2172

Kong, X., Duan, Y., Sang, Y., Li, Y., Zhang, H., Liang, Y., et al. (2019a). LncRNA-CDC6 promotes breast cancer progression and function as ceRNA to target CDC6 by sponging microRNA-215. *J. Cell. Physiol.* 234 (6), 9105–9117. doi:10.1002/jcp.27587

Lan, W., Dong, Y., Chen, Q., Zheng, R., Liu, J., Pan, Y., et al. (2022). Kgancda: Predicting circRNA-disease associations based on knowledge graph attention network. *Brief. Bioinform.* 23 (1), bbab494. doi:10.1093/bib/bbab494

Lan, W., Lai, D., and Chen, Q. (2020). Ldicdl: LncRNA-disease association identification based on collaborative deep learning[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19, 1715–1723. doi:10.1109/TCBB.2020.3034910

Li, G., Luo, J., Liang, C., Xiao, Q., Ding, P., and Zhang, Y. (2019). Prediction of LncRNA-disease associations based on network consistency projection. *Ieee Access* 7, 58849–58856. doi:10.1109/access.2019.2914533

Liang, Y., Song, X., Li, Y., Chen, B., Zhao, W., Wang, L., et al. (2020). Retraction note to: LncRNA BCRT1 promotes breast cancer progression by targeting miR-1303/PTBP3 axis. *Mol. Cancer* 19 (1), 131–220. doi:10.1186/s12943-022-01576-y

Liang, Y., Wu, Y., and Zhang, Z. (2022a). Hyb4mC: A hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction[J]. *BMC Bioinforma.* 23 (1), 1–18. doi:10.1186/s12859-022-04789-6

Liang, Y., Zhang, Z. Q., Liu, N. N., Wu, Y. N., Gu, C. L., and Wang, Y. L. (2022b). Magcnse: Predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinforma.* 23 (1), 1–22. doi:10.1186/s12859-022-04715-w

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, T., Han, Z., Li, H., Zhu, Y., Sun, Z., and Zhu, A. (2018). LncRNA DLEU1 contributes to colorectal cancer progression via activation of KPNA3. *Mol. Cancer* 17 (1), 1–13. doi:10.1186/s12943-018-0873-2

Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., et al. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nat. Commun.* 7 (1), 12791–12813. doi:10.1038/ncomms12791

Peng, L. H., Chen, Y. Q., Ma, N., and Chen, X. (2017). Narrmda: Negative-aware and rating-based recommendation algorithm for miRNA-disease association prediction. *Mol. Biosyst.* 13 (12), 2650–2659. doi:10.1039/c7mb00499k

Peng, L. H., Sun, C. N., Guan, N. N., Qiang, J., and Chen, X. (2018). Hnmda: Heterogeneous network-based miRNA-disease association prediction. *Mol. Genet. Genomics* 293 (4), 983–995. doi:10.1007/s00438-018-1438-1

Peng, L. H., Tian, X. F., Shen, L., Kuang, M., Li, T. B., Tian, G., et al. (2020a). Identifying effective antiviral drugs against SARS-CoV-2 by drug repositioning through virus-drug association prediction. *Front. Genet.* 11, 577387. doi:10.3389/fgene.2020.577387

Peng, L., Shen, L., Liao, L., Liu, G., and Zhou, L. (2020b). Rnmfmda: A microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Front. Microbiol.* 11, 592430. doi:10.3389/fmicb.2020.592430

Peng, L., Wang, C., Tian, X., Zhou, L., and Li, K. (2021). Finding lncRNA-protein interactions based on deep learning with dual-net neural architecture[J]. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* doi:10.1109/TCBB.2021.3116232

Peng, L. H., Shen, L., Xu, J. L., Tian, X. F., Liu, F. X., Wang, J. J., et al. (2021b). Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures[J]. *Sci. Rep.* 11 (1), 1–11.

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022a). Cell-cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: Data resources and computational strategies. *Brief. Bioinform.* 23 (4), bbac234. doi:10.1093/bib/bbac234

Peng, L. H., Tan, J. W., Tian, X. F., and Zhou, L. Q. (2022b). EnANNDeep: An ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models[J]. *Interdiscip. Sci. Comput. Life Sci.*, 1–24. doi:10.1007/s12539-021-00483-y

Qiao, K., Ning, S., Wan, L., Wu, H., Wang, Q., Zhang, X., et al. (2019). LINC00673 is activated by YY1 and promotes the proliferation of breast cancer cells via the miR-515-5p/MARK4/Hippo signaling pathway. *J. Exp. Clin. Cancer Res.* 38 (1), 418–515. doi:10.1186/s13046-019-1421-7

Sarrafzadeh, S., Geranpayeh, L., and Ghafouri-Fard, S. (2017). Expression analysis of long non-coding PCAT-1in breast cancer. *Int. J. Hematol. Oncol. Stem Cell Res.* 11 (3), 185–191.

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., and Sharma, K. K. (2010). Various types and management of breast cancer: An overview. *J. Adv. Pharm. Technol. Res.* 1 (2), 109–126.

Shen, L., Liu, F. X., Huang, L., Liu, G. Y., Zhou, L. Q., and Peng, L. H. (2022). VDA-RWLRLS: An anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Shi, F., Xiao, F., Ding, P., Qin, H., and Huang, R. (2016). Long noncoding RNA highly up-regulated in liver cancer predicts unfavorable outcome and regulates metastasis by MMPs in triple-negative breast cancer. *Arch. Med. Res.* 47 (6), 446–453. doi:10.1016/j.arcmed.2016.11.001

Shi, Z., Luo, Y., Zhu, M., Zhou, Y., Zheng, B., Wu, D., et al. (2019). Expression analysis of long non-coding RNA HAR1A and HAR1B in HBV-induced hepatocullular carcinoma in Chinese patients. *Lab. Med.* 50 (2), 150–157. doi:10.1093/labmed/lmy055

Shi, Z., Zhang, H., Jin, C., Quan, X., and Yin, Y. (2021). A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. *BMC Bioinforma.* 22 (1), 136–220. doi:10.1186/s12859-021-04073-z

Siegel, R., and Naishadhamjemal, D. A. (2013). Cancer statistics, 2013. *Ca. Cancer J. Clin.* 63, 11–30. doi:10.3322/caac.21166

Sledge, G. W., Mamounas, E. P., Hortobagyi, G. N., Burstein, H. J., Goodwin, P. J., and Wolff, A. C. (2014). Past, present, and future challenges in breast cancer treatment. *J. Clin. Oncol.* 32 (19), 1979–1986. doi:10.1200/JCO.2014.55.4139

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief. Bioinform.* 23 (4), bbac266. doi:10.1093/bib/bbac266

Sun, W., Wu, Y., Yu, X., Liu, Y., Song, H., Xia, T., et al. (2013). Decreased expression of long noncoding RNA AC096655.1-002 in gastric cancer and its clinical significance. *Tumour Biol.* 34 (5), 2697–2701. doi:10.1007/s13277-013-0821-0

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., et al. (2017). Risk factors and preventions of breast cancer. *Int. J. Biol. Sci.* 13 (11), 1387–1397. doi:10.7150/ijbs.21635

Tang, W., Lu, G., and Ji, Y. (2022). Long non-coding RNA PCAT1 sponges miR-134-3p to regulate PITX2 expression in breast cancer[J]. *Mol. Med. Rep.* 25 (3), 1–10.

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2021). Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front. Immunol.* 11, 3824. doi:10.3389/fimmu.2020.603615

Tian, X., Shen, L., Gao, P., Huang, L., Liu, G., Zhou, L., et al. (2022). Discovery of potential therapeutic drugs for COVID-19 through logistic matrix factorization with kernel diffusion. *Front. Microbiol.* 13, 13. doi:10.3389/fmicb.2022.740382

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat. Rev. Drug Discov.* 12 (6), 433–446. doi:10.1038/nrd4018

Waks, A. G., and Winer, E. P. (2019). Breast cancer treatment: A review. *Jama* 321 (3), 288–300. doi:10.1001/jama.2018.19323

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26 (13), 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, J., Chen, X., Hu, H., Yao, M., Song, Y., Yang, A., et al. (2021b). PCAT-1 facilitates breast cancer progression via binding to RACK1 and enhancing oxygen-independent stability of HIF-1α. *Mol. Ther. - Nucleic Acids* 24, 310–324. doi:10.1016/j.omtn.2021.02.034

Wang, M. N., You, Z. H., Wang, L., Li, L. P., and Zheng, K. (2021a). Ldgrnmf: LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* 424, 236–245. doi:10.1016/j.neucom.2020.02.062

Wang, N., Zhong, C., Fu, M., Li, L., Wang, F., Lv, P., et al. (2019). Long non-coding RNA HULC promotes the development of breast cancer through regulating LYPD1 expression by sponging miR-6754-5p. *Onco. Targets. Ther.* 12, 10671–10679. doi:10.2147/OTT.S226040

Wu, Y., Yang, X., and Chen, Z. (2019). m6A-induced lncRNA RP11 triggers the dissemination of colorectal cancer cells via upregulation of Zeb1[J]. *Mol. cancer* 18 (1), 1–16.

Xi, Y., and Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Transl. Oncol.* 14 (10), 101174. doi:10.1016/j.tranon.2021.101174

Xing, Z., Park, P. K., Lin, C., and Yang, L. (2015). LncRNA BCAR4 wires up signaling transduction in breast cancer. *RNA Biol.* 12 (7), 681–689. doi:10.1080/15476286.2015.1053687

Xu, W., Zhou, G., Wang, H., Liu, Y., Chen, B., Chen, W., et al. (2020). Circulating lncRNA SNHG11 as a novel biomarker for early diagnosis and prognosis of colorectal cancer. *Int. J. Cancer* 146 (10), 2901–2912. doi:10.1002/ijc.32747

Yang, J., Grünewald, S., and Wan, X. F. (2013). Quartet-net: A quartet-based method to reconstruct phylogenetic networks[J]. *Mol. Biol.* 30 (5), 1206–1217.

Yang, J., Peng, S., and Zhang, B. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases[J]. *Geroscience* 42 (1), 353–372.

Zhang, H., Jiang, L., Zhong, S., Li, J., Sun, D., Hou, J., et al. (2021b). The role of long non-coding RNAs in drug resistance of cancer. *Clin. Genet.* 99 (1), 84–92. doi:10.1111/cge.13800

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021a). Using network distance analysis to predict lncRNA–miRNA interactions. *Interdiscip. Sci. Comput. Life Sci.* 13 (3), 535–545. doi:10.1007/s12539-021-00458-z

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict MicroRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18 (2), 405–415. doi:10.1109/tcbb.2019.2931546

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). Sflln: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Inf. Sci. (N. Y).* 497, 189–201. doi:10.1016/j.ins.2019.05.017

Zhao, Q., Yang, Y., Ren, G., and Fan, C. (2019). Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans. Nanobioscience* 18 (4), 578–584. doi:10.1109/TNB.2019.2922214

Zhou, L., Duan, Q., Tian, X., Tang, J., and Peng, L. H. (2021a). LPI-HyADBS: A hybrid framework for lncRNA-protein interaction prediction integrating feature selection and classification[J]. *BMC Bioinforma.* 22 (1), 1–31.

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021b). LPI-deepGBDT: A multiple-layer deep framework based on gradient boosting decision trees for lncRNA-protein interaction identification. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8

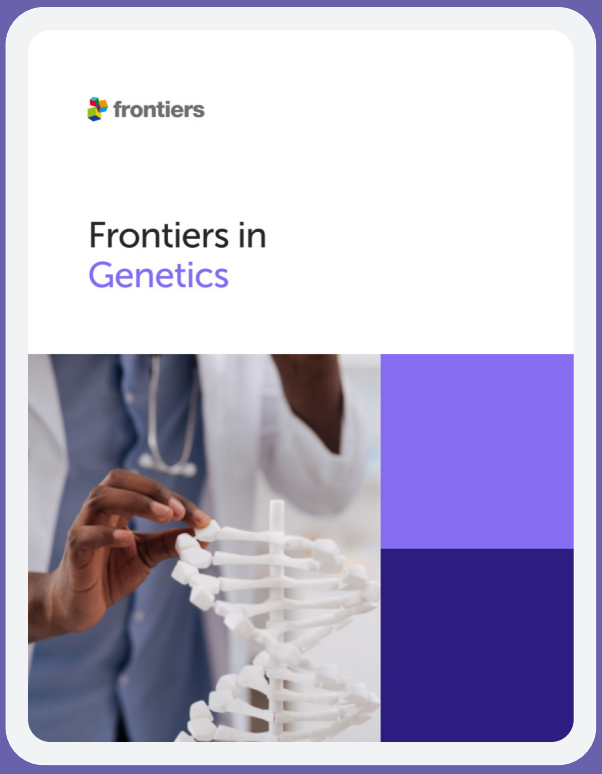
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact