frontiers | Research Topics

Application of network theoretic approaches in biology

Edited by
Rinku Sharma, Josh Clevenger, Mallana Gowdra Mallikarjuna, Sudeepto Bhattacharya and Manish Kumar Pandey

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-3167-9
DOI 10.3389/978-2-8325-3167-9

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Application of network theoretic approaches in biology

Topic editors

Rinku Sharma — Channing Division of Network Medicine, Brigham and Women's Hospital, United States

Josh Clevenger — HudsonAlpha Institute for Biotechnology, United States

Mallana Gowdra Mallikarjuna — Indian Agricultural Research Institute (ICAR), India

Sudeepto Bhattacharya — Shiv Nadar University, India

Manish Kumar Pandey — International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India

Citation

Sharma, R., Clevenger, J., Mallikarjuna, M. G., Bhattacharya, S., Pandey, M. K., eds. (2023). *Application of network theoretic approaches in biology*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3167-9

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.

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 21 July 2023
DOI 10.3389/fgene.2023.1250548

OPEN ACCESS

EDITED AND REVIEWED BY
Richard D. Emes,
Nottingham Trent University,
United Kingdom

*CORRESPONDENCE
Mallana Gowdra Mallikarjuna,
mg.mallikarjuna@icar.gov.in
mgrpatil@yahoo.com

RECEIVED 30 June 2023
ACCEPTED 10 July 2023
PUBLISHED 21 July 2023

CITATION
Mallikarjuna MG, Pandey MK, Sharma R, Clevenger J and Bhattacharya S (2023), Editorial: Application of network-theoretic approaches in biology. *Front. Genet.* 14:1250548. doi: 10.3389/fgene.2023.1250548

Editorial: Application of network-theoretic approaches in biology

Mallana Gowdra Mallikarjuna [1]*, Manish Kumar Pandey[2], Rinku Sharma[1,3], Josh Clevenger[4] and Sudeepto Bhattacharya[5]

[1]Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, India, [2]The International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India, [3]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States, [4]HudsonAlpha Institute for Biotechnology, Huntsville, AL, United States, [5]Department of Mathematics, Shiv Nadar University, Greater Noida, India

KEYWORDS

gene regulatory network, co-expression analyses, network theoretic approaches, biology, protein–protein interaction, systems biology

Editorial on the Research Topic
Application of network theoretic approaches in biology

Introduction

Biological complexity explicitly occurs through non-linear interactions mostly entangled in nature. This complexity comprises many interactions among entities (*viz.*, genes, proteins, metabolites, and species) at various spatial and temporal scales as complex adaptive systems showing characteristic features like self-organisation, modularity, emergence, non-linear interactions, collective response, and adaptation. The theory of complex networks provides an appropriate formal framework for modelling of such complex systems in order to obtain meaningful insights into biological complexity at the local or gene family level (Mallikarjuna et al., 2020; Mallikarjuna et al., 2022) and at the global scale (Sharma et al., 2021). The ocean of biological data generated by high-throughput technologies in the current genomics era have led to the application of various network-theoretic empirical investigations, in which the formal framework is used to obtain meaningful insights into system complexity. Our effort to pool studies on network-theoretic approaches in biology to the understanding of biological complexity has resulted in the compilation of ten research studies in the current Research Topic entitled *Application of network theoretic approaches in biology*, which are broadly categorised and highlighted under the following headings.

Methods and tools

The availability of various user-friendly approaches and software applications has expanded the use of network-theoretic approaches to understand the complex biological

process at individual and systems levels. Four articles in the current Research Topic deal with web tools and methods. A user-friendly web tool for the construction of gene regulatory networks, known as the "Consensus Approach for Gene Regulatory Network Construction" (CAGNC), was developed using the R programming language. CAGNC provides a network file with the edge scores representing significant interactions between each gene pair (Sarkar et al.). A new method, PolyReco, was developed to provide a reference model for processing, with automatic labelling of collinear regions and recognition of polyploidy events (Wang et al.). Lee et al. propose a hybrid intelligence platform, StarGazer (https://github.com/AstraZeneca/StarGazer), which provides an interactive dashboard that allows rapid searching for potential novel drug targets and the use of repositioning strategies via the Streamlit tool. Co-expression studies aid in the discovery of network patterns, functional module identification, and trait-linked marker mining at the system level. Finally, Xiao et al. present a knowledge-injected semi-supervised learning (KISL) method (https://github.com/Mowonhoo/KISL.git) for the identification of outstanding modules in a co-expression network. The KISL approach utilises *a priori* biological information and semi-supervised clustering to solve the issues present in contemporary clustering approaches, such as weighted gene co-expression network analysis (WGCNA).

Plant and microbial systems

Two articles on plant systems are included in the current Research Topic, the first reporting on an investigation of chloride channels (CLCs) and the second presenting a study on the basic pentacysteine6 transcription factor. CLCs are known to regulate the pH of Golgi networks in plants. Here, an effort was made to identify the *CLC* gene family members in the recently sequenced wheat gnome (Fecht-Bartenbach et al., 2007). A total of 23 *CLCs* were identified in the wheat genome and exhibited a functional response to low-nitrogen and salt stresses (Mao et al.). Furthermore, genome-wide co-expression analysis in *Arabidopsis thaliana* indicated a key regulatory BPC6 regulating responses to various abiotic stresses (Zhang et al.).

In the domain of microbial systems, an agglomerative method consisting of complex network analysis and flux balance analyses (FBAs) was employed to examine the energy-intensive aromatic amino acid biosynthesis pathway (tryptophan, tyrosine, and phenylalanine) in 29 free-living bacteria and archaea species. The study identified several common hubs between the connected and the whole-genome networks, showing that the connected pathway network can act as a proxy for the whole-genome network in prokaryotes (Priya and Sinha).

Human systems

At present, the utilization of network-theoretic approaches plays a significant role in unravelling intricate regulatory patterns and hubs within the fields of disease genomics and systems biology in humans (Barabási et al., 2011). In this Research Topic, one such study demonstrated the application of network-theoretic approaches to the identification of herbal medicines that act on immune cell infiltration and immune- and ferroptosis-associated gene expression levels to treat valvular atrial fibrillation. The study concluded that the herbs with rich curcumin content and resveratrol biochemical compounds (*viz*., Rhizoma Curcumae Longae and *Curcuma kwangsiensis*) mitigate myocardial fibrosis to improve valvular atrial fibrillation by modulating the TGFβ/Smad signalling pathway (Jiang et al.). Co-expression analysis is most widely employed to reveal highly synergistic sets of genes, functional modules, and hub genes at a systems level. Here, co-expression analysis of human *Candida* infection revealed the important modules and eight hub genes (*JUN*, *ATF3*, *VEGFA*, *SLC2A1*, *HK2*, *PTGS2*, *PFKFB3*, and *KLF6*) that were found to be enriched with hypoxia, angiogenesis, vasculogenesis, hypoxia-induced signalling, cancer, diabetes, and transplant-related disease pathways mediating host–pathogen interactions (Naik and Mohammed). Furthermore, co-expression studies with four molecular subtypes of breast cancer (*viz*., luminal A, luminal B, Her2, and basal) showed no correlations between copy number variations (CNVs) and the co-expression pattern of the genomic region 8q24.3 (Hernández-Gómez et al.).

Conclusion and perspectives

In conclusion, our Research Topic presents various statistical methods and tools expanding the utility of network-theoretic approaches. Other articles demonstrate the application of various network approaches in developing our understanding of biological phenomena in plant, microbial, and human systems. Nevertheless, for large-scale applications and utilisation of network approaches in biology, there is a further need to undertake some of the following measures, although this is not an exhaustive list. First, more efforts should be made to develop biologist-friendly servers and tools for various types of network analysis, which can allow us to derive meaningful information from the ocean of *omics* data. Second, there is a need for the development of system-specific network approaches in order to understand species interactions from complex ecological and evolutionary perspectives. Finally, validation of major hubs through genetic and in-depth network-theoretic approaches could demonstrate the biological significance of network-theoretic studies in biology.

Author contributions

All the authors of this editorial have made a substantial, direct, and intellectual contribution to the draft and approved it for publication.

Funding

The work was supported by the Network Project on Computational Biology and Agricultural Bioinformatics (Agril.Edn.14(44)/2014-A&P), the Early Career Research Award scheme (ECR/2017/000675), and SERB-NPDF (PDF/2020/001158), Science and Engineering Research Board (SERB), Government of India.

Acknowledgments

We thank all the authors for their contributions to this Research Topic, and we also appreciate all reviewers for their valuable time and constructive comments on the submitted manuscripts.

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

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* 12 (1), 56–68. doi:10.1038/nrg2918

Fecht-Bartenbach, J. V. D., Bogner, M., Krebs, M., Stierhof, Y.-D., Schumacher, K., and Ludewig, U. (2007). Function of the anion transporter *atclc*-d in the *trans*-golgi network. *Plant J.* 50 (3), 466–474. doi:10.1111/j.1365-313X.2007.03061.x

Mallikarjuna, M. G., Thirunavukkarasu, N., Sharma, R., Shiriga, K., Hossain, F., Bhat, J. S., et al. (2020). Comparative transcriptome analysis of iron and zinc deficiency in maize (*Zea mays* L). *Plants* 9 (12), 1812. doi:10.3390/plants9121812

Mallikarjuna, M. G., Sharma, R., Veeraya, P., Tyagi, A., Rao, A. R., Chandappa, L. H., et al. (2022). Evolutionary and functional characterisation of glutathione peroxidases showed splicing mediated stress responses in maize. *Plant Physiol. biochem.* 178, 40–54. doi:10.1016/j.plaphy.2022.02.024

Sharma, R., Upadhyay, S., Bhattacharya, S., and Singh, A. (2021). Abiotic stress-responsive mirna and transcription factor-mediated gene regulatory network in *Oryza sativa*: Construction and structural measure study. *Front. Genet.* 12, 618089. doi:10.3389/fgene.2021.618089

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 November 2021
doi: 10.3389/fgene.2021.745827

A Web Tool for Consensus Gene Regulatory Network Construction

Chiranjib Sarkar[1]*, *Rajender Parsad*[2], *Dwijesh C. Mishra*[2] and *Anil Rai*[2]

[1]ICAR-Indian Agricultural Research Institute, New Delhi, India, [2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

Gene regulatory network (GRN) construction involves various steps of complex computational steps. This step-by-step procedure requires prior knowledge of programming languages such as R. Development of a web tool may reduce this complexity in the analysis steps which can be easy accessible for the user. In this study, a web tool for constructing consensus GRN by combining the outcomes obtained from four methods, namely, correlation, principal component regression, partial least square, and ridge regression, has been developed. We have designed the web tool with an interactive and user-friendly web page using the php programming language. We have used R script for the analysis steps which run in the background of the user interface. Users can upload gene expression data for constructing consensus GRN. The output obtained from analysis will be available in downloadable form in the result window of the web tool.

Keywords: web tool, PHP, fisher's weighted method, consensus approach, gene regulatory network

OPEN ACCESS

Edited by:
Josh Clevenger, HudsonAlpha Institute for Biotechnology, United States

Reviewed by:
Min Li, Central South University, China
Juan Wang, Inner Mongolia University, China

***Correspondence:**
Chiranjib Sarkar
cschiranjib9@gmail.com

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

Received: *22 July 2021*
Accepted: *19 October 2021*
Published: *24 November 2021*

Citation:
Sarkar C, Parsad R, Mishra DC and Rai A (2021) A Web Tool for Consensus Gene Regulatory Network Construction. Front. Genet. 12:745827. doi: 10.3389/fgene.2021.745827

1 INTRODUCTION

Gene regulatory network (GRN) construction is important for understanding complex biological processes. GRNs are represented as the nodes connected with edges where the nodes indicate the genes and each edge indicates the strength of the relationship between the genes. GRNs are constructed from high-dimensional gene expression data containing thousands of genes with expression values at different conditions or experiments. It is a computationally challenging task for analyzing high-dimensional gene expression data in a stepwise workflow. Constructing a GRN from gene expression data involves various steps of data analysis. The steps involved in GRN construction required use of computational techniques. Prior knowledge of the programming language is required for analyzing gene expression data as well as network construction. There are different statistical methods proposed for inferring GRN from high-dimensional expression data, and these methods are implemented using different R packages available in the CRAN depository. Some of the proposed statistical methods are implemented with online web tools. R packages like "BNArray" (Chen et al., 2006), "minet" (Meyer et al., 2008), "dna" (Gill et al., 2014), and "ENA" (Allen, 2014) are implemented based on the Bayesian network, mutual information, differential network analysis methods, and ensemble network aggregation, respectively. Instead of executing a script for each step of GRN construction, web tool development may provide easy accessibility to the user. There are some web tools developed for GRN construction like MIDER (Villaverde et al., 2014), NetworkAnalyst (Zhou et al., 2019), CoExpNetViz (Tzfadia et al., 2016), and GeNeCK (Zhang et al., 2019). For easy accessibility and to provide a more user-friendly procedure, we have introduced a web tool for constructing consensus GRN. It allows users to provide their own gene expression data to get the significant edges and nodes of the GRN. In our web tool, we have used Fisher's weighted method for combining the output of GRN obtained from correlation, principal component regression (PCR), partial least square (PLS), and ridge regression (Sarkar et al., 2020). The data analysis part of

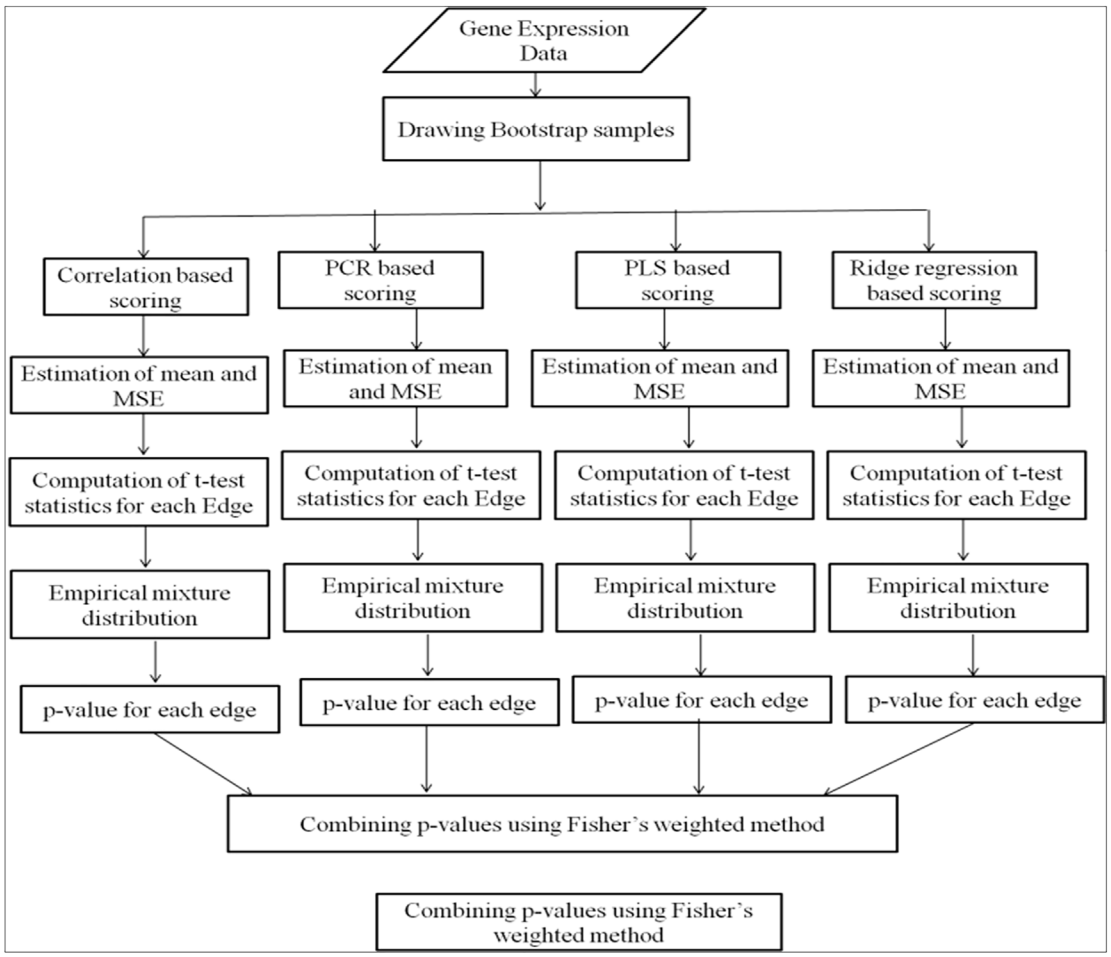


FIGURE 1 | Workflow of web tool for consensus GRN construction.

computing the edge score from correlation, PCR, PLS, and ridge regression has been written in R programming language. The web pages were designed using the HTML and php languages with a user-friendly interface. Users can provide the input file in Microsoft Excel format, and the output of significant edges in each step will also be provided in Excel format.

2 PROGRAM DESCRIPTION AND METHODS

Our developed web tool mainly follows three steps—data uploading, data analysis, and combining the outputs of four methods. The input data of gene expressions can be provided in comma separated value (.csv) file format or in Microsoft excel format containing the list of genes in rows and the conditions or various experiments in columns. The user-uploaded input data are renamed with the date and time of data uploading to avoid repetition in the uploaded file name. The edge scores are computed using four methods, *i.e.*, correlation, PCR, PLS, and ridge regression methods. Probability values are computed for edges from the mixture distribution of edge scores obtained from each method. The probability values are combined using Fisher's weighted method (**Figure 1**). Different steps of analysis are done using R

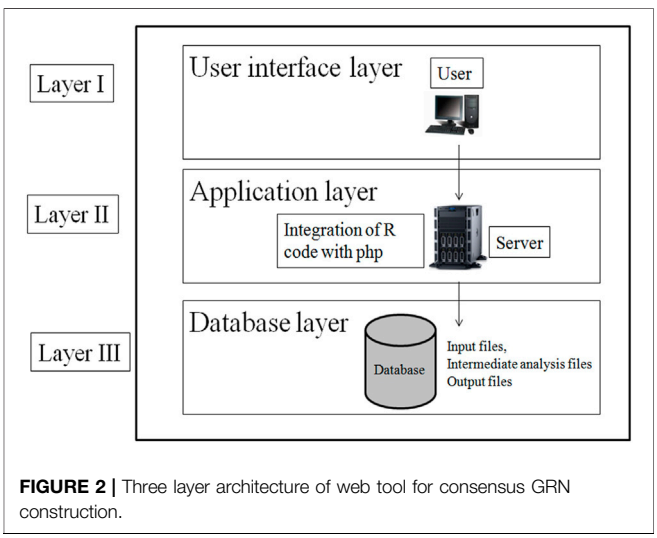


FIGURE 2 | Three layer architecture of web tool for consensus GRN construction.

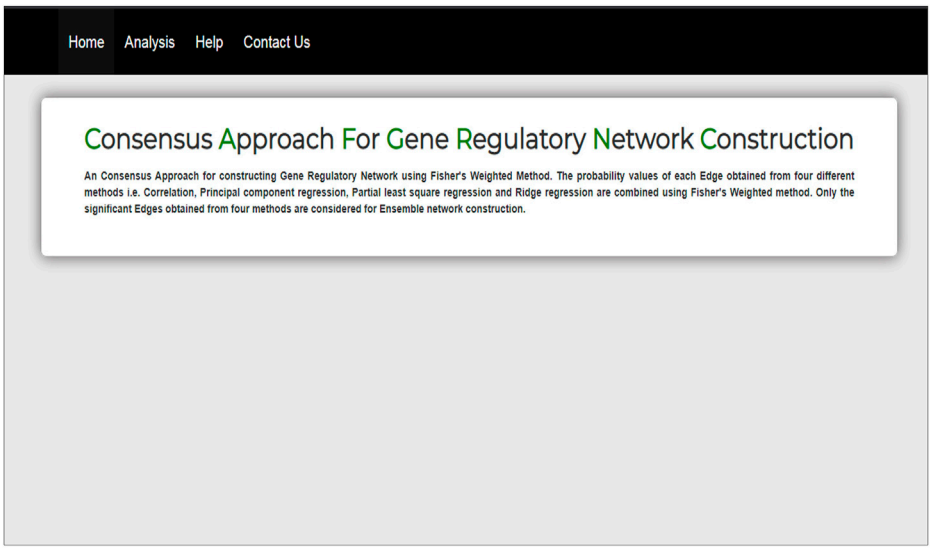


FIGURE 3 | Interface of homepage of web tool.

programming. Few R packages like "dna" and "fdrtool" are used in writing the R script for the analysis. The outputs of the analysis are available in downloadable format in the result tab. Each output file contains the names of the interacting genes and the connectivity score.

2.1 Design of the Web Tool

The web tool has been designed using standard three-layer web architecture (**Figure 2**). The three layers of web architecture are:

- Layer I—user interface layer (UIL)
- Layer II—application layer (APL)
- Layer III—database layer (DBL)

2.1.1 User interface layer

The UIL for the web tool was developed using HTML (Hyper Text Markup Language), CSS, and JavaScript. The UIL consists of forms to interact with users. In UIL, users can upload the gene expression dataset in excel format and download the result file.

2.1.2 Application Layer

The APL of the web tool has been designed using php and R code. The R script for constructing GRN has been integrated with php for analysis of gene expression data. The R script is executed in the background of the web tool which is not visible to the user.

2.1.3 Database Layer

The DBL has been designed as server side file storage. This layer stores the user-provided input file, the intermediate files generated in R script execution, and the final result file. Intermediate files are like files containing a pairwise scoring matrix from four individual methods: file containing *p*-value, fdr value, and F_w score.

The php scripts and R scripts are given in the **Supplementary File**.

2.2 Data Analysis

The expression values of genes in the input data file are considered for computing the connectivity score of each pair of genes using correlation, PCR, PLS, and ridge regression. Bootstrap samples are drawn from the input dataset. The "Sample" function has been used to draw bootstrap samples in R script. For each bootstrap sample, the connectivity score is computed using the four methods. The probability values of pair of genes are computed to measure the statistical significance of the connectivity of gene pair. The probabilities of gene pairs are obtained from the mixture distribution of the connectivity scores of all possible pairs of genes (Efron, 2004).

The correlation-based connectivity score (Gill et al., 2010) is:

$$S_{ik} = \frac{x_i^T x_k}{\sqrt{\left(x_i^T x_i\right)\left(x_k^T x_k\right)}} \tag{1}$$

where x_j and x_k are the standardized expression values of the ith and kth genes, respectively, and S_{ik} is the connectivity score between the i^{th} and k^{th} genes.

The PCR-based connectivity score (Pihur et al., 2008) is:

$$\left[s_{g1}, \ldots, s_{g,g-1}, s_{g,g+1}, \ldots s_{gp}\right]^T = V\hat{\beta}_g \tag{2}$$

where s_{gp} is the connectivity score between the *g*th and *p*th genes and V is the matrix of eigen vectors computed from gene expression values.

The PLS-Based Connectivity Scoring Is

$$\hat{s}_{ik} = \frac{\sum_{l=1}^{v} \hat{\beta}_{il} c_{ik}^{(l)} + \sum_{l=1}^{v} \hat{\beta}_{kl} c_{ik}^{(l)}}{2} \tag{3}$$

where

$$\hat{\beta}_{il} = \left(t_i^{(l)^T} t_i^{(l)}\right)^{-1} t_i^{(l)^T} x_i$$

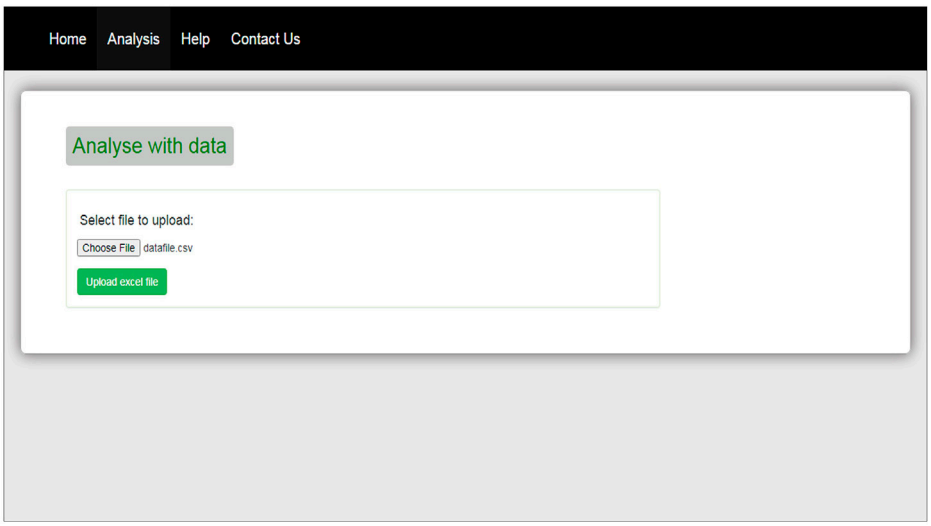


FIGURE 4 | The upload option in analysis tab of web tool.

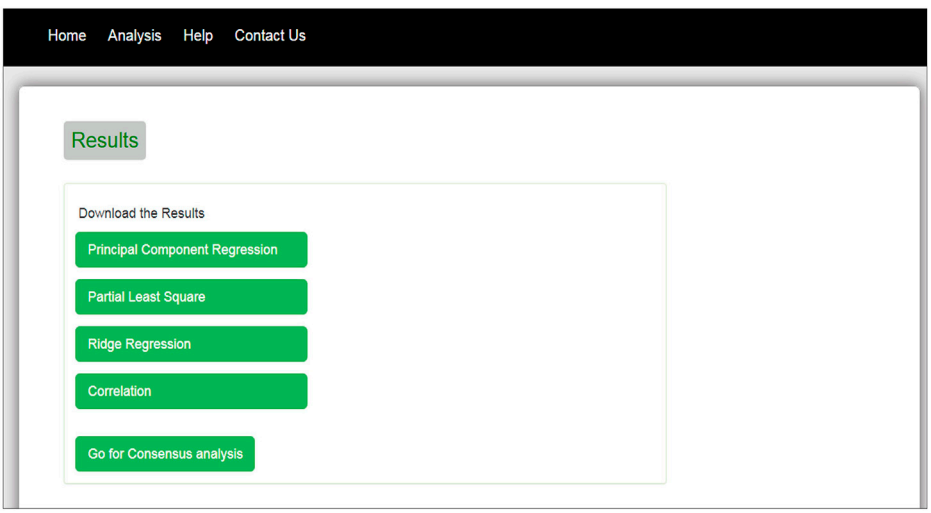


FIGURE 5 | Download tab of results obtained from four methods.

$$t_i^{(l)} = \sum_{k \neq i}^{p} c_{ik}^{(l)} X_k^{(l)}$$

$$c_{ik}^{(l)} = \frac{X^{(l)^T} x_i}{\sqrt{x_i^T X^{(l)} X^{(l)^T} x_i}}$$

The ridge regression-based connectivity score (Gill et al., 2010) is:

$$\left[s_{g,1}, \ldots, s_{g,g-1}, s_{g,g+1}, \ldots, s_{g,p}\right]^T = \left(\tilde{X}_g^T \tilde{X}_g + \lambda I\right)^{-1} \tilde{X}_g x_g \quad (4)$$

where s_{gp} is the connectivity score between the gth and pth genes.

The computation of the connectivity scores was implemented using the "dna" R package.

The mean and standard error (SE) are calculated as (Sarkar et al., 2020):

$$\bar{s}_{ik} = \frac{\sum_{i \neq k}^{n} \sum_{j=1}^{B} s_{ik_j}}{B} \quad (5)$$

$$Se = \frac{1}{\sqrt{B-1}} \sqrt{\sum_{i \neq k}^{n} \sum_{j=1}^{B} \left(s_{ik_j} - \bar{s}_{ik}\right)^2} \quad (6)$$

where B is the number of Bootstrap samples.

The computed t-test statistic is as follows:

$$t = \frac{\bar{s}_{ik}}{Se} \quad (7)$$

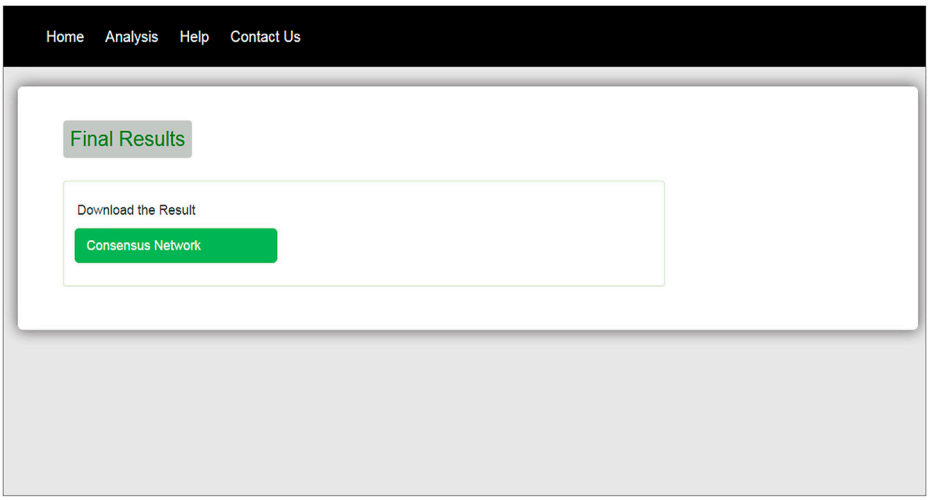


FIGURE 6 | Download tab for final result.

For the correlation-based scoring method, the *t*-test statistic is computed as follows:

$$t = \frac{\bar{s}_{ik}\sqrt{n-2}}{\sqrt{1-\bar{s}_{ik}^2}} \tag{8}$$

The t-statistic values are used for mixture distribution estimation using the "fdrtool" R package (Klaus and Strimmer, 2015).

The *p*-values are combined using Fisher's weighted method (Hedges and Olkin, 2014) following the steps as given in Sarkar et al. (2020):

$$F_w = -2\ln\left(p_1 \times p_2 \times p_3 \times p_4\right) \tag{9}$$

2.3 Implementation

The interface of our web tool has four tabs "Home," "Analysis," "Help," and "Contact Us" (**Figure 3**). The "Analysis" tab has an option to upload gene expression data (**Figure 4**). The input file format of gene expression values should be in comma-separated values (csv) or Excel with genes in rows and conditions in columns. The output files are available in Excel format in the download tab of each method (**Figure 5**). The output file consists of edges, connectivity scores of edges, fdr values, and *p*-values of each edge. The *p*-values of edges computed from the four methods are combined using Fisher's weighted method, and the combined result is available in downloadable format (**Figure 6**). The final output file contains the lists of the significant edges with F-score. The final result file contains the edges for consensus GRN.

DISCUSSION

In this study, a web tool named "Consensus Approach for Gene Regulatory Network Construction" for GRN construction has been developed which provides the network file containing the edge scores of significant interactions of gene pairs. The output file can be visualized using network visualization tools like Cytoscape. In our web tool, we provide the output file containing all the score and statistic values obtained from four individual methods which can also be visualized in Cytoscape. The web tool is easy to use in that it does not require any prior knowledge of R programming and computational steps. It will be very easy for users to construct GRN from gene expression data.

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

CS is carried out the whole work and prepared the manuscript. RP helped in conceptualization, writing-review and editing. DM helped in writing-reviewing and editing. AR helped in conceptualization.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.745827/full#supplementary-material

REFERENCES

Allen, J. D. (2014). *ENA : Ensemble Network Aggregation*. Welthandelsplatz: The Comprehensive R Archive Network (CRAN). R package Version 1.3-0.

Chen, X., Chen, M., and Ning, K. (2006). BNArray: an R Package for Constructing Gene Regulatory Networks from Microarray Data by Using Bayesian Network. *Bioinformatics* 22 (23), 2952–2954. doi:10.1093/bioinformatics/btl491

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing. *J. Am. Stat. Assoc.* 99 (465), 96–104. doi:10.1198/016214504000000089

Gill, R., Datta, S., and Datta, S. (2010). A Statistical Framework for Differential Network Analysis from Microarray Data. *BMC Bioinformatics* 11, 95. doi:10.1186/1471-2105-11-95

Gill, R., Datta, S., Datta, S., and Datta, S. (2014). Dna: An R Package for Differential Network Analysis. *Bioinformation* 10 (4), 233–234. doi:10.6026/97320630010233

Hedges, L. V., and Olkin, I. (2014). *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press.

Klaus, B., and Strimmer, K. (2015). *Fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism*. Welthandelsplatz.

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). Minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics* 9, 461. doi:10.1186/1471-2105-9-461

Pihur, V., Datta, S., and Datta, S. (2008). Reconstruction of Genetic Association Networks from Microarray Data: a Partial Least Squares Approach. *Bioinformatics* 24 (4), 561–568. doi:10.1093/bioinformatics/btm640

Sarkar, C., Parsad, R., Mishra, D. C., and Rai, A. (2020). An Ensemble Approach for Gene Regulatory Network Study in rice Blast. *J. Crop Weed* 16 (3), 1–8. doi:10.22271/09746315.2020.v16.i3.1358

Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. (2016). CoExpNetViz: Comparative Co-expression Networks Construction and Visualization Tool. *Front. Plant Sci.* 6, 1194. doi:10.3389/fpls.2015.01194

Villaverde, A. F., Ross, J., Morán, F., and Banga, J. R. (2014). MIDER: Network Inference with Mutual Information Distance and Entropy Reduction. *PloS one* 9 (5), e96732. doi:10.1371/journal.pone.0096732

Zhang, M., Li, Q., Yu, D., Yao, B., Guo, W., Xie, Y., et al. (2019). GeNeCK: a Web Server for Gene Network Construction and Visualization. *BMC bioinformatics* 20 (1), 12–17. doi:10.1186/s12859-018-2560-0

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: a Visual Analytics Platform for Comprehensive Gene Expression Profiling and Meta-Analysis. *Nucleic Acids Res.* 47 (W1), W234–W241. doi:10.1093/nar/gkz240

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 in Genetics

ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/fgene.2022.846795

Genome-Wide Identification and Functional Characterization of the Chloride Channel TaCLC Gene Family in Wheat (*Triticum aestivum* L.)

Peijun Mao, Yonghang Run, Hanghui Wang, Changdong Han, Lijun Zhang, Kehui Zhan, Haixia Xu* and Xiyong Cheng*

Co-construction State Key Laboratory of Wheat and Maize Crop Science, Collaborative Innovation Center of Henan Grain Crops, College of Agronomy, Henan Agricultural University, Zhengzhou, China

OPEN ACCESS

Edited by:
Manish Kumar Pandey, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India

Reviewed by:
Xiaojun Nie, Northwest A&F University, China
Yuehui Tang, Zhoukou Normal University, China
Parviz Heidari, Shahrood University of Technology, Iran

****Correspondence:***
Haixia Xu
hauxhx@163.com
Xiyong Cheng
xyc634@163.com

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

Received: *31 December 2021*
Accepted: *11 February 2022*
Published: *16 March 2022*

Citation:
Mao P, Run Y, Wang H, Han C, Zhang L, Zhan K, Xu H and Cheng X (2022) Genome-Wide Identification and Functional Characterization of the Chloride Channel TaCLC Gene Family in Wheat (Triticum aestivum L.). Front. Genet. 13:846795. doi: 10.3389/fgene.2022.846795

In plants, chloride channels (CLC) are involved in a series of specific functions, such as regulation of nutrient transport and stress tolerance. Members of the wheat *Triticum aestivum* L. CLC (TaCLC) gene family have been proposed to encode anion channels/transporters that may be related to nitrogen transportation. To better understand their roles, TaCLC family was screened and 23 *TaCLC* gene sequences were identified using a Hidden Markov Model in conjunction with wheat genome database. Gene structure, chromosome location, conserved motif, and expression pattern of the resulting family members were then analyzed. Phylogenetic analysis showed that the TaCLC family can be divided into two subclasses (I and II) and seven clusters (-a, -c1, -c2, -e, -f1, -f2, and -g2). Using a wheat RNA-seq database, the expression pattern of TaCLC family members was determined to be an inducible expression type. In addition, seven genes from seven different clusters were selected for quantitative real-time PCR (qRT-PCR) analysis under low nitrogen stress or salt stress conditions, respectively. The results indicated that the gene expression levels of this family were up-regulated under low nitrogen stress and salt stress, except the genes of TaCLC-c2 cluster which were from subfamily -c. The yeast complementary experiments illustrated that *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* all had anion transport functions for NO_3^- or Cl^-, and compensated the hypersensitivity of yeast GEF1 mutant strain YJR040w (*Δgef1*) in restoring anion-sensitive phenotype. This study establishes a theoretical foundation for further functional characterization of *TaCLC* genes and provides an initial reference for better understanding nitrate nitrogen transportation in wheat.

Keywords: wheat (*Triticum aestivum* L.), CLC, gene family, nitrate nitrogen, functional characterization

INTRODUCTION

Nitrogen is the most important nutrient element for plant growth and development. The composition of more complex compounds such as proteins, nucleic acids and enzymes, which are crucial for plant functioning, is inseparable from nitrogen itself (Lv et al., 2021). Despite its benefits, excessive nitrogen application brings serious negative effects, including ground water nitrate pollution and eutrophication of rivers and lakes. According to available statistical data, 25 million

tons of nitrogen fertilizer are applied annually in China, three times the world average. However, the utilization efficiency of nitrogen fertilizer has only reached 30% of the applied amount (Ju and Zhang, 2017). Given this, improving nitrogen use efficiency is of great significance to crops, as better efficiency will reduce nitrogen fertilizer pollution in surrounding ecosystems.

The main forms of nitrogen absorbed and used by most plants are nitrate nitrogen (NO_3^-) and ammonium nitrogen (NH_4^+). In plants, ammonium nitrogen absorption is mainly regulated by the ammonium transporter (AMT) genes (Li et al., 2017). Conversely, nitrate nitrogen uptake is mainly regulated by four types of NO_3^- transporters: low affinity nitrate transporter NRT1, high affinity nitrate transporter NRT2, chloride channel protein CLC, and slow anion channel related homologue SLAC1/SLAH (Liu R. et al., 2020). Of these, the chloride channel (CLC) gene family is widely distributed across a variety of archaea, microbial fungi, mammals, and plants (Park et al., 2017). The members of the CLC family were first discovered in *Torpedo California* (electric ray fish) by White and Miller in 1979 (White and Miller, 1979). The first family gene was isolated from marine ray (*Torpedo marmorata*) by Jentsch in 1990 and named *CLC-0* (Jentsch et al., 1990). Later, it was discovered that the *CLC* family genes also existed in most plants, including *Arabidopsis thaliana*, *Nicotiana tabacum*, *Oryza sativa* (L.), *Poncirus trifoliata* (L.) Raf., *Zea mays* (L.), and *Glycine max* (Lurin et al., 1996; Diedhiou and Golldack, 2006; Lv et al., 2009; Wang et al., 2015; Wei et al., 2015). Plants CLC proteins play important roles in turgor maintenance, stomatal movement, ion homeostasis, as well as enhancing drought and salt tolerance, and increasing nitrate accumulation (Wei et al., 2015; Zhang et al., 2018; Liu C. et al., 2020).

Structurally, the identified *CLC* genes all have a highly conserved voltage-gated chloride channel (Voltage-gate CLC) domain and two conserved Cystathionine beta synthase (CBS) domains (Xing et al., 2020). At present, the function and classification of Arabidopsis *CLC* family genes have been extensively and deeply studied. Specifically, members of the AtCLC gene family have been divided into two subclasses (I and II) and 7 subfamilies (-a, -b, -c, -d, -e, -f, and -g) (Nedelyaeva et al., 2020). Subclass I is mainly composed of AtCLC-a/-b/-c/-d/-g subfamily, while subclass II is composed of AtCLC-e/-f subfamily (Nedelyaeva et al., 2020). In addition, it was found that subclass I contained the patterns GxGIPE (I), GKxGPxxH (II), and PxxGxLF (III). Comparatively, subclass II did not contain any of the above conserved motifs. When x in the conserved region of the gene GxGIPE (I) is a proline (P, Pro) residue, NO_3^- is preferentially transported. However, when x is a serine (S, Ser) residue, Cl^- is preferentially transported. Past work has also shown that when the x in the conserved region (II) is a conservative-gated glutamate (E, Glu) and the fourth residue in the conserved region (III) is proton glutamate (E, Glu) residue, the protein function is a CLC antiporter rather than a CLC channel (Subba et al., 2021). The *CLC* family genes are not only comprised of channel proteins for anions such as chloride ion and nitrate ion, but also plays vital functional roles in regulating stomatal movement, maintaining both the potential balance and proton gradient in plant cell, transporting and accumulating nutrients in plants (Zifarelli and Pusch, 2010). In Arabidopsis, the *AtCLC-a* gene is located on the vacuolar membrane of plant vacuole and functions as a NO_3^-/H^+ exchanger (De Angeli et al., 2006; Wege et al., 2014). The C-terminus of *AtCLC-a* can be combined with ATP and nitrate/proton alkynol to regulate the specific accumulation of nitrate in the vacuole (De Angeli et al., 2009). *AtCLC-b* is located on the vacuolar membrane as the second vacuolar NO_3^-/H^+ exchanger (Whiteman et al., 2008; von der Fecht-Bartenbach et al., 2010). *AtCLC-c* is involved in stomatal movement and associated with salt tolerance (Harada et al., 2004; Jossier et al., 2010). *AtCLC-d* mediates the transport of anions such as Cl^- or NO_3^-, and regulates the luminal pH of the Golgi network (von der Fecht-Bartenbach et al., 2007). *AtCLC-e* and *AtCLC-f* are related to the thylakoid and Golgi membranes, respectively (Marmagne et al., 2007). *AtCLC-a*, *AtCLC-b*, and *AtCLC-d* play important roles in regulating root elongation (Moradi et al., 2015). *OsCLC1* improves rice drought tolerance and increases yield (Um et al., 2018). The overexpression of the maize gene *ZmCLC-d* in *Arabidopsis thaliana* allows for better tolerance to cold, drought and salt stresses by an increased germination rate, root length, plant survival rate, antioxidant enzyme activities, and a reduced accumulation of Cl^- in transgenic plants (Wang et al., 2015).

Wheat (*Triticum aestivum* L.) serves as the staple food for 30% of global population, and is an important cereal crop with a high demand for nitrogen fertilizer to enable the grain protein accumulation (Zörb et al., 2018). As a preferred nitrate crop, wheat mainly absorbs nitrate nitrogen as its nitrogen source. Studies showed that wheat biomass under conditions of either single ammonium nitrogen or single nitrate nitrogen was lower than when combined ammonium nitrogen and nitrate nitrogen were used (Ijato et al., 2021). Single ammonium nitrogen has a toxic effect on wheat, while nitrate nitrogen alleviates part of this toxic effect (Wang et al., 2016). The *CLC* genes have been shown to be related to the transportation of nitrate. However, the functional roles of the *Triticum aestivum* L. *CLC* (*TaCLC*) genes remain less well known. In this study, 23 wheat *TaCLC* genes were identified by a genome-wide search using released wheat genome data. We analyzed the phylogeny as well as conserved motifs in TaCLC proteins, gene structures, and expression patterns under stress condition. In addition, the part of *TaCLC* genes were functionally characterized in yeast mutant. Our study will lay a preliminary foundation for future research into the functions of *TaCLCs* gene.

MATERIALS AND METHODS

Identification of *Triticum aestivum* L. CLC Gene Family Members in Wheat

Wheat genome (*Triticum_aestivum*.IWGSC.dna.toplevel.fa, 2021), GFF3 file (*Triticum_aestivum*.IWGSC.49.gff3, 2021), and protein sequences (*Triticum_aestivum*.IWGSC.pep.all.fa, 2021) were downloaded from the ensembl plant website (http://plants.ensembl.org/index.html).

The Pfam protein family database (http://pfam.xfam.org/) was used to search the ID number of the CLC gene family and its

distribution across species (Finn et al., 2006). The Pfam accession number ID PF00654 was then used to search the sequences with default parameter settings in the Emsembl Plant database (http://plants.ensembl.org/index.html, *Triticum aestivum* IWGSCv1.1) (Cheng et al., 2018). The Blastp program of TBtools software was used to compare the amino acid sequence of wheat genome (http://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/triticum_aestivum/pep/) with the protein sequence of *AtCLC* genes (**Supplementary Table S1**) and *AtCLC* used as the reference sequence (Chen et al., 2020). Then the *TaCLC* candidate genes obtained by HMM search and Blastp were compared. The ID of the same gene obtained by search retained only the longest transcript sequence. Incomplete gene sequences without either the initial or termination codon and mutation sequences were removed. The remaining sequences were detected using the CDD tool in NCBI (https://www.ncbi.nlm.nih.gov/) and any sequences with incomplete voltage-gated CLC domains were discarded. The final genes were members of the wheat TaCLC gene family. The gene length, encoded amino acid length, intron number, exon number and other biological information were downloaded from the Emsembl Plant database. The TaCLC sequence of wheat was submitted to blast online program in Ensembl Plant to find homologous genes in *Arabidopsis thaliana genome* (TAIR10.1). The GFF annotation information of *TaCLCs* gene was downloaded from the Emsembl Plant database. Subcellular localization of each *TaCLC* gene was predicted using WolfPSORT (https://wolfpsort.hgc.jp/) online tools. The gene structure was analyzed using TBtools software. We identified 33 conserved motifs of TaCLC based on the HMM logo in the Pfam database. The conserved motifs of TaCLC protein were identified by MEME (http://meme-suite.org/). The number of motifs setting was 33 and the motif length was set to 2–200 aa. Default parameters were used for all the programmes unless otherwise stated.

Phylogenetic Analysis of *Triticum aestivum* L. CLC Proteins

The protein sequences of AtCLC, OsCLC and GmCLC were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/), RiceDate (https://www.ricedata.cn/), and SoyBase (https://www.soybase.org/), respectively. All these protein sequences are listed in **Supplementary Table S1**. The Clustal W tool in MAGA7.0 software was used to compare the reported CLC protein sequences of *Arabidopsis thaliana* (At), *Oryza sativa* L. (Os), *Glycine max* (Gm) and the newly identified wheat TaCLC protein sequences (Kumar et al., 2016). A phylogenetic tree was constructed using the neighbor-joining (NJ) method with the bootstrap replicates setting parameter of 1000.

Physical Location on Chromosomes, Collinearity Analysis, and Transmembrane Structure Prediction of Encoding Protein

The physical, chromosomal position of each *TaCLC* gene was obtained from the IWGSC RefSeq V1.0 database (https://urgi.versailles.inra.fr/). The physical map was plotted using MapInspect software according to the starting position and chromosome length of each gene. One Step MCScanX program of TBtools software was used to compare and integrate the whole genome sequence and gene structure annotation information of wheat and the E-value parameter setting was 10E-05 (Chen et al., 2020). In order to further predict the interspecific evolution mechanism of TaCLC family members, we also used TBtools software to integrate and compare the whole wheat genome sequence and gene structure annotation with *Arabidopsis thaliana*, *Oryza sativa* L. and *Triticum dicoccoides*. The data were from Emsembl Plant database (http://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/) and the E-value parameter setting was 10E-05. Dual Systeny Plot program and Circle Gene View of TBtools software were used to visualize the collinearity results of *TaCLC* gene sequences. PROTTER (http://wlab.ethz.ch/protter/start/) online tools were used to analyze the transmembrane structure of *TaCLCs* gene-encoded proteins.

Expression Prediction of *Triticum aestivum* L. CLC Genes in Various Wheat Tissues

Using the wheat expression database ExpWheat (https://wheat.pw.usda.gov/WheatExp/), the expression patterns of *TaCLC* family genes in wheat plant roots, stems, leaves, panicles, grains, and other tissues were obtained. The expression of genes at different stages was also predicted. A heat map of *TaCLCs* gene expression was drawn by Heml software.

Plant Materials, Growth Conditions, and Stress Conditions

Wheat (*Triticum aestivum* L. cv. Yunong 804) seeds were planted for 5 days in culture dishes containing water in a greenhouse at 22°C with a 16 h light/8 h dark photoperiod. The 5-day-old wheat seedlings were transplanted into a nutrient solution (**Supplementary Table S2**), which was replaced every 3 days until the seedlings reached the three-leaf stage. The plantlets were then subjected to low nitrogen (0.4 mM NH_4NO_3) or salt stress (100 mM NaCl) treatment. Roots and shoots were collected at 0, 2, 6, 12, and 24 h after treatment with 0.4 mM NH_4NO_3. Whole plantlets were collected at 0, 1, 3, 6, 9, 12, and 24 h after treatment with 100 mM NaCl. Each treatment was conducted using three independent biological replicates and samples were collected from three plants for each treatment at each replication. All plant samples were immediately collected and frozen in liquid nitrogen prior to storing at −80°C for further RNA isolation.

Quantitative Real-Time PCR Validation

To understand the expression pattern of *TaCLC* genes under short-term stress of different environment conditions, we selected one gene from each cluster of the TaCLC gene family. In total, seven genes were screened to validate expression levels. Specific fluorescence quantitative primers (**Supplementary Table S3**) were designed using Primer 5 software (http://www.premierbiosoft.com/index.html). Total RNA was extracted from different treatment groups using TransZOL (TransGen

Biotech, Beijing, China) according to the manufacturer's instructions. Then, cDNA was synthesized using PrimeScript™ RT reagent Kit with gDNA Eraser (Takara, Dalian, China). Reverse transcription cDNA was used as template for amplification with Hieff UNICON® qPCR SYBR Green Master Mix (YEASEN Biotechnology, Shanghai, China) in Quantstudio™5 (Thermo Fisher, Shanghai, China). Quantitative real-time PCR (qRT-PCR) thermocycling conditions were as follows: 95°C for 5 min, followed by 40 cycles of 95°C for 10 s and 60°C for 30 s, and a final extension at 72°C for 5 min. The expression levels of six genes (TaCLC-c2 cluster genes were removed because they were not expressed and the relative expression cannot be calculated) selected from the TaCLC gene family under different stress conditions were calculated using the $2^{-\Delta\Delta ct}$ method. The wheat *β-actin* gene was used as an internal reference control. The average of three independent biological replicates was used for all data analysis.

Functional Complementation Experiment in Yeast

Yeast mutant strain YJR040w (*Δgef1*; MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; YJR040w: kanMX4) lacks the only CLC family protein—GEF1—in *Saccharomyces Cerevisiae* and is sensitive to low nitrogen medium. To further illustrate the function of *TaCLC* genes, we used the EUROSCARF (http://www.euroscarf.de/index.php?name=News) yeast mutant YJR040w that heterologously expressed the TaCLCs. The primer sequences of *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* used for gene amplification are listed in **Supplementary Table S3**. The selected *TaCLCs* gene was cloned into the yeast expression vector p416 by ClonExpress® II One Step Cloning Kit (Vazyme Biotech Co., Ltd., Nanjing, China). The constructed recombinant plasmid was then transformed into YJR040w using the lithium acetate transformation method and uniformly plated on a SD/-Ura Broth (Coolaber Technology Co., Ltd., Beijing, China) solid plate. The monoclonal was selected and verified by PCR with p416-F/p416-R (**Supplementary Table S3**). The monoclonal containing the target band was transferred to liquid SD/-Ura medium and cultured to OD_{600} = 0.6. These initial yeast cultures were used to prepare a series of diluents (10^{-1}). From each gradient, 5 μl samples of each diluted culture were plated into YTD medium (1% yeast extract/2% tryptone/2% dextrose), YTD medium supplemented with 1 M KCl, 1 M NaCl, or 1 M KNO_3, respectively. The plates were incubated at 30°C for 3–5 days and photographed.

RESULTS

Identification of *Triticum aestivum* L. CLC Family Genes in Wheat

We used the voltage-gated chloride channel protein conserved domain PF00654 as a search sequence in conjunction with a wheat genome database to perform alignment and to remove incomplete conserved domains. A total of 23 non-redundant *TaCLC* candidate gene sequences were identified and named based on the orthologous relationships of the rice family genes (**Table 1**). Information about the chromosome on which the gene was located was indicated by chromosomal arm symbols on the gene/protein name. When multiple gene sequences of the wheat TaCLC family members formed the same cluster as a certain gene as that seen in rice, a number was added after the gene name to distinguish these two genes, such as *TaCLC-c1-3DS-1* and *TaCLC-c1-3DS-2*.

There were five *TaCLC* family members, all of which had homologous genes on the A, B, and D genomes including TaCLC-a, -c2, -f1, -f2, and -g2. However, TaCLC-c1 cluster and TaCLC-e cluster members contained only two homologous sequences. Comparatively, the TaCLC-c1 cluster member had one gene on the A genome and two genes on the D genome. The amino acid length of the TaCLC proteins ranged from 573 aa (TaCLC-g2-2AL) to 822 aa (TaCLC-c1-3DS-2). TaCLC proteins contained 8–12 transmembrane regions, of which 14 TaCLC proteins had 11 transmembrane regions and one TaCLC proteins (TaCLC-c1-3DS) had the least transmembrane regions with 8. The results of the visualized protein topology diagram are shown in **Supplementary Figure S1**.

Phylogenetic and Structural Analyses of *Triticum aestivum* L. CLC Proteins

In this study, the CLC family protein sequences of 23 *Triticum aestivum* (TaCLC), 7 *Arabidopsis thaliana* (AtCLC), 8 *Oryza sativa* (OsCLC), and 8 *Glycine max* (GmCLC) were aligned to construct a phylogenetic tree. This was performed using the neighbor-joining method with 1,000 bootstrap replication (**Figure 1**). According to the clustering criteria of rice and Arabidopsis, wheat TaCLC proteins were divided into seven clusters including TaCLC -a, -c1, -c2, -e, -f1, -f2 and -g2. Among these, the cluster -a was the largest group and contained six genes; comparatively, cluster -e was the smallest and only contained two members. The remaining five clusters all contained only three genes.

Structural analysis showed that cluster -a, cluster -c1, cluster -c2, and cluster -g2 all contained typical conserved regions GxGIPE (I), GKxGPxxH (II), and PxxGxLF (III). Cluster -e and cluster -f1/-f2 did not have this typical structure (**Supplementary Figure S2**). These results indicated that wheat *TaCLC* genes were divided into subclass I and subclass II according to the conserved structures of the *CLC* genes.

Chromosomal Distribution of *Triticum aestivum* L. CLC Genes

TaCLC genes were distributed on 12 chromosomes of wheat (**Figure 2**). The physical locations of *TaCLC* genes are shown in **Table 1**. The 23 *TaCLC* genes were unevenly distributed on chromosomes, of which 3A, 3D, 6A, 6B, and 6D chromosomes had three gene members. The distribution of the *TaCLC* genes in the A (8), B (7) and D (8) subgenomes was relatively balanced. This was consistent with the fact that nearly half of the family members have three homologous genes.

TABLE 1 | Main information of 23 wheat TaCLC genes.

Gene Name	Ensembl ID	Gene length	Amino acid length	Intron No.	Exon No.	Location		SL	TMS	Chromosome	Accession of Arabidopsis
						Start	End				
TaCLC-a-6AS-1	TraesCS6A02G098500.2	3,036	796	3	4	65681325	65684847	PM	11	6A	*AT3G27170 (AtCLC-b)*
TaCLC-a-6AS-2	TraesCS6A02G098600.1	2,813	778	3	4	65797706	65800777	PM	11	6A	*AT5G40890 (AtCLC-a)*
TaCLC-a-6BS-1	TraesCS6B02G126400.1	3,161	787	4	5	121811989	121815556	PM	11	6B	*AT3G27170 (AtCLC-b)*
TaCLC-a-6BS-2	TraesCS6B02G126800.1	2,870	784	3	4	122290496	122293648	PM	11	6B	*AT3G27170 (AtCLC-b)*
TaCLC-a-6DS-1	TraesCS6D02G084300.1	3,146	787	4	5	49218043	49221583	PM	11	6D	*AT3G27170 (AtCLC-b)*
TaCLC-a-6DS-2	TraesCS6D02G084000.2	2,735	784	3	4	48866564	48869557	PM	11	6D	*AT3G27170 (AtCLC-b)*
TaCLC-c1-3AS	TraesCS3A02G125300.1	3,435	805	7	8	100885506	100893368	PM	12	3A	*AT5G33280 (AtCLC-c)*
TaCLC-c1-3DS-1	TraesCS3D02G126600.1	3,913	805	7	8	84568670	84582503	PM	12	3D	*AT5G33280 (AtCLC-c)*
TaCLC-c1-3DS-2	TraesCS3D02G126700.1	1,722	573	3	4	84587555	84589918	PM	8	3D	*AT5G33280 (AtCLC-c)*
TaCLC-c2-3AL	TraesCS3A02G390100.1	2,919	795	7	8	638325536	638329362	PM	11	3A	*AT5G33280 (AtCLC-c)*
TaCLC-c2-3B	TraesCS3B02G418700.1	2,964	795	7	8	655435367	655439266	PM	11	3B	*AT5G33280 (AtCLC-c)*
TaCLC-c2-3DL	TraesCS3D02G379600.1	2,862	794	7	8	496503737	496507536	PM	11	3D	*AT5G33280 (AtCLC-c)*
TaCLC-e-3AL	TraesCS3A02G253600.3	2,838	717	6	7	474962330	474970320	PM	11	3A	*AT4G35440(AtCLC-e)*
TaCLC-e-3B	TraesCS3B02G285500.1	2,415	717	6	7	457013924	457026915	PM	11	3B	*AT4G35440(AtCLC-e)*
TaCLC-f1-6AL	TraesCS6A02G283600.3	3,223	781	8	9	514703319	514709411	PM	9	6A	*AT1G55620 (AtCLC-f)*
TaCLC-f1-6BL	TraesCS6B02G312100.1	2,136	711	7	8	559340524	559348729	PM	9	6B	*AT1G55620 (AtCLC-f)*
TaCLC-f1-6DL	TraesCS6D02G264100.1	3,246	785	8	9	372824290	372830240	PM	9	6D	*AT1G55620 (AtCLC-f)*
TaCLC-f2-7AS	TraesCS7A02G240700.2	2,776	743	8	9	216343576	216349884	PM	9	7A	*AT1G55620 (AtCLC-f)*
TaCLC-f2-7BS	TraesCS7B02G136300.1	2,785	743	8	9	168313998	168321034	PM	10	7B	*AT1G55620 (AtCLC-f)*
TaCLC-f2-7DS	TraesCS7D02G239700.3	2,454	764	8	9	204246408	204253040	PM	10	7D	*AT1G55620 (AtCLC-f)*
TaCLC-g2-2AL	TraesCS2A02G517500.3	2,097	822	8	9	740847366	740855721	PM	11	2A	*AT5G33280 (AtCLC-g)*
TaCLC-g2-2BL	TraesCS2B02G546000.1	3,033	817	8	9	742813858	742822299	PM	11	2B	*AT5G33280 (AtCLC-g)*
TaCLC-g2-2DL	TraesCS2D02G519000.2	3,035	818	8	9	608915455	608923751	PM	11	2D	*AT5G33280 (AtCLC-g)*

SL, subcellular location; TMS, transmembrane segments; PM, plasma membrane; Accession of Arabidopsis, Ensembl Plant database accession number of TaCLC genes' ortholog in Arabidopsis thaliana.

There were tandem repeat events in the cluster -a and cluster -c1, and they appeared on chromosomes in the form of tandem gene clusters. For example, *TaCLC-a-6AS-1* and *TaCLC-a-6AS-2*, *TaCLC-a-6BS-1* and *TaCLC-a-6BS-2*, *TaCLC-a-6DS-1* and *TaCLC-a-6DS-2*, and *TaCLC-c1-3DS-1* and *TaCLC-a-3DS-2* all formed different tandem gene clusters on chromosomes 6A, 6B, 6D, and 3D, respectively. This may also be the reason the number of TaCLC members was more than that of other species.

Gene Collinearity Analysis

The results of intraspecies collinearity analysis showed that the 23 wheat *TaCLC* genes constituted 15 pairs of collinear genes (**Figure 3**). Among them, four genes including *TaCLC-a-6AS-2*, *TaCLC-a-6BS-2*, *TaCLC-a-6DS-2*, and *TaCLC-c1-6DS-2* had no collinearity with other *TaCLC* family genes and had tandem repeats. There was no collinearity between homologous genes *TaCLC-f1-6BL* and *TaCLC-f1-6DL*, between *TaCLC-a-6BS-1* and *TaCLC-a-6DS-1*, and no gene of cluster -c1 on the 3B chromosome. This also indicated that

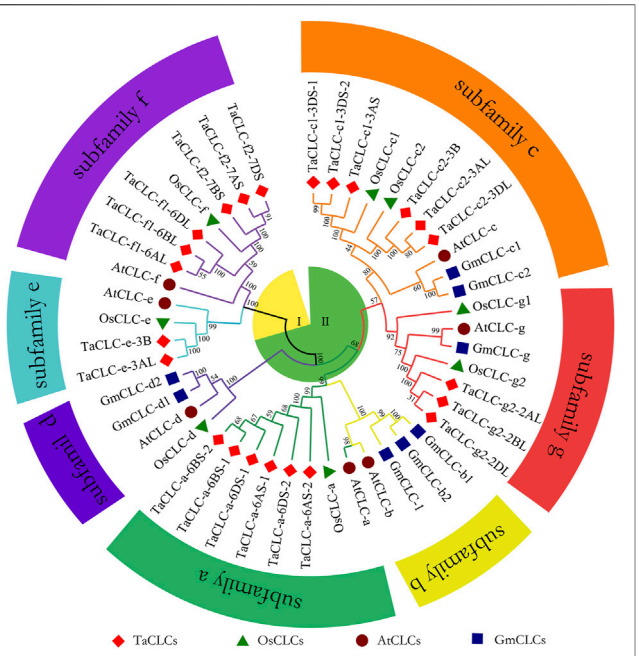


FIGURE 1 | The phylogenetic tree of the CLC family of wheat, rice, soybean and Arabidopsis. The proteins belonging to each of four species are represented by different shapes and colors. The TaCLC family proteins were divided into seven subfamilies (-a, -b, -c, -d, -e, -f, and -g) and two subclasses (I and II), which are indicated with different shapes and colors of lines. The CLC proteins loci of wheat, rice, soybean, and Arabidopsis are listed in **Supplementary Table S1**. Ta, *Triticum aestivum*; Os, *Oryza sativa*; Gm, *Glycine max*; At, *Arabidopsis thaliana*.

homologous fragment loss occurred simultaneously during evolution of the genes.

The results of collinearity analysis among species showed that *TaCLC* genes had collinearity with genes in genome of *Arabidopsis thaliana*, *Oryza sativa* L. and *Triticum dicoccoides* (**Figure 4**; **Supplementary Table S4**). The number of collinear gene pairs was different between species. Only two *TaCLC* genes between the genome of *Triticum aestivum* L. and *Arabidopsis thaliana* had collinearity, namely *TaCLC-f2-7BS* and *TaCLC-f2-7DS* (**Figure 4A**). The number of collinearity gene pairs in *Triticum aestivum* L. and *Oryza sativa* L. was 13, and the collinearity genes were mainly cluster -c2 and subfamily -e, -f, and -g (**Figure 4B**). *Triticum aestivum* L. and *Triticum dicoccoides* had 35 collinear gene pairs, which was the largest number (**Figure 4C**). Among them, 19 genes in *TaCLC* had collinearity with *Triticum dicoccoides* genes. There were 7, 6 and 6 *TaCLC* genes in the three subgenomes of A, B, and D of wheat, respectively, which had a collinearity relationship with the *Triticum dicoccoides* genes. This illustrated that the distribution of these 19 genes in the genome was uniform, and they were in a relatively conservative state during the evolution process from *Triticum dicoccoides* to *Triticum aestivum* L. Both *TaCLC-f2-7BS* and *TaCLC-f2-7DS* existed in the collinear gene pairs between *Triticum aestivum* L. and *Arabidopsis thaliana*, *Oryza sativa* L. and *Triticum dicoccoides*, indicating that these two genes may be ubiquitous in monocotyledons and dicotyledons. They formed before species differentiation and evolved for a longer time.

Triticum aestivum L. CLCs Gene Structure and Conservative Motif Analysis

TBtools software was used to compare the CDS sequences of 23 *TaCLCs* with their corresponding genome sequences. A visual structure map of the *TaCLC* genes was then obtained (**Figure 5A**). *TaCLC* genes contained 3–8 introns and 4–9 exons. The introns and exons of *TaCLC-a-6AS-1*, *TaCLC-a-6AS-2*, *TaCLC-a-6BS-2*, *TaCLC-a-6DS-2*, and *TaCLC-c1-3DS-2* had the smallest number among all identified *TaCLC* genes with 3 introns and 4 exons. *TaCLC-f1-6AL*, *TaCLC-f1-6DL*, *TaCLC-f2-7AS*, *TaCLC-f2-7BS*, *TaCLC-f2-7DS*, *TaCLC-g2-2AL*, *TaCLC-g2-2BL*, and *TaCLC-g2-2DL* had the largest number, with 8 introns and 9 exons. Exon and intron number among the same cluster of homologous genes also varied. For example, *TaCLC-f1-6AL* contained 8 introns and 9 exons, whereas *TaCLC-f1-6BL* contained 7 introns and 8 exons.

Using online tools to analyze conservative protein motifs, the motif structure of the same cluster of homologous genes was determined to be basically similar (**Figure 5B**). These 33 motifs were evenly distributed in 7 clusters, and the number of motifs in each cluster was also similar (**Supplementary Table S5**). The conserved motifs of TaCLC proteins were divided into two categories. The relatively conserved motifs in cluster -a and cluster -c1, -c2, and -g2 were different from those in cluster -e and cluster -f1 and -f2, indicating that the conserved motifs of the whole TaCLC gene family could be divided into two cases.

Predictive Analysis of *Triticum aestivum* L. CLCs Gene Expression at Different Developmental Stages of Wheat

The expression patterns of *TaCLC* genes in wheat tissues (roots, stems, leaves, spikes, and grains) at different developmental stages were analyzed using available wheat RNA-seq databases (**Figure 6**). The developmental stages were as follows: z10 (seedling), z13 (three leaves), z23 (three tillers), z30 (spike at 1 cm), z32 (two nodes), z39 (meiosis), z65 (anthesis), z71 (2 DAA), z75 (14 DAA) and z85 (30 DAA). Of the 23 genes, the expression data on 21 genes were obtained from RNA-seq databases. The only missing genes were *TaCLC-c1-3DS-2* and *TaCLC-c2-3DL*. *TaCLC-c2-3AL*, *TaCLC-c2-3B*, *TaCLC-a-6BS-1*, *TaCLC-a-6DS-1*, and *TaCLC-f1-6BL* had low expression level in all detected periods and tissues. This was especially true for the z75 (14 DAA) period, which had very low gene expression. The expression levels of *TaCLC-c1-3AS* in roots of z10 (seedling stage), z13 (three leaves stage), z39 (meiosis stage), and leaves of z71 (2 DAA stage), as well as *TaCLC-c1-3DS-1* in leaves of z13 (three-leaves stage) were all higher in all examined *TaCLC* genes.

TaCLC-a-6BS-2 had the highest expression level in cluster -a, while all the genes of cluster -c2 were not expressed. The expression level of *TaCLC-c1-3AS* was higher than that of *TaCLC-c1-3DS-1* during all the stages except for the leaves of z23 (three tillers) in cluster -c1. In cluster -e, the expression levels

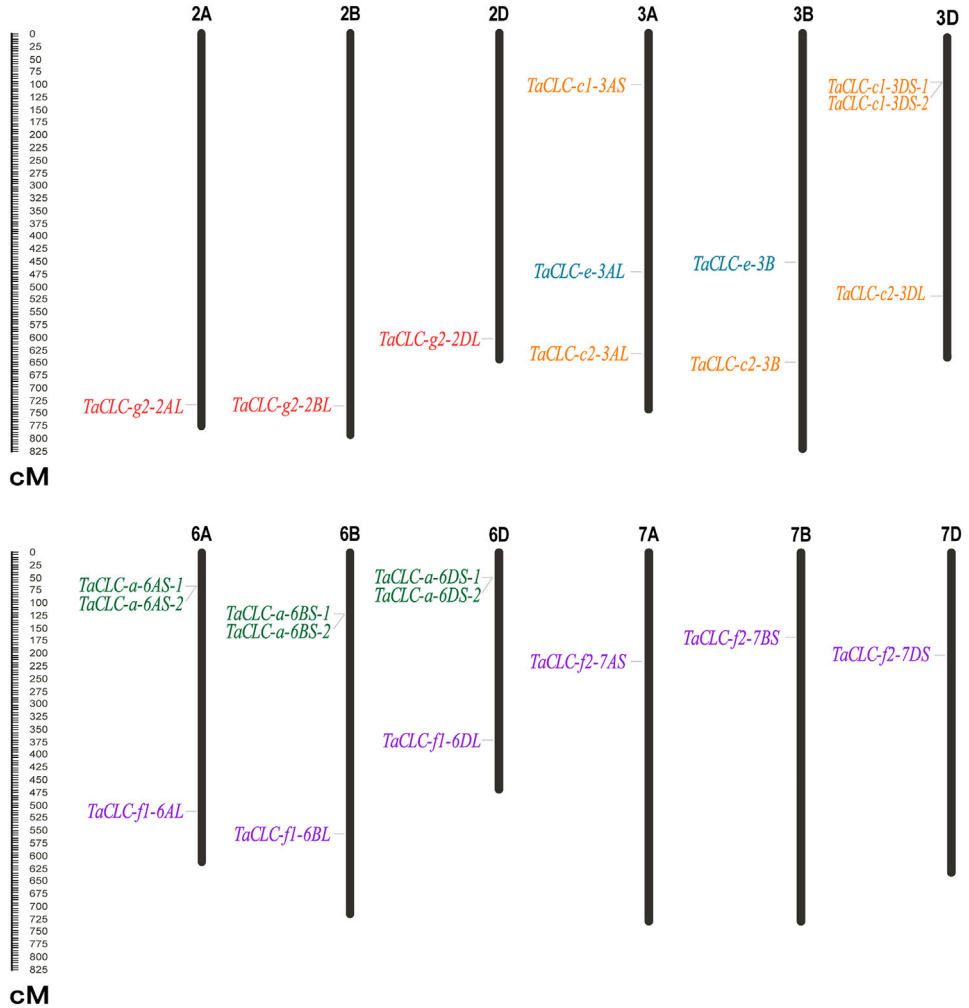


FIGURE 2 | Distribution of *TaCLC* genes on chromosomes. The names of each chromosome are shown above. The gene names are indicated on the left and the starting positions are on the right side of the chromosomes. The *TaCLC* genes in each subfamily are specified by the same color. The chromosome lengths are shown in Mb (Millions of bases).

of *TaCLC-e-3AL* and *TaCLC-e-3B* were similar in all stages and tissues except in leaves of z10 (seedling) and z13 (three leaves). Among the three cluster -f1 genes, the expression level of *TaCLC-f1-6AL* was the highest and expressed in all tissues at each stage, while the expression level of *TaCLC-f2-7BS* was the highest among the three genes of cluster -f2. In cluster -g2, the expression levels of *TaCLC-g2-2AL* and *TaCLC-g2-2DL* genes in leaves of z23 (three tillers) and z71 (2 DAA) stages were higher than those of *TaCLC-g2-2BL*.

Expression Analysis of *Triticum aestivum* L. CLCs Gene Under Low Nitrogen Stress or Salt Stress

According to the predictive analysis of gene expression heat map, we selected *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, *TaCLC-e-3AL*, *TaCLC-f1-6AL*, *TaCLC-f2-7BS*, *TaCLC-g2-2DL*, and *TaCLC-c2-3AL* from each cluster to analyze the expression patterns under low nitrogen stress or salt stress using qRT-PCR. The gene expression of this family of genes was determined to be up-regulated under conditions of low nitrogen stress or salt stress except for *TaCLC-c2-3AL* (**Figure 7**). This finding was consistent with the data from the wheat RNA-seq database that the genes of cluster -c2 were not expressed across all stages.

The expression patterns of different genes in different parts were also diverse under low nitrogen stress (**Figures 7A,B**). Among them, the expression of the *TaCLC-a-6AS-1* gene was at its highest level after 2 h in the shoot and at its highest level at 6 h in the root tissues when compared with the untreated tissues. Under conditions of low nitrogen stress, expression patterns in the shoot tissues were mainly divided into three types (**Figure 7A**). The expression levels of the other five genes were the highest after 6 h of low nitrogen stress, except for the *TaCLC-a-6AS-1* gene. The expression patterns of *TaCLC-e-3AL*, *TaCLC-f1-6AL*, and *TaCLC-f2-7BS* genes were all down-

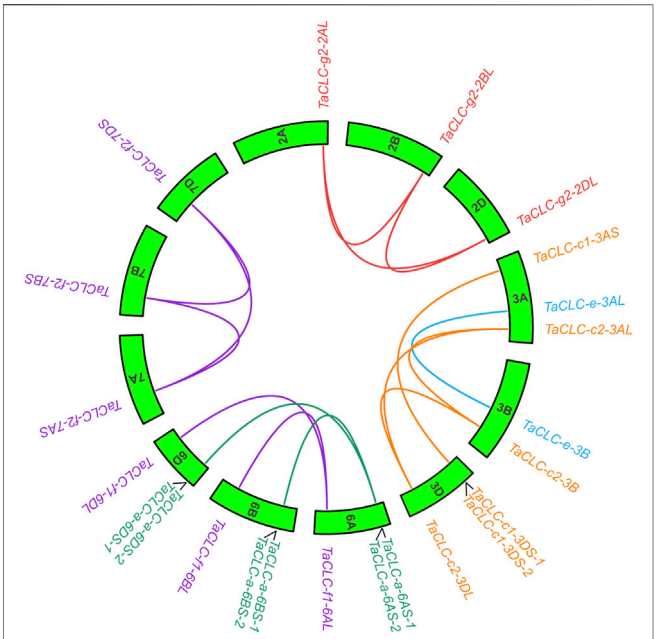


FIGURE 3 | Collinearity analysis of *TaCLC* genes. *TaCLC* genes in five subfamilies (-a, -c, -e, -f, and -g) are represented by different colors of lines. Wheat chromosomes are represented by green boxes with chromosome names inside.

regulated and then up-regulated. Comparatively, the expression patterns of the *TaCLC-c1-3AL* and *TaCLC-g2-2AL* genes were slowly up-regulated and then down-regulated. The expression pattern of the *TaCLC-a-6AS-1* gene was completely different from that of the other five genes, indicating that *TaCLC-a-6AS-1* belonged to a gene type that was stress-induced and involved rapid up-regulation. Notably, its expression level remained higher than when without stress. In the root parts and under conditions of low nitrogen stress, the expression patterns of the other five genes were first up-regulated and then down-regulated. This occurred across all genes except for the *TaCLC-g2-2DL* gene. The expression of *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, *TaCLC-e-3AL*, and *TaCLC-f2-7BS* all reached their highest respective levels at 6 h after low nitrogen stress treatment (**Figure 7B**). Only *TaCLC-f1-6AL* reached its highest expression level at 2 h; the expression pattern of the *TaCLC-g2-2DL* gene was first down-regulated, up-regulated, and finally down-regulated.

As shown in **Figure 7C**, all six genes reached their highest respective expression levels after 24 h salt stress treatment. The relative expression levels of *TaCLC-c1-3AS*, *TaCLC-e-3AL*, and *TaCLC-f2-7BS* were all higher than those of the other genes. The other five genes showed a down-regulated expression pattern under transient salt stress, and began to show an upward tendency at approximately 6 h treatment, except for the *TaCLC-c1-3AS* gene. Three genes—*TaCLC-a-6AS-1*, *TaCLC-e-3AL*, and *TaCLC-g2-2DL*—were all down-regulated, and finally up-regulated across 6–24 h. The *TaCLC-f1-6AL* and *TaCLC-f2-7BS* genes were both down-regulated to their lowest levels at 3 h and then up-regulated to their highest level at 24 h. The expression of *TaCLC-c1-3AS* did not change much before 1 h, and reached its lowest level at 3 h. After a brief up-regulation

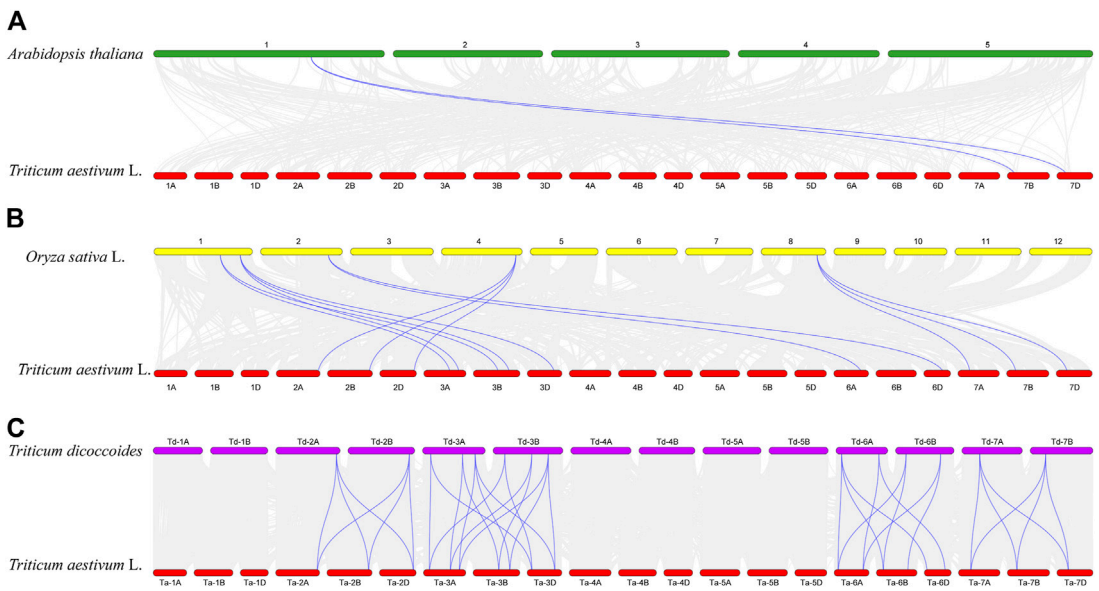


FIGURE 4 | Collinearity analysis of *TaCLC* genes between *Triticum aestivum* L. and *Arabidopsis thaliana* **(A)**, *Oryza sativa* L. **(B)**, and *Triticum dicoccoides* **(C)**. The chromosomes of wheat, Arabidopsis, rice, and emmer are expressed in red, green, yellow and purple, respectively. The gray line represents all the collinearity gene pairs between the two species, and the blue bright line represents the collinearity gene pair between the *TaCLC* genes of wheat and the corresponding species. The names of each chromosome are shown above and below.

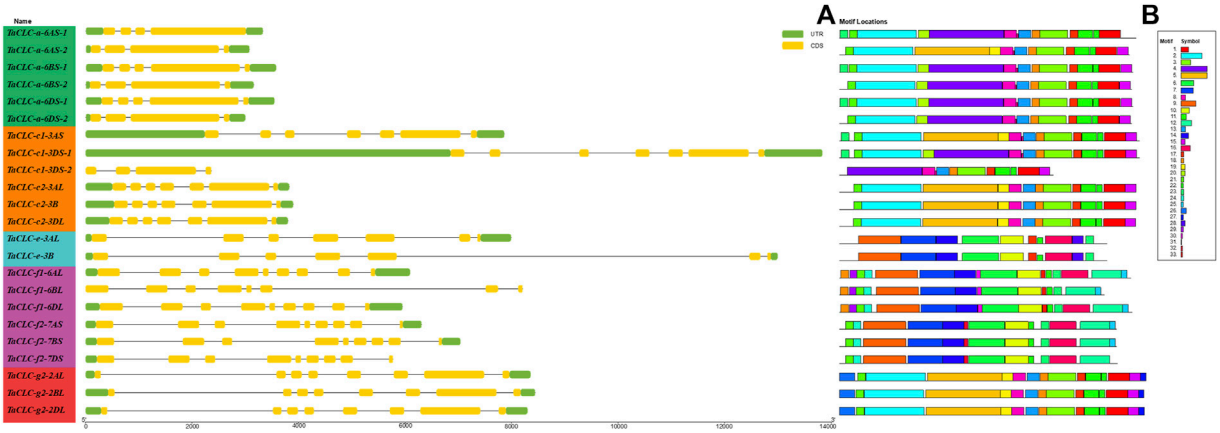


FIGURE 5 | Gene structure and conservation motifs of *TaCLC* gene. **(A)** Gene structure. *TaCLC* genes are displayed in the order based on phylogenetic analysis of their proteins. Introns, exons and noncoding regions are represented with black lines, yellow boxes and green boxes, respectively. **(B)** Conserved motifs of TaCLC proteins. Thirty three motifs were identified and are marked with different colors.

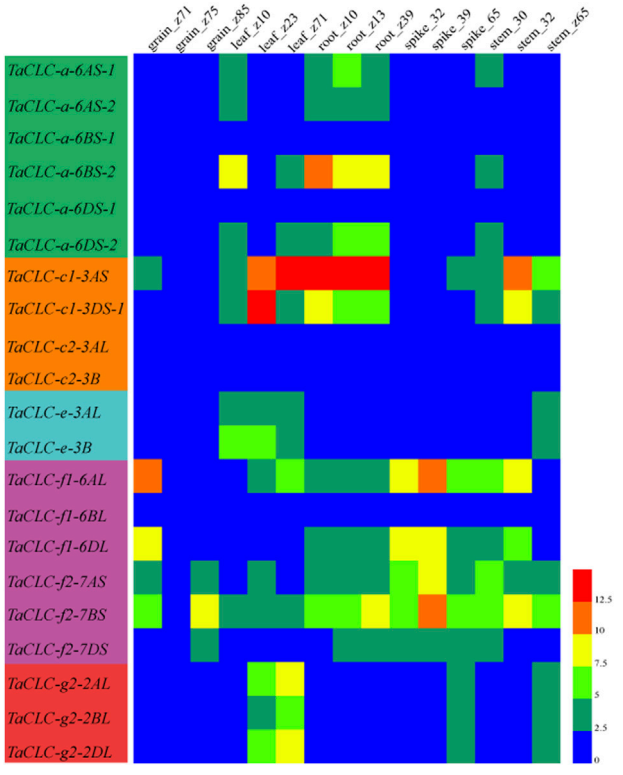


FIGURE 6 | Heat map of *TaCLC* genes expressed at different developmental stages in wheat tissues. The values were obtained from the RNA-seq data of Chinese spring wheat in different tissues and periods provided by WheatExp database (Pearce et al., 2015). The values of 0–12.5 represent the lowest to the highest expression levels of the detected genes.

from 3 to 6 h, its expression began to be down-regulated from 6 to 12 h. Its expression was finally up-regulated to its highest level from 12 to 24 h.

Functional Complementation of *Triticum aestivum* L. CLC Members in Yeast Mutant

The budding yeast *S. cerevisiae* has been shown to be an excellent model for studying ion transport properties and physiological function of ion homeostasis (Xu et al., 2013). The existence of mutant strains lacking their own transport systems has provided an efficient tool for the molecular study of transporters from higher eukaryotes upon their expression in yeast cells (Xu et al., 2008). In *S. cerevisiae*, the *GEF1* gene encodes a single putative CLC chloride channel/transporter. The corresponding mutant strain—*Δgef1*—lacks GEF1 gene and is sensitive to extracellular cations (Lv et al., 2009). Studies have shown that the *AtCLC-c* gene compensates for the inhibited growth of *Δgef1* deletion mutant yeast cells on YTD medium containing high concentrations of either NaCl or KCl (Lv et al., 2009). The growth of *Δgef1* cells is not significantly different from that of wild type cells, indicating that the toxicity of these salts on the growth of *Δgef1* cells is more related to the properties of anions than the properties of cations (Lv et al., 2009). In this study, we investigated the function of *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* using the yeast strain YJR040w (*Δgef1*). The AtCLC-c-p416 recombinant plasmid vector was transferred into the yeast strain YJR040w and used as a positive control, while the p416 vector was transferred into the yeast strain YJR040w and used as a negative control. The results illustrated that the three genes *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* also compensated for the inhibitory effects of the GEF1 deletion mutant yeast YJR040w cells on the growth of YTD medium containing either 1 M NaCl or 1 M KCl (**Figure 8**). All the test transgenic CLC genes partially restored the growth function of mutant yeast cells YJR040w on YTD medium containing 1 M KNO_3, indicating that *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* in wheat and *AtCLC-c*

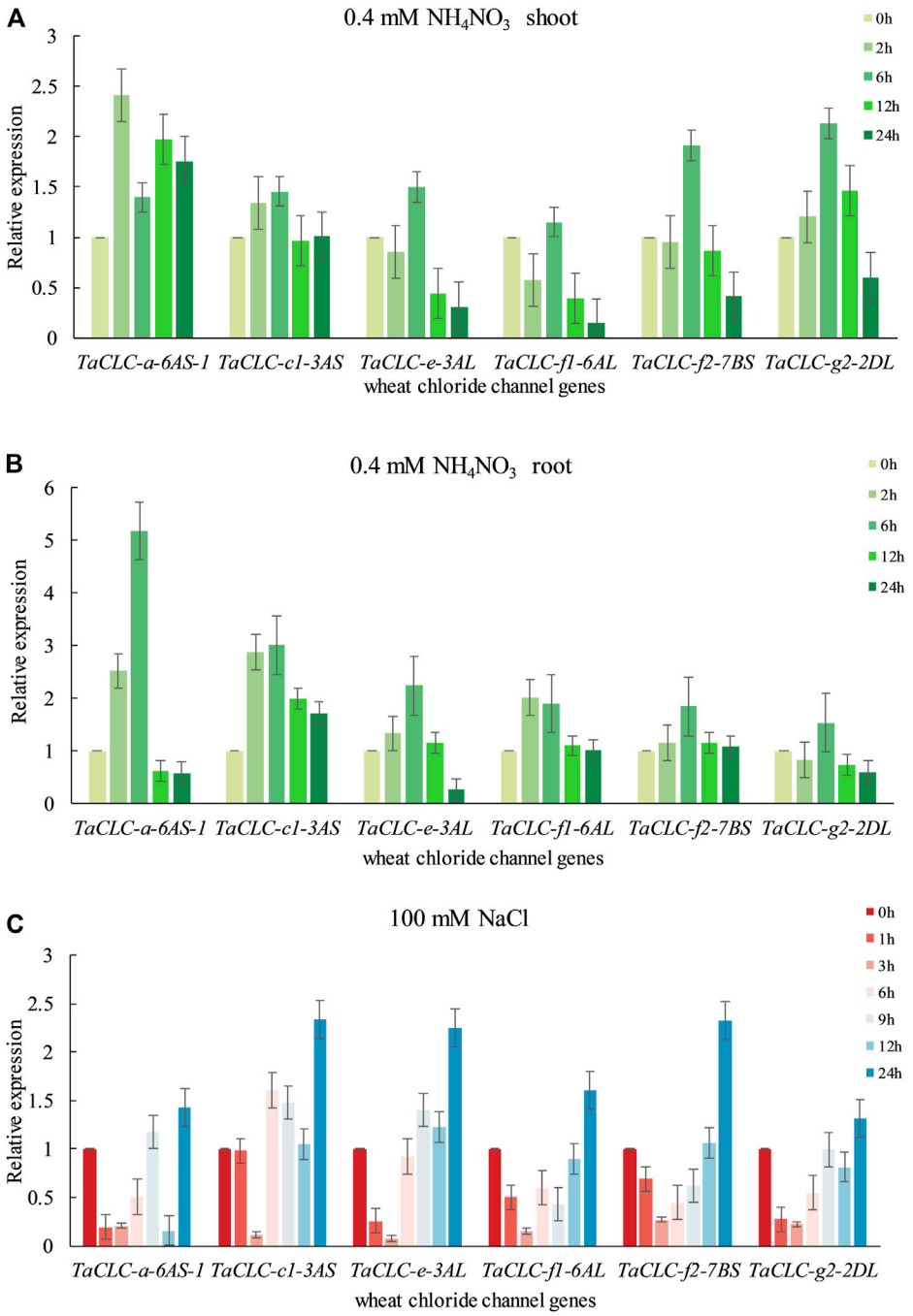


FIGURE 7 | Quantitative real-time PCR validation of *TaCLC* genes in wheat seedlings under different stresses. **(A)** Expression of *TaCLC* genes in the wheat shoot tissue after 0.4 mM NH_4NO_3 treatment for 0, 2, 6, 12 and 24 h. **(B)** Expression of *TaCLC* genes in the wheat root after 0.4 mM NH_4NO_3 treatment for 0, 2, 6, 12 and 24 h. **(C)** Expression of *TaCLC* genes in wheat seedlings after 100 mM NaCl treatment for 0, 1, 3, 6, 9, 12 and 24 h.

in Arabidopsis had certain NO_3^- transport ability. These results showed that the three selected TaCLC genes compensated for the fact that the GEF1 mutant YJR040w blocks transport of Cl^- or NO_3^- after the deletion of the CLC genes. The proteins encoded by these three genes of *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* exhibited anion transport activity of Cl^- or NO_3^-.

DISCUSSION

The main nitrogen sources for plants are either nitrate nitrogen or ammonium nitrogen. Under natural field conditions, the content of nitrate nitrogen in soil is much higher than that of ammonium nitrogen. Therefore, understanding the efficient absorption mechanism(s) of nitrate nitrogen in crops would provide a

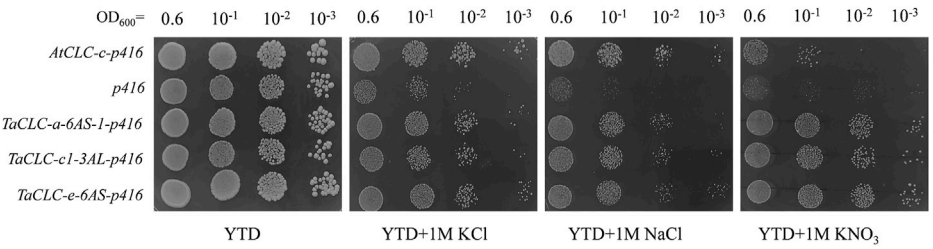


FIGURE 8 | Yeast *Δgef1* mutant complementation experiment. Yeast transformed with empty vector (p416) was used as a negative control and *AtCLC-c* was used as a positive control. The tested YTD (1% yeast extract/2% tryptone/2% dextrose) media supplemented with different concentrations of compounds (1 M KCl/1 M NaCl/ 1 M KNO_3) are indicated under each panel. Serial dilutions (10^{-1}) of yeast culture were plated.

theoretical basis for improving crop absorption—potentially by using biotechnological approaches (Meng et al., 2019). Currently, there remains little information on the CLC gene family in plants, with information relegated to only some species like *Arabidopsis thaliana*, *Glycine max*, *Oraza sativa*, and *Zea mays* (Diedhiou and Golldack, 2006; Lv et al., 2009; Wang et al., 2015; Wei et al., 2019). For wheat, there remains no previous work regarding a genome-wide identification and analysis of wheat *TaCLC* family genes. With the successful completion of genome sequence of wheat, it is now possible to identify members of the TaCLC gene family at the whole genome level (Appels et al., 2018). According to the homologous relationship of *CLC* genes in different species, a total of 23 *TaCLC* genes in seven clusters were identified using the wheat genome database. Their homologous relationships, gene structure, chromosomal localization, and expression pattern were subsequently analyzed. These results will lay the foundation for the functional study of *TaCLC* genes and provide a theoretical reference for understanding nitrate transport in wheat.

Across species, the name of the CLC protein family remains confusing because different researchers have updated the data at different stages. For example, in addition to *ZmCLC1*, *ZmCLC2*, and *ZmCLC3* genes in maize, there are also *ZmCLC-a, -b, -c, -d* and other genes (Yu et al., 2017). Moreover, the rice *OsCLC* genes used in this study were originally named *OsCLC 1–7*. After comparing, we corrected the name to *OsCLC-a*, *OsCLC-c1*, and other more consistent gene names (**Supplementary Table S1**). In this study the *TaCLC* genes were named according to the Arabidopsis classification criteria (-a, -b, -c, -d, -e, -f, -g). During the structural analysis of the TaCLC protein, we found that the TaCLC-a/-c1/-c2/-g2 contained a common conserved domain that was also found in other species. The x residue in the conserved region (I) of the TaCLC-a cluster was proline, while in the TaCLC-c1/-c2/-g2 it was serine. Therefore, we speculated that the function of the TaCLC-a cluster of genes was similar to that of the CLC-a cluster of genes in other species. Moreover, NO_3^- was preferentially transported, which may indicate the gene's role as a NO_3^-/H^+ exchanger. The TaCLC-c1/-c2/-g2 cluster genes have a higher affinity for Cl^- and preferentially transport Cl^- (Nedelyaeva et al., 2019). This hypothesis is also consistent with the up-regulation of the *TaCLC-a-6AS-1* gene in our qRT-PCR results.

In plants, the functional research of *CLC* genes has mainly focused on the physiological and molecular regulatory functions of either Cl^- or NO_3^- absorption, transportation, and chloride (salt) tolerance (Li et al., 2006; Nakamura et al., 2006; Wei et al., 2013; Wong et al., 2013; Liao et al., 2018; Nedelyaeva et al., 2018; Nedelyaeva et al., 2019). Therefore, genetic research into the CLC-a/-c cluster is more extensive. In Arabidopsis, studies have found that *AtCLC-e* and *AtCLC-a* also have interconnected transporters in the nitrate assimilation pathway (Monachello et al., 2009). In addition, *AtCLC-a* plays a different role in the regulation of guard cell expansion in that *AtCLC-a* promotes anion accumulation during light-induced guard cell expansion and stomata opening (Wege et al., 2014). The N-terminus of *AtCLC-a* is phosphorylated by *AtSnRK 2.6* (*AtOST1*, *At4g33950*), which mediates stoma closure by excluding anions (Wege et al., 2014). *AtPP2A-C5* (*At1g69960*) interacts with *AtCLC* family genes, and *AtPP2A-5C* overexpression in plants increases the activity of *AtCLC-c*. This increases the ability of Cl^- to enter into vacuoles and improves salt tolerance of plants (Hu et al., 2017). Based on the hypothesis of homologous sequence alignment, it was speculated that the *TaSnRK 2.6* (*TaOST1, TraesCS5D02G081700*) and TaCLC-a cluster of genes, as well as the *TaPP2A-5C* (*TraesCS6D02G1714000*) and TaCLCs should be related.

Among plants, only three *PtCLC* genes have been identified in *Populus trichocarpa* (Yu et al., 2017), and 22 *GhCLC* genes in upland cotton (*Gossypium hirsutum* L.) (Liu X. et al., 2020). In most cases, the number of *CLC* genes identified in other plants is 5–7, while the number of *TaCLC* genes identified in wheat is up to 23 genes. The physical location and collinearity analysis of the gene on its corresponding chromosome indicated that there was a tandem repeat in the evolution of wheat *TaCLC* genes, which might be the reason why there are more *TaCLC* genes in wheat than other species. In this study, the TaCLC-b/-d subfamily genes were not found in the wheat genome, indicating that the fragment and gene loss of the homologous gene in this family may have simultaneously occurred. It may also be the artificial deletion caused by the incompleteness of the conserved domain in the previous analysis of 34 *CLC* genes annotated in the Emsembl Plant database. For example, the CD-search results of proteins encoded by the *AtCLC-b* and *AtCLC-d* genes in NCBI did not have a complete voltage gated CLC domain (PF00654), so we removed TaCLC-b/-d in the identification of the wheat TaCLC gene family. At the same time, we also named the 11 removed genes which had incomplete domains according to the

phylogenetic tree (**Supplementary Figure S3**; **Supplementary Table S6**). In the *CLC* genes identified in both *Camellia sinensis* and *Nicotiana tabacum*, the number of motifs used was 20 and 24, respectively (Zhang et al., 2018; Xing et al., 2020). According to the HMM logo in the Pfam database, the motifs number of TaCLCs was determined to be 33 and the structure type was similar to the motif's position structural type of the CLC which can be divided into two types. In addition, two different motif position structure types were also found in tobacco, which indicated that the functions of the *CLC* genes have become diversified. RNA-seq database analysis showed that the function and expression of each gene were different. Notably, even homologous genes played different functions at different development stages, such as *TaCLC-a-6AS-1*, *TaCLC-a-6BS-1*, and *TaCLC-a-6DS-1*. It seems that evolutionary pressures can extend the members of the gene family (Abdullah et al., 2021; Musavizadeh et al., 2021). Moreover, mutations in coding sites and promoter regions can affect the function of members of the gene family (Faraji et al., 2021).

To analyze the expression patterns of the TaCLC gene family under low nitrogen stress or salt stress, seven genes from each cluster of the TaCLC gene family were selected for fluorescence qRT-PCR. The results showed that the family genes were up-regulated under short-term low nitrogen stress or salt stress. This was true except for the TaCLC-c2 cluster genes, which basically had no expression across the various periods of stress. The expression pattern of the CLC gene family in most species has been focused on across different tissue types as well as during long-term abiotic stress (Lv et al., 2009; Jossier et al., 2010; Wei et al., 2015; Zhang et al., 2018; Liu X. et al., 2020). However, the expression patterns of *TaCLC* at each time point were different from those of other species. For instance, *AtCLCa-d* had a relatively stable expression abundance at all developmental stage. The expression level of *AtCLC-e* was equivalent to that of *AtCLC-f*, but was significantly lower than that of subclass I genes (Wei et al., 2015). In tobacco, all expressed *NtCLC* genes had low expression levels in the roots. After 7 days of salt stress (300 mM NaCl), the expression levels of multiple *NtCLC* genes were all significantly up-regulated (Zhang et al., 2018). In pomegranate (*Punica granatum*), the expression level of the *PgCLC* genes under salt stress was high in leaves and low in roots. Moreover, *PgCLC* genes have been shown to affect the accumulation of Cl^-, SO_4^{2-}, and NO_3^- in pomegranate tissues under salt stress (Liu C. et al., 2020). Our results showed that all *TaCLC* genes were expressed under low nitrogen stress conditions over the short term. For low nitrogen stress, the response of the *TaCLC-a-6AS-1* gene revealed it had the highest expression level. The relative expression levels of the *TaCLC-c1-3AS*, *TaCLC-e-3AL*, and *TaCLC-g2-2DL* genes were all higher under salt stress. These results were consistent with the characteristics that cluster -a had NO_3^- transport capacity and clusters -c1/-g2 had Cl^- transport capacity in the CLC gene family. Based on the results of qRT-PCR analysis, we speculated that *TaCLC* genes (except the genes of TaCLC-c2 cluster) could respond to anion deficiency stresses.

The budding yeast, *S. cerevisiae* allows for large-scale, genome-wide analyses in a fast and economically efficient manner. Work with *S. cerevisiae* allows for the discovery and/or characterization of many aspects of ion transporter function (Locascio et al., 2019). Studies have shown that *GEF1*, as a chloride channel gene, maintains the charge balance in yeast cells. Through its acidic interior, the cation can be localized in either the internal organs or vacuoles of the cell, thus playing a role in cation detoxification (Gaxiola et al., 1998; Li et al., 2015; Nagatsuka et al., 2017). The mutant yeast *Δgef1* lacks the chloride ion channel gene and is blocked when transporting Cl^- in the intracellular vesicles (vacuoles or Golgi apparatus) of yeast cells (Hechenberger et al., 1996). It also has hypersensitivity to several extracellular cations. At present, *Arabidopsis thaliana*, *Glycine max*, *Oryza sativa* L., and *Suaeda altissima* (L.) have been studied using yeast mutants (Nakamura et al., 2006; Lv et al., 2009; Nedelyaeva et al., 2019; Wei et al., 2019). In this study, various chlorides, sulfates, and nitrates were respectively added to SD, SR, YPEG, YPG, YPD, or YTD media to determine whether *TaCLC* genes inhibit the sensitivity of yeast mutant cells to metal cations and to determine their anion transport function. In these experiments, we found that the growth of *Δgef1* yeast cells with YTD medium as the culture substrate was most suitable for YJR040w mutation. This stood in contrast to using SD, SR, YPEG, YPG, or YPD media as the culture substrate.

Previous studies showed that *AtCLC-c* gene had the ability to transport Cl^- (Lv et al., 2009). In our study, we found that when transformed one of the genes *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL*, transgenic yeast mutant strains had a strong ability to transport Cl^- compared with control yeast strain, albeit they did not reach the tolerance of *AtCLC-c* gene transformation (**Figure 8**). It indicated that a large number of *TaCLC* genes may simultaneously play roles in the transport of anions such as Cl^- or NO_3^-. For a single *TaCLC* gene, its anion transport capacity was not very strong. In addition, *AtCLC-c* gene-transformed yeast mutant strains have not been studied in a medium containing KNO_3. In our study, we found that *AtCLC-c* gene-transformed yeast mutant strains could transport NO_3^- and inhibit the cation hypersensitivity in yeast GEF1 mutants (**Figure 8**). There was no difference in the growth of *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* in YTD medium containing 1 M KNO_3, and the three genes had similar NO_3^- transport ability without exogenous Cl^- interference. According to the regularity of the conserved motifs of the CLC and the analysis of the TaCLC protein sequences, there may be differences in preferential transport ability of NO_3^- or Cl^- among *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* in the case of exogenous Cl^- interference.

CONCLUSION

In this study, a genome-wide identification of *CLC* genes in wheat was performed and 23 *TaCLC* genes were identified. The gene structure, chromosomal location, conserved motif and expression pattern of the members of the family were then analyzed. The family was divided into two main subclasses (I and II) and seven clusters (-a, -c1, -c2, -e, -f1, -f2, and -g2). The 23 *TaCLC* genes of

the family were unevenly distributed on wheat chromosomes and some genes in the cluster had tandem duplication. *TaCLC* gene expression was illustrated using qRT-PCR, and the results showed that the expression pattern of this gene family was induced by low nitrogen stress or salt stress except for *TaCLC-c2* which was from subfamily -c. The function of some *TaCLC* genes was studied using yeast mutant strains. The results of yeast mutant complementation experiments showed that *TaCLC-a-6AS-1*, *TaCLC-c1-3AS*, and *TaCLC-e-3AL* all had anion transport functions for NO_3^- or Cl^- and compensated the hypersensitivity of yeast GEF1 mutants in restoring anion-sensitive phenotype. Collectively, these results provide theoretical reference for studying the response of *TaCLC* family genes to low nitrogen stress and the physiological functions of anion transport in wheat.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

HX and XC conceived and designed the experiments. PM, YR, and HW conducted the experiments. CH performed the expression level examination of the stressed plants. PM, LZ and KZ analyzed the data. PM wrote the paper. HX revised the article. All authors read and approved the final manuscript.

FUNDING

This work was funded by the Henan Natural Science Foundation (202300410217) and Henan Science and Technology Research Project (202102110178).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.846795/full#supplementary-material

REFERENCES

Abdullah, Faraji, S., Mehmood, F., Malik, H. M. T., Ahmed, I., Heidari, P., et al. (2021). The GASA Gene Family in Cacao (*Theobroma cacao*, Malvaceae): Genome Wide Identification and Expression Analysis. *Agronomy* 11 (7), 1425. doi:10.3390/agronomy11071425

Appels, R., Eversole, K., Appels, R., Eversole, K., Feuillet, C., Keller, B., et al. (2018). Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome. *Science* 361 (6403), 661. doi:10.1126/science.aar7191

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13 (8), 1194–1202. doi:10.1016/j.molp.2020.06.009

Cheng, X., Liu, X., Mao, W., Zhang, X., Chen, S., Zhan, K., et al. (2018). Genome-Wide Identification and Analysis of HAK/KUP/KT Potassium Transporters Gene Family in Wheat (*Triticum aestivum* L.). *Int. J. Mol. Sci.* 19 (12), 3969. doi:10.3390/ijms19123969

De Angeli, A., Monachello, D., Ephritikhine, G., Frachisse, J. M., Thomine, S., Gambale, F., et al. (2006). The Nitrate/Proton Antiporter AtCLCa Mediates Nitrate Accumulation in Plant Vacuoles. *Nature* 442 (7105), 939–942. doi:10.1038/nature05013

De Angeli, A., Moran, O., Wege, S., Filleur, S., Ephritikhine, G., Thomine, S., et al. (2009). ATP Binding to the C Terminus of the *Arabidopsis thaliana* Nitrate/Proton Antiporter, AtCLCa, Regulates Nitrate Transport into Plant Vacuoles. *J. Biol. Chem.* 284 (39), 26526–26532. doi:10.1074/jbc.M109.005132

Diédhiou, C. J., and Golldack, D. (2006). Salt-Dependent Regulation of Chloride Channel Transcripts in Rice. *Plant Sci.* 170 (4), 793–800. doi:10.1016/j.plantsci.2005.11.014

Faraji, S., Heidari, P., Amouei, H., Filiz, E., Abdullah, P., and Poczai, P. (2021). Investigation and Computational Analysis of the Sulfotransferase (SOT) Gene Family in Potato (*Solanum tuberosum*): Insights into Sulfur Adjustment for Proper Development and Stimuli Responses. *Plants* 10 (12), 2597. doi:10.3390/plants10122597

Fecht-Bartenbach, J. v. d., Bogner, M., Krebs, M., Stierhof, Y.-D., Schumacher, K., and Ludewig, U. (2007). Function of the Anion Transporter At*CLC*-d in the *Trans*-golgi Network. *Plant J.* 50 (3), 466–474. doi:10.1111/j.1365-313X.2007.03061.x

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., et al. (2006). Pfam: Clans, Web Tools and Services. *Nucleic Acids Res.* 34, D247–D251. doi:10.1093/nar/gkj149

Gaxiola, R. A., Yuan, D. S., Klausner, R. D., and Fink, G. R. (1998). The Yeast CLC Chloride Channel Functions in Cation Homeostasis. *Proc. Natl. Acad. Sci.* 95 (7), 4046–4050. doi:10.1073/pnas.95.7.4046

Harada, H., Kuromori, T., Hirayama, T., Shinozaki, K., and Leigh, R. A. (2004). Quantitative Trait Loci Analysis of Nitrate Storage in Arabidopsis Leading to an Investigation of the Contribution of the Anion Channel Gene, *AtCLC-c*, to Variation in Nitrate Levels. *J. Exp. Bot.* 55 (405), 2005–2014. doi:10.1093/jxb/erh224

Hechenberger, M., Schwappach, B., Fischer, W. N., Frommer, W. B., Jentsch, T. J., and Steinmeyer, K. (1996). A Family of Putative Chloride Channels from *Arabidopsis* and Functional Complementation of a Yeast Strain with a *CLC* Gene Disruption. *J. Biol. Chem.* 271 (52), 33632–33638. doi:10.1074/jbc.271.52.33632

Hu, R., Zhu, Y., Wei, J., Chen, J., Shi, H., Shen, G., et al. (2017). Overexpression of *PP2A-C5* that Encodes the Catalytic Subunit 5 of Protein Phosphatase 2A in Arabidopsis Confers Better Root and Shoot Development under Salt Conditions. *Plant Cell Environ.* 40 (1), 150–164. doi:10.1111/pce.12837

Ijato, T., Porras-Murillo, R., Ganz, P., Ludewig, U., and Neuhäuser, B. (2021). Concentration-Dependent Physiological and Transcriptional Adaptations of Wheat Seedlings to Ammonium. *Physiol. Plantarum* 171 (3), 328–342. doi:10.1111/ppl.13113

Jentsch, T. J., Steinmeyer, K., and Schwarz, G. (1990). Primary Structure of *Torpedo marmorata* Chloride Channel Isolated by Expression Cloning in *Xenopus* Oocytes. *Nature* 348 (6301), 510–514. doi:10.1038/348510a0

Jossier, M., Kroniewicz, L., Dalmas, F., Le Thiec, D., Ephritikhine, G., Thomine, S., et al. (2010). The Arabidopsis Vacuolar Anion Transporter, AtCLCc, is Involved in the Regulation of Stomatal Movements and Contributes to Salt Tolerance. *Plant J.* 64 (4), 563–576. doi:10.1111/j.1365-313X.2010.04352.x

Ju, X.-t., and Zhang, C. (2017). Nitrogen Cycling and Environmental Impacts in upland Agricultural Soils in North China: A Review. *J. Integr. Agric.* 16 (12), 2848–2862. doi:10.1016/s2095-3119(17)61743-x

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Li, B., Lu, M.-Q., Wang, Q.-Z., Shi, G.-Y., Liao, W., and Huang, S.-S. (2015). Raman Spectra Analysis for Single Mitochondrias after Apoptosis Process of Yeast Cells Stressed by Acetic Acid. *Chin. J. Anal. Chem.* 43 (5), 643–650. doi:10.1016/s1872-2040(15)60824-6

Li, T., Liao, K., Xu, X., Gao, Y., Wang, Z., Zhu, X., et al. (2017). Wheat Ammonium Transporter (AMT) Gene Family: Diversity and Possible Role in Host-Pathogen Interaction with Stem Rust. *Front. Plant Sci.* 8, 1. doi:10.3389/fpls.2017.01637

Li, W.-Y. F., Wong, F.-L., Tsai, S.-N., Phang, T.-H., Shao, G., and Lam, H.-M. (2006). Tonoplast-located *GmCLC1* and *GmNHX1* from Soybean Enhance NaCl Tolerance in Transgenic Bright Yellow (BY)-2 Cells. *Plant Cell Environ.* 29 (6), 1122–1137. doi:10.1111/j.1365-3040.2005.01487.x

Liao, Q., Zhou, T., Yao, J.-y., Han, Q.-f., Song, H.-x., Guan, C.-y., et al. (2018). Genome-Scale Characterization of the Vacuole Nitrate Transporter *Chloride Channel (CLC)* Genes and Their Transcriptional Responses to Diverse Nutrient Stresses in Allotetraploid Rapeseed. *Plos One* 13 (12), e0208648. doi:10.1371/journal.pone.0208648

Liu, C., Zhao, Y., Zhao, X., Dong, J., and Yuan, Z. (2020). Genome-wide Identification and Expression Analysis of the *CLC* Gene Family in Pomegranate (*Punica granatum*) Reveals its Roles in Salt Resistance. *BMC Plant Biol.* 20 (1), 1. doi:10.1186/s12870-020-02771-z

Liu, R., Jia, T., Cui, B., and Song, J. (2020). The Expression Patterns and Putative Function of Nitrate Transporter 2.5 in Plants. *Plant Signaling Behav.* 15 (12), 1815980. doi:10.1080/15592324.2020.1815980

Liu, X., Pi, B., Pu, J., Cheng, C., Fang, J., and Yu, B. (2020). Genome-Wide Analysis of Chloride Channel-Encoding Gene Family Members and Identification of CLC Genes that Respond to Cl−/salt Stress in upland Cotton. *Mol. Biol. Rep.* 47 (12), 9361–9371. doi:10.1007/s11033-020-06023-z

Locascio, A., Andrés-Colás, N., Mulet, J. M., and Yenush, L. (2019). *Saccharomyces cerevisiae* as a Tool to Investigate Plant Potassium and Sodium Transporters. *Int. J. Mol. Sci.* 20 (9), 2133. doi:10.3390/ijms20092133

Lurin, C., Geelen, D., Barbier-Brygoo, H., Guern, J., and Maurel, C. (1996). Cloning and Functional Expression of a Plant Voltage-dependent Chloride Channel. *Plant Cell* 8 (4), 701–711. doi:10.1105/tpc.8.4.701

Lv, Q.-d., Tang, R.-j., Liu, H., Gao, X.-s., Li, Y.-z., Zheng, H.-q., et al. (2009). Cloning and Molecular Analyses of the *Arabidopsis thaliana* Chloride Channel Gene Family. *Plant Sci.* 176 (5), 650–661. doi:10.1016/j.plantsci.2009.02.006

Lv, X., Zhang, Y., Hu, L., Zhang, Y., Zhang, B., Xia, H., et al. (2021). Low-Nitrogen Stress Stimulates Lateral Root Initiation and Nitrogen Assimilation in Wheat: Roles of Phytohormone Signaling. *J. Plant Growth Regul.* 40 (1), 436–450. doi:10.1007/s00344-020-10112-5

Marmagne, A., Vinauger-Douard, M., Monachello, D., de Longevialle, A. F., Charon, C., Allot, M., et al. (2007). Two Members of the *Arabidopsis* CLC (Chloride Channel) Family, AtCLCe and AtCLCf, are Associated With Thylakoid and Golgi Membranes, Respectively. *J. Exp. Bot.* 58 (12), 3385–3393. doi:10.1093/jxb/erm187

Meng, L., Dong, J.-x., Wang, S.-s., Song, K., Ling, A.-f., Yang, J.-g., et al. (2019). Differential Responses of Root Growth to Nutrition with Different Ammonium/nitrate Ratios Involve Auxin Distribution in Two Tobacco Cultivars. *J. Integr. Agric.* 18 (12), 2703–2715. doi:10.1016/s2095-3119(19)62595-5

Monachello, D., Allot, M., Oliva, S., Krapp, A., Daniel-Vedele, F., Barbier-Brygoo, H., et al. (2009). Two Anion Transporters AtCLCa and AtCLCe Fulfil Interconnecting but Not Redundant Roles in Nitrate Assimilation Pathways. *New Phytol.* 183 (1), 88–94. doi:10.1111/j.1469-8137.2009.02837.x

Moradi, H., Elzenga, T., and Lanfermeijer, F. (2015). Role of the *AtCLC* Genes in Regulation of Root Elongation in Arabidopsis. *Iran. J. Gen. Plant Breed.* 4, 1–8.

Musavizadeh, Z., Najafi-Zarrini, H., Kazemitabar, S. K., Hashemi, S. H., Faraji, S., Barcaccia, G., et al. (2021). Genome-Wide Analysis of *Potassium Channel* Genes in Rice: Expression of the *OsAKT* and *OsKAT* Genes Under Salt Stress. *Genes* 12 (5), 784. doi:10.3390/genes12050784

Nagatsuka, Y., Kiyuna, T., Kigawa, R., Sano, C., and Sugiyama, J. (2017). Prototheca Tumulicola sp. nov., a Novel Achlorophyllous, Yeast-Like Microalga Isolated from the Stone Chamber interior of the *Takamatsuzuka tumulus*. *Mycoscience* 58 (1), 53–59. doi:10.1016/j.myc.2016.09.005

Nakamura, A., Fukuda, A., Sakai, S., and Tanaka, Y. (2006). Molecular Cloning, Functional Expression and Subcellular Localization of Two Putative Vacuolar Voltage-Gated Chloride Channels in rice (*Oryza sativa* L.). *Plant Cell Physiol.* 47 (1), 32–42. doi:10.1093/pcp/pci220

Nedelyaeva, O. I., Shuvalov, A. V., and Balnokin, Y. V. (2020). Chloride Channels and Transporters of the CLC Family in Plants. *Russ. J. Plant Physiol.* 67 (5), 767–784. doi:10.1134/s1021443720050106

Nedelyaeva, O. I., Shuvalov, A. V., Karpichev, I. V., Beliaev, D. V., Myasoedov, N. A., Khalilova, L. A., et al. (2019). Molecular Cloning and Characterisation of SaCLCa1, a Novel Protein of the Chloride Channel (CLC) Family from the Halophyte *Suaeda Altissima* (L.) Pall. *J. Plant Physiol.* 240, 152995. doi:10.1016/j.jplph.2019.152995

Nedelyaeva, O. I., Shuvalov, A. V., Mayorova, O. V., Yurchenko, A. A., Popova, L. G., Balnokin, Y. V., et al. (2018). Cloning and Functional Analysis of *SaCLCc1*, a Gene Belonging to the Chloride Channel Family (*CLC*), from the Halophyte *Suaeda altissima* (L.) Pall. *Dokl. Biochem. Biophys.* 481 (1), 186–189. doi:10.1134/s1607672918040026

Park, E., Campbell, E. B., and MacKinnon, R. (2017). Structure of a CLC Chloride Ion Channel by Cryo-Electron Microscopy. *Nature* 541 (7638), 500–505. doi:10.1038/nature20812

Pearce, S., Vazquez-Gross, H., Herin, S. Y., Hane, D., Wang, Y., Gu, Y. Q., et al. (2015). WheatExp: An RNA-Seq Expression Database for Polyploid Wheat. *BMC Plant Biol.* 15, 1. doi:10.1186/s12870-015-0692-1

Subba, A., Tomar, S., Pareek, A., and Singla-Pareek, S. L. (2021). The Chloride Channels: Silently Serving the Plants. *Physiol. Plant.* 171 (4), 688–702. doi:10.1111/ppl.13240

Um, T. Y., Lee, S., Kim, J.-K., Jang, G., and Choi, Y. D. (2018). Chloride Channel 1 Promotes Drought Tolerance in rice, Leading to Increased Grain Yield. *Plant Biotechnol. Rep.* 12 (4), 283–293. doi:10.1007/s11816-018-0492-9

von der Fecht-Bartenbach, J., Bogner, M., Dynowski, M., and Ludewig, U. (2010). CLC-b-Mediated NOFormula/H+ Exchange across the Tonoplast of Arabidopsis Vacuoles. *Plant Cell Physiol.* 51 (6), 960–968. doi:10.1093/pcp/pcq062

Wang, F., Gao, J., Tian, Z., Liu, Y., Abid, M., Jiang, D., et al. (2016). Adaptation to Rhizosphere Acidification Is a Necessary Prerequisite for Wheat (*Triticum aestivum* L.) Seedling Resistance to Ammonium Stress. *Plant Physiol. Biochem.* 108, 447–455. doi:10.1016/j.plaphy.2016.08.011

Wang, S., Su, S. Z., Wu, Y., Li, S. P., Shan, X. H., Liu, H. K., et al. (2015). Overexpression of Maize Chloride Channel Gene *ZmCLC-d* in *Arabidopsis thaliana* Improved its Stress Resistance. *Biol. Plant* 59 (1), 55–64. doi:10.1007/s10535-014-0468-8

Wege, S., De Angeli, A., Droillard, M.-J., Kroniewicz, L., Merlot, S., Cornu, D., et al. (2014). Phosphorylation of the Vacuolar Anion Exchanger AtCLCa is Required for the Stomatal Response to Abscisic Acid. *Sci. Signal.* 7 (333), 1. doi:10.1126/scisignal.2005140

Wei, P., Che, B., Shen, L., Cui, Y., Wu, S., Cheng, C., et al. (2019). Identification and Functional Characterization of the Chloride Channel Gene, *GsCLC-c2* from Wild Soybean. *BMC Plant Biol.* 19, 1. doi:10.1186/s12870-019-1732-z

Wei, Q. J., Gu, Q. Q., Wang, N. N., Yang, C. Q., and Peng, S. A. (2015). Molecular Cloning and Characterization of the Chloride Channel Gene Family in Trifoliate Orange. *Biol. Plant* 59 (4), 645–653. doi:10.1007/s10535-015-0532-z

Wei, Q., Liu, Y., Zhou, G., Li, Q., Yang, C., and Peng, S.-a. (2013). Overexpression of *CsCLCc*, a Chloride Channel Gene from *Poncirus trifoliata*, Enhances Salt Tolerance in *Arabidopsis*. *Plant Mol. Biol. Rep.* 31 (6), 1548–1557. doi:10.1007/s11105-013-0592-1

White, M. M., and Miller, C. (1979). A Voltage-Gated Anion Channel from the Electric Organ of *Torpedo californica*. *J. Biol. Chem.* 254 (20), 10161–10166. doi:10.1016/s0021-9258(19)86687-5

Whiteman, S.-A., Serazetdinova, L., Jones, A. M. E., Sanders, D., Rathjen, J., Peck, S. C., et al. (2008). Identification of Novel Proteins and Phosphorylation Sites in a Tonoplast Enriched Membrane Fraction of *Arabidopsis thaliana*. *Proteomics* 8 (17), 3536–3547. doi:10.1002/pmic.200701104

Wong, T. H., Li, M. W., Yao, X. Q., and Lam, H. M. (2013). The GmCLC1 Protein from Soybean Functions as a Chloride Ion Transporter. *J. Plant Physiol.* 170 (1), 101–104. doi:10.1016/j.jplph.2012.08.003

Xing, A., Ma, Y., Wu, Z., Nong, S., Zhu, J., Sun, H., et al. (2020). Genome-Wide Identification and Expression Analysis of the CLC Superfamily Genes in Tea Plants (*Camellia sinensis*). *Funct. Integr. Genom.* 20 (4), 497–508. doi:10.1007/s10142-019-00725-9

Xu, H., Jiang, X., Zhan, K., Cheng, X., Chen, X., Pardo, J. M., et al. (2008). Functional Characterization of a Wheat Plasma Membrane Na+/H+ Antiporter in Yeast. *Arch. Biochem. Biophys.* 473 (1), 8–15. doi:10.1016/j.abb.2008.02.018

Xu, Y., Zhou, Y., Hong, S., Xia, Z., Cui, D., Guo, J., et al. (2013). Functional Characterization of a Wheat NHX Antiporter Gene TaNHX2 that Encodes a K+/H+ Exchanger. *Plos One* 8 (11), e78098. doi:10.1371/journal.pone.0078098

Yu, B., Liu, Z., and Wei, P. (2017). Research Progresses on Whole Genome Discovery and Function Analysis of CLC Homologous Genes in Plants Based on Salt Stress. *J. Nanjing Agric. Univ.* 40 (2), 187–194. doi:10.7685/jnau.201701020

Zhang, H., Jin, J., Jin, L., Li, Z., Xu, G., Wang, R., et al. (2018). Identification and Analysis of the Chloride Channel Gene Family Members in Tobacco (*Nicotiana tabacum*). *Gene* 676, 56–64. doi:10.1016/j.gene.2018.06.073

Zifarelli, G., and Pusch, M. (2010). CLC Transport Proteins in Plants. *Febs Lett.* 584 (10), 2122–2127. doi:10.1016/j.febslet.2009.12.042

Zörb, C., Ludewig, U., and Hawkesford, M. J. (2018). Perspective on Wheat Yield and Quality with Reduced Nitrogen Supply. *Trends Plant Sci.* 23 (11), 1029–1037. doi:10.1016/j.tplants.2018.08.012

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: 20 April 2022
doi: 10.3389/fgene.2022.842387

PolyReco: A Method to Automatically Label Collinear Regions and Recognize Polyploidy Events Based on the K_S Dotplot

*Fushun Wang[1,2†], Kang Zhang[3,4,5†], Ruolan Zhang[1], Hongquan Liu[6], Weijin Zhang[1], Zhanxiao Jia[1] and Chunyang Wang[3,4]**

[1]Department of Information Science and Technology, Hebei Agricultural University, Baoding, China, [2]Hebei Key Laboratory of Agricultural Big Data, Baoding, China, [3]Department of Life Science, Hebei Agricultural University, Baoding, China, [4]State Key Laboratory of North China Crop Improvement and Regulation, Hebei Agricultural University, Baoding, China, [5]Hebei Key Laboratory of Plant Physiology and Molecular Pathology, Hebei Agricultural University, Baoding, China, [6]Department of Urban and Rural Construction, Hebei Agricultural University, Baoding, China

OPEN ACCESS

Edited by:
Manish Kumar Pandey,
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India

Reviewed by:
Victor A. Albert,
University at Buffalo, United States
Aamir W. Khan,
University of Missouri, United States

***Correspondence:**
Chunyang Wang
shmwcy@hebau.edu.cn

†These authors share first authorship between correspondence and specialty section

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

Received: 23 December 2021
Accepted: 21 March 2022
Published: 20 April 2022

Citation:
Wang F, Zhang K, Zhang R, Liu H, Zhang W, Jia Z and Wang C (2022) PolyReco: A Method to Automatically Label Collinear Regions and Recognize Polyploidy Events Based on the K_S Dotplot. Front. Genet. 13:842387. doi: 10.3389/fgene.2022.842387

Polyploidization plays a critical role in producing new gene functions and promoting species evolution. Effective identification of polyploid types can be helpful in exploring the evolutionary mechanism. However, current methods for detecting polyploid types have some major limitations, such as being time-consuming and strong subjectivity, etc. In order to objectively and scientifically recognize collinearity fragments and polyploid types, we developed PolyReco method, which can automatically label collinear regions and recognize polyploidy events based on the K_S dotplot. Combining with whole-genome collinearity analysis, PolyReco uses DBSCAN clustering method to cluster K_S dots. According to the distance information in the *x*-axis and *y*-axis directions between the categories, the clustering results are merged based on certain rules to obtain the collinear regions, automatically recognize and label collinear fragments. According to the information of the labeled collinear regions on the *y*-axis, the polyploidization recognition algorithm is used to exhaustively combine and obtain the genetic collinearity evaluation index of each combination, and then draw the genetic collinearity evaluation index graph. Based on the inflection point on the graph, polyploid types and related chromosomes with polyploidy signal can be detected. The validation experiments showed that the conclusions of PolyReco were consistent with the previous study, which verified the effectiveness of this method. It is expected that this approach can become a reference architecture for other polyploid types classification methods.

Keywords: clustering, collinearity fragment, polyploidy, DBSCAN, chromosome

INTRODUCTION

Studying the process of polyploidization is essential for the in-depth understanding of evolutionary laws (Marcet-Houben and Gabaldón 2015), and exploring the stability and chromosome rearrangement of the genome. Polyploidization of gymnosperms and almost all angiosperms are considered to be the main reason for the diversity of land plants (Li et al., 2016; Hao et al., 2017). Polyploidy can produce a large number of duplicated genes in the genome (Wang et al., 2018). These

genes may play an important role in functional evolution, environmental adaptation, and new species formation (Wang et al., 2017; Wang et al., 2019). The recombination of some homologous chromosomes after polyploidization often causes the instability of the genome structure, and processes such as chromosome breakage and fusion often occur, which can lead to large-scale duplicated gene loss in the genome (Wang et al., 2007; Wang et al., 2011). If two species have a common ancestor, after polyploidization, although there will be differences between the genomes, the two species still have a relatively close relationship. This close relationship can be expressed in the form of collinearity. The more complete the collinearity fragment, the closer relationship between the two species is. Exploring the collinearity between species has big significance in understanding the origin of species and the evolution of the genome.

Cheng et al. (2019) drew a K_S dotplot of homologous genes within Spirogloea muscicola, and found that it had recently experienced a whole genome triplication event. Through collinearity analysis, Wang et al., 2011 found that *Brassica rapa* and *Arabidopsis thaliana* experienced a whole genome triplication event. By constructing a phylogenetic gene tree, Dong et al. (2021) revealed that a whole-genome duplication event occurred in Magnoliales and Laurales. Xu et al. (2020) by drawing a phylogenetic gene tree, found that Scutellaria baicalensis and Scutellaria barbata had a whole-genome duplication event about 13.28 million years ago. Yan et al., 2021 by drawing the distribution graph of the synonymous substitution rate (K_S), found two WGD events in Juglans mandshurica and Juglans regia. By drawing the distribution graph of the synonymous substitution rate (K_S).

Although the K_S distribution graph combined with the molecular clock (Miyata et al., 1980) can calculate the doubling time, it is a challenge to determine the collinearity information between the chromosomes. In addition, the above method, which injects prior knowledge, manually marks the collinear area by observing the atlas, and then recognizes the polyploidization through the combination of the regions. This kind of recognition method has low recognition efficiency, high dependence on prior knowledge, as well as strong subjectivity, and easy to introduce human error. Due to the lack of objective evaluation criteria, the identification of the polyploid types of is still very challenging. In terms of the types of polyploidization and the choice of chromosomes, the same atlas will cause different personal perceptions. This deviation will affect the subsequent research on chromosome rearrangement (Zhang et al., 2021). Therefore, we develop a computational model PolyReco to accurately identify and characterize some polyploid types in atlas.

Considering only K_S values for identifying polyploid types might be insufficient, we add the gene positional information in PolyReco. Genes are aligned in sequential order on each of the chromosomes, so incorporating the gene positional information on chromosomes with K_S values will likely increase the accuracy of polyploid type classification. In this study, sequence comparisons were performed based on the whole genome data of *Vitis vinifera* and *Salix sinopurpurea*, combined with whole genome collinearity analysis, to obtain the summary data of

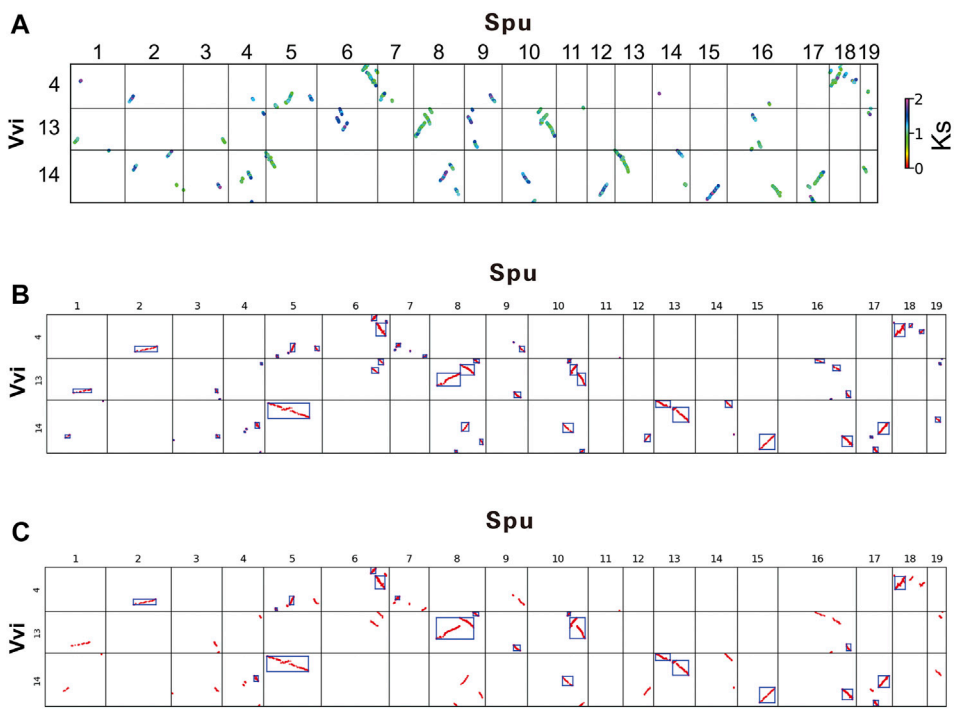


FIGURE 1 | (A) K_S dotplot between Salix sinopurpurea and Vitis vinifera genome homologous genes **(B)** DBSCAN cluster recognition effect figure **(C)** Automatic label result of collinearity fragments based on DBSCAN.

TABLE 1 | Partial data of grape chromosome 13 cluster.csv file.

chr1	chr2	id	l_x	l_y	num	r_x	r_y	y1-y2	x1-x2
1	13	1	145	354	32	257	228	126	112
4	13	1	1,266	1,153	10	1,310	1,097	56	44
6	13	1	838	1,267	17	913	1,097	170	75
6	13	2	735	994	26	834	840	154	99
8	13	1	592	1,267	56	671	1,155	112	79

TABLE 2 | Partial data of grape chromosome 13 combine.csv file.

chr1	chr2	id	l_x	l_y	num	r_x	r_y	Δy	Δx
8	13	1	592	1,267	56	671	1,155	112	79
8	13	2	78	1,089	239	605	446	643	527
9	13	1	378	257	63	478	89	168	100
10	13	1	1,327	1,267	65	1,426	1,157	110	99
10	13	2	1,419	1,088	276	1,934	457	631	515

homology information and K_S values between genomes. PolyReco comprehensively utilizes digital image processing technology and DBSCAN method, and realizes the automatic recognition and labeling of the collinear region based on the K_S dotplot of homologous genes. The model uses the collinear area as the unit and combines the combination strategy to construct the combination evaluation standard. According to the performance of the chromosome combination, determine the specific polyploidization and draw the combined figure of the polyploidization. This study aims to develop a polyploidization classification tool, which has the potential to take chromosome position information into account with the K_S values for boosting polyploid type prediction performance.

MATERIALS AND METHODS

Data Sources

With the whole genome data of *Salix sinopurpurea* (Spu) and *Brassica rapa* (Bra) as the main research materials, comparative genomics was used to compare the collinearity between *Salix sinopurpurea* and the reference genome *Vitis vinifera* (Vvi), *Brassica rapa* and the reference genome *Arabidopsis thaliana* (Ath).

Genomes and their gene annotations of *Salix sinopurpurea* and *Vitis vinifera* were downloaded from Joint Genome Institute. Download the required documents for *Brassica rapa* and *Arabidopsis thaliana* at http://brassicadb.org/and http://www.arabidopsis.org/shangxia, respectively.

Preprocessing of the Data Sources

Due to the huge amount of original genome data, in order to extract target data from the genome sequence and annotation files, the downloaded genome data is processed with a custom python script to obtain the blast results, which is convenient for subsequent research and analysis. Screen the original data of species genomes, information was extracted from the genome annotation files, which include chromosome number, gene start and end positions, gene transcription direction, and gene ID information, and then rename the gene ID and the number of the genes was given in order of their appearance on chromosomes. Map the gene ID in the CDS sequence and protein sequence file to the new ID of the corresponding gene in the genome annotation file. Label the processed genomic data with a unified naming method.

Homologous Sequence Alignment

Blastp was used to explore to align genomic sequences of different species. Screen out gene pairs with the expected value (E-value) not greater than 10–5 and score evaluation (Score) higher than 100, so that the subsequent genome collinearity analysis results are more reliable.

Draw the K_S Dotplot of Homologous Genes

The WGDI (Sun et al., 2021) use MAFFT (Wong, Suchard, and Huelsenbeck 2008) or MUSCLE (Edgar 2004) to perform multiple sequence alignment, and calculates the synonymous substitution rate using the yn00 (Yang et al., 2000) or ng86 (Nei and Gojobori 1986) program of the PAML package. Finally, the visualization is realized by extracting block, and then output blockinfo file.

Collinear Fragment Labeling Method Based on Clustering

In this paper, the input for DBSCAN requires the blockinfo file generated by WGDI and the chromosome length information (len file) of the two species. By setting the epsilon (eps) and minimum points (MinPts), cluster analysis is performed on the collinearity fragments in the K_S dotplot. The collinear region was then obtained from the clustering results combined with certain rules for merging. And then realizes the automatic identification and labeling of the collinear region. The comparison result of a chromosome of the target species and a chromosome of the reference species is shown as a cell on the K_S dotplot, that is, a comparison unit.

The K_S dotplot between *Salix sinopurpurea* and *Vitis vinifera* genome homologous genes drawn by wgdi (**Figure 1A**). The horizontal axis represents the chromosome of the target species (*Salix sinopurpurea*) and the vertical represents the chromosome of the reference species (*Vitis vinifera*). On the K_S dotplot, the chromosome number of *Salix sinopurpurea* is shown from left to right, and the chromosome number of *Vitis vinifera* is shown from top to bottom. The K_S value ranges from 0.00 to 2.00. As shown in the figure, different colored points correspond to different K_S values. It can be observed in **Figure 1A**, in addition to the clear and complete homologous fragments of grape chromosome 4 with *Salix sinopurpurea* chromosomes 6 and 18, it also has fuzzy and unclear homologous fragments with *Salix sinopurpurea* chromosomes 1, 2 and 4. The reason why these fragments are unclear and incomplete is that they are doubled by the whole genome triplication events shared by older dicots. The collinearity of the homologous fragments produced by the whole genome triplication events shared by ancient dicotyledons is far inferior to that of the whole genome duplication events shared by the Salicaceae. The specific manifestation is that the K_S value is significantly large, belongs to the blue-purple system, scarce and fragmented seriously. The

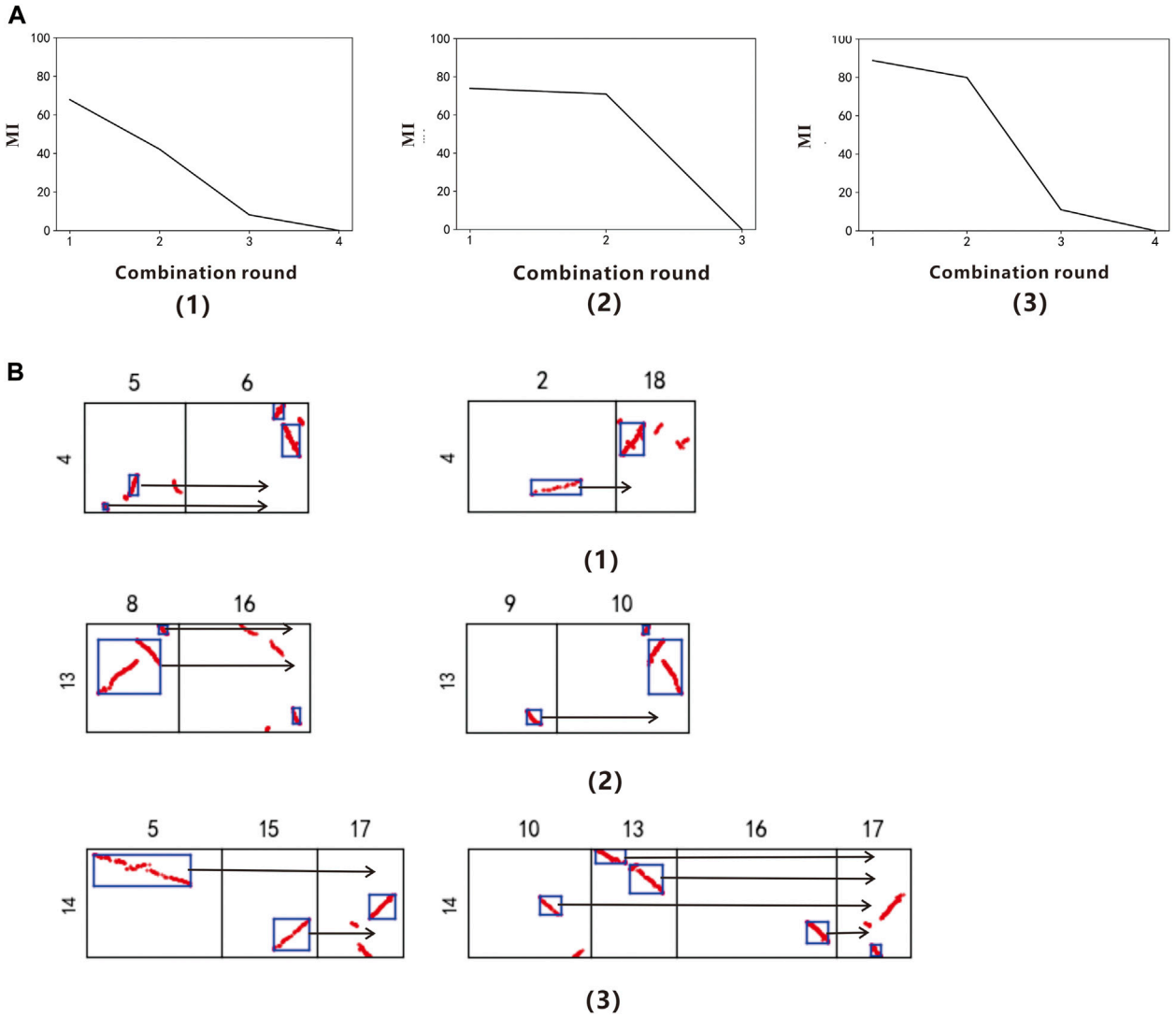


FIGURE 2 | (A) Collinearity evaluation index line chart of Vitis vinifera Chr.4, Chr.13, and Chr.14 **(B)** Combination figure of Salix sinopurpurea polyploidy. **(A,B)** correspond to each other. (1) The grape chromosome 4 (2) The grape chromosome 13 (3) The grape chromosome 14.

results showed the structural similarities and differences between genomes. The generated data and pictures provide references for follow-up research.

Using the DBSCAN algorithm, by setting the eps to 50 and the MinPts to 3, cluster the K_S dotplot between *Salix sinopurpurea* and *Vitis vinifera* genome homologous genes. The algorithm outputs the clustering result figure (**Figure 1B**), in which each category is represented by a rectangular box. The DBSCAN can cluster out complete collinearity fragments, in grape chromosome 14 and *Salix sinopurpurea* chromosome 5, as well as in grape chromosome 14 and *Salix sinopurpurea* chromosome 15. It also can identify fragmented collinearity fragments in grape chromosome 4 and *Salix sinopurpurea* chromosome 5, grape chromosome 13 and *Salix sinopurpurea* chromosome 16. These will accurately reflect the relationship between the collinearity fragments and improve the subsequent combination effect.

The model sorts the category in the same comparison unit from top to bottom to generate ID, and calculates the number of homologous gene points in the box (num), the length of the box ($y1$-$y2$), and the width of the box ($x1$-$x2$). The model then generates cluster.csv files that contain the target species chromosome number (chr1), the reference species chromosome number (chr2), ID, and the horizontal and vertical coordinates of the upper left corner point are l_x and l_y, respectively, num, the horizontal and vertical coordinates of the lower right point are r_x and r_y, respectively, $y1$-$y2$, and $x1$-$x2$. Part of the data in the cluster.csv file of grape chromosome 13 is shown in **Table 1**. Among them, chromosome 13 and chromosome 1 form a class. The coordinates of the upper left corner of this class are 145, 354, and the coordinates of the lower right corner are 257, 228. The number of homologous genes contained is

TABLE 3 | *Vitis vinifera* Chr. 13 result.csv file.

chr1	chr2	id	l_x	l_y	num	r_x	r_y	Sumy	Δy	Δx	Comro
8	13	2	78	1,089	239	605	446	947	643	527	1
8	13	1	592	1,267	56	671	1,155	947	112	79	1
16	13	1	1,163	283	42	1,235	91	947	192	72	1
10	13	2	1,419	1,088	276	1,934	457	909	631	515	2
10	13	1	1,327	1,267	65	1,426	1,157	909	110	99	2
9	13	1	378	257	63	478	89	909	168	100	2

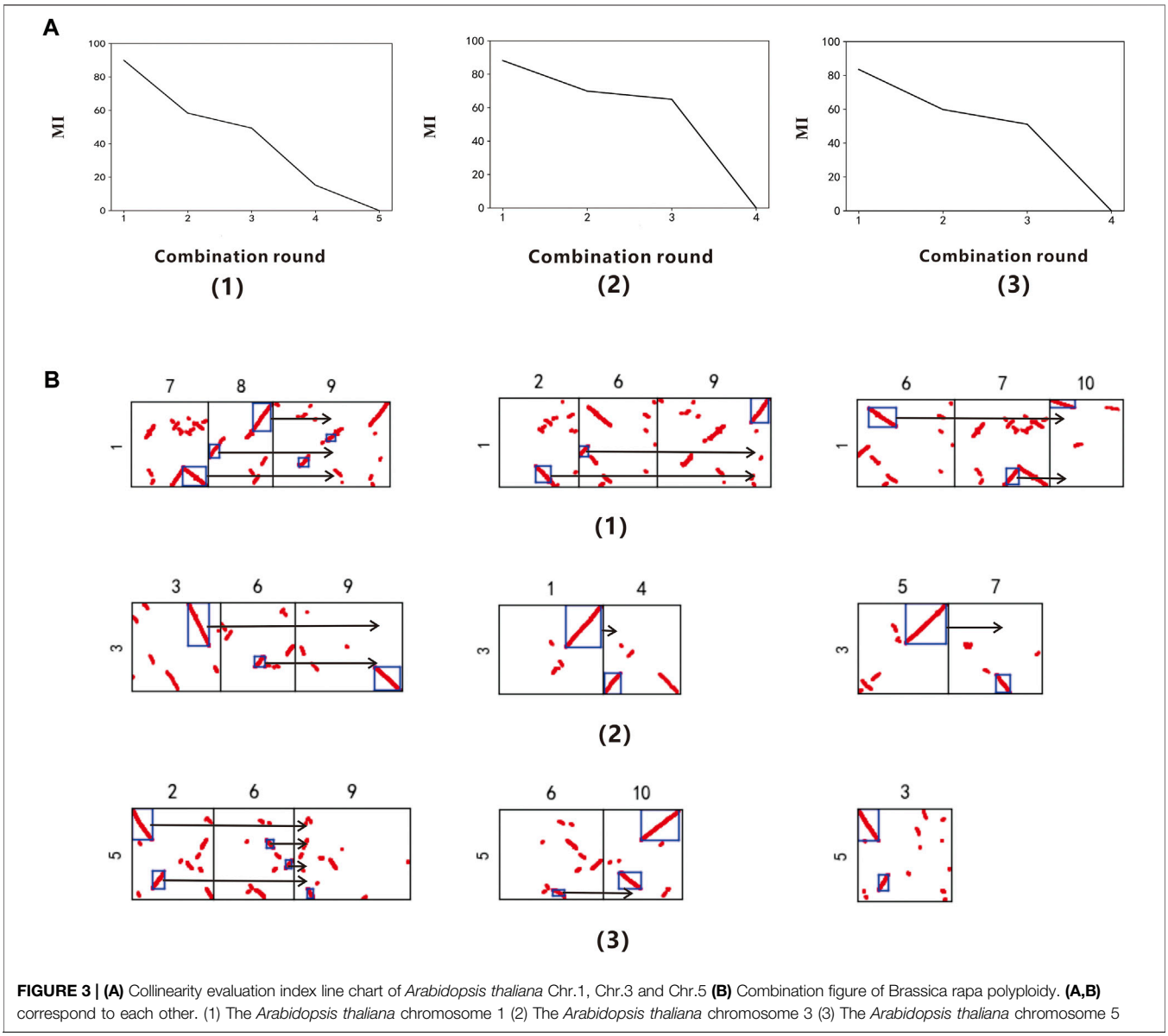


FIGURE 3 | (A) Collinearity evaluation index line chart of *Arabidopsis thaliana* Chr.1, Chr.3 and Chr.5 **(B)** Combination figure of Brassica rapa polyploidy. **(A,B)** correspond to each other. (1) The *Arabidopsis thaliana* chromosome 1 (2) The *Arabidopsis thaliana* chromosome 3 (3) The *Arabidopsis thaliana* chromosome 5

32. The length of the cluster box is 126 coordinate lengths, and the width is 112 coordinate lengths.

In order to perform a combined analysis on the identified collinearity fragments, the model read the cluster.csv file generated by clustering. The model uses *y*_gap and *x*_gap, which represents the gap in the longitudinal and horizontal directions of adjacent collinear segments, as the basis for judging overlap. The location information of the gene is combined to set the parameters gap and Slen. In the comparison unit, the parameter gap represents the mean value of *y*_gap. Through a series of experiments and

TABLE 4 | *Arabidopsis thaliana* Chr. 3 result.csv file.

chr1	chr2	id	l_x	l_y	num	r_x	r_y	Sumy	Δy	Δx	Comro
3	3	1	2,919	5,436	857	4,001	2,763	4,793	2,673	1,082	1
9	3	1	3,322	1,458	827	4,386	8	4,793	1,450	1,064	1
6	3	1	1,571	2,132	260	2,021	1,462	4,793	670	450	1
1	3	1	2,581	5,431	1,036	3,948	2,863	3,799	2,568	1,367	2
4	3	1	3	1,231	501	631	0	3,799	1,231	628	2
5	3	1	1,955	5,432	1,374	3,668	3,021	3,533	2,411	1,713	3
7	3	1	1,555	1,130	373	1,998	8	3,533	1,122	443	3

continuous optimization of parameter selection, it is finally determined that 1/6 of the corresponding chromosome length of the target species is the value of parameter Slen. For all collinearity fragments whose num is greater than the specified value, one condition is that the collinearity fragments do not overlap, another is overlap. In the first condition, collinear fragments will be merged if $0 \leq y\_gap \leq gap$ and $0 \leq x\_gap \leq$ Slen. And in the second condition, there is overlap in the *y*-axis direction, they will be merged when it meets $0 \leq x\_gap <$ Slen; if there is overlap in the *x*-axis direction, when it meets $0 \leq y\_gap \leq$ gap, merge them. Finally, the model output the clustering result graph (**Figure 1C**), in which the merged result is marked with a rectangular box, and the combine.csv file is generated.

The content of the combine.csv file is the same as the cluster.csv file. **Table 2** shows some data of the combine.csv file. Among them, chromosome 13 and chromosome 8 form two classes. The coordinates of the upper left corner of the first class are 592, 1,264, and the lower right corner are 671, 1,155. The number of homologous genes contained in this class is 56. The length of the labeled box Δy is 112 coordinate length, and the width Δx is 79 coordinate length; the upper left corner coordinate of the second class are 78, 1,089, and the lower right corner are 605, 446. The number of homologous genes contained in this class is 239, Δy is 643 coordinate length, Δx is 537.

Polyploidy Recognition Algorithm

In order to determine the polyploid types and related chromosomes of the species, we develop the polyploidy recognition algorithm. The algorithm read the generated combine.csv file, look for the labeled box with the largest Δy, mark it, and then look up and down to find the labeled box with the length less than it. The result.csv file will be built by adopting exhaustive above process. Among them, comro represents the combination round, sumy represents the sum of Δy in same combination round. In order to determine the specific polyploidy of the species, the gene collinearity evaluation index line chart is drawn. The horizontal is the combined round, and the vertical is the corresponding gene collinearity evaluation index. The significant inflection point in the line chart represents the corresponding polyploid type. The gene collinearity evaluation index (MI) was calculated by dividing the cumulative collinearity fragments length to the corresponding chromosome length of the reference species in the len file to describe the performance of the polyploidy in the corresponding combination round, and the larger its value, the better the performance.

$$MI = \frac{\sum_{m=1}^{n} \Delta y_m}{len_i}$$

Where Δy_m and n are the length and number of collinearity fragments in the same combination round, respectively; len_i is the corresponding chromosome length of the reference species, i is the corresponding chromosome number.

After determining the combination round, output the combined result graph. To describe the analysis of input, output and fetching the final results of the analysis, we made pseudo code. The pseudo code of the chromosome collinearity fragment labeling and polyploidy recognition algorithm is shown as below for a better understanding of the context and better assess the relevance of this paper.

Algorithm 1. Chromosome collinearity fragment labeling and polyploidy recognition algorithm.

Chromosome collinearity fragment labeling and polyploidy recognition algorithm

Input: the csv file generated by WGDI, chr1 list, chr2 list

Output: cluster.csv, combine.csv, result.csv, line chart, combination chart

```
x=len(che1); y= len(che2)
initial eps, MinPts
clustering K_S botplot by DBSCAN
output cluster.csv
for (i=1, j=1;i<=x, j<=y; i++, j++ )
    initial gap, gap_wide
    merge the clustering box
    update cluster.csv
output combine.csv
for (m=1; m<=y; m++)
    find the longest segment
    search up and down to merge
    update result.csv
output result.csv
calculate MI
output line chart
output combination chart
```

RESULTS

Salix sinopurpurea Polyploidy Recognition

Using PolyReco to objectively determine the polyploid types of *Salix sinopurpurea*. The model use *Vitis vinifera* as the reference

genome to identify the target species *Salix sinopurpurea* polyploid type, and read the data of *Vitis vinifera* chromosome 4, 13 and 14. The DBSCAN algorithm obtains the collinearity fragments, and get the combine.csv file. Then using the polyploidy recognition algorithm to combine the labeled boxes exhaustively, and get the gene collinearity evaluation index table of each chromosome in different combination rounds (**Supplementary Table S1**). In the colinearity evaluation index line chart (**Figure 2A**), we can find that chromosomes 4, 13, and 14 of *Vitis vinifera* have obvious inflection points when the combined round is 2. Therefore, it is determined that the *Salix sinopurpurea* has a whole genome duplication event recently. This conclusion can be found in Wang et al., 2011.

After determining the specific polyploidy, we can output the information of the labeled box participating in the polyploidy, and obtain the result.csv file of *Vitis vinifera* chromosomes 4, 13, and 14, in which the data of chromosome 13 is shown in **Table 3**.

According to the result.csv file, output the combined figure of the *Salix sinopurpurea* polyploidy (**Figure 2B**). When the combination round is 2, get the two groups with the highest scores, among which the chromosomes 5 and 6 of the *Salix sinopurpurea* can be combined into a relatively complete chromosome 4 of *Vitis vinifera*, the corresponding MI is 67.87%; the chromosomes 2 and 18 of the *Salix sinopurpurea* can be combined into a relatively complete chromosome 4 of *Vitis vinifera*, and the corresponding MI is 42.19%. The chromosomes 8 and 16 of the *Salix sinopurpurea* can be combined into a relatively complete *Vitis vinifera* chromosome 13 with MI of 73.93%; the chromosomes 9 and 10 of the *Salix sinopurpurea* can be combined into a relatively complete *Vitis vinifera* chromosome 4, MI is 70.96%. The chromosomes 5, 15, and 17 of *Salix sinopurpurea* can be combined into a relatively complete *Vitis vinifera* chromosome 14 with MI of 88.74%; the chromosomes 10, 13, 16, and 17 of *Salix sinopurpurea* can be combined into a relatively complete *Vitis vinifera* chromosome 14, MI is 79.88%.

Brassica rapa Polyploidy Recognition

In order to further verify the universality of the method, according to the procedure in 3.1, the model uses *Arabidopsis thaliana* as the reference genome to identify the polyploidy type of the target species *Brassica rapa*. Through reading the data of chromosomes 1, 3 and 5 of *Arabidopsis thaliana*, we finally get the gene collinearity evaluation index table of each chromosome in different combination rounds (**Supplementary Table S2**). In the collinearity evaluation index line chart (**Figure 3A**), it can be found that chromosomes 1, 3, and 5 of *Arabidopsis thaliana* had an obvious turning point when the combination round was 3. Therefore, it is determined that the *Brassica rapa*. had a whole genome triplication event recently. This conclusion can be found in Wang (2011a).

After determining the specific polyploidy, the result.csv file of *Arabidopsis thaliana* chromosomes 1, 3, and 5 is obtained, in which the data of chromosome 3 is shown in **Table 4**.

According to the result.csv file, output the combined figure of *Brassica rapa* polyploidy (**Figure 3B**). When the combination round is 3, get the three groups with the highest scores, among which the chromosomes 7, 8, and 9 of the *Brassica rapa* can be combined into a relatively complete chromosome 1 of *Arabidopsis thaliana*, the corresponding MI is 89.88%; the chromosomes 2, 6, and 9 of the *Brassica rapa* can be combined into a relatively complete chromosome 1 of *Arabidopsis thaliana*, and the corresponding MI is 58.31%; the chromosomes 6, 7, and 10 of the *Brassica rapa* can be combined into a relatively complete chromosome 1 of *Arabidopsis thaliana*, and the corresponding MI is 49.38%. The chromosomes 3, 6, and 9 of the *Brassica rapa* can be combined into a relatively complete *Arabidopsis thaliana* chromosome 3 with MI of 88.16%; the chromosomes 1 and 4 of the *Brassica rapa* can be combined into a relatively complete *Arabidopsis thaliana* chromosome 3 with MI of 69.87%; the chromosomes 5 and 7 of the *Brassica rapa* can be combined into a relatively complete *Arabidopsis thaliana* chromosome 3 with MI of 64.98%. The chromosomes 2, 6, and 9 of the *Brassica rapa* can be combined into a relatively complete *Arabidopsis thaliana* chromosome 5, MI is 83.54%; the chromosomes 6 and 10 of the *Brassica rapa* can be combined into a relatively complete *Arabidopsis thaliana* chromosome 5, MI is 59.84%; the chromosomes 3 of the *Brassica rapa* can be combined into a relatively complete *Arabidopsis thaliana* chromosome 5, MI is 51.17%.

In **Figure 3B**, there are two seemingly identical collinearity fragments among the four fragments formed by *Arabidopsis thaliana* chromosome 3 and *Brassica rapa* chromosome 4, with the naked eye. But only one is labeled and used, because it has a small number of homologous genes. So this method can break through the limitations of the human eye, and find chromosome fragments with strong collinearity, as well as provide a basis for objective judgment of polyploidy.

DISCUSSION

The previous study has mostly used to observe the atlas with prior knowledge to identify the polyploid types of the species. This method has some major limitations, such as low efficiency, high dependence on prior knowledge, strong subjectivity, lack of objective evaluation criteria, and easy introduction of human error. In this paper, digital image processing technology was used to identify polyploid types based on clustering algorithms. The K_S dotplot of homologous genes was used as the research object, and the DBSCAN method was used to cluster. Then we can obtain the collinear fragments and automatically label collinear region. According to the gene collinearity evaluation index line chart of each combination, the model can determine the polyploid type and related chromosome combination. The study mainly focused on developing a polyploidization recognition algorithm and providing the method to speed up the evolutionary laws of gene structure associated with polyploidy research. PolyReco involves more than a simple labels of collinear regions, but also gives the polyploidy types through the collinearity evaluation index line chart and related chromosomes at the end. Compared with MCScanX (Wang et al., 2012), PolyReco labels the specific gene segments involved in the polyploidy events and improves the recognition efficiency of polyploidy. Compared to traditional methods, PolyReco reduces the dependence on prior knowledge, solves the limitations of the human eye in visual space, comply with artificial logic analysis and reasoning process.

Moreover, the PolyReco can not only provides an effective method for large-scale rapid identification of genome polyploidy but also has important application value in distant hybrid breeding (Rabanus-Wallace et al., 2021).

In summary, the proposed PolyReco provides a reference model for processing automatically label collinear regions and recognize polyploidy. However, the K_S dotplot is sensitive to the size of the parameter Eps. When a large value is used for Eps, the fragmented collinearity segments are easy to cluster together. On the contrary, it is easy to separate continuous fragments so that complete collinearity fragments cannot be clustered. In the next step, we expect to study the DBSCAN clustering method based on adaptive Eps to further optimize the clustering effect.

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

FW and KZ conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper. RZ performed the experiments, analyzed the data, prepared figures and/or tables, and authored drafts of the paper. HL designed the experiments and analyzed the data. WZ and ZJ analyzed the data, prepared figures and/or tables. CW conceived and designed the experiments, authored or reviewed drafts of the paper. Manuscript is approved by all authors for publication.

FUNDING

This work was supported by Science and Technology Research Project of Hebei Province Colleges and Universities (No. QN2020421); the Scientific Research Project of Introducing Talents of Hebei Agricultural University (No. YJ201944); the Innovative Research Group Project of Hebei Natural Science Foundation (Grant No. C2020204111); the International Science and Technology Cooperation base Special Project of Hebei (Grant No. 20592901D); National Natural Science Foundation of China (31901864).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.842387/full#supplementary-material

REFERENCES

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., et al. (2019). Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* 179 (5), 1057–1067. doi:10.1016/j.cell.2019.10.019

Dong, S., Liu, M., Liu, Y., Chen, F., Yang, T., Chen, L., et al. (2021). The Genome of Magnolia Biondii Pamp. Provides Insights into the Evolution of Magnoliales and Biosynthesis of Terpenoids. *Hortic. Res.* 8 (1), 38. doi:10.1038/s41438-021-00471-9

Edgar, R. C. (2004). MUSCLE: a Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics* 5, 113. doi:10.1186/1471-2105-5-113

Hao, M., Li, A., Shi, T., Luo, J., Zhang, L., Zhang, X., et al. (2017). The Abundance of Homoeologue Transcripts Is Disrupted by Hybridization and Is Partially Restored by Genome Doubling in Synthetic Hexaploid Wheat. *Bmc Genomics* 18, 149. doi:10.1186/s12864-017-3558-0

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., and De Smet, R. (2016). Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. *Plant Cell* 28 (2), 326–344. doi:10.1105/tpc.15.00877

Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *Plos Biol.* 13 (8), e1002220. doi:10.1371/journal.pbio.1002220

Miyata, T., Yasunaga, T., and Nishida, T. (1980). Nucleotide Sequence Divergence and Functional Constraint in mRNA Evolution. *Proc. Natl. Acad. Sci. U.S.A.* 77 (12), 7328–7332. doi:10.1073/pnas.77.12.7328

Nei, M., and Gojobori, T. (1986). Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Mol. Biol. Evol.* 3 (5), 418–426. doi:10.1093/oxfordjournals.molbev.a040410

Rabanus-Wallace, M. T., Hackauf, B., Mascher, M., Lux, T., Wicker, T., Gundlach, H., et al. (2021). Chromosome-scale Genome Assembly Provides Insights into rye Biology, Evolution and Agronomic Potential. *Nat. Genet.* 53 (4), 564–573. doi:10.1038/s41588-021-00807-0

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., et al. (2021). WGDI: A User-Friendly Toolkit for Evolutionary Analyses of Whole-Genome Duplications and Ancestral Karyotypes. *bioRxiv*. doi:10.1101/2021.04.29.441969

Wang, J., Sun, P., Li, Y., Liu, Y., Yang, N., Yu, J., et al. (2018). An Overlooked Paleotetraploidization in Cucurbitaceae. *Mol. Biol. Evol.* 35 (1), 16–26. doi:10.1093/molbev/msx242

Wang, J., Sun, P., Li, Y., Liu, Y., Yu, J., Ma, X., et al. (2017). Hierarchically Aligning 10 Legume Genomes Establishes a Family-Level Genomics Platform. *Plant Physiol.* 174 (1), 284–300. doi:10.1104/pp.16.01981

Wang, J., Yuan, J., Yu, J., Meng, F., Sun, P., Li, Y., et al. (2019). Recursive Paleohexaploidization Shaped the Durian Genome. *Plant Physiol.* 179 (1), 209–219. doi:10.1104/pp.18.00921

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The Genome of the Mesopolyploid Crop Species Brassica Rapa. *Nat. Genet.* 43 (10), 1035–1039. doi:10.1038/ng.919

Wang, X., Tang, H., Bowers, J. E., Feltus, F. A., and Paterson, A. H. (2007). Extensive Concerted Evolution of rice Paralogs and the Road to Regaining independence. *Genetics* 177 (3), 1753–1763. doi:10.1534/genetics.107.073197

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity. *Nucleic Acids Res.* 40 (7), e49. doi:10.1093/nar/gkr1293

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment Uncertainty and Genomic Analysis. *Science* 319 (5862), 473–476. doi:10.1126/science.1151532

Xu, Z., Gao, R., Pu, X., Xu, R., Wang, J., Zheng, S., et al. (2020). Comparative Genome Analysis of Scutellaria Baicalensis and Scutellaria Barbata Reveals the Evolution of Active Flavonoid Biosynthesis. *Genomics, Proteomics & Bioinformatics* 18 (3), 230–240. doi:10.1016/j.gpb.2020.06.002

Yan, F., Xi, R. M., She, R. X., Chen, P. P., Yan, Y. J., Yang, G., et al. (2021). Improved De Novo Chromosome-level Genome Assembly of the Vulnerable walnut Tree Juglans Mandshurica Reveals Gene Family Evolution and Possible Genome Basis of Resistance to Lesion Nematode. *Mol. Ecol. Resour.* 21 (6), 2063–2076. doi:10.1111/1755-0998.13394

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000). Codon-substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics* 155 (1), 431–449. doi:10.1093/genetics/155.1.431

Zhang, Y., Wang, F. S., Zhang, Z. K., Jia, Z. X., and Wang, C. Y. (2021). Music Emotion Recognition Method Based on Multi Feature Fusion. *Ijart* 13, 1–22. doi:10.1504/ijart.2021.10043883

Zhao, M. H. (2019). "Comparative Genomics and Bioiformatics Research into Salicaceae Genomes," (Tangshan, China: North China University of Science and Technology). Dissertation.

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: 31 May 2022
doi: 10.3389/fgene.2022.868015

StarGazer: A Hybrid Intelligence Platform for Drug Target Prioritization and Digital Drug Repositioning Using Streamlit

*Chiyun Lee[1†], Junxia Lin[2†], Andrzej Prokop[3], Vancheswaran Gopalakrishnan[4], Richard N. Hanna[5], Eliseo Papa[6], Adrian Freeman[7], Saleha Patel[7], Wen Yu[8], Monika Huhn[9], Abdul-Saboor Sheikh[1], Keith Tan[10], Bret R. Sellman[4], Taylor Cohen[4], Jonathan Mangion[1], Faisal M. Khan[8], Yuriy Gusev[2] and Khader Shameer[8]**

[1]Data Science and Artificial Intelligence, BioPharmaceuticals R&D, AstraZeneca, Cambridge, United Kingdom, [2]Georgetown University, Washington, DC, United States, [3]Biometrics, Oncology R&D, AstraZeneca, Warsaw, Poland, [4]Discovery Microbiome, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, United States, [5]Early Respiratory and Immunology, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, United States, [6]Research Data and Analytics, R&D IT, AstraZeneca, Cambridge, United Kingdom, [7]Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, United Kingdom, [8]Data Science and Artificial Intelligence, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, United States, [9]Biometrics and Information Sciences, BioPharmaceuticals R&D, AstraZeneca, Mölndal, Sweden, [10]Neuroscience, BioPharmaceuticals R&D, AstraZeneca, Cambridge, United Kingdom

Target prioritization is essential for drug discovery and repositioning. Applying computational methods to analyze and process multi-omics data to find new drug targets is a practical approach for achieving this. Despite an increasing number of methods for generating datasets such as genomics, phenomics, and proteomics, attempts to integrate and mine such datasets remain limited in scope. Developing hybrid intelligence solutions that combine human intelligence in the scientific domain and disease biology with the ability to mine multiple databases simultaneously may help augment drug target discovery and identify novel drug-indication associations. We believe that integrating different data sources using a singular numerical scoring system in a hybrid intelligent framework could help to bridge these different omics layers and facilitate rapid drug target prioritization for studies in drug discovery, development or repositioning. Herein, we describe our prototype of the StarGazer pipeline which combines multi-source, multi-omics data with a novel target prioritization scoring system in an interactive Python-based Streamlit dashboard. StarGazer displays target prioritization scores for genes associated with 1844 phenotypic traits, and is available via https://github.com/AstraZeneca/StarGazer.

Keywords: multi-omics, target prioritization, drug discovery, repositioning, data integration, streamlit, stargazer, hybrid intelligence

OPEN ACCESS

Edited by:
Rinku Sharma, Harvard Medical School, United States

Reviewed by:
Rajesh Kumar Pathak, Chung-Ang University, South Korea
Sezen Vatansever, Icahn School of Medicine at Mount Sinai, United States

****Correspondence:***
Khader Shameer
shameer.khader@astrazeneca.com

[†]These authors contributed equally to this work

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

Received: *01 February 2022*
Accepted: *29 April 2022*
Published: *31 May 2022*

Citation:
Lee C, Lin J, Prokop A, Gopalakrishnan V, Hanna RN, Papa E, Freeman A, Patel S, Yu W, Huhn M, Sheikh A-S, Tan K, Sellman BR, Cohen T, Mangion J, Khan FM, Gusev Y and Shameer K (2022) StarGazer: A Hybrid Intelligence Platform for Drug Target Prioritization and Digital Drug Repositioning Using Streamlit. Front. Genet. 13:868015. doi: 10.3389/fgene.2022.868015

INTRODUCTION

Drug repositioning has been rapidly gaining attention in the drug discovery domain during the past decade (Xue et al., 2018). Drug repositioning/repurposing describes the act of identifying alternative uses for a drug beyond the scope of its original indication, regardless of whether it has been FDA-approved or has failed in clinical trials (Pushpakom et al., 2019). The reasons for investing into drug repositioning are very numerous indeed.

Traditionally, a standard drug development cycle is estimated to take around 10 years and requires billions of dollars of investment, notwithstanding the still disappointingly high failure rate at clinical trials (Li et al., 2016). In light of these problems, drug repositioning holds potential to drastically reduce the time and money needed to bring a drug to the market: it has been estimated to reduce the time by half and cut costs by 5-fold when compared to developing a new drug from scratch (Shameer et al., 2015). These factors alone highlight the appealing opportunity to bring medicines to patients faster, and potentially into areas of unmet therapeutic demand. Moreover, it allows for the existing arsenal of approved drugs to be more broadly utilized, and for the opportunity to salvage some costs involved in the development of drugs that failed in clinical trials. Finally, the sheer variety in successful and promising repositioning strategies to date speaks to the potential for unearthing profound biological links between different diseases, driving paradigm shifts in our approach to modern medicine (Lee and Bhakta, 2021).

Drug target prioritization is an essential step for repositioning as it aims to highlight the potential drug targets for a particular disease. Applying computational methods to analyze and process multi-omics data is an effective approach for achieving this (Ashburn and Thor, 2004; Glicksberg et al., 2014; Shameer et al., 2018a; Pushpakom et al., 2019; Guo et al., 2021; Rapicavoli et al., 2022). Whilst there is now a vast wealth of biochemical and biomedical data in the current era of high-throughput omics technology, our ability to integrate and interpret these data has lagged behind and is presenting a great challenge in disease biology (Shameer et al., 2015). While machine learning approaches are generally used to develop tools to integrate, analyze and interpret multi-omics data, it remains a challenge that mere automation of predicting biological insights might overrepresent hypotheses that cannot be validated using function test experiments (Hodos et al., 2016; Peters et al., 2017). In such a scenario, we recommend the application of a hybrid intelligence platform that enables visual intelligence, quick search, contextual interpretations with quantitative approaches as a way to address this problem. Hybrid intelligence systems have been developed to address challenging problems in biomedicine, including remote patient diagnosis (Abu-Doleh et al., 2012; Li et al., 2014a; Akata et al., 2020; Guo et al., 2021; Weissler et al., 2021). However, such approaches are not readily available to address challenges in data integration and mining associated with drug target prioritization and drug repositioning.

Data from genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS) have been used for drug target prioritization (Ferrero and Agarwal, 2018). Whilst GWAS aim to identify associations between genetic variants with a single phenotype, PheWAS interrogate numerous phenotypic traits at once (Denny et al., 2010). As of 06 October 2021, the EMBL-EBI GWAS catalog collates associations from 5,370 studies that, in total, identified more than 290,000 associations. The utility of this GWAS dataset can be further amplified by narrowing down the genes of interest to only those with known drug indications (Sanseau et al., 2012). Importantly, a three-step strategy for drug repositioning using PheWAS data has already been proposed (Rastegar-Mojarad et al., 2015): (Xue et al., 2018)—identify all genes with known associations with the phenotypic trait of interest using PheWAS data; (Pushpakom et al., 2019);—identify all drugs with associations with the previously identified genes using data from DrugBank; and (Li et al., 2016)—return all the drugs identified in the previous step as candidates for repositioning for the original phenotypic trait of interest. Others have gone further by incorporating a combination of data from GWASs (Khosravi et al., 2019), expression profile analysis (Lau and So, 2020), functional annotation, biological network analysis, and gene-set association (Reay and Cairns, 2021).

Taken together, these data highlight the potential of using GWAS and PheWAS data for drug target prioritization. However, the field is still young, and integrating disparate data sources remains relatively limited in scope (Gallo et al., 2021). We hypothesize that integrating multimodal data sources using a singular numerical scoring system could accelerate the discovery and prioritization of drug targets. In light of this, we present our interactive dashboard, StarGazer, which aims to address these challenges by integrating three different datatypes (i.e., disease-target association, target druggability, and target protein-protein interaction) into a novel scoring system, utilizing real-time API calls and Python-based Streamlit technology. While these types of datasets have been used for numerous repositioning studies separately (Liu et al., 2014; Khaladkar et al., 2017; Hermawan et al., 2020; Wijetunga et al., 2020; Adikusuma et al., 2021; Attique et al., 2021; Ghoussaini et al., 2021; Portelli et al., 2021; Tan et al., 2021; Varghese and Majumdar, 2022; Zhao et al., 2022), StarGazer represents the first ever integration of the PheWAS catalog, Open Targets, STRING and Pharos, all of which are well-curated, well-studied, open access databases. Furthermore, computational repositioning studies focus largely on singular diseases, phenotypes or drugs, but StarGazer is equipped for flexible investigation into any of the 1,844 phenotypes and traits within the dashboard. Much of the data is up-to-date with the latest science, as it is loaded in real-time before it is analyzed in real time. StarGazer's drug target prioritization mode allows for rapid identification of potential drug targets for a disease of interest, also providing immediate analysis of various aspects surrounding drug development, such as druggability and the nature of the target-disease association. In addition to this target prioritization feature, we anticipate that StarGazer's ability to display all phenotypes associated with genes or gene variants of interest in an easily digestible manner to be of great value to exploratory or analytical workflows. Furthermore, StarGazer's other features include the support of initial

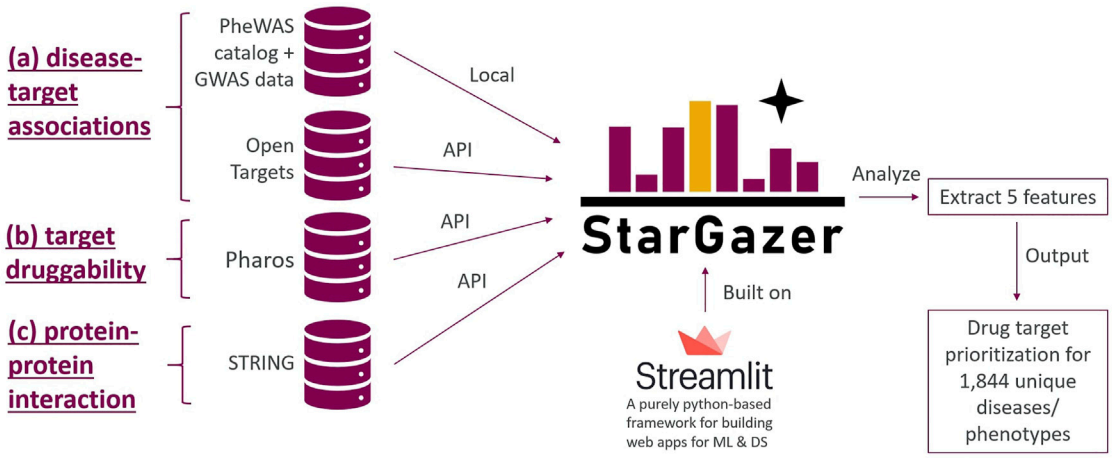


FIGURE 1 | The StarGazer drug target prioritization framework considers the following five features for each of the 1844 diseases in StarGazer's disease list (Xue et al., 2018):—the odds ratios of association between targets and phenotypic variants of interest from GWAS and PheWAS data (Pushpakom et al., 2019);—the target-disease association scores from Open Targets (Li et al., 2016);—the druggability data of genes of interest from Pharos (Shameer et al., 2015);—the degree of nodes in protein-protein interaction networks of genes of interest from STRING; and (Lee and Bhakta, 2021)—the presence of the gene variant of interest in both PheWAS and GWAS datasets. All data, except the PheWAS and GWAS data, are loaded in real-time by API calls and therefore present the latest evidence for drug repositioning strategies. The above five features are then integrated to provide a singular numerical StarGazer score which quantifies the drug repositioning potential of a gene. StarGazer is built on the Python-based Streamlit platform, which is largely used for building sleek and modern web applications for machine-learning and data science.

discoveries by interrogating the precise contribution of evidence from each data source.

DATA

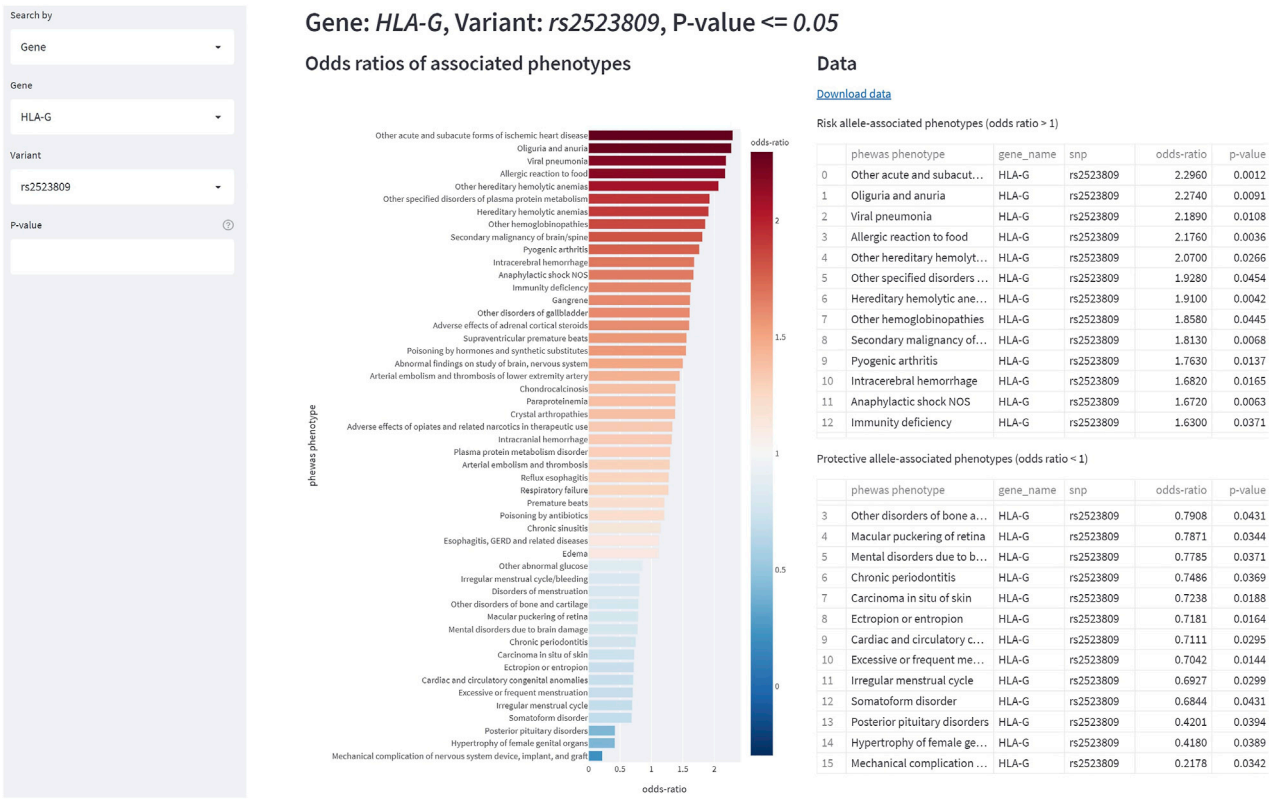
Disease-target data are acquired from the PheWAS catalog (https://phewascatalog.org/phewas) and OpenTargets (https://genetics.opentargets.org/). The latest PheWAS catalog was created in 2013 by generating odds ratios of association between 3,144 SNPs identified in GWASs and 1,358 phenotypes derived from the electronic medical records of 13,835 individuals of European ancestry, and the data is loaded locally (Denny et al., 2013). The list of phenotypic variants from the PheWAS catalog as well as from the GWASs within the PheWAS catalog were aggregated and filtered to remove duplicates, producing a list of 1844 phenotypic traits which StarGazer uses for subsequent analysis. OpenTargets version 22.02 is the latest version at the time of writing, and provides 7,980,448 target-disease association scores extracted from 21 public databases containing diverse forms of evidence, from genetic and drug associations to text mining and animal model data amongst others (Ochoa et al., 2021). Data from OpenTargets is acquired in real-time via API calls.

Target druggability data are acquired in real-time via API calls through Pharos (https://pharos.nih.gov/) to access the Target Central Resource Database (TCRD) (Sheils et al., 2021). The TCRD categorizes 20,412 targets, at the time of writing, into four groups of increasing druggability evidence: Tdark, Tbio, Tchem, and Tclin. A variety of evidence is integrated for classification, such as data from ChEMBL (Mendez et al., 2019), Guide to Pharmacology (Armstrong et al., 2019), DrugCentral (Avram et al., 2021), and antibodypedia (Kiermer, 2008), amongst many more, as well as gene ontology and text-mining analysis. Tclin genes are already targets of approved drugs, whilst Tchem genes have drugs with evidence of sufficient activity against the gene. Tbio genes have weak evidence for druggability, and Tdark genes have an unknown level of druggability.

Protein-protein interaction data are acquired in real-time via API calls from the STRING database (https://string-db.org/). STRING version 11.5 contains data of 20,052,394,042 protein-protein interactions from 14,094 organisms, of which only human genes and orthologous genes were used in StarGazer (Szklarczyk et al., 2021), which were analyzed using the Python package, pyvis. Gene ontology enrichment analysis is also performed by STRING.

METHODS

StarGazer was built using Streamlit (https://streamlit.io/), a relatively new Python-based tool for developing web applications for machine learning and data science. It enables data scientists to build web applications purely from Python scripts quickly and seamlessly. The Streamlit dashboard allows for local files to be loaded, as well as data to be requested from databases via real-time API calls. The StarGazer drug target prioritization framework considers the following five features for each disease (**Figure 1**): (Xue et al., 2018)—the odds ratios of association between targets and phenotypic variants of interest from GWAS and PheWAS data; (Pushpakom et al., 2019);—the target-disease association scores from Open Targets; (Li et al.,



	phewas phenotype	gene_name	snp	odds-ratio	p-value
0	Other acute and subacut…	HLA-G	rs2523809	2.2960	0.0012
1	Oliguria and anuria	HLA-G	rs2523809	2.2740	0.0091
2	Viral pneumonia	HLA-G	rs2523809	2.1890	0.0108
3	Allergic reaction to food	HLA-G	rs2523809	2.1760	0.0036
4	Other hereditary hemolyt…	HLA-G	rs2523809	2.0700	0.0266
5	Other specified disorders …	HLA-G	rs2523809	1.9280	0.0454
6	Hereditary hemolytic ane…	HLA-G	rs2523809	1.9100	0.0042
7	Other hemoglobinopathies	HLA-G	rs2523809	1.8580	0.0445
8	Secondary malignancy of…	HLA-G	rs2523809	1.8130	0.0068
9	Pyogenic arthritis	HLA-G	rs2523809	1.7630	0.0137
10	Intracerebral hemorrhage	HLA-G	rs2523809	1.6820	0.0165
11	Anaphylactic shock NOS	HLA-G	rs2523809	1.6720	0.0063
12	Immunity deficiency	HLA-G	rs2523809	1.6300	0.0371

	phewas phenotype	gene_name	snp	odds-ratio	p-value
3	Other disorders of bone a…	HLA-G	rs2523809	0.7908	0.0431
4	Macular puckering of retina	HLA-G	rs2523809	0.7871	0.0344
5	Mental disorders due to b…	HLA-G	rs2523809	0.7785	0.0371
6	Chronic periodontitis	HLA-G	rs2523809	0.7486	0.0369
7	Carcinoma in situ of skin	HLA-G	rs2523809	0.7238	0.0188
8	Ectropion or entropion	HLA-G	rs2523809	0.7181	0.0164
9	Cardiac and circulatory c…	HLA-G	rs2523809	0.7111	0.0295
10	Excessive or frequent me…	HLA-G	rs2523809	0.7042	0.0144
11	Irregular menstrual cycle	HLA-G	rs2523809	0.6927	0.0299
12	Somatoform disorder	HLA-G	rs2523809	0.6844	0.0431
13	Posterior pituitary disorders	HLA-G	rs2523809	0.4201	0.0394
14	Hypertrophy of female ge…	HLA-G	rs2523809	0.4180	0.0389
15	Mechanical complication …	HLA-G	rs2523809	0.2178	0.0342

FIGURE 2 | The StarGazer interface after searching "*HLA-G*" in Gene mode. At $p = 0.05$, the first allele returned is rs11206510. The color-coded bar chart shows the odds ratio of association of the allele with each phenotype. The table on the right is the same data tabulated which can be downloaded as a csv file. The StarGazer Variant mode is similar in appearance.

2016);—the druggability data of genes of interest from Pharos; (Shameer et al., 2015);—the degree of nodes in protein-protein interaction networks of genes of interest from STRING; and (Lee and Bhakta, 2021)—the presence of the gene variant of interest in both PheWAS and GWAS datasets. Each gene was analyzed with respect to each of these five features, and five scores were computed corresponding to each of the above features. These five scores were then normalized to ensure equal maximum contribution, before summing the five normalized scores to obtain an overall score (i.e., the StarGazer score) which has a maximum score of 1. The targets were then ranked in descending order to facilitate target prioritization.

Processing of Disease-Target Data

Analysis of the PheWAS and GWAS odds ratios involved identifying risk associations where the odds ratio ≥1 (i.e., more associated with the occurrence of the phenotype), and protective associations where the odds ratio <1 (i.e., more associated with the non-occurrence of the disease). In the risk allele-based target prioritization, odds ratios were taken as they were. However, in protective-allele-based target prioritization, odds ratios were subtracted by 1, as the lower ratio implies higher magnitude of association. An average value was taken for odds ratios from multiple studies of the same gene, before normalizing to generate the feature score. Another feature score was generated by determining if the gene target was present in both the PheWAS and GWAS datasets, assigning a score of 1 for the PheWAS-GWAS intersection score, which is otherwise 0. Finally, the target-disease association feature scores from OpenTargets were values between 0 and 1, calculated in a similar manner as the PheWAS catalogue analysis.

Processing of Target Druggability Data

For analysis of the druggability data from Pharos, the number of distinct druggability levels that a target has was counted, with the exception of Tdark, e.g., a target with Tbio, Tclin, and Tdark labels is scored 2 (1 + 1 + 0). These scores were then normalized against the highest druggability feature score of each gene.

Processing of Protein-Protein Interaction Data

The degree of the node in the protein-protein interaction networks from STRING is the number of proteins directly connected to the target node via functional associations, which include experimentally confirmed interactions, predicted interactions and text mining data. Node degrees were computed for each gene in a network and calculated as a ratio of the highest node degree in that network, as a gene with higher interactivity within a STRING network is more likely to be

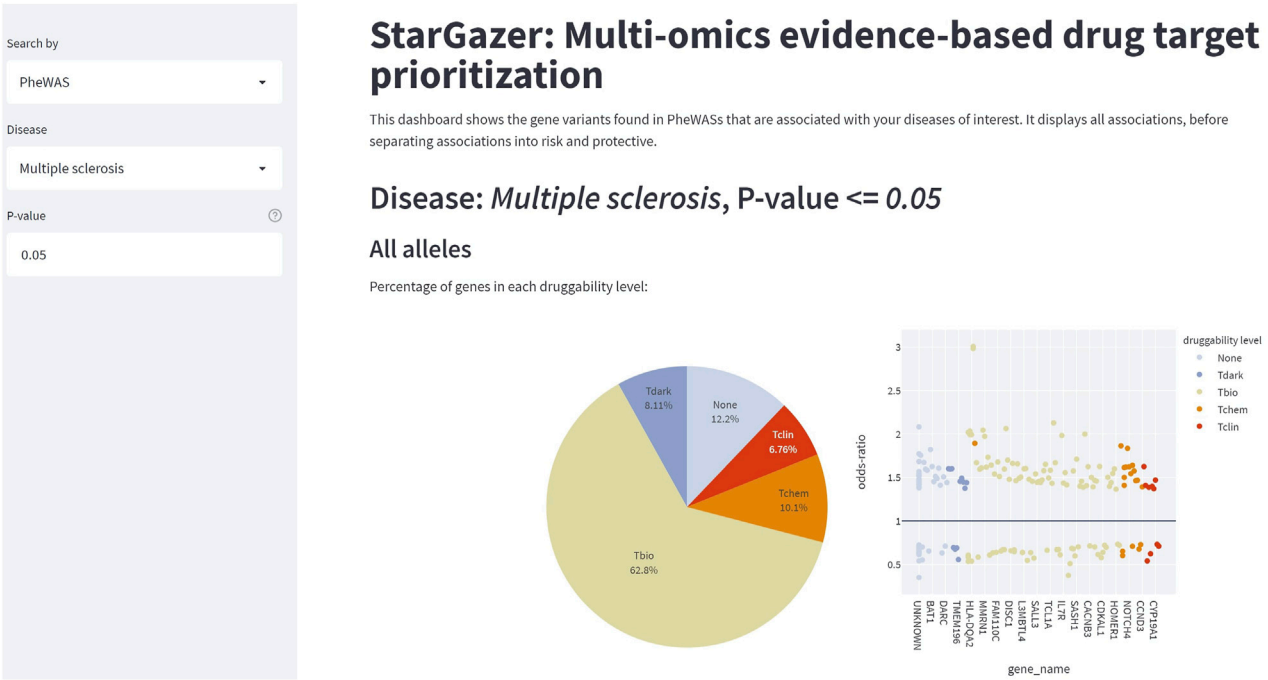


FIGURE 3 | The StarGazer interface after searching "Multiple sclerosis" in PheWAS mode. At p = 0.05, 7.37% of genes with associations with multiple sclerosis were categorized as Tclin, i.e., already targets of FDA-approved drugs. The distribution of genes in each druggability level is shown by pie chart and scatter plot, the latter of which also showing the odds ratios of each allele of each gene. Some gene names are not shown. This data is re-analyzed to show only risk alleles, or only protective alleles. Tabulated data can be visualized and downloaded. The StarGazer modes, GWAS and GWAS-PheWAS Union, are similar in appearance.

biologically underpinning the molecular pathway that contributes to a phenotype. The calculation of node degrees scores this way also reduces effects of false positive interactions.

RESULTS

The StarGazer dashboard (https://github.com/AstraZeneca/StarGazer) offers eight modes of data exploration for drug target prioritization using the data analyzed as described in Methods. The modern yet simple interface allows for rapid navigation without the need for specialist training or programming experience. StarGazer allows users to search by genes or gene variants which displays all associated phenotypic variants ranked by odds ratio graphically, as well as in tabular format (**Figure 2**). Red bars indicate an odds ratio of greater than 1 (i.e., risk association), whilst blue bars indicate less than 1 (i.e., protective association). Users can also search by the PheWAS, GWAS, and GWAS-PheWAS Union modes of exploration, which returns odds ratios of all variants of genes associated with the phenotype of interest from the respective datasets, as well as their corresponding druggability levels (**Figure 3**). When searching in the GWAS-PheWAS Intersection mode, only variants with associations identified in both GWAS and PheWAS datasets are shown (**Figure 4**). For these variants, the dashboard also provides association odds ratios, druggability data, protein-protein interaction networks and gene ontology enrichment analysis for the disease of interest (**Figure 5**). Finally, when users search by disease target prioritization, the overall StarGazer score is shown for each gene with association with the disease of interest (**Figure 6**). Contextual information on any of these genes can be found immediately using the build-in NCBI search tool. For each of these exploration modes, users can also modify the p-value to only display associations of desired statistical significance assigned by the origin data source.

Use Case: StarGazer for Understanding Complex Diseases

In the following case study, we posed as someone who was simply curious about the possible mechanistic causes of insomnia, and consequently adopted a more exploratory workflow. As insomnia is a complex and relatively understudied disorder, we set the p-value to a less stringent 0.05 to prevent issues in, for example, study sample size or sensitivity from masking any potentially true associations. This returned a list of 106 genes with associations with insomnia, 62 of which had at least one risk-associated allele, and 46 had at least one protection-associated allele (**Table 1**). After searching on NCBI, there were three genes found to have significant relevance to insomnia. *DISC1* encodes a scaffold protein which is involved in brain development, and its mutations have been implicated in schizophrenia and other psychiatric disorders (Dahoun et al., 2017); *MAOA* encodes a mitochondrial oxidative deaminase targeting amines such as

FIGURE 4 | The StarGazer interface after searching "Type 2 diabetes" in GWAS-PheWAS Intersection mode. At p = 0.05, 23 SNPs were identified to have associations in both PheWASs and GWASs. Top left: pie chart displaying the proportion of SNPs that were identified in either PheWAS or GWAS datasets, or in both datasets. Top right: pie chart displaying druggability information of the genes of these SNPs. Tclin in red implies genes already have drugs targeting them available on the market, whilst Tchem, Tbio, Tdark, and None, indicate progressively decreasing levels of druggability. Bottom left: scatter plot highlighting individually reported odds-ratios of associations of SNPs from various GWASs. Bottom right: a protein-protein interaction network constructed from the genes of alleles detected in both GWASs and the PheWAS catalog. The gene ontology enrichment analysis feature is not shown in the figure.

dopamine, norepinephrine, and serotonin, and mutations in the gene can result in Brunner syndrome, a psychiatric and sleep disorder (Brunner et al., 2007); *MEIS1* is a HOX gene thought to have a pleiotropic effect on chronic insomnia disorder, and have possible association with restless leg syndrome (Sarayloo et al., 2019). We also found genes with a variety of functions and unclear links with insomnia. Tumor suppressor genes, *CMTM7* (Li et al., 2014b), *NKAPL* (Okuda et al., 2015) and *ATM* (encoding ATM checkpoint kinase) (Shiloh and Ziv, 2013) may allude to aberrant DNA damage responses contributing to insomnia, and indeed, there are several reports of links between DNA damage and sleep in the literature (Carroll et al., 2016; Zada et al., 2021). HLA isoforms indicate a potential immunity-related cause of insomnia (Choo, 2007). *In vitro* mutants in vesicular trafficking protein, dynamin-1, have impaired ability to recycle neurotransmitter at synapses (Chung et al., 2010), providing a more obvious potential link with insomnia. Finally, genes with noticeably pleiotropic effect were also found to have a high StarGazer score. One such example is estrogen receptor (*ESR1*), important for gestation in women but is in addition expressed in many non-reproductive tissues in both sexes, as it has roles more broadly in growth and metabolism (Barros and Gustafsson, 2011). Not only is estrogen receptor linked with breast cancer but also with osteoporosis (Gennari et al., 2007), and thus makes for a peculiar hit on the StarGazer analysis. Although additional investigations are required to ascertain the link between these genes and phenotypes, it is exciting to hypothesize about the underlying molecular mechanisms. This is especially the case for insomnia, a disorder of sleep which is a biological process we still have a relatively poor understanding of.

DISCUSSION

StarGazer is a novel application built for rapid investigation of drug repositioning strategies. It combines multi-source, multi-

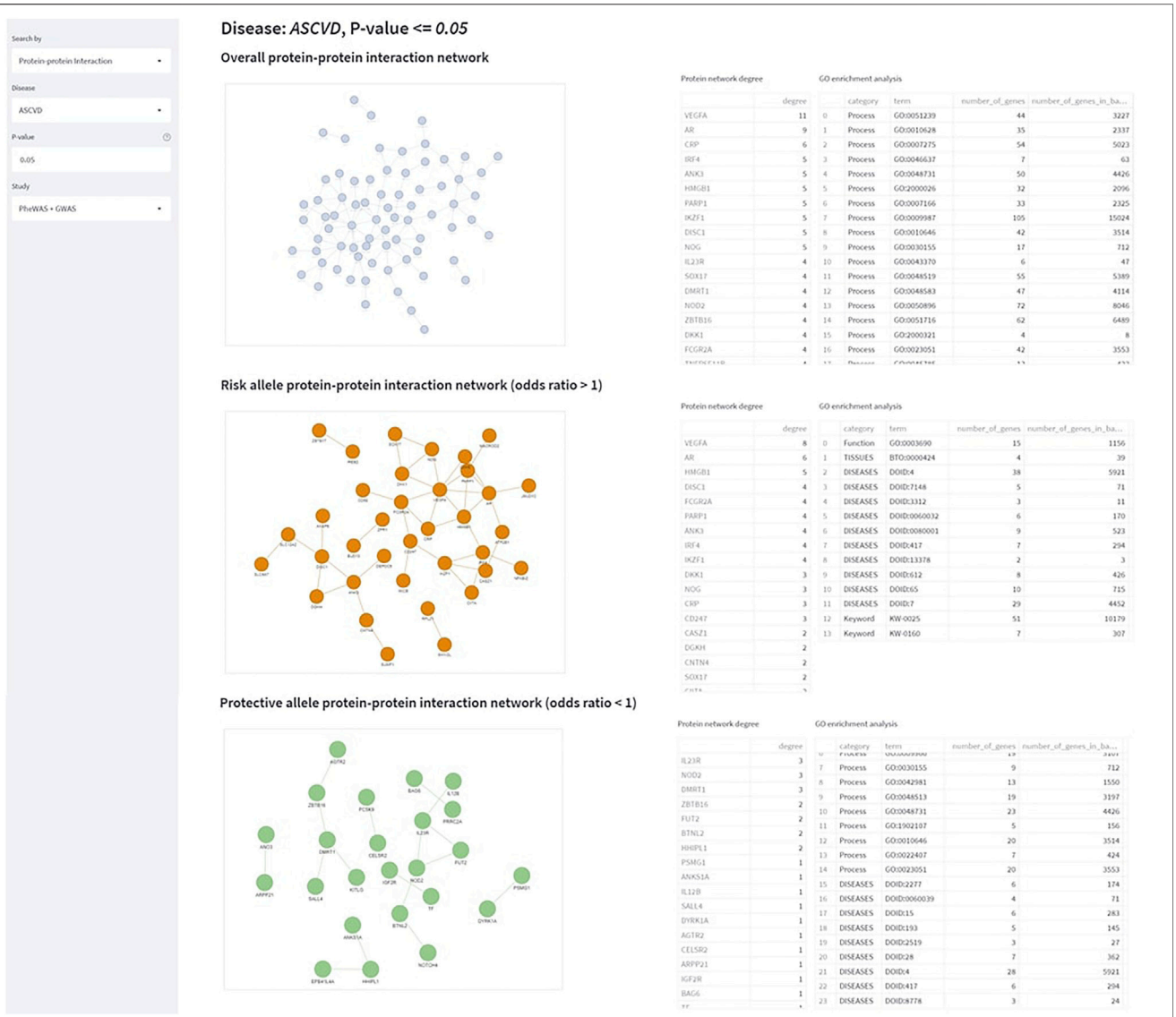
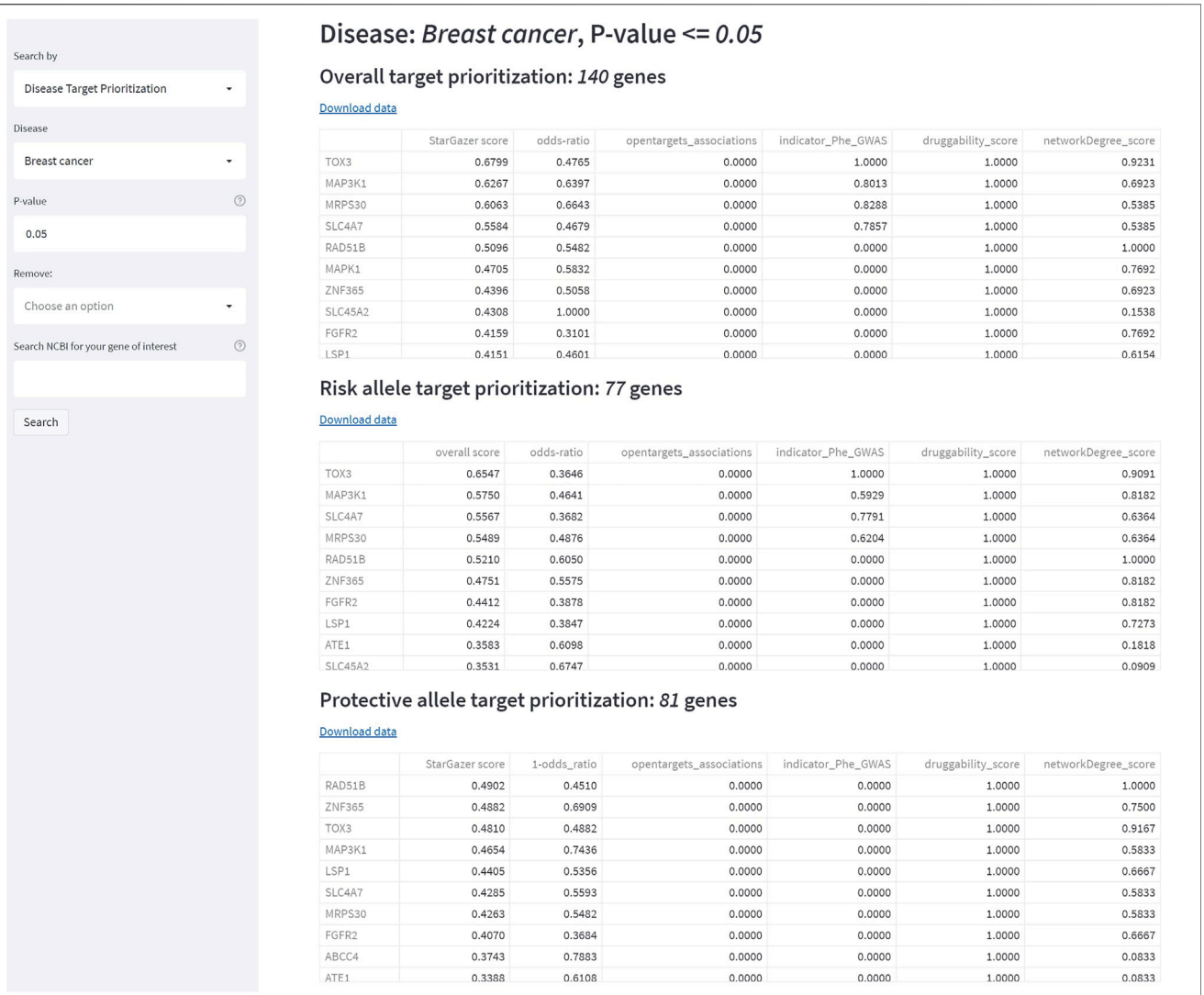


FIGURE 5 | The StarGazer interface after searching "ASCVD" in Protein-protein interaction mode. Protein-protein interaction networks are shown of all alleles, risk alleles, and protective alleles. The node degree of the genes of these alleles are computed, and gene ontology enrichment analysis is performed on the right.

omics data with a novel target prioritization scoring system in an interactive Python-based Streamlit dashboard. StarGazer analyzes and integrates disease-target associations, druggability data, and protein-protein interaction data before extracting five features from the data to create an overall StarGazer score for every potential target associated with StarGazer's curated list of 1844 phenotypic variants.

StarGazer is adapted to facilitate exploration of the human biology landscape from a birds-eye view, allowing rapid digestion of information from PheWASs/GWASs, which otherwise contains many tens of thousands of complex multivariate datapoints. Streamlit, as a user interface package adapted for complex data visualization and user interactivity, was considered to be a well-suited technology for such a task. Indeed, the importance of the flexibility in visualization methods, and live data retrieval and analysis is becoming increasingly clear, with their applications ever-expanding (Badgeley et al., 2016; Moosavinasab et al., 2016).

We demonstrate the utility in integrating several omics datasets and returning easy-to-interpret analysis metrics in an interactive dashboard. One can easily imagine the power of such a strategy as we incorporate state-of-the-art, machine learning-based, multi-omics integration techniques, as well as a wider variety of high quality data. In an era where the speed at which we can generate data is accelerating at a higher rate than we can analyze it, we anticipate that integrative scores and visualization tools will grow increasingly essential in biology, and that we must begin to break away from the more rigid, single-use analysis framework that forms the modern paradigm for analyzing not just GWAS and PheWAS data (Diogo et al., 2018; Ferrero and



	StarGazer score	odds-ratio	opentargets_associations	indicator_Phe_GWAS	druggability_score	networkDegree_score
TOX3	0.6799	0.4765	0.0000	1.0000	1.0000	0.9231
MAP3K1	0.6267	0.6397	0.0000	0.8013	1.0000	0.6923
MRPS30	0.6063	0.6643	0.0000	0.8288	1.0000	0.5385
SLC4A7	0.5584	0.4679	0.0000	0.7857	1.0000	0.5385
RAD51B	0.5096	0.5482	0.0000	0.0000	1.0000	1.0000
MAPK1	0.4705	0.5832	0.0000	0.0000	1.0000	0.7692
ZNF365	0.4396	0.5058	0.0000	0.0000	1.0000	0.6923
SLC45A2	0.4308	1.0000	0.0000	0.0000	1.0000	0.1538
FGFR2	0.4159	0.3101	0.0000	0.0000	1.0000	0.7692
LSP1	0.4151	0.4601	0.0000	0.0000	1.0000	0.6154

	overall score	odds-ratio	opentargets_associations	indicator_Phe_GWAS	druggability_score	networkDegree_score
TOX3	0.6547	0.3646	0.0000	1.0000	1.0000	0.9091
MAP3K1	0.5750	0.4641	0.0000	0.5929	1.0000	0.8182
SLC4A7	0.5567	0.3682	0.0000	0.7791	1.0000	0.6364
MRPS30	0.5489	0.4876	0.0000	0.6204	1.0000	0.6364
RAD51B	0.5210	0.6050	0.0000	0.0000	1.0000	1.0000
ZNF365	0.4751	0.5575	0.0000	0.0000	1.0000	0.8182
FGFR2	0.4412	0.3878	0.0000	0.0000	1.0000	0.8182
LSP1	0.4224	0.3847	0.0000	0.0000	1.0000	0.7273
ATE1	0.3583	0.6098	0.0000	0.0000	1.0000	0.1818
SLC45A2	0.3531	0.6747	0.0000	0.0000	1.0000	0.0909

	StarGazer score	1-odds_ratio	opentargets_associations	indicator_Phe_GWAS	druggability_score	networkDegree_score
RAD51B	0.4902	0.4510	0.0000	0.0000	1.0000	1.0000
ZNF365	0.4882	0.6909	0.0000	0.0000	1.0000	0.7500
TOX3	0.4810	0.4882	0.0000	0.0000	1.0000	0.9167
MAP3K1	0.4654	0.7436	0.0000	0.0000	1.0000	0.5833
LSP1	0.4405	0.5356	0.0000	0.0000	1.0000	0.6667
SLC4A7	0.4285	0.5593	0.0000	0.0000	1.0000	0.5833
MRPS30	0.4263	0.5482	0.0000	0.0000	1.0000	0.5833
FGFR2	0.4070	0.3684	0.0000	0.0000	1.0000	0.6667
ABCC4	0.3743	0.7883	0.0000	0.0000	1.0000	0.0833
ATE1	0.3388	0.6108	0.0000	0.0000	1.0000	0.0833

FIGURE 6 | The StarGazer interface after searching "Breast cancer" in Disease Target Prioritization mode. At $p = 0.05$, 140 genes are returned to have association with breast cancer. Genes are ranked in StarGazer score, which describes how suitable a gene is for drug repositioning. The subsequent five columns are the individual scores of the five features extracted from all of the data that contribute to the StarGazer score. Data are separated into all alleles, risk alleles, then protective alleles, and can be downloaded as csv files.

Agarwal, 2018; Robinson et al., 2018; Lau and So, 2020), but multi-omics data in general (Subramanian et al., 2020).

StarGazer has been built with the goal of pushing multi-omics integration towards upward scalability by providing users with immediate access to contextual information on genes of potential interest by automatically performing several steps of follow-up analysis on all genes - this saves a considerable amount of time from performing speculative follow-up analysis. These follow-up analysis steps are completed in bulk through the processing of the single-omic layers, which removes the need for users to analyze every gene separately for various properties and then later compare the results to make sense of the evidence. Not only does integrating single-omic layers increase the speed of exploratory data analysis, but it also provides additional value from combining multiple pieces of evidence as opposed to focusing on individual single high-confidence pieces of information, especially when the different types of data are likely to have an intimate biological relationship, e.g., combining a gene's DNA, RNA and protein information together is likely to be more valuable than analyzing them independently as they are functionally coupled. This approach may be our best strategy for uncovering complex and profound relationships and hence, the phrase "the whole is greater than the sum of its parts" holds particularly true in the context of multi-omics data analysis. A more integrated strategy may also be more useful in helping us understand the genetic basis of complex diseases driven by genes and gene variants with pleiotropic functions or effects. Applying the latest ideas on pleiotropy in biological systems to future work may allow us to obtain a more complete understanding of genome-phenome relationships and

TABLE 1 | Top 30 hits from Disease Target Prioritization mode analysis of "Insomnia" using StarGazer.

Gene Name	StarGazer Score	Odds-Ratio	OpenTargets Associations	Indicator Phe/GWAS	Druggability Score	Network Degree Score
HLA-DRB1	0.456	0.725	0.000	0.000	1.000	0.556
ESR1	0.433	0.655	0.009	0.000	0.500	1.000
GRIN2B	0.407	0.756	0.000	0.000	0.500	0.778
MEIS1	0.395	0.251	1.000	0.000	0.500	0.222
MAOA	0.344	0.888	0.000	0.000	0.500	0.333
DNM1	0.337	0.630	0.000	0.000	0.500	0.556
HLA-DQB1	0.320	0.653	0.000	0.000	0.500	0.444
BMP4	0.307	0.591	0.000	0.000	0.500	0.444
ATM	0.293	0.188	0.000	0.000	0.500	0.778
CMTM7	0.288	0.941	0.000	0.000	0.500	0.000
NKAPL	0.288	0.938	0.000	0.000	0.500	0.000
GRIA1	0.286	0.263	0.000	0.000	0.500	0.667
TOMM40	0.280	0.677	0.000	0.000	0.500	0.222
NR5A2	0.278	0.668	0.000	0.000	0.500	0.222
HDAC9	0.276	0.768	0.000	0.000	0.500	0.111
MS4A6A	0.271	0.631	0.000	0.000	0.500	0.222
DISC1	0.267	0.166	0.000	0.000	0.500	0.667
ST6GAL1	0.265	0.716	0.000	0.000	0.500	0.111
SLC22A3	0.264	0.600	0.000	0.000	0.500	0.222
EFNA5	0.264	0.600	0.000	0.000	0.500	0.222
NRGN	0.263	0.706	0.000	0.000	0.500	0.111
DRD2	0.263	0.000	0.150	0.000	0.500	0.667
RNASET2	0.263	0.591	0.000	0.000	0.500	0.222
FGFR2	0.263	0.257	0.000	0.000	0.500	0.556
UBE2L3	0.262	0.701	0.000	0.000	0.500	0.111
YDJC	0.260	0.690	0.000	0.000	0.500	0.111
CDC42BPB	0.260	0.690	0.000	0.000	0.500	0.111
LAMP3	0.258	0.791	0.000	0.000	0.500	0.000
ARG1	0.254	0.660	0.000	0.000	0.500	0.111
CCND3	0.251	0.643	0.000	0.000	0.500	0.111

thus drive novel discoveries previously inaccessible in the biomedical field (Shameer et al., 2021).

Limitations

This should, of course, highlight to the reader the current co-dependence between broader exploratory analytical approaches, such as StarGazer, with those that possess stronger statistical power, aimed at target confirmation at the cost of breadth and fewer omics layers, and of course, experimental confirmation. Moving forwards, we should hope that the field develops more sophisticated strategies for these types of analysis. All in all, we anticipate StarGazer to be potentially useful in providing insights into many types of biological pathways, as long as the molecular perturbations that are linked with disease lie close to the genetic level. Whilst it is easy to imagine StarGazer's utility for studying diseases caused by variants of proteins or nucleic acids due to their more direct connection to genome-level information, studying metabolic disorders of carbohydrates and lipids would be possible but more difficult.

We wish to highlight that, although the barrier to entry for multi-omics data analysis is low, there seems yet a limitless space for improvement in the field at the time of writing. In the future, we aim to incorporate gene ontology terms enrichment analysis, gene semantic similarity, and gene expression data into our target prioritization framework, and improve on the implementation of protein-protein interaction networks (Shameer et al., 2016; Peters et al., 2017). Whilst the current version of StarGazer extracts several features for target-disease associations, the assessment of target druggability uses only one dataset to generate one feature. Although the knowledge-based classification of the genome that Pharos provides is very high quality data, it is less indicative of future potential developments as it reflects only the current status of the druggability landscape of human biology. Therefore, more predictive datasets, such as computational docking predictions using structural data from molecular techniques or even AI-based computational prediction, may provide more robust insight into the future (Baek et al., 2021; Jumper et al., 2021).

StarGazer's use of API calls allows for the majority of its data to be updated automatically with the latest relevant studies, aside from the PheWAS catalog which was performed in 2013—it would be invaluable if a similar study was repeated to include the GWAS data which was generated during the decade that has elapsed since the original effort. Furthermore, a variety of machine learning strategies have been applied to multi-omics data analysis and show great promise in assisting precision medicine and repositioning (Shameer et al., 2018b; Nicora et al., 2020; Reel et al., 2021), and is therefore an area we are interested in developing for StarGazer. Another avenue for future development is to improve on the standardization of clinical terms between the different datasets, which is a problem not unique to StarGazer but found ubiquitously in healthcare-related work (Wears, 2015; Beck et al., 2019). This problem manifested itself as data from OpenTargets

being underrepresented in the overall StarGazer score. We hypothesize that using a combination of standardized codes for clinical terms, such as ICD-9/-10 (https://www.cdc.gov/nchs/icd/icd9.htm, https://www.cdc.gov/nchs/icd/icd10.htm), and EFO (https://www.ebi.ac.uk/efo/faq.html), would help with this problem, as well as further curate our list of 1844 phenotypic variants. Currently, the code for installation can be found on GitHub (https://github.com/AstraZeneca/StarGazer).

CONCLUSION

We have created StarGazer (https://github.com/AstraZeneca/StarGazer), an interactive dashboard that facilitates rapid investigation of potential novel drug targets and repositioning strategies. It integrates three different types of data (disease-target data, target druggability data, and protein-protein interaction data) from four different knowledgebases (the PheWAS catalog, OpenTargets, Pharos, and STRING) to extract five features that are then processed to return a singular normalized "StarGazer" score. All genes with associations with any of the 1844 phenotypic variants in the StarGazer disease list are then ranked in suitability for drug repositioning strategies for the disease of interest.

We demonstrate the utility in integrating several omics datasets to return easy-to-interpret analysis metrics in an interactive dashboard. One can easily imagine the power of such a strategy as we incorporate machine learning techniques as well as a wider variety of high quality data. It is anticipated that such integrative analysis strategies will become commonplace as biomedical data science grows to explore more multi-disciplinary and multi-omic datasets. Integrative scores and visualization tools for high dimensional data will become essential as we navigate science in this era where we are generating data at a such an enormous pace, thus we have positioned StarGazer to push multi-omics integration towards upward scalability.

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 below: https://github.com/AstraZeneca/StarGazer.

AUTHOR CONTRIBUTIONS

CL and JL developed the project with critical inputs from VG, AF, MH, KT, BS, TC, JM, and YG. MH led app hosting, and AP and RNH conducted validation analysis. KS led and oversaw the project. The original manuscript was written by CL and JL, with additions and edits provided by VG, EP, AF, SP, WY, A-SS, KT, BS, TC, JM, FMK, YG, and KS. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to acknowledge Anshul Kanakia for insightful discussions, and Stefano Borini for assistance with GitHub.

REFERENCES

Abu-Doleh, A. A., Al-Jarrah, O. M., and Alkhateeb, A. (2012). Protein Contact Map Prediction Using Multi-Stage Hybrid Intelligence Inference Systems. *J. Biomed. Inf.* 45 (1), 173–183. doi:10.1016/j.jbi.2011.10.008

Adikusuma, W., Irham, L. M., Chou, W.-H., Wong, H. S.-C., Mugiyanto, E., Ting, J., et al. (2021). Drug Repurposing for Atopic Dermatitis by Integration of Gene Networking and Genomic Information. *Front. Immunol.* 12, 724277. doi:10.3389/fimmu.2021.724277

Akata, Z., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 18–28. doi:10.1109/mc.2020.2996587

Armstrong, J. F., Faccenda, E., Harding, S. D., Pawson, A. J., Southan, C., Sharman, J. L., et al. (2019). The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: Extending Immunopharmacology Content and Introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res.* 48, D1006–D1021. doi:10.1093/nar/gkz951

Ashburn, T. T., and Thor, K. B. (2004). Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discov.* 3 (8), 673–683. doi:10.1038/nrd1468

Attique, Z., Ali, A., Hamza, M., al-Ghanim, K. A., Mehmood, A., Khan, S., et al. (2021). In-silico Network-Based Analysis of Drugs Used against COVID-19: Human Well-Being Study. *Saudi J. Biol. Sci.* 28 (3), 2029–2039. doi:10.1016/j.sjbs.2021.01.006

Avram, S., Bologa, C. G., Holmes, J., Bocci, G., Wilson, T. B., Nguyen, D.-T., et al. (2021). DrugCentral 2021 Supports Drug Discovery and Repositioning. *Nucleic Acids Res.* 49 (D1), D1160–D1169. doi:10.1093/nar/gkaa997

Badgeley, M. A., Shameer, K., Glicksberg, B. S., Tomlinson, M. S., Levin, M. A., McCormick, P. J., et al. (2016). EHDViz: Clinical Dashboard Development Using Open-Source Technologies. *BMJ Open* 6 (3), e010579. doi:10.1136/bmjopen-2015-010579

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* 373 (6557), 871–876. doi:10.1126/science.abj8754

Barros, R. P. A., and Gustafsson, J.-Å. (2011). Estrogen Receptors and the Metabolic Network. *Cell Metab.* 14 (3), 289–299. doi:10.1016/j.cmet.2011.08.005

Beck, T., Shorter, T., and Brookes, A. J. (2019). GWAS Central: a Comprehensive Resource for the Discovery and Comparison of Genotype and Phenotype Data from Genome-wide Association Studies. *Nucleic Acids Res.* 48, D933–D940. doi:10.1093/nar/gkz895

Brunner, H. G. (2007). "MAOA Deficiency and Abnormal Behaviour: Perspectives on an Assocation," in *Novartis Foundation Symposia [Internet]*. Editors G. R. Bock and J. A. Goode (Chichester, UK: John Wiley & Sons), 155–167. Available at: https://onlinelibrary.wiley.com/doi/10.1002/9780470514825.ch9. doi:10.1002/9780470514825.ch9

Carroll, J. E., Cole, S. W., Seeman, T. E., Breen, E. C., Witarama, T., Arevalo, J. M. G., et al. (2016). Partial Sleep Deprivation Activates the DNA Damage Response (DDR) and the Senescence-Associated Secretory Phenotype (SASP) in Aged Adult Humans. *Brain, Behav. Immun.* 51, 223–229. doi:10.1016/j.bbi.2015.08.024

Choo, S. Y. (2007). The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Med. J.* 48 (1), 11. doi:10.3349/ymj.2007.48.1.11

Chung, C., Barylko, B., Leitz, J., Liu, X., and Kavalali, E. T. (2010). Acute Dynamin Inhibition Dissects Synaptic Vesicle Recycling Pathways that Drive

Spontaneous and Evoked Neurotransmission. *J. Neurosci.* 30 (4), 1363–1376. doi:10.1523/jneurosci.3427-09.2010

Dahoun, T., Trossbach, S. V., Brandon, N. J., Korth, C., and Howes, O. D. (2017). The Impact of Disrupted-In-Schizophrenia 1 (DISC1) on the Dopaminergic System: a Systematic Review. *Transl. Psychiatry* 7 (1), e1015. doi:10.1510.1038/tp.2016.282

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic Comparison of Phenome-wide Association Study of Electronic Medical Record Data and Genome-wide Association Study Data. *Nat. Biotechnol.* 31 (12), 1102–1111. doi:10.1038/nbt.2749

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: Demonstrating the Feasibility of a Phenome-wide Scan to Discover Gene-Disease Associations. *Bioinformatics* 26 (9), 1205–1210. doi:10.1093/bioinformatics/btq126

Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., et al. (2018). Phenome-wide Association Studies across Large Population Cohorts Support Drug Target Validation. *Nat. Commun.* 9 (1), 4285. doi:10.1038/s41467-018-06540-3

Ferrero, E., and Agarwal, P. (2018). Connecting Genetics and Gene Expression Data for Target Prioritisation and Drug Repositioning. *BioData Min.* 11 (1), 7. doi:10.1186/s13040-018-0171-y

Gallo, K., Goede, A., Eckert, A., Moahamed, B., Preissner, R., and Gohlke, B.-O. (2021). PROMISCUOUS 2.0: a Resource for Drug-Repositioning. *Nucleic Acids Res.* 49 (D1), D1373–D1380. doi:10.1093/nar/gkaa1061

Gennari, L., Merlotti, D., Valleggi, F., Martini, G., and Nuti, R. (2007). Selective Estrogen Receptor Modulators for Postmenopausal Osteoporosis. *Drugs & Aging* 24 (5), 361–379. doi:10.2165/00002512-200724050-00002

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., et al. (2021). Open Targets Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics. *Nucleic Acids Res.* 49 (D1), D1311–D1320. doi:10.1093/nar/gkaa840

Glicksberg, B. S., Li, L., Cheng, W-Y., Shameer, K., Hakenberg, J., Castellanos, R., et al. (2014). "An Integrative Pipeline for Multi-Modal Discovery of Disease Relationships," in *Biocomputing 2015 [Internet]* (Hawaii, USA: World Scientific), 407–418. Available from: http://www.worldscientific.com/doi/abs/10.1142/9789814644730_0039.

Guo, Z., Shen, Y., Wan, S., Shang, W., and Yu, K. (2021). Hybrid Intelligence-Driven Medical Image Recognition for Remote Patient Diagnosis in Internet of Medical Things. *IEEE J. Biomed. Health Inf.* doi:10.1109/jbhi.2021.3139541

Hermawan, A., Putri, H., and Utomo, R. Y. (2020). Functional Network Analysis Reveals Potential Repurposing of β-blocker Atenolol for Pancreatic Cancer Therapy. *DARU J. Pharm. Sci.* 28 (2), 685–699. doi:10.1007/s40199-020-00375-4

Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P., and Dudley, J. T. (2016). In Silico methods for Drug Repurposing and Pharmacology. *WIREs Mech. Dis.* 8 (3), 186–210. doi:10.1002/wsbm.1337

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Khaladkar, M., Koscielny, G., Hasan, S., Agarwal, P., Dunham, I., Rajpal, D., et al. (2017). Uncovering Novel Repositioning Opportunities Using the Open Targets Platform. *Drug Discov. Today* 22 (12), 1800–1807. doi:10.1016/j.drudis.2017.09.007

Khosravi, A., Jayaram, B., Goliaei, B., and Masoudi-Nejad, A. (2019). Active Repurposing of Drug Candidates for Melanoma Based on GWAS, PheWAS and a Wide Range of Omics Data. *Mol. Med.* 25 (1), 30. doi:10.1186/s10020-019-0098-x

Kiermer, V. (2008). Antibodypedia. *Nat. Methods* 5 (10), 860. doi:10.1038/nmeth1008-860

Lau, A., and So, H.-C. (2020). Turning Genome-wide Association Study Findings into Opportunities for Drug Repositioning. *Comput. Struct. Biotechnol. J.* 18, 1639–1650. doi:10.1016/j.csbj.2020.06.015

Lee, C., and Bhakta, S. (2021). The Prospect of Repurposing Immunomodulatory Drugs for Adjunctive Chemotherapy against Tuberculosis: A Critical Review. *Antibiotics* 10 (1), 91. doi:10.3390/antibiotics10010091

Li, H., Li, J., Su, Y., Fan, Y., Guo, X., Li, L., et al. (2014). A Novel 3p22.3 Gene CMTM7 Represses Oncogenic EGFR Signaling and Inhibits Cancer Cell Growth. *Oncogene* 33 (24), 3109–3118. doi:10.1038/onc.2013.282

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2016). A Survey of Current Trends in Computational Drug Repositioning. *Brief. Bioinform* 17 (1), 2–12. doi:10.1093/bib/bbv020

Li, S., Kang, L., and Zhao, X. M. (2014). A Survey on Evolutionary Algorithm Based Hybrid Intelligence in Bioinformatics. *Biomed. Res. Int.* 2014, 362738. doi:10.1155/2014/362738

Liu, Z., Borlak, J., and Tong, W. (2014). Deciphering miRNA Transcription Factor Feed-Forward Loops to Identify Drug Repurposing Candidates for Cystic Fibrosis. *Genome Med.* 6 (12), 94. doi:10.1186/s13073-014-0094-2

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2019). ChEMBL: towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* 47 (D1), D930–D940. doi:10.1093/nar/gky1075

Moosavinasab, S., Patterson, J., Strouse, R., Rastegar-Mojarad, M., Regan, K., Payne, P. R. O., et al. (2016). 'RE:fine Drugs': an Interactive Dashboard to Access Drug Repurposing Opportunities. *Database* 2016, baw083. doi:10.1093/database/baw083

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., and Bellazzi, R. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Front. Oncol.* 10, 1030. doi:10.3389/fonc.2020.01030

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., et al. (2021). Open Targets Platform: Supporting Systematic Drug-Target Identification and Prioritisation. *Nucleic Acids Res.* 49 (D1), D1302–D1310. doi:10.1093/nar/gkaa1027

Okuda, H., Kiuchi, H., Takao, T., Miyagawa, Y., Tsujimura, A., Nonomura, N., et al. (2015). A Novel Transcriptional Factor Nkapl Is a Germ Cell-specific Suppressor of Notch Signaling and Is Indispensable for Spermatogenesis. *PLOS ONE* 10 (4), e0124293. doi:10.1371/journal.pone.0124293

Peters, L. A., Perrigoue, J., Mortha, A., Iuga, A., Song, W.-m., Neiman, E. M., et al. (2017). A Functional Genomics Predictive Network Model Identifies Regulators of Inflammatory Bowel Disease. *Nat. Genet.* 49 (10), 1437–1449. doi:10.1038/ng.3947

Portelli, M. A., Rakkar, K., Hu, S., Guo, Y., and Adcock, I. M. (2021). Translational Analysis of Moderate to Severe Asthma GWAS Signals into Candidate Causal Genes and Their Functional, Tissue-dependent and Disease-Related Associations. *Front. Allergy* 2, 738741. doi:10.3389/falgy.2021.738741

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug Repurposing: Progress, Challenges and Recommendations. *Nat. Rev. Drug Discov.* 18 (1), 41–58. doi:10.1038/nrd.2018.168

Rapicavoli, R. V., Alaimo, S., Ferro, A., and Pulvirenti, A. (2022). "Computational Methods for Drug Repurposing," in *Computational Methods for Precision Oncology [Internet]*. Editor A. Laganà (Cham: Springer International Publishing), 119–141. Available at: https://link.springer.com/10.1007/978-3-030-91836-1_7. doi:10.1007/978-3-030-91836-1_7

Rastegar-Mojarad, M., Ye, Z., Kolesar, J. M., Hebbring, S. J., and Lin, S. M. (2015). Opportunities for Drug Repositioning from Phenome-wide Association Studies. *Nat. Biotechnol.* 33 (4), 342–345. doi:10.1038/nbt.3183

Reay, W. R., and Cairns, M. J. (2021). Advancing the Use of Genome-wide Association Studies for Drug Repurposing. *Nat. Rev. Genet.* 22 (10), 658–671. doi:10.1038/s41576-021-00387-z

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739

Robinson, J. R., Denny, J. C., Roden, D. M., and Van Driest, S. L. (2018). Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records. *Clin. Transl. Sci.* 11 (2), 112–122. doi:10.1111/cts.12522

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., et al. (2012). Use of Genome-wide Association Studies for Drug Repositioning. *Nat. Biotechnol.* 30 (4), 317–320. doi:10.1038/nbt.2151

Sarayloo, F., Dion, P. A., and Rouleau, G. A. (2019). MEIS1 and Restless Legs Syndrome: A Comprehensive Review. *Front. Neurol.* 10, 935. doi:10.3389/fneur.2019.00935

Shameer, K., Glicksberg, B. S., Badgeley, M. A., Johnson, K. W., and Dudley, J. T. (2021). Pleiotropic Variability Score: A Genome Interpretation Metric to Quantify Phenomic Associations of Genomic Variants. *bioRxiv*. doi:10.1101/2021.07.18.452819

Shameer, K., Glicksberg, B. S., Hodos, R., Johnson, K. W., Badgeley, M. A., and Readhead, B. (2018). Systematic Analyses of Drugs and Disease Indications in RepurposeDB

Reveal Pharmacological, Biological and Epidemiological Factors Influencing Drug Repositioning. *Briefings Bioinforma.* 19 (4), 656–678. doi:10.1093/bib/bbw136

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., and Sengupta, P. P. (2018). Machine Learning in Cardiovascular Medicine: Are We There yet? *Heart* 104 (14), 1156–1164. doi:10.1136/heartjnl-2017-311198

Shameer, K., Readhead, B., and T. Dudley, J. (2015). Computational and Experimental Advances in Drug Repositioning for Accelerated Therapeutic Stratification. *Ctmc* 15 (1), 5–20. doi:10.2174/1568026615666150112103510

Shameer, K., Tripathi, L. P., Kalari, K. R., Dudley, J. T., and Sowdhamini, R. (2016). Interpreting Functional Effects of Coding Variants: Challenges in Proteome-Scale Prediction, Annotation and Assessment. *Brief. Bioinform* 17 (5), 841–862. doi:10.1093/bib/bbv084

Sheils, T. K., Mathias, S. L., Kelleher, K. J., Siramshetty, V. B., Nguyen, D.-T., Bologa, C. G., et al. (2021). TCRD and Pharos 2021: Mining the Human Proteome for Disease Biology. *Nucleic Acids Res.* 49 (D1), D1334–D1346. doi:10.1093/nar/gkaa993

Shiloh, Y., and Ziv, Y. (2013). The ATM Protein Kinase: Regulating the Cellular Response to Genotoxic Stress, and More. *Nat. Rev. Mol. Cell Biol.* 14 (4), 197–210. doi:10.1038/nrm3546

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and its Application. *Bioinform Biol. Insights* 14, 117793221989905. doi:10.1177/1177932219899051

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074

Tan, X., Gong, L., Li, X., Zhang, X., Sun, J., Luo, X., et al. (2021). Promethazine Inhibits Proliferation and Promotes Apoptosis in Colorectal Cancer Cells by Suppressing the PI3K/AKT Pathway. *Biomed. Pharmacother.* 143, 112174. doi:10.1016/j.biopha.2021.112174

Varghese, R., and Majumdar, A. (2022). A New Prospect for the Treatment of Nephrotic Syndrome Based on Network Pharmacology Analysis. *Curr. Res. Physiology* 5, 36–47. doi:10.1016/j.crphys.2021.12.004

Wears, R. L. (2015). Standardisation and its Discontents. *Cogn. Tech. Work* 17 (1), 89–94. doi:10.1007/s10111-014-0299-6

Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., et al. (2021). The Role of Machine Learning in Clinical Research: Transforming the Future of Evidence Generation. *Trials* 22 (1), 537. doi:10.1186/s13063-021-05489-x

Wijetunga, I., McVeigh, L. E., Charalambous, A., Antanaviciute, A., Carr, I. M., Nair, A., et al. (2020). Translating Biomarkers of Cholangiocarcinoma for Theranosis: A Systematic Review. *Cancers* 12 (10), 2817. doi:10.3390/cancers12102817

Xue, H., Li, J., Xie, H., and Wang, Y. (2018). Review of Drug Repositioning Approaches and Resources. *Int. J. Biol. Sci.* 14 (10), 1232–1244. doi:10.7150/ijbs.24612

Zada, D., Sela, Y., Matosevich, N., Monsonego, A., Lerer-Goldshtein, T., Nir, Y., et al. (2021). Parp1 Promotes Sleep, Which Enhances DNA Repair in Neurons. *Mol. Cell* 81 (24), 4979–4993. e7. doi:10.1016/j.molcel.2021.10.026

Zhao, K., Shi, Y., and So, H.-C. (2022). Prediction of Drug Targets for Specific Diseases Leveraging Gene Perturbation Data: A Machine Learning Approach. *Pharmaceutics* 14 (2), 234. doi:10.3390/pharmaceutics14020234

Conflict of Interest: Authors CL, AP, VG, RNH, EP, AF, SP, WY, MH, A-SS, KT, BS, TC, JM, FMK, and KS are or were employed by AstraZeneca.

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

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fgene.2022.886860

OPEN ACCESS

EDITED BY
Rinku Sharma,
Brigham and Women's Hospital,
United States

REVIEWED BY
Yanlin Zhang,
Second Affiliated Hospital of Soochow University, China
Emanuele Micaglio,
IRCCS San Donato Polyclinic, Italy

*CORRESPONDENCE
Fengxia Lin,
szlinfx@163.com
Xiao Ke,
kexiao@email.szu.edu.cn

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

RECEIVED 01 March 2022
ACCEPTED 06 September 2022
PUBLISHED 28 September 2022

CITATION
Jiang F, Zhang W, Lu H, Tan M, Zeng Z, Song Y, Ke X and Lin F (2022), Prediction of herbal medicines based on immune cell infiltration and immune- and ferroptosis-related gene expression levels to treat valvular atrial fibrillation. *Front. Genet.* 13:886860. doi: 10.3389/fgene.2022.886860

Prediction of herbal medicines based on immune cell infiltration and immune- and ferroptosis-related gene expression levels to treat valvular atrial fibrillation

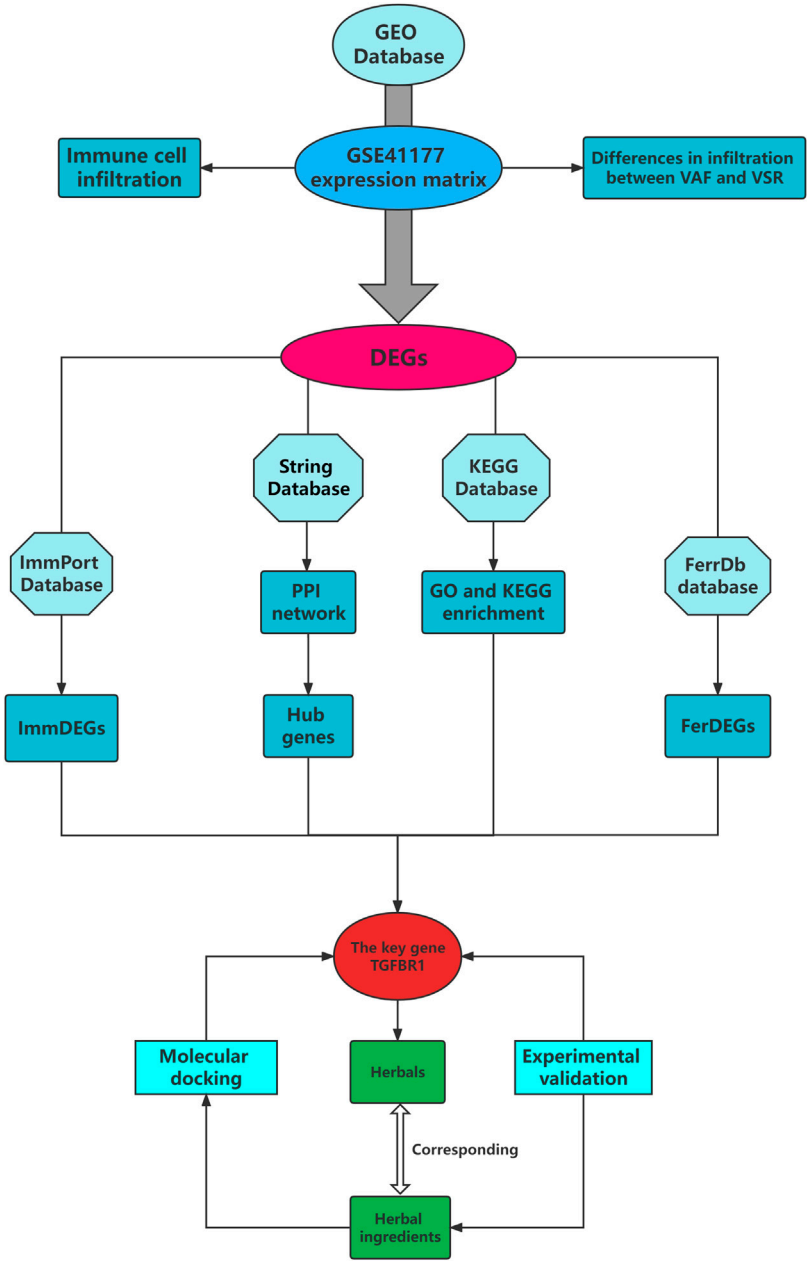
Feng Jiang[1], Weiwei Zhang[1], Hongdan Lu[1], Meiling Tan[2], Zhicong Zeng[1], Yinzhi Song[1], Xiao Ke[3]* and Fengxia Lin[1]*

[1]Cardiology Department, Affiliated Baoan TCM Hospital, Guangzhou University of Traditional Chinese Medicine, Shenzhen, China, [2]Wenhua Community Health Service Center, Shenzhen Luohu Hospital Group, Shenzhen, China, [3]Department of Cardiology, Fuwai Hospital, Chinese Academy of Medical Sciences, Shenzhen(Shenzhen Sun Yat-sen Cardiovascular Hospital), Shenzhen, China

Inflammatory immune response is apparently one of the determinants of progressive exacerbation of valvular atrial fibrillation(VAF). Ferroptosis, an iron-dependent modality of regulated cell death, is involved in the immune regulation of cardiovascular disease. However, the relevant regulatory mechanisms of immune infiltration and ferroptosis in VAF have been less studied. In the current study, a highly efficient system for screening immunity- and ferroptosis-related biomarkers and immunomodulatory ability of herbal ingredients has been developed with the integration of intelligent data acquisition, data mining, network pharmacology, and computer-assisted target fishing. VAF patients showed higher infiltration of neutrophils and resting stage dendritic cells, while VSR patients showed higher infiltration of follicular helper T cells. In addition, six (e.g., PCSK2) and 47 (e.g., TGFBR1) ImmDEGs and one (SLC38A1) and four (TGFBR1, HMGB1, CAV1, and CD44) FerDEGs were highly expressed in patients with valvular sinus rhythm (VSR) and VAF, respectively. We further identified a core subnetwork containing 34 hub genes, which were intersected with ImmDEGs and FerDEGs to obtain the key gene TGFBR1. Based on TGFBR1, 14 herbs (e.g., Fructus zizyphi jujubae, Semen Juglandis, and Polygonum cuspidatum) and six herbal ingredients (curcumin, curcumine, D-glucose, hexose, oleovitamin A, and resveratrol) were predicted. Finally, TGFBR1 was found to dock well with curcumin and resveratrol, and it was further verified that curcumin and resveratrol could significantly reduce myocardial fibrosis. We believe that herbs rich in curcumin and resveratrol such as Rhizoma curcumae longae and Curcuma kwangsiensis, mitigate myocardial fibrosis to improve VAF by modulating the TGFβ/Smad signaling pathway. This strategy provides a prospective approach systemically characterizing phenotype-target-herbs relationships based on the tissue-specific biological functions in VAF and brings us new insights into the searching lead compounds from Chinese herbs.

KEYWORDS

immune cell infiltration, ferroptosis, atrial fibrillation, herbal medicine, prediction



GRAPHICAL ABSTRACT

Introduction

Atrial fibrillation (AF), a common cardiovascular disorder, shows considerably high prevalence across the world, with age being the most important risk factor (Kornej et al., 2020). AF (Table 1) is the most common persistent arrhythmia (Chiang et al., 2013) and an important contributor to stroke, which is the second leading cause of death worldwide (Pistoia et al., 2016). AF has been found to be present in approximately 10% patients with stroke at the time of the attack (Freedman et al., 2017). In fact, considering gaps in monitoring, this percentage is bound to be higher. Haeusler *et al.* on continuous surveillance detected AF in >30% patients with cryptogenic stroke (Haeusler et al., 2018). It is notable that cardiogenic stroke is more severe than other stroke subtypes (Kamel and Healey, 2017). AF is a significant contributor to cardiovascular mortality (Hohendanner et al., 2018), such

TABLE 1 Nonstandard Abbreviations and Acronyms.

Full name	Abbreviation
Atrial fibrillation	AF
Valvular atrial fibrillation	VAF
valvular heart disease	VHD
Valvular sinus rhythm	VSR
differentially expressed genes	DEGs
Immune-related DEGs	ImmDEGs
Ferroptosis-related DEGs	FerDEGs
Protein–protein interaction	PPI
Biological process	BP
Cellular component	CC
Molecular function	MF

as myocardial infarction and heart failure (Ruddox et al., 2017; Carlisle et al., 2019). AF and heart failure reportedly co-exist in up to 30% patients owing to numerous shared pathophysiological mechanisms that facilitate the maintenance of each condition (Prabhu et al., 2017). AF can be divided into valvular and non-valvular AF, the former is typically associated with worse prognosis. Valvular atrial fibrillation (VAF) is one of the common clinical manifestations of valvular heart disease (VHD), and VAF can in turn exacerbate VHD (Gaborit et al., 2005). The timing of intervention in asymptomatic patients with VHD remains controversial, interventions are usually initiated when a decline in exercise capacity is observed or when there is shortness of breath (Baumgartner et al., 2020). Consequently, the risk of death always persists when patients develop severe VAF symptoms (e.g., panic palpitations and restricted activity). Anticoagulation therapy is the most basic method to treat VAF, but treatment efficacy becomes limited with disease progress (Lip et al., 2019). Valve replacement is another commonly used treatment method, but complications such as re-thrombosis and recurrent AF pose a challenge. Valvular sinus rhythm (VSR) represents the early stage of VHD, and as the disease progresses, it evolves into VAF, which is one of the most severe stages of VHD. The pathogenesis of AF remains poorly understood, inflammatory signals are apparently one of the determinants of progressive exacerbation of AF (Nattel et al., 2020). The accumulation of immune cells, such as macrophages, in atrial tissue mediates inflammatory responses, resulting in atrial electrophysiology remodeling (Sun et al., 2016). This inflammatory pathological response increases the incidence of AF, and a mutually reinforcing vicious circle is created (Hu et al., 2015). In addition, ferroptosis plays a potential role in AF. Ferroptosis, an iron-dependent modality of regulated cell death, is distinctly different from cell death mechanisms such as apoptosis, necrosis, and autophagy (Li J. et al., 2020; Tang et al., 2021). Ferroptosis and inflammatory responses promote each other (Sun et al., 2020). Inhibition of ferroptosis has been reported to reduce susceptibility to frequent excessive alcohol consumption-induced AF (Dai et al., 2022). However, only few studies have explored inflammatory responses and mechanisms underlying ferroptosis in VAF. Therefore, we aimed to detect differences in immune cell infiltration and immune- and ferroptosis-related gene expression levels in patients with VAF. Our core goal was to determine how to delay VAF progression. Herbal medicines, a natural treasure trove, contain dozens or even hundreds of ingredients; their mechanisms of action often involve multiple pathways and are thus complex. Numerous herbal medicines have been proven to be effective to prevent and treat cardiovascular diseases (e.g., hypertension) in several

TABLE 2 List of all software and websites used in this study.

Name	Entrance
GEO database	https://www.ncbi.nlm.nih.gov/geo/
R soft and main plug-in package	Version: R 4.1.1; Package: limma, clusterprofiler
ImmPort database	https://www.immport.org/home
String databse	https://cn.string-db.org/
Cytoscape	Version: Cytoscape_v3.9.0; Plug-in: Degree
FerrDb database	http://www.zhounan.org/ferrdb/
KEGG Mapper–Color	https://www.kegg.jp/kegg/mapper/color.htmlv
HERB database	http://herb.ac.cn/
PubChem database	https://pubchem.ncbi.nlm.nih.gov/
ChemOffice	Chem3D 19.0
Uniprot database	https://www.uniprot.org/
PDB database	https://www.rcsb.org/
Autodock vina	Autodock vina 1.1.2

TABLE 3 List of primers for Real-time PCR.

Target	Primer	Sequence (5–3′)
VIMENTIN	FP	CTGCTTCAAGACTCGGTGGAC
	RP	ATCTCCTCCTCGTACAGGTCG
Collagen I	FP	AAGTCACCGAGAGAATTGTCAC
	RP	AGAGAGCCTGTCTTAGCATATCC
α-SAM′	FP	GGACGTACAACTGGTATTGTGC
	RP	TCGGCAGTAGTCACGAAGGA

randomized controlled trials (Hao et al., 2017). With recent advancements in technologies, methods such as high-throughput sequencing have been widely adopted to study the active ingredients of herbal medicines and to identify target genes regulated by them.

Herein we used the Gene Expression Omnibus (GEO) database to obtain information pertaining to local gene expression profiles of patients with VAF and VSR and compared differences in immune cell infiltration and immune- and ferroptosis-related gene expression levels, from the data thus collated, we sought to predict effective herbal medicines to treat VAF.

Materials and methods

Gene expression profile of patients with VAF and VSR

Gene expression profile of patients with VAF and VSR was obtained by searching the GEO database (Table 2); gene IDs were collected and then converted into gene symbols.

Analyses of immune cell infiltration and differentially expressed genes (DEGs)

The CIBERSORT deconvolution method was used to study immune cell infiltration. The gene expression profiles were normalized and screened for DEGs using the limma R package based on the cutoff criteria of $|\log FC| \geq 1$ and adjP value ≤0.05.

Immune-related DEGs (ImmDEGs) and ferroptosis related DEGs (FerDEGs)

In addition to immune cell infiltration analysis, we studied the differential expression of immune-and ferroptosis-related genes in patients with VAF and VSR. Immune- and ferroptosis-related genes were separately identified from the ImmPort and FerrDb databases, respectively; subsequently, they were intersected with DEGs to obtain a list of ImmDEGs and FerDEGs, respectively.

Protein-protein interaction (PPI), hub genes, and enrichment analyses

We used the STRING database to subject DEGs to PPI analysis and top 30 genes were filtered based on the MCODE plugin of Cytoscape, these genes were considered to be hub genes. DEGs were also subjected Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses using the R package clusterprofiler (cutoff: $p \leq 0.05$ and $q \leq 0.05$). For the enrichment results, in addition to visualizing them as bubble plots, DEGs were tagged in the interested enrichment pathway of by using the color tool of the KEGG database.

Key genes and herbal medicine prediction

A key gene was defined as a gene that was a ImmDEG, FerDEG, and hub geneenriched in an immune-related pathway. Based on the identified key genes, we reverse predicted target herbal medicines and ingredients using the HERB database.

Molecular docking for validation

The protein structure of key genes encoded were downloaded from the PDB database and the structure of predicted herbal ingredients required from Pubchem database. Using Autodock vina tools to molecularly dock the key genes with herbal ingredients and the model of lowest binding free energy was regarded as the best bond way.

Experimental design

The HL-1 cells were used to construct the AF model (Hu et al., 2021). The HL-1 cell line was purchased from Shanghai (TongPai, China), used for *in vitro* research and cultivated in DMEM containing 10% foetal bovine serum (Gibco, MA, United States) and 0.1 mM norepinephrine in a 37 °C cell incubator with 5% CO2. Prior to each experiment, HL-1 cells were inoculated in six-well plates and treated as described below when cells reached 70% confluency. Normal control group (NC): HL-1 cells were cultured in DMEM for 48 h. AngII group (AG): HL-1 cells were incubated with 200 nM AngII for 48 h. Curcumin group (CG): HL-1 was first incubated with curcumin for 2 h in a concentration gradient (0, 5, 25, 50, 100, 250, 400, 1000ug/mL) and then 200 nM AngII was added for 48 h. Resveratrol group (RG): HL-1 was first incubated with Resveratrol for 2 h in a concentration gradient (0, 10, 50, 100, 200, 800, 1600ug/mL) and then 200 nM AngII was added for 48 h.

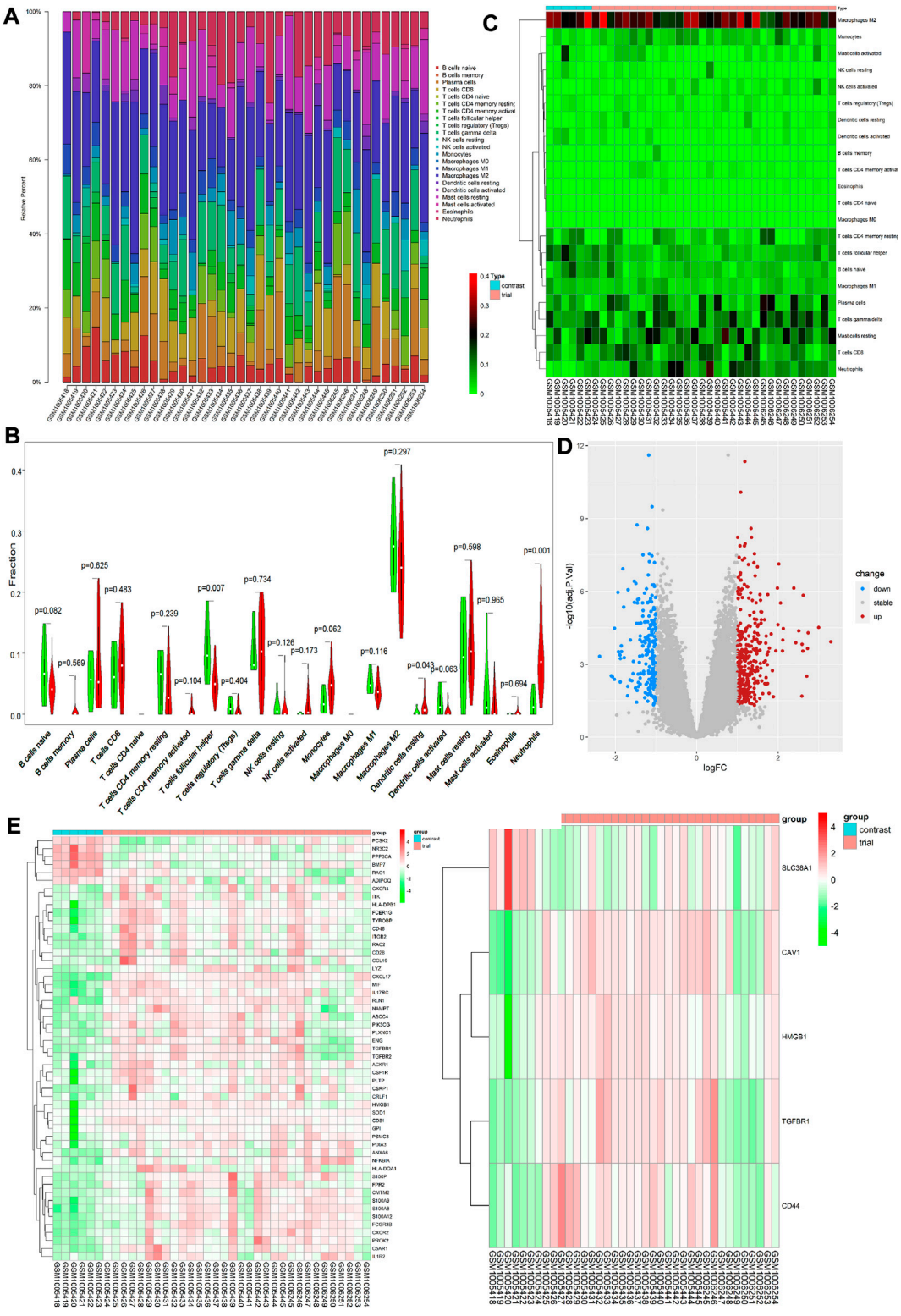


FIGURE 1
(A) Each bar represents a sample, and each color represents a type of immune cell. Area of the color represents the percentage of immune cell infiltration responsible for total immune cell infiltration. **(B)** Each column represents a sample, and each row represents a type of immune cell. Color transition from green to red represents an increase in immune cell infiltration level. **(C)** Red and green violin columns represent patients with VAF and VSR, respectively. The vertical axis represents the ratio of immune cell infiltration responsible for total immune cell infiltration. *p* value, obtained
(*Continued*)

FIGURE 1
using the Wilcoxon test, represents the difference between the immune cell infiltration level in patients with VSR and VAF. **(D)** Upregulated DEGs are highlighted in red and downregulated DEGs in blue. Criteria: |logFC| ≥ 1 and adjP value ≤0.05. **(E)** Expression levels of 53 ImmDEGs are shown; the darker the red color, the higher the expression level, and the darker the green color, the lower the expression level. **(F)** Expression levels of five FerDEGs are shown; the darker the red color, the higher the expression level, and the darker the green color, the lower the expression level. Contrast group = patients with VSR; trial group = patients with VAF.

Assays of CCK8

The growth status of each group of cells was detected by CCK8 and the effect of each herbal medicine on the viability of HL-1 cells was counted and observed to select the optimal concentration for drug intervention. The method was as follows: approximately 5 × 103 cells were cultiviated in 96 well-plates. Cells were incubated with the CCK8 reagent (10ul) for 2 h at 37°C, follow by observation at an absorbance of 450 nm of light by a Thermomax microplate reader(Molecular devices, CA, United States).

Detection of qPCR

Total RNA was extracted from cultivated cells by using Trizol regents. cDNA was synthesized by using the EvoM-MLV Kits. RT-PCR was performed using 2X SYBR Green qPCR Master Mix (K1070-500, APExBIO, US) on a CFX96 Real-Time PCR Detection System (Bio-Rad Laboratories) following the manufacturer's protocol, and analyzed by delta-delta-CT method and given as ratio compared with vehicle control. The following optimized conditions were used: 95 °C for 30 s, 95 °C for 5 s, and 40 cycles at 60 °C for 5 s. The levels of mRNA were normalized in relevance to endogenous GAPDH, and the expression of target genes was analyzed by the method of 2-ΔΔct. The above experiments were repeated three times independently. The primer sequences used in this study are listed in Table 3. All experiments were performed in triplicate.

Statistical analyses

Data were expressed as mean ± standard error of the mean (SEM). Graphpad Prism 9.3.1 software was used to perform unpaired Student's t-tests to analyse differences in quantitative variables between groups and to construct statistical histograms. p value ≤ 0.05 was considered as indicating statistically significant differences.

Results

Gene expression profiles

We downloaded the gene expression matrix of GSE41177 (Yeh et al., 2013) (GPL570 platform) from the GEO database. This data matrix contains gene expression level data for patients with VAF(32 samples) and VSR(6 samples), with samples from both the left atria and pulmonary vein and the surrounding left atrial junction.

Immune cell infiltration and DEGs

On analysis of immune cell infiltration, we initially found differences between patients with VAF and those with VSR (see Figures 1A,C). Subsequently, performing Wilcoxon test, we found that neutrophils (p = 0.001) and resting stage dendritic cells (p = 0.043) were highly expressed in patients with VAF, while follicular helper T cells (p = 0.007) were highly expressed in those with VSR (Figure 1B). We could also identify 585 DEGs: 210 genes were down- and 375 were upregulated (Figure 1D).

ImmDEGs and FerDEGs

On analyzing immune-related gene expression levels, we identified 53 ImmDEGs: six of them (e.g., PCSK2) were highly expressed in patients with VSR and 47 (e.g., TGFBR1, IL1R2, and CD48) were highly expressed in those with VAF(Figure 1E). Similarly, on analyzing ferroptosis-related gene expression levels, we identified five FerDEGs: one of them, i.e., SLC38A1, was highly expressed in patients with VSR and four (TGFBR1, HMGB1, CAV1, and CD44) were highly expressed in those with VAF (see Figure 1F).

PPI network construction, hub gene selection, and enrichment analysis

STRING database was used to perform the PPI analysis of DEGs with the medium confidence ≥0.4 and the top three

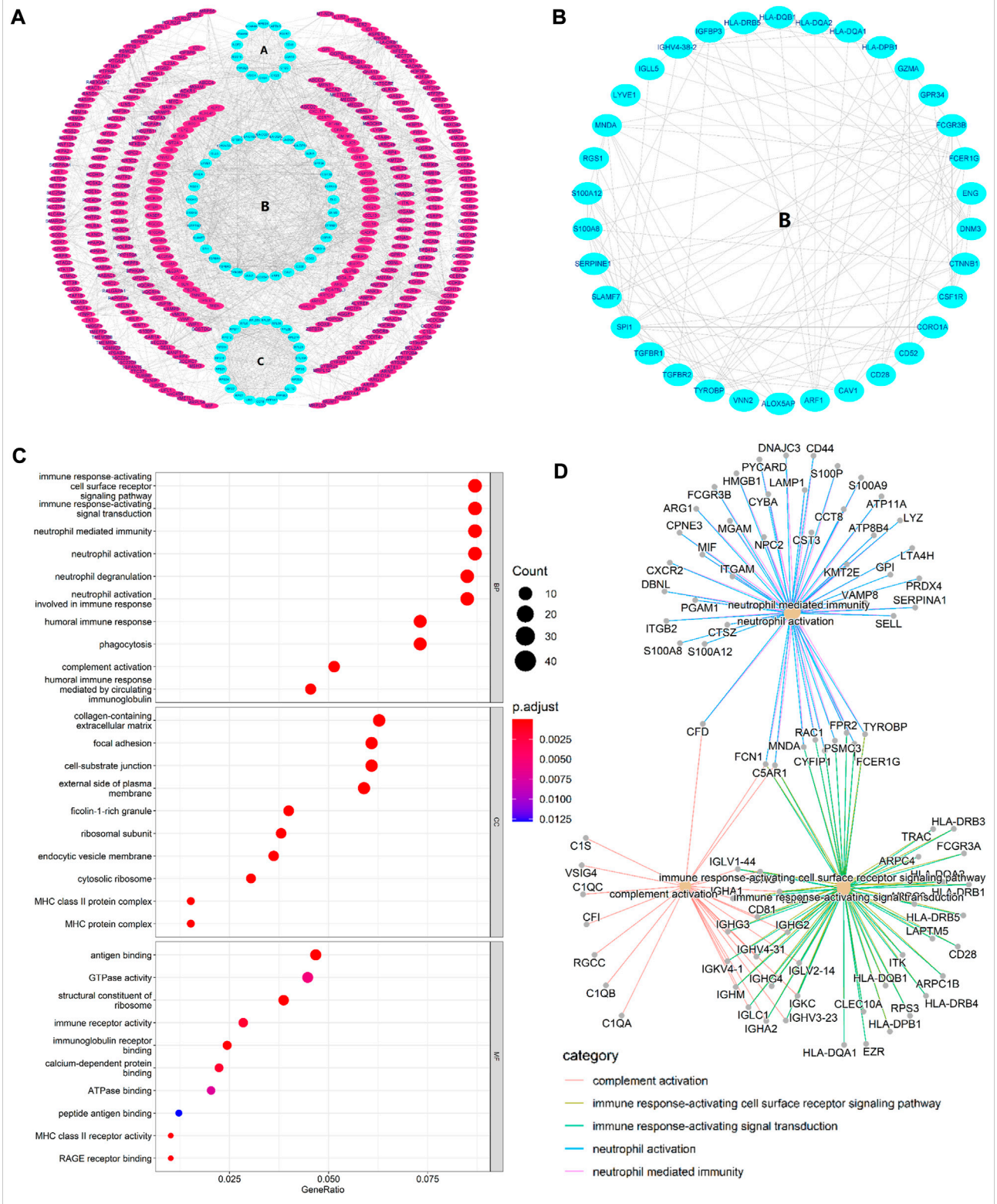


FIGURE 2
(A) PPI network. The three innermost circles represent the top three core subnetworks **(A–C)**. In the network, each node represents a DEG, and the edges represent the correlation between the nodes. In general, the more edges a node had, the greater the role of that node. **(B)** This map shows the core subnetwork **(B)**. Each node represents a hub gene, and the edges represent the correlation between the nodes. In general, the more edges a node had, the greater the role of that node. **(C)** Top 10 GO enrichment results. The horizontal axis represents the gene ratio, i.e., the ratio of the number of DEGs to number of total genes. Dot size is proportional to the gene ratio, and dot color from blue to red that the adjusted *p* value is smaller. **(D)** Association among the top five immune-related biological processes was identified by analyzing co-enriched DEGs.

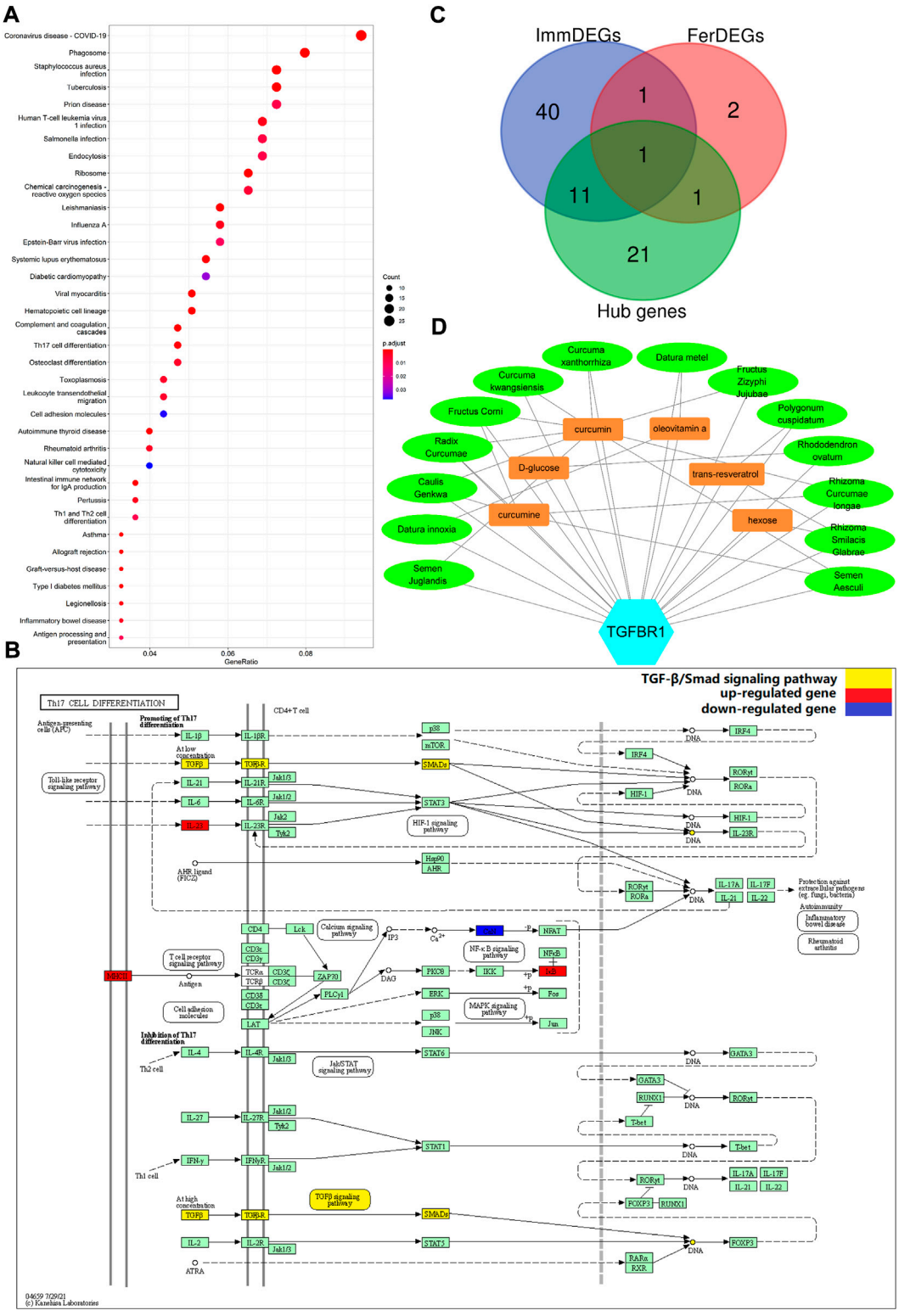


FIGURE 3
(A) Enriched KEGG pathways (n = 36) are shown. The horizontal axis represents the ratio of the number of DEGs enriched in a pathway to the total number of genes in the pathway. Dot size is proportional to the gene-ration, and dot color from blue to red implies that the adjusted *p* value is smaller. **(B)** Th17 cell differentiation pathway. Red nodes represent upregulated DEGs, blue nodes represent downregulated DEGs; the pathway marked in yellow is the TGFβ/Smad signaling pathway. **(C)** TGFBR1 at the intersection of ImmDEGs, FerDEGs, and hub genes. TGFBR1 and
(*Continued*)

FIGURE 3
CAV1 were the intersecting genes between hub genes and FerDEGs, andCD28, ENG, S100A12, HLA-DPB1, HLA-DQA1, TGFBR1, TGFBR2, TYROBP, CSF1R, FCGR3B, FCER1G, and S100A8 were the intersecting genes between hub genes and ImmDEGs. **(D)** This network displays the correspondence between herbal medicines and herbal ingredients. The blue hexagon represents our key genes, TGFBR1. Brown rectangles represent the six predicted herbal ingredients and green ovals represent the 14 predicted herbal medicines. The lines between the herbal medicines and herbal ingredients show that they have some correspondence.

TABLE 4 List of 14 herbal medicines and six herbal ingredients predicted in this study.

Herbal ingredients	Herbal medicines		
curcumin	Curcuma kwangsiensis	Caulis Genkwa	Curcuma xanthorrhiza
	Radix Curcumae	Fructus Corni	Semen Aesculi
	Semen Juglandis	Rhizoma Curcumae longae	
curcumine	Rhizoma Curcumae longae	Radix Curcumae	Fructus Corni
	Semen Aesculi	Caulis Genkwa	Semen Juglandis
resveratrol	Polygonum cuspidatum	Rhizoma Smilacis Glabrae	
oleovitamin a	Datura metel	Datura innoxia	
D-glucose	Rhododendron ovatum		
Hexose	Rhizoma Smilacis Glabrae		

core subnetworks were screened using the MCODE plugin of Cytoscape (Figure 2A), and hub genes were subsequently selected. The largest core subnetwork, subnetwork B, which contains 34 hub genes (Figure 2B). GO enrichment analysis(see Figure 2C) revealed that DEGs were enriched in 209 biological processes (GO-BP) mainly associated with, for example, immune response-activating cell surface receptor signaling pathway and lymphocyte-mediated immunity. DEGs were also enriched in 62 cellular components (GO-CC), and the major categories represented included, for example, MHC class II protein complex and immunoglobulin complex; similarly, DEGs were enriched in 12 molecular functions(GO-MF), with the major categories being structural constituent of ribosome and immunoglobulin receptor binding. In addition, we explored the connection among immune-related BP by analyzing co-enriched DEGs found that some DEGs, such as FCN1 and FPR2, were co-enriched in multiple BP; they were observed to participate in neutrophil-mediated immunity, immune response-activating cell surface receptor signaling pathway, as well as complement activation (Figure 2D). With regard to KEGG pathway enrichment analysis, 36 pathways, such as Th17 cell differentiation and endocytosis, were enriched (Figure 3A). We chose the TGFβ signaling pathway for further analysis, which is important for Th17 cell differentiation (Figure 3B).

Key genes and predicted herbal medicines

On intersecting ImmDEGs, FerDEGs, and hub genes, we obtained two common genes: TGFBR1 and HMGB1 (Figure 3C). We selected TGFBR1 as the key gene after comprehensive analyses of pertinent immune-related KEGG pathway, and 14 herbs (*Fructus zizyphi jujubae*, *Curcuma kwangsiensis*, Semen Juglandis, *Polygonum cuspidatum*, *Curcuma xanthorrhiza*, *Rhizoma curcumae longae*, *Rhododendron ovatum*, *Datura metel*, *Datura innoxia*, Fructus Corni, *Semen aesculi*, *Rhizoma Smilacis Glabrae*, *Radix Curcumae*, and *Caulis genkwa*) and six ingredients (curcumin, curcumine, D-glucose, hexose, oleovitamin A, and resveratrol) were consequently predicted (Table 4). There was a correspondence between herbs and herbal medicines ingredients, such as curcumin, in *Rhizoma curcumae longae* and *Curcuma kwangsiensis*, Semen Juglandis and *Radix Curcumae*, amongst others. A visual network diagram (Figure 3D) was constructed to clearly present the relationship between TGFBR1 and herbs/ingredients.

Molecular docking

Our molecular docking results showed that curcumin and resveratrol docked well with TGFBR1, while D-glucose and oily vitamin A did not bind very tightly to TGFBR1(Figures 4A–D). Unfortunately, we were failed to complete the molecular docking of

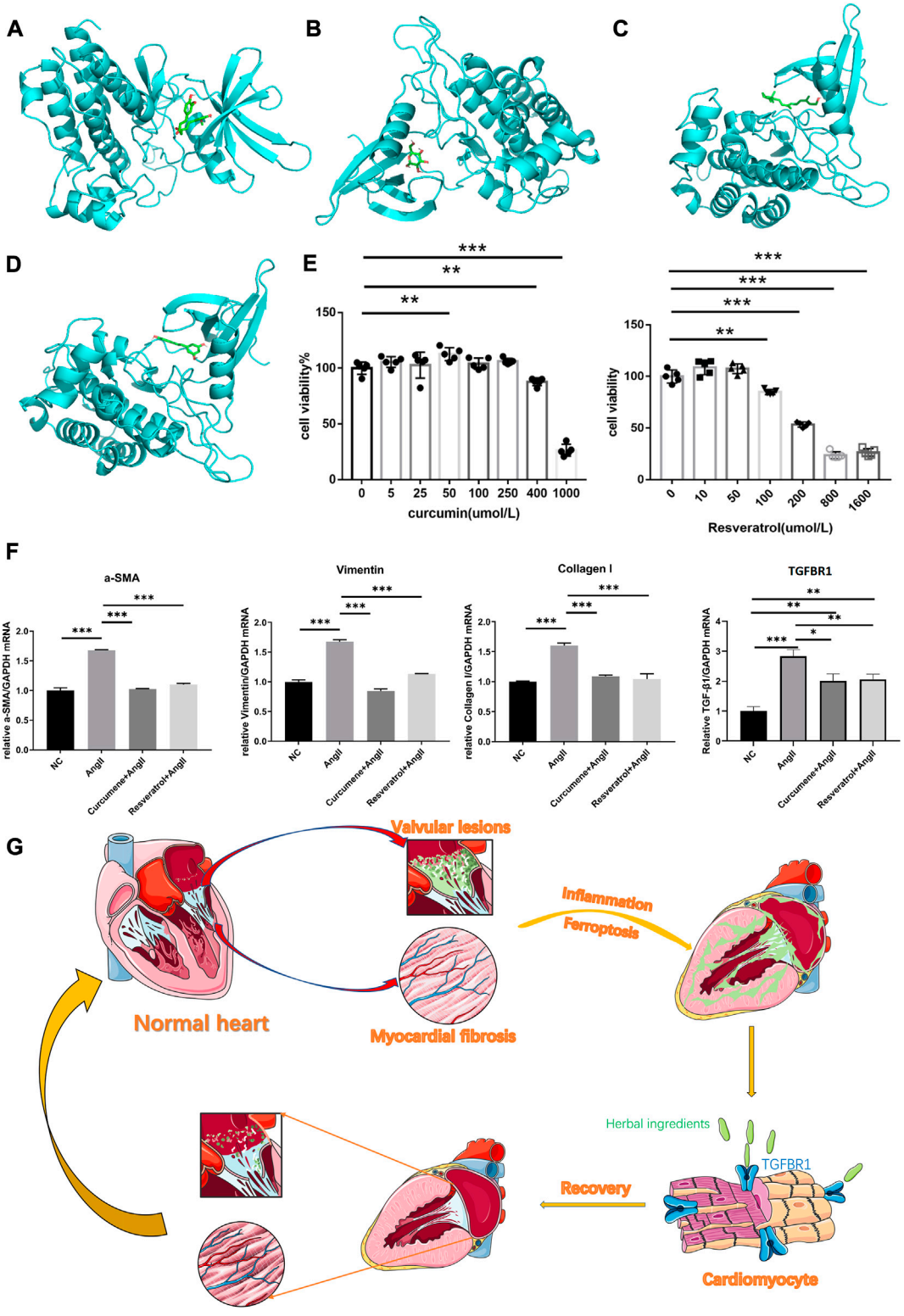


FIGURE 4
(A) The best binding model of curcumin to TGFBR1 with a minimum binding free energy of 8.6 kcal/mol. **(B)** The best binding model of D-glucose to TGFBR1 with a minimum binding free energy of 5.7 kcal/mol. **(C)** The best binding model of oleovitamin A to TGFBR1 with a minimum binding free energy of 7.2 kcal/mol. **(D)** The best binding model of resveratrol to TGFBR1 with a minimum binding free energy of 8.5 kcal/mol **(E)** CCK-8 assays was used to measure the viability of HL-1 cells. **(F)** Quantitative reverse transcription-PCR was used to measure the key gene expression levels. Data are shown as mean ± standard error of the mean. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **(G)** The potential underlying mechanism map.

curcumin and hexose to TGFBR1 due to the unavailability of the 2D or 3D structures of curcumine and Hexose. Usually, lower binding free energy results in higher binding model stability. Apparently, it was easy to find that curcumin and resveratrol bound most strongly to TGFBR1, so we selected these two active ingredients as the target components for subsequent CCK8 and qPCR experiments.

Assays of CCK8

The CCK8 results showed that the intervening concentration of curcumin increased HL-1 cell activity at 50umol/L, did not affect HL-1 cell viability at the remaining low to medium concentrations (≤250umol/L), and significantly inhibited HL-1 activity at high concentrations (≥400umol/L). Resveratrol had no significant effect on HL-1 cell activity at low concentrations (≤50umol/L) and significantly inhibited HL-1 cell activity at high concentrations (≥100umol/L) (see Figure 4E). Therefore, both curcumin and resveratrol were selected at a concentration of 50umol/L in subsequent qPCR experiments.

Detection of qPCR

At this intervention level, the qPCR results showed that the expression of TGFBR1, vimentin, α-SMA and collagen I was significantly lower in NC, CG and RG than in AG. Surprisingly, vimentin, α-SMA and collagen I in CG and RG were not significantly different from NC. However, although TGFBR1 was significantly lower in CG and RG compared to the AG group, it was still higher compared to NC (Figure 4F).

Discussion

We herein found that patients with VAF and VSR showed differences in gene expression and immune cell infiltration levels. In comparison to patients with VSR, 375 up- and 210 down-regulated genes were identified in those with VAF. Further, in local cardiac tissue, patients with VAF showed higher infiltration of neutrophils and resting stage dendritic cells, while those with VSR showed higher infiltration of follicular helper T cells. Neutrophil-mediated inflammatory responses are involved in a variety of cardiovascular diseases (e.g., AF, myocardial infarction, heart failure), which is mainly associated with neutrophil extracellular traps (NETs) that recruit other inflammatory cells such as macrophages to amplify the inflammatory response and promotes collagen synthesis in cardiac tissue leading to fibrosis (Warnatsch et al., 2015; Döring et al., 2020; He et al., 2021; Ling and Xu, 2021).

Dendritic cells are most powerful antigen-presenting cells derived from bone marrow and are essential for stimulating adaptive immunity produced by T cells as well as an important bridge between innate and adaptive immunity, and has been found to be associated with cardiac valve tissue inflammation (Skowasch et al., 2005; Waisman et al., 2017; Collin and Bigley, 2018). Dendritic cells in damaged heart tissue can also secrete inflammatory factors and directly activate fibroblasts to proliferate (Lee et al., 2018), a process that promotes the production of collagen fibres. Resting stage Dendritic cells are predominantly found in peripheral tissues and are specifically responsible for antigen capture rather than antigen presentation, which reserving them for the future initiation of T cell-mediated immune responses (van Duivenvoorde et al., 2006; Tiberio et al., 2018). High infiltration of resting Dendritic cells in VAF without activation may indicate that the local tissue immune response is not strong, persistent low levels of inflammation may be an explanation. Unfortunately, Chronic inflammation leads to tissue damage and this damage process is usually accompanied by fibrotic repair, thus creating a vicious cycle of inflammatory damage and fibrotic repair, which eventually leads to continuous cardiac fibrosis that is closely associated with the development of AF (Abe et al., 2018; Smolgovsky et al., 2021). Follicular helper T cells are a specific subset of T cells that are essential for germinal centre formation, differentiation and maturation of B-cell (Choi and Crotty, 2021). This sort of T cells are usually found in inflamed tissues of secondary lymphoid and non-lymphoid organs and provide auxiliary support to B cells such as stimulating them to produce antibodies (Hutloff, 2018; Yoshitomi and Ueno, 2021). It has been reported that such cells are significantly associated with pulmonary fibrosis, skin fibrosis and systemic sclerosis, and that the main mechanism may be related to immune disorders leading to excessive accumulation of antibodies to form inflammatory fibrotic repair after immune damage (Clark, 2018; Beurier et al., 2021; Zhang et al., 2021). However, it is still very poorly studied in cardiac tissue fibrosis, and only very few studies have reported finding that this cell is associated with the cardiac inflammatory response such as in heart transplants, where it enhances the function of B cells to promote a chronic inflammatory response (Wang Y. et al., 2020). Interestingly, in the VSR group, there was a high infiltration of Follicular helper T cells but not B cells, suggesting at least that the accumulation of antibodies formed during the VSR period was not too high, possibly reflecting, to some extent, only mild fibrosis in the heart tissue during this period. Thus, the immune infiltration findings were more prone to suggest chronic inflammation in both VSR and VAF, with post-inflammatory damage followed by fibrotic repair throughout the evolution of VSR to VAF.

Furthermore, we discovered 47 ImmDEGs that were highly expressed in patients with VAF (e.g., TGFBR1, IL1R2, and CD48) and six that were lowly expressed (e.g., PCSK2). Four FerDEGs

(TGFBR1, CAV1, HMGB1, and CD44) were also highly expressed in the VAF group, whereas one (SLC38A1) was lowly expressed. Interestingly, TGFBR1 and HMGB1, two intersecting genes between ImmDEGs and FerDEGs, were both highly expressed in the VAF group. The crosstalk between immune response and ferroptosis was well established and it has been investigated for the treatment of tumours such as using activation of $CD8^+$ to induce ferroptosis in tumour cells (Tang et al., 2020). The crosstalk between immune cells and ferroptosis can occur in three ways: by the immune cells themselves produce ferroptosis when immune disorder; by tissue cells where ferroptosis is recognised by immune cells and produces an inflammatory clearance response; above both are simultaneously exist (Chen et al., 2021; Yao et al., 2021). In essence, both of way are inflammatory responses, with a sustained inflammatory response leading to further tissue damage and subsequent repair, this repair process that inevitably involves increased secretion and even accumulation of collagen fibres leading to tissue fibrosis. It is clear that this mechanism is likely to be present in the process of evolution from VSR to VAF.

On further analyses, we found that TGFBR1 was the intersecting gene between not only ImmDEGs and FerDEGs but also hub genes. Therefore, we herein considered TGFBR1 as the key gene. TGFBR1, a pleiotropic cytokine, plays a pivotal role in immune response and mediates a vicious cycle of inflammation and tissue fibrosis (Bonniaud et al., 2005; Esebanmen and Langridge, 2017). In fact, TGFBR1 is also involved in tissue fibrosis during ferroptosis. Li et al. reported that ferroptosis inhibitor liproxstatin-1 alleviates radiation-induced lung fibrosis via TGFBR1 downregulation (Li et al., 2019). An increasing body of evidence indicates that AF development is associated with atrial myocardial fibrosis, which presumably underlies the pathology of this persistent arrhythmia (Dzeshka et al., 2015; Jalife and Kaur, 2015; Sohns and Marrouche, 2020). In cardiac tissue, TGFBR1 is evidently involved in the process of tissue fibrosis that can cause VAF and it has been found to cause or exacerbate AF by promoting atrial tissue fibrosis (Khalil et al., 2017; Liu Y. et al., 2021). Wang et al. suggested that quercetin alleviates AF by inhibiting fibrosis of atrial tissues through inhibiting the TGF-β/Smads signaling pathway (Wang et al., 2021). Khalil et al. reported that TGFBR1 participates in tissue fibrosis primarily via the TGFβ-/Smad signaling pathway (Khalil et al., 2017). In traditional Chinese medicine, some herbs, such as Taohong Siwu, have been also observed to significantly attenuate myocardial fibrosis by inhibiting fibrosis proliferation and collagen deposition via this pathway (Tan et al., 2021). In the present study, we also found TGFβ-/Smad signaling pathway to be significantly enriched as a sub-pathway of Th17 cell differentiation. Based on TGFBR1, we predicted six herbal ingredients and 14 herbal medicines. Some of the herbal ingredients identified herein reportedly alleviate tissue fibrosis by modulating the TGF-β/Smad signaling pathway; for example, curcumin has been reported to attenuate pulmonary, hepatic, and renal interstitial fibrosis (Saidi et al., 2019; Wang Z. et al., 2020; Kong et al., 2020). Moreover, curcumin has been found to be effective for treating cardiovascular diseases, such as heart failure, myocardial infarction, atherosclerosis (Li H. et al., 2020), and and it can significantly inhibit the duration of atrial fibrillation episodes, attenuate cardiac fibrosis (Yue et al., 2021). However, there are fewer studies on curcumin's anti-fibrotic effects through its action on TGFBR1. Therefore, using *in vitro* models and qPCR assays, we found that curcumin significantly reduced the expression of TGFBR1 and fibrosis indicators that Vimentin, α-SMA, collagen I are common indicators of myocardial fibrosis (Ma et al., 2018; Liu M. et al., 2021), which tentatively confirmed the potential of this substance to improve VAF by interfering with TGFBR1 to reduce atrial tissue fibrosis. Curcumin has also been shown to exert anti-inflammatory effects by inhibiting neutrophil infiltration (Antoine et al., 2013), which may also be one of mechanism for reducing myocardial fibrosis.

Resveratrol, another popular herbal ingredient, has also shown good efficacy in the treatment of several cardiovascular diseases (Baczko and Light, 2015; Chong et al., 2015; Yousefian et al., 2019). Resveratrol can effectively improve atrial fibrillation by inhibiting NADPH oxidase and ion channels (Barangi et al., 2018), and like curcumin, it also acts as an anti-inflammatory agent by inhibiting neutrophil activation (Tsai et al., 2019), and attenuates myocardial ischemia-reperfusion injury by inhibiting ferroptosis (Li et al., 2022). However, direct evidence that resveratrol ameliorate atrial myocardial fibrosis is still lacking, so we evaluated their efficacy by intervening in fibrotic HL-1 cells with resveratrol. Same as curcumin, the qPCR results confirmed that resveratrol significantly reduced the expression of TGFBR1 as well as the indicator of fibrosis including Vimentin, α-SMA and collagen I in HL-1 induced by AngII, which given a robust evidence for this potential candidate to act as a treatment for AVF. Therefore, curcumin and resveratrol have great potential to improve AF by acting on TGFBR1 expression to reduce myocardial fibrosis. Oleovitamin A and D-glucose are readily available from daily foods, they have been less studied in AF and VHD, and our molecular docking results suggest that they do not bind very strongly to TGFBR1, and thus we believe they might have less potential to treat VAF.

Conclusion

We believe that herbs rich in curcumin, resveratrol, such as *Rhizoma curcumae longae*, *Curcuma xanthorrhiza*, and *Caulis genkwa*, attenuate myocardial fibrosis to alleviate VAF by acting on TGFBR1 (see Figure 4G for the potential underlying mechanism), they seem to be effective treatment strategy for VAF.

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

FL and XK conceived the idea, designed the study. ZZ and YS contributed to the revision of the manuscript draft. FJ and WZ conducted the data analysis and visualization. HL and MT performed the data interpretation and literature collection. FJ and FL wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (82004320), the Natural Science Foundation of Guangdong Province of China (2021A1515011095, 2022A1515011710, 2022A1515010679), the Science and Technology Project of Shenzhen City of China (JCYJ20190807115201653), Shenzhen Bao'an Chinese Medicine Hospital Research Program (BAZYY20220702) and Technology Innovation department of Baoan (2021JD103).

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

Abe, I., Teshima, Y., Kondo, H., Kaku, H., Kira, S., Ikebe, Y., et al. (2018). Association of fibrotic remodeling and cytokines/chemokines content in epicardial adipose tissue with atrial myocardial fibrosis in patients with atrial fibrillation. *Heart rhythm.* 15, 1717–1727. doi:10.1016/j.hrthm.2018.06.025

Antoine, F., Simard, J. C., and Girard, D. (2013). Curcumin inhibits agent-induced human neutrophil functions *in vitro* and lipopolysaccharide-induced neutrophilic infiltration *in vivo*. *Int. Immunopharmacol.* 17, 1101–1107. doi:10.1016/j.intimp.2013.09.024

Baczko, I., and Light, P. E. (2015). Resveratrol and derivatives for the treatment of atrial fibrillation. *Ann. N. Y. Acad. Sci.* 1348, 68–74. doi:10.1111/nyas.12843

Barangi, S., Hayes, A. W., and Karimi, G. (2018). The more effective treatment of atrial fibrillation applying the natural compounds; as NADPH oxidase and ion channel inhibitors. *Crit. Rev. Food Sci. Nutr.* 58, 1230–1241. doi:10.1080/10408398.2017.1379000

Baumgartner, H., Iung, B., and Otto, C. M. (2020). Timing of intervention in asymptomatic patients with valvular heart disease. *Eur. Heart J.* 41, 4349–4356. doi:10.1093/eurheartj/ehaa485

Beurier, P., Ricard, L., Eshagh, D., Malard, F., Siblany, L., Fain, O., et al. (2021). TFH cells in systemic sclerosis. *J. Transl. Med.* 19, 375. doi:10.1186/s12967-021-03049-0

Bonniaud, P., Margetts, P. J., Ask, K., Flanders, K., Gauldie, J., and Kolb, M. (2005). TGF-beta and Smad3 signaling link inflammation to chronic fibrogenesis. *J. Immunol.* 175, 5390–5395. doi:10.4049/jimmunol.175.8.5390

Carlisle, M. A., Fudim, M., DeVore, A. D., and Piccini, J. P. (2019). Heart failure and atrial fibrillation, like fire and fury. *JACC. Heart Fail.* 7, 447–456. doi:10.1016/j.jchf.2019.03.005

Chen, X., Kang, R., Kroemer, G., and Tang, D. (2021). Ferroptosis in infection, inflammation, and immunity. *J. Exp. Med.* 218, e20210518. doi:10.1084/jem.20210518

Chiang, C. E., Zhang, S., Tse, H. F., Teo, W. S., Omar, R., and Sriratanasathavorn, C. (2013). Atrial fibrillation management in Asia: From the Asian expert forum on atrial fibrillation. *Int. J. Cardiol.* 164, 21–32.

Choi, J., and Crotty, S. (2021). Bcl6-Mediated transcriptional regulation of follicular helper T cells (TFH). *Trends Immunol.* 42, 336–349. doi:10.1016/j.it.2021.02.002

Chong, E., Chang, S. L., Hsiao, Y. W., Singhal, R., Liu, S. H., Leha, T., et al. (2015). Resveratrol, a red wine antioxidant, reduces atrial fibrillation susceptibility in the failing heart by PI3K/AKT/eNOS signaling pathway activation. *Heart rhythm.* 12, 1046–1056. doi:10.1016/j.hrthm.2015.01.044

Clark, R. A. (2018). That's not helping-T follicular helper cells drive skin fibrosis. *Sci. Immunol.* 3, eaat6417. doi:10.1126/sciimmunol.aat6417

Collin, M., and Bigley, V. (2018). Human dendritic cell subsets: An update. *Immunology* 154, 3–20. doi:10.1111/imm.12888

Dai, C., Kong, B., Qin, T., Xiao, Z., Fang, J., Gong, Y., et al. (2022). Inhibition of ferroptosis reduces susceptibility to frequent excessive alcohol consumption-induced atrial fibrillation. *Toxicology* 465, 153055. doi:10.1016/j.tox.2021.153055

Döring, Y., Libb, P., and Soehnlein, O. (2020). Neutrophil extracellular traps participate in cardiovascular diseases: Recent experimental and clinical insights. *Circ. Res.* 126, 1228–1241. doi:10.1161/CIRCRESAHA.120.315931

Dzeshka, M. S., Lip, G. Y., Snezhitskiy, V., and Shantsila, E. (2015). Cardiac fibrosis in patients with atrial fibrillation: Mechanisms and clinical implications. *J. Am. Coll. Cardiol.* 66, 943–959. doi:10.1016/j.jacc.2015.06.1313

Esebanmen, G. E., and Langridge, W. H. R. (2017). The role of TGF-beta signaling in dendritic cell tolerance. *Immunol. Res.* 65, 987–994. doi:10.1007/s12026-017-8944-9

Freedman, B., Camm, J., Calkins, H., Healey, J. S., Rosenqvist, M., Wang, J., et al. (2017). Screening for atrial fibrillation: A Report of the AF-SCREEN International Collaboration. *Circulation* 135, 1851–1867. doi:10.1161/CIRCULATIONAHA.116.026693

Gaborit, N., Steenman, M., Lamirault, G., Le Meur, N., Le Bouter, S., Lande, G., et al. (2005). Human atrial ion channel and transporter subunit gene-expression remodeling associated with valvular heart disease and atrial fibrillation. *Circulation* 112, 471–481. doi:10.1161/CIRCULATIONAHA.104.506857

Haeusler, K. G., Tutuncu, S., and Schnabel, R. B. (2018). Detection of atrial fibrillation in cryptogenic stroke. *Curr. Neurol. Neurosci. Rep.* 18, 66. doi:10.1007/s11910-018-0871-1

Hao, P., Jiang, F., Cheng, J., Ma, L., Zhang, Y., and Zhao, Y. (2017). Traditional Chinese medicine for cardiovascular disease: Evidence and potential mechanisms. *J. Am. Coll. Cardiol.* 69, 2952–2966. doi:10.1016/j.jacc.2017.04.041

He, L., Liu, R., Yue, H., Zhu, G., Fu, L., Chen, H., et al. (2021). NETs promote pathogenic cardiac fibrosis and participate in ventricular aneurysm formation after ischemia injury through the facilitation of perivascular

fibrosis. *Biochem. Biophys. Res. Commun.* 583, 154–161. doi:10.1016/j.bbrc.2021.10.068

Hohendanner, F., Heinzel, F. R., Blaschke, F., Pieske, B. M., Haverkamp, W., Boldt, H. L., et al. (2018). Pathophysiological and therapeutic implications in patients with atrial fibrillation and heart failure. *Heart fail. Rev.* 23, 27–36. doi:10.1007/s10741-017-9657-9

Hu, H. J., Wang, X. H., Liu, Y., Zhang, T. Q., Chen, Z. R., Zhang, C., et al. (2021). Hydrogen Sulfide ameliorates Angiotensin II-induced atrial fibrosis progression to atrial fibrillation through inhibition of the Warburg effect and Endoplasmic Reticulum stress. *Front. Pharmacol.* 12, 690371. doi:10.3389/fphar.2021.690371

Hu, Y. F., Chen, Y. J., Lin, Y. J., and Chen, S. A. (2015). Inflammation and the pathogenesis of atrial fibrillation. *Nat. Rev. Cardiol.* 12, 230–243. doi:10.1038/nrcardio.2015.2

Hutloff, A., (2018). T follicular helper-like cells in inflamed non-lymphoid tissues. *Front. Immunol.* 23 (9), 1707. doi:10.3389/fimmu.2018.01707

Jalife, J., and Kaur, K. (2015). Atrial remodeling, fibrosis, and atrial fibrillation. *Trends cardiovasc. Med.* 25, 475–484. doi:10.1016/j.tcm.2014.12.015

Kamel, H., and Healey, J. S. (2017). Cardioembolic stroke. *Circ. Res.* 120, 514–526. doi:10.1161/CIRCRESAHA.116.308407

Khalil, H., Kanisicak, O., Prasad, V., Correll, R. N., Fu, X., Schips, T., et al. (2017). Fibroblast-specific TGF-beta-Smad2/3 signaling underlies cardiac fibrosis. *J. Clin. Invest.* 127, 3770–3783. doi:10.1172/JCI94753

Kong, D., Zhang, Z., Chen, L., Huang, W., Zhang, F., Wang, L., et al. (2020). Curcumin blunts epithelial-mesenchymal transition of hepatocytes to alleviate hepatic fibrosis through regulating oxidative stress and autophagy. *Redox Biol.* 36, 101600. doi:10.1016/j.redox.2020.101600

Kornej, J., Börschel, C. S., Benjamin, E. J., and Schnabel, R. B. (2020). Epidemiology of atrial fibrillation in the 21st Century: Novel methods and new insights. *Circ. Res.* 127, 4–20.

Lee, J. S., Jeong, S. J., Kim, S., Chalifour, L., Yun, T. J., Miah, M. A., et al. (2018). Conventional dendritic cells Impair Recovery after myocardial infarction. *J. Immunol.* 201, 1784–1798. doi:10.4049/jimmunol.1800322

Li, H., Sureda, A., Devkota, H. P., Pittalà, V., Barreca, D., Silva, A. S., et al. (2020). Curcumin, the golden spice in treating cardiovascular diseases. *Biotechnol. Adv.* 38, 107343. doi:10.1016/j.biotechadv.2019.01.010

Li, J., Cao, F., Yin, H. L., Huang, Z. J., Lin, Z. T., Mao, N., et al. (2020). Ferroptosis: Past, present and future. *Cell Death Dis.* 11, 88. doi:10.1038/s41419-020-2298-2

Li, T., Tan, Y., Ouyang, S., He, J., and Liu, L. (2022). Resveratrol protects against myocardial ischemia-reperfusion injury via attenuating ferroptosis. *Gene* 808, 145968. doi:10.1016/j.gene.2021.145968

Li, X., Duan, L., Yuan, S., Zhuang, X., Qiao, T., and He, J. (2019). Ferroptosis inhibitor alleviates Radiation-induced lung fibrosis (RILF) via down-regulation of TGF-β1. *J. Inflamm.* 16, 11. doi:10.1186/s12950-019-0216-0

Ling, S., and Xu, J. W. (2021). NETosis as a pathogenic factor for heart failure. *Oxid. Med. Cell. Longev.* 2021, 6687096. doi:10.1155/2021/6687096

Lip, G. Y. H., Collet, J. P., Caterina, R., Fauchier, L., Lane, D. A., Larsen, T. B., et al. (2019). Antithrombotic therapy in atrial fibrillation associated with valvular heart disease: A joint consensus document from the European heart rhythm association (EHRA) and European Society of Cardiology working group on thrombosis, endorsed by the ESC working group on valvular heart disease, cardiac arrhythmia Society of Southern Africa (CASSA), heart rhythm Society (HRS), Asia Pacific heart rhythm Society (APHRS), South African heart (SA heart) association and Sociedad Latinoamericana de Estimulacion Cardiaca y Electrofisiologia (SOLEACE). *Europace* 19, 1757–1758. doi:10.1093/europace/eux240

Liu, M., López de Juan Abad, B., and Cheng, K. (2021). Cardiac fibrosis: Myofibroblast-mediated pathological regulation and drug delivery strategies. *Adv. Drug Deliv. Rev.* 173, 504–519. doi:10.1016/j.addr.2021.03.021

Liu, Y., Yin, Z., Xu, X., Liu, C., Duan, X., Song, Q., et al. (2021). Crosstalk between the activated Slit2-Robo1 pathway and TGF-β1 signalling promotes cardiac fibrosis. *Esc. Heart Fail.* 8, 447–460. doi:10.1002/ehf2.13095

Ma, Z. G., Yuan, Y. P., Wu, H. M., Zhang, X., and Tang, Q. Z. (2018). Cardiac fibrosis: New insights into the pathogenesis. *Int. J. Biol. Sci.* 14, 1645–1657. doi:10.7150/ijbs.28103

Nattel, S., Heijman, J., Zhou, L., and Dobrev, D. (2020). Molecular Basis of atrial fibrillation Pathophysiology and therapy: A Translational perspective. *Circ. Res.* 127, 51–72. doi:10.1161/CIRCRESAHA.120.316363

Pistoia, F., Sacco, S., Tiseo, C., Degan, D., Ornello, R., and Carolei, A. (2016). The Epidemiology of atrial fibrillation and stroke. *Cardiol. Clin.* 34, 255–268. doi:10.1016/j.ccl.2015.12.002

Prabhu, S., Voskoboinik, A., Kaye, D. M., and Kistler, P. M. (2017). Atrial fibrillation and heart failure - cause or effect? *Heart Lung Circ.* 26, 967–974. doi:10.1016/j.hlc.2017.05.117

Ruddox, V., Sandven, I., Munkhaugen, J., Skattebu, J., Edvardsen, T., and Otterstad, J. E. (2017). Atrial fibrillation and the risk for myocardial infarction, all-cause mortality and heart failure: A systematic review and meta-analysis. *Eur. J. Prev. Cardiol.* 24, 1555–1566. doi:10.1177/2047487317715769

Saidi, A., Kasabova, M., Vanderlynden, L., Wartenberg, M., Kara-Ali, G. H., Marc, D., et al. (2019). Curcumin inhibits the TGF-β1-dependent differentiation of lung fibroblasts via PPARγ-driven upregulation of cathepsins B and L. *Sci. Rep.* 9, 491. doi:10.1038/s41598-018-36858-3

Skowasch, D., Schrempf, S., Wernert, N., Steinmetz, M., Jabs, A., Tuleta, I., et al. (2005). Cells of primarily extra-valvular origin in degenerative aortic valves and bioprostheses. *Eur. Heart J.* 26, 2576–2580. doi:10.1093/eurheartj/ehi458

Smolgovsky, S., Ibeh, U., Tamayo, T. P., and Alcaide, P. (2021). Adding insult to injury - inflammation at the heart of cardiac fibrosis. *Cell. Signal.* 77, 109828. doi:10.1016/j.cellsig.2020.109828

Sohns, C., and Marrouche, N. F. (2020). Atrial fibrillation and cardiac fibrosis. *Eur. Heart J.* 41, 1123–1131. doi:10.1093/eurheartj/ehz786

Sun, Y., Chen, P., Zhai, B., Zhang, M., Xiang, Y., Fang, J., et al. (2020). The emerging role of ferroptosis in inflammation. *Biomed. Pharmacother.* 127, 110108. doi:10.1016/j.biopha.2020.110108

Sun, Z., Zhou, D., Xie, X., Wang, S., Wang, Z., Zhao, W., et al. (2016). Cross-talk between macrophages and atrial myocytes in atrial fibrillation. *Basic Res. Cardiol.* 111, 63. doi:10.1007/s00395-016-0584-z

Tan, Z., Jiang, X., Zhou, W., Deng, B., Cai, M., Deng, S., et al. (2021). Taohong siwu decoction attenuates myocardial fibrosis by inhibiting fibrosis proliferation and collagen deposition via TGFBR1 signaling pathway. *J. Ethnopharmacol.* 270, 113838. doi:10.1016/j.jep.2021.113838

Tang, D., Chen, X., Kang, R., and Kroemer, G. (2021). Ferroptosis: Molecular mechanisms and health implications. *Cell Res.* 31, 107–125. doi:10.1038/s41422-020-00441-1

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

Tiberio, L., Del Prete, A., Schioppa, T., Sozio, F., Bosisio, D., and Sozzani, S. (2018). Chemokine and chemotactic signals in dendritic cell migration. *Cell. Mol. Immunol.* 15, 346–352. doi:10.1038/s41423-018-0005-3

Tsai, Y. F., Chen, C. Y., Chang, W. Y., Syu, Y. T., and Hwang, T. L. (2019). Resveratrol suppresses neutrophil activation via inhibition of Src family kinases to attenuate lung injury. *Free Radic. Biol. Med.* 145, 67–77. doi:10.1016/j.freeradbiomed.2019.09.021

van Duivenvoorde, L. M., van Mierlo, G. J., Boonman, Z. F., and Toes, R. E. (2006). Dendritic cells: Vehicles for tolerance induction and prevention of autoimmune diseases. *Immunobiology* 211, 627–632. doi:10.1016/j.imbio.2006.05.014

Waisman, A., Lukas, D., Clausen, B. E., and Yogev, N. (2017). Dendritic cells as gatekeepers of tolerance. *Semin. Immunopathol.* 39, 153–163. doi:10.1007/s00281-016-0583-z

Wang, H., Jiang, W., Hu, Y., Wan, Z., Bai, H., Yang, Q., et al. (2021). Quercetin improves atrial fibrillation through inhibiting TGF-β/Smads pathway via promoting MiR-135b expression. *Phytomedicine.* 93, 153774. doi:10.1016/j.phymed.2021.153774

Wang, Y., Liu, Z., Wu, J., Li, F., Li, G., and Dong, N. (2020). Profiling circulating T follicular helper cells and their effects on B cells in post-cardiac transplant recipients. *Ann. Transl. Med.* 8, 1369. doi:10.21037/atm-20-3027

Wang, Z., Chen, Z., Li, B., Zhang, B., Du, Y., Liu, Y., et al. (2020). Curcumin attenuates renal interstitial fibrosis of obstructive nephropathy by suppressing epithelial-mesenchymal transition through inhibition of the TLR4/NF-κB and PI3K/AKT signalling pathways. *Pharm. Biol.* 58, 828–837. doi:10.1080/13880209.2020.1809462

Warnatsch, A., Ioannou, M., Wang, Q., and Papayannopoulos, V. (2015). Inflammation. Neutrophil extracellular traps license macrophages for cytokine production in atherosclerosis. *Science* 349, 316–320. doi:10.1126/science.aaa8064

Yao, Y., Chen, Z., Zhang, H., Chen, C., Zeng, M., Yunis, J., et al. (2021). Author Correction: Selenium-GPX4 axis protects follicular helper T cells from ferroptosis. *Nat. Immunol.* 22, 1599. doi:10.1038/s41590-021-01063-4

Yeh, Y. H., Kuo, C. T., Lee, Y. S., Lin, Y. M., Nattel, S., Tsai, F. C., et al. (2013). Region-specific gene expression profiles in the left atria of patients with valvular atrial fibrillation. *Heart rhythm.* 10, 383–391. doi:10.1016/j.hrthm.2012.11.013

Yoshitomi, H., and Ueno, H. (2021). Shared and distinct roles of T peripheral helper and T follicular helper cells in human diseases. *Cell. Mol. Immunol.* 18, 523–527. doi:10.1038/s41423-020-00529-z

Yousefian, M., Shakour, N., Hosseinzadeh, H., Hayes, A. W., Hadizadeh, F., and Karimi, G. (2019). The natural phenolic compounds as modulators of NADPH oxidases in hypertension. *Phytomedicine* 55, 200–213. doi:10.1016/j.phymed.2018.08.002

Yue, H., Zhao, X., Liang, W., Qin, X., Bian, L., He, K., et al. (2021). Curcumin, novel application in reversing myocardial fibrosis in the treatment for atrial fibrillation from the perspective of transcriptomics in rat model. *Biomed. Pharmacother.* 146, 112522. doi:10.1016/j.biopha.2021.112522

Zhang, N., Li, X., Wang, J., Wang, J., Li, N., Wei, Y., et al. (2021). Galectin-9 regulates follicular helper T cells to inhibit humoral autoimmunity-induced pulmonary fibrosis. *Biochem. Biophys. Res. Commun.* 534, 99–106. doi:10.1016/j.bbrc.2020.11.097

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 November 2022
DOI 10.3389/fgene.2022.917636

OPEN ACCESS

EDITED BY
Mallana Gowdra Mallikarjuna,
Indian Agricultural Research Institute (ICAR), India

REVIEWED BY
Shiyu Song,
Nanjing University, China
Luis Antonio Pérez-García,
Autonomous University of San Luis Potosí, Mexico

*CORRESPONDENCE
Akram Mohammed,
amoham18@uthsc.edu

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

RECEIVED 11 April 2022
ACCEPTED 08 November 2022
PUBLISHED 22 November 2022

CITATION
Naik S and Mohammed A (2022), Coexpression network analysis of human candida infection reveals key modules and hub genes responsible for host-pathogen interactions.
Front. Genet. 13:917636.
doi: 10.3389/fgene.2022.917636

Coexpression network analysis of human candida infection reveals key modules and hub genes responsible for host-pathogen interactions

Surabhi Naik[1] and Akram Mohammed[2]*

[1]Department of Surgery, James D. Eason Transplant Institute, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States, [2]Center for Biomedical Informatics, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States

Invasive fungal infections are a significant reason for morbidity and mortality among organ transplant recipients. Therefore, it is critical to investigate the host and candida niches to understand the epidemiology of fungal infections in transplantation. *Candida albicans* is an opportunistic fungal pathogen that causes fatal invasive mucosal infections, particularly in solid organ transplant patients. Therefore, identifying and characterizing these genes would play a vital role in understanding the complex regulation of host-pathogen interactions. Using 32 RNA-sequencing samples of human cells infected with *C. albicans*, we developed WGCNA coexpression networks and performed DESeq2 differential gene expression analysis to identify the genes that positively correlate with human candida infection. Using hierarchical clustering, we identified 5 distinct modules. We studied the inter- and intramodular gene network properties in the context of sample status traits and identified the highly enriched genes in the correlated modules. We identified 52 genes that were common in the most significant WGCNA turquoise module and differentially expressed genes in human endothelial cells (HUVEC) infection vs. control samples. As a validation step, we identified the differentially expressed genes from the independent Candida-infected human oral keratinocytes (OKF6) samples and validated 30 of the 52 common genes. We then performed the functional enrichment analysis using KEGG and GO. Finally, we performed protein-protein interaction (PPI) analysis using STRING and CytoHubba from 30 validated genes. We identified 8 hub genes (*JUN, ATF3, VEGFA, SLC2A1, HK2, PTGS2, PFKFB3, and KLF6*) that were enriched in response to hypoxia, angiogenesis, vasculogenesis, hypoxia-induced signaling, cancer, diabetes, and transplant-related disease pathways. The discovery of genes and functional pathways related to the immune system and gene coexpression and differential gene expression analyses may serve as novel diagnostic markers and potential therapeutic targets.

KEYWORDS

host-pathogen interaction, correlation network, RNA-sequencing, immune response, candida albican, WGCNA, transplantation

Introduction

Solid organ transplant (SOT) patients are exposed to various complications, e.g., invasive fungal infection and organ failure, which are the major challenge in SOT and affect the morbidity and mortality in transplant patients. The most prevalent invasive fungal infection in SOT is Candidiasis, which includes about 60% of infections, followed by aspergillosis accounts for up to 25% of fungal infections (Shoham and Marr, 2012).

An opportunistic fungal pathogen, *Candida albicans*, is part of healthy human gut microbiota. However, when immunity is compromised or suppressed, particularly in organ transplant individuals, AIDS patients, chemotherapy-treated patients, and neonates, the mucosal layer becomes more susceptible to fatal invasive *C. albicans* infections such as candidiasis (Sangeorzan et al., 1994; Rhodus et al., 1997; Revankar et al., 1998; Redding et al., 1999; Willis et al., 1999), (Sobel, 1985). *C. albicans* can switch from an avirulent commensal yeast form to a virulent invasive hyphal form in which hyphae invade through the mucosal layer and disseminate/propagate through the blood, infecting other organs as well as developing multidrug resistance (Klepser, 2006; Cowen et al., 2015; Arendrup and Patterson, 2017; Pendleton et al., 2017; Nishimoto et al., 2020). In the process of *C. albicans* infection, the first site of host-pathogen interactions is epithelial and endothelial cells (Barker et al., 2008; Liu et al., 2015a). The development of invasive fungal diseases relies on the synergy between the host immune response and fungal virulence. Comprehensive network analysis is vital to understanding the regulatory network and rewiring to respond to these infections.

Recent efforts have been made for the functional and molecular characterization of *C. albicans* genes using RNA sequencing (Wu et al., 2016; Brown et al., 2019; Romo et al., 2019; Zhang et al., 2019; de Vries et al., 2020; Thomas et al., 2020; Xu et al., 2020). Numerous studies suggest gene biomarkers as potential therapeutic targets and diagnostic markers in various fungal infections (Dix et al., 2015; Huppler et al., 2017; Díez et al., 2021; Hamam et al., 2021). Weighted gene correlation network analysis (WGCNA) has been widely used in disease diagnosis (Liu et al., 2017; Liang et al., 2018; Tang et al., 2018; Li et al., 2019; Yin et al., 2019), physiology (Kadarmideen et al., 2011; Zuo et al., 2018; Chen et al., 2019), drug targets (Puniya et al., 2013; Maertens et al., 2018), and cross-species (Mueller et al., 2017) but has never been applied in the context of candida pathogenesis (Thomas et al., 2020). Therefore, we developed a novel approach to identify host-pathogen interactions in *C. albicans* and humans.

In this work, we applied WGCNA to analyze 32 RNA-seq samples from *in vitro* infection of *C. albicans* on human endothelial and oral epithelial cells after 1.5, 5, and 8-h of infection and controls. We identified 5 modules in human endothelial cells (HUVEC) human cell lines in infection vs. control status and separately identified differentially expressed genes (DEG). We reported the common genes across the two methods (WGCNA and DEG). We then validated a subset of genes using differential gene expression analysis of candida-infected human cell lines OKF6. Finally, we performed protein-protein interaction network analysis and identified hub genes that could be novel targets to investigate *C. albicans* infection in humans. Through these central genes' biological and molecular functions, we gained insights into the signaling pathways previously not correlated with the fungal pathogen-host response and other diseases.

Materials and methods

Data collection

All processed gene expression datasets were collected from publicly available NCBI Gene Expression Omnibus GSE56093 (Liu et al., 2015a). The raw sequence data was aligned to the human and candida reference genomes separately by Liu et al. (Liu et al., 2015a), and the resultant count matrices were utilized for the WGCNA and DEG analyses. This dataset was comprised of 88 samples from *in vitro* and *in vivo* experiments. Of those, we only utilized 32 *in vitro* samples of human cell lines (endothelial and epithelial) infected with *C. albicans* (SC5314 and WO1 strains) and their controls at three different time points. More information is given in Supplementary Table S1. The overall methodology steps are shown in Figure 1.

Data normalization and transformation

We performed normalization on the RPKM (reads per kilobase of transcript, per million mapped reads) values using the GCRMA limma package (Gautier et al., 2004) by first removing features with counts <10 in 90% of the samples, as these could be a potential cause of the noise. Then, we performed and compared three data transformation techniques, logarithmic, regularized logarithmic, and variance stabilizing transformation (Lin et al., 2008), to stabilize the variance across sample mean values. We chose the regularized log transformation due to stability (Supplementary Figure S1).

Weighted gene coexpression network framework

We constructed the weighted gene coexpression network using the R WGCNA package (Langfelder and Horvath, 2008). The normalized data were used as input for network construction and gene module detection. It uses correlation to find functional modules of the highly correlated gene networks. First, we evaluated the soft threshold power (β)

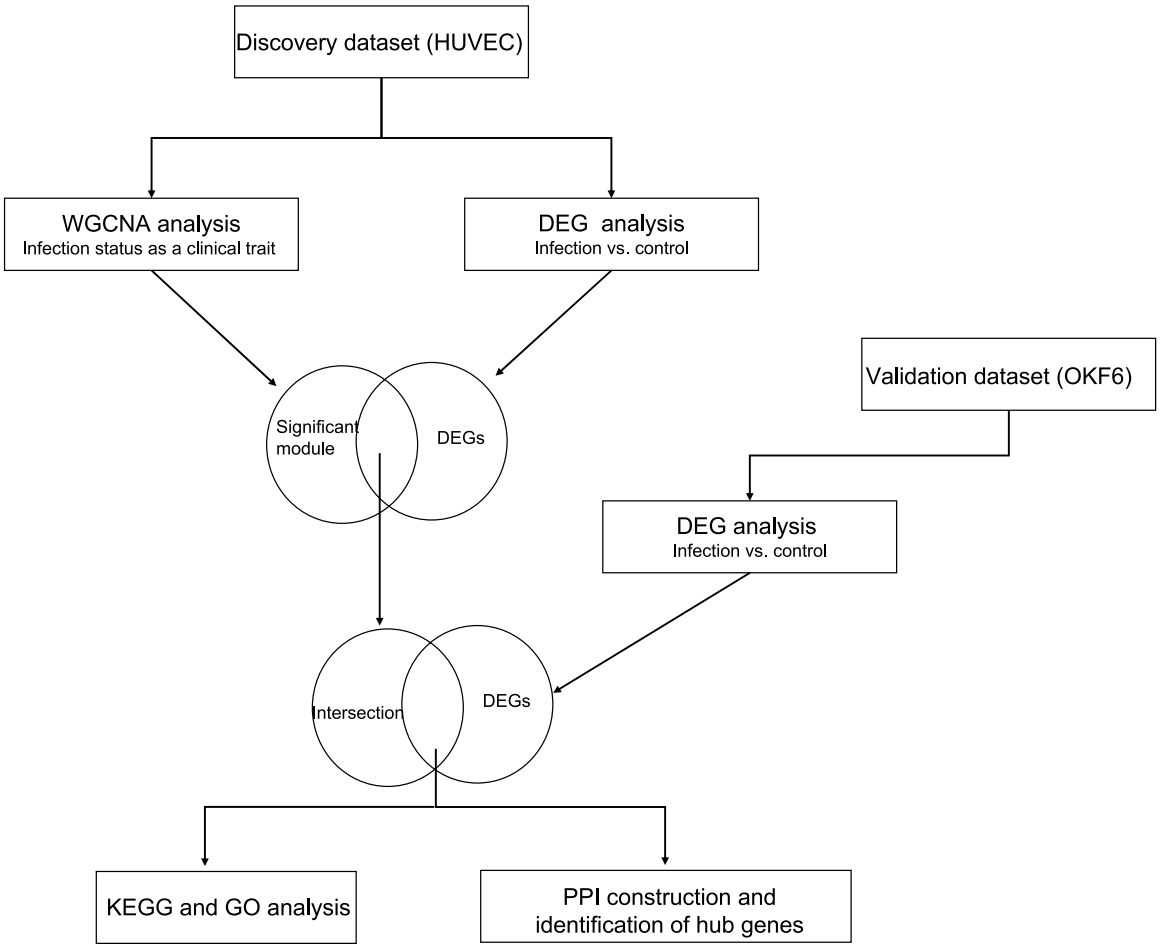


FIGURE 1
The schematic representation of the overall methodology: The discovery dataset was analyzed using two independent methods (WGCNA and DESeq2). Their intersecting genes were overlapped with the DEG list from the independent validation set to build the PPI network and identify the hub genes.

to convert coexpression into weight with a scale-free topology index of 0.9. We chose soft threshold powers of 8 to calculate the correlations between the adjacent genes (Supplementary Figure S2). Pearson correlations between each gene pair were calculated. We then converted this adjacency matrix into a topological overlap matrix (TOM) to define gene clusters that show the amount of overlap in shared neighbors of the gene network. The dissimilarity measure was determined for hierarchical clustering and module detection. Modules of clusters of genes with high topological overlap were selected using a dynamic tree-cut algorithm. Several modules were identified, and the modules with similar expression levels were merged by calculating their eigengenes corresponding to their correlations. We further determined the association of these modules with the external traits. We identified the genes with high gene significance (GS) and module membership (MM) in the turquoise and blue modules in HUVEC data. Last, intramodular connectivity was analyzed in human modules using MTR>0.35 and p-value < 0.05. All the categorical variables were binarized for the analyses.

Identification of differentially expressed genes

Differentially expressed genes were identified using DESeq2 R Bioconductor package (Love et al., 2014). We used raw counts that were fed to the DESeq2 since it corrects for library size. The variance stabilizing transformations (VST) function estimated the sample differences (Lin et al., 2008). The statistical significance for the differentially expressed genes was set to q-value < 0.05 and log2 fold change (log2FC) > 1.

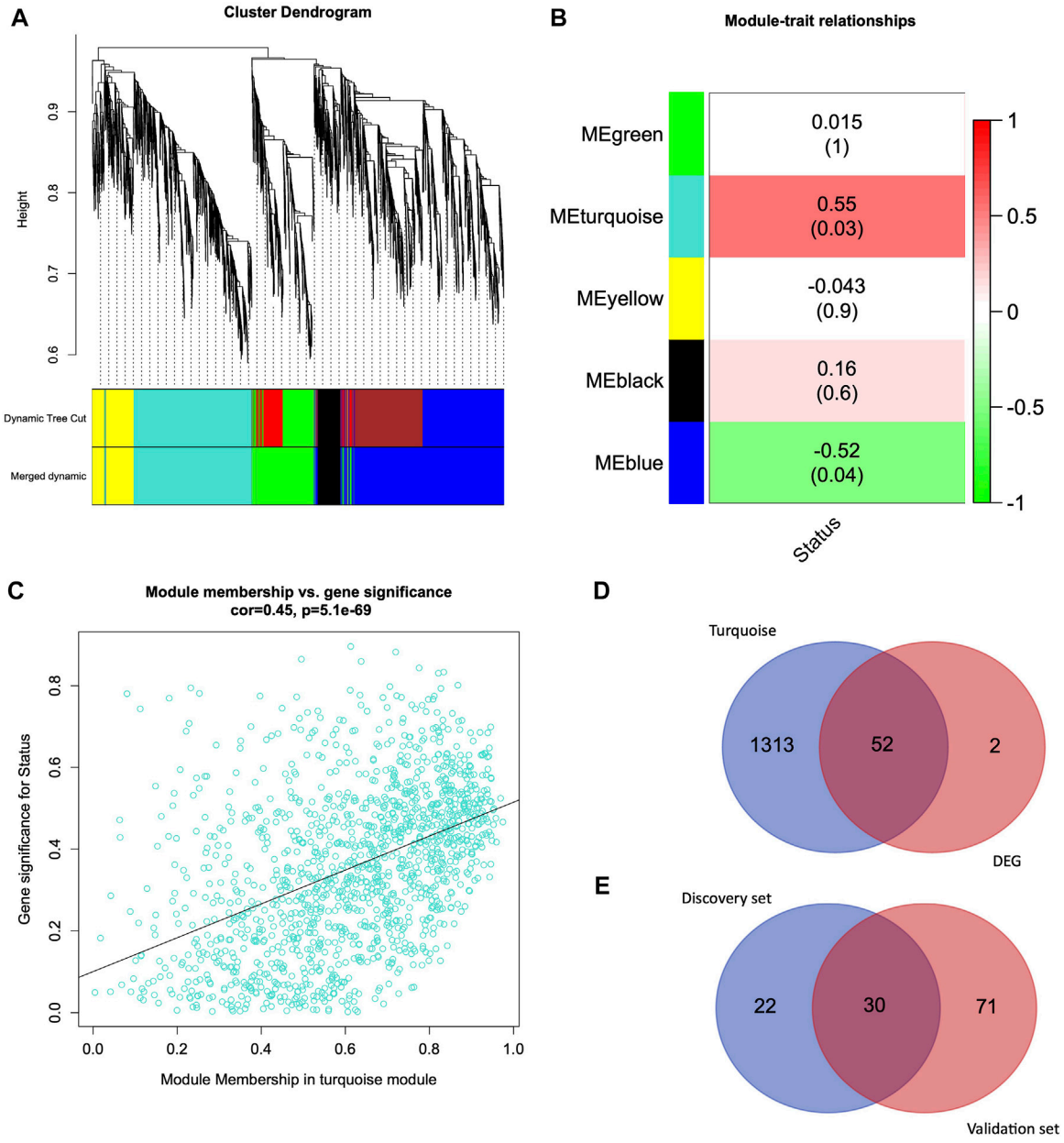


FIGURE 2
Weighted Gene Coexpression Network Analysis and Venn Diagram **(A)** Hierarchical clustering of 4,669 genes from HUVEC discovery dataset **(B)** Module-trait relationship exhibiting associations of module eigengenes with the clinical trait (infection status). **(C)** Relationship between turquoise module membership (MM) and gene significance (GS). **(D)** Venn diagram representing the overlapping genes from the turquoise module genes and differentially expressed genes. **(E)** Venn diagram representing the common genes between the discovery set genes (overlapping genes from WGCNA turquoise module and DEG genes) and differentially expressed genes from the validation set.

Functional enrichment analysis of genes

We performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses to study the role of the genes and identify their biological functions and pathways. Gene Ontology analysis was performed to determine the biological process. We considered an adjusted *p*-value threshold of ≤ 0.05 and a minimum gene count of 3 for the KEGG pathways and GO functional terms. As the contribution of all the genes is not the same, we identified hub genes and further investigated their function.

Statistical analysis and data visualization

The R programming language (Horgan, 2012) was used to normalize the RNA-seq data. We conducted Fisher's exact tests to identify the statistically significant Gene Ontology terms and functional classes. Enrichment analysis based on a hypergeometric test was implemented, and Benjamini Hochberg multiple testing was used to correct the *p*-value. Data visualization to show differentially expressed genes between infected and uninfected groups for top selected genes was plotted using the complex Heatmap function in R. The data visualization was performed using the cluster profiler package in R (Yu et al., 2012).

Protein-protein interaction network analysis

The validated genes are uploaded into the STRING database, and high confidence interaction score ≥ 0.7 was used to reduce false-positive interactions (Bozhilova et al., 2019). The resultant network output was loaded into Cytoscape. CytoHubba (Chin et al., 2014) was used with the Maximal Clique Centrality (MCC) algorithm to discover the hub genes in the PPI network (Li and Xu, 2019).

Results

Network construction and module identification

Weighted Gene Correlation Network Analysis was conducted on HUVEC data. We performed hierarchical clustering of genes using a topological overlap matrix and merged modules with similar expression profiles (Figure 2A). Each leaf corresponds to a gene, and branches correspond to the cluster of highly coexpressed genes. After cutting tree branches, we identified five different modules, turquoise, yellow, black, blue, and green, with 1,365, 459, 261, 1829, and 755 genes in HUVEC (Supplementary Table S2). A total of 4,669 genes were identified from the HUVEC data set, and in each module, the number of genes ranged between 261 and 1829.

Module association with external traits

We further analyzed the module trait relationship (MTR) between the module eigengene and clinical traits, where each cell represents the correlation strength (red is positively correlated, and green is negatively correlated) with their corresponding *p*-value (Figure 2B). We demonstrate that some module eigengenes are highly correlated with infection (status traits). We observed that the turquoise ($r = 0.55$, $p = 0.03$) and blue ($r = -0.52$, $p = 0.04$) modules were highly correlated with the infection status in HUVEC cells. Since the turquoise module, with 1,365 genes, is the most significantly correlated with the clinical trait, we focused on this module for further analysis.

Intramodular connectivity using gene significance and module membership

We quantified genes with high significance for the trait status of HUVEC and high module membership by comparing their similarities in every module. There was a highly significant correlation between gene significance and module membership in the turquoise module. Figure 2C represents the correlation between turquoise module membership and gene significance ($r = 0.45$, $p = 5.1e\text{-}69$).

Differentially expressed genes and intersection with WGCNA

We used DESeq2 as a second independent method on the entire HUVEC dataset to identify 54 genes that were differentially expressed in the HUVEC (infection vs. control) samples (q-value < 0.05 and log2FC > 1). From the WGCNA analysis, we identified 1,365 genes in the most significantly correlated turquoise module (Figure 2B). When we further investigated the intersection of WGCNA and DEGs, 52 genes were common between the turquoise module and the DEG list (Figure 2D). The list of turquoise module genes, DEGs, and intersecting genes is given in Supplementary Table S4.

Validation of candidate genes

In order to validate these 52 common genes, we utilized a validation dataset comprised of candida-infected human oral keratinocytes (OKF6 cell line). We performed the differential gene expression analysis on infection vs. control and identified 101 DEGs. When we overlapped these 101 genes with 52 common genes from the discovery dataset, we found 30 genes that were differentially expressed in the OKF6 validation dataset (Figure 2E and Supplementary Table S3). The following are the 30 validated genes: *SLC2A1, ATF3, JUN, KDM7A, DUSP1, PTGS2, NAB2, PIM1, MAFF, ADM, PFKFB3, KLF6, BNIP3, CSRNP1, VEGFA, ENO2, ANKRD37, PPP1R15A, KDM3A, ANGPTL4, BHLHE40, ARRDC3, SLC2A3, KLF7, DDIT4, ERRFI1, KLF4, FOSL2, EFNA1, and HK2.* The list of genes from the discovery set, validation set, and intersecting genes are given in Supplementary Table S4.

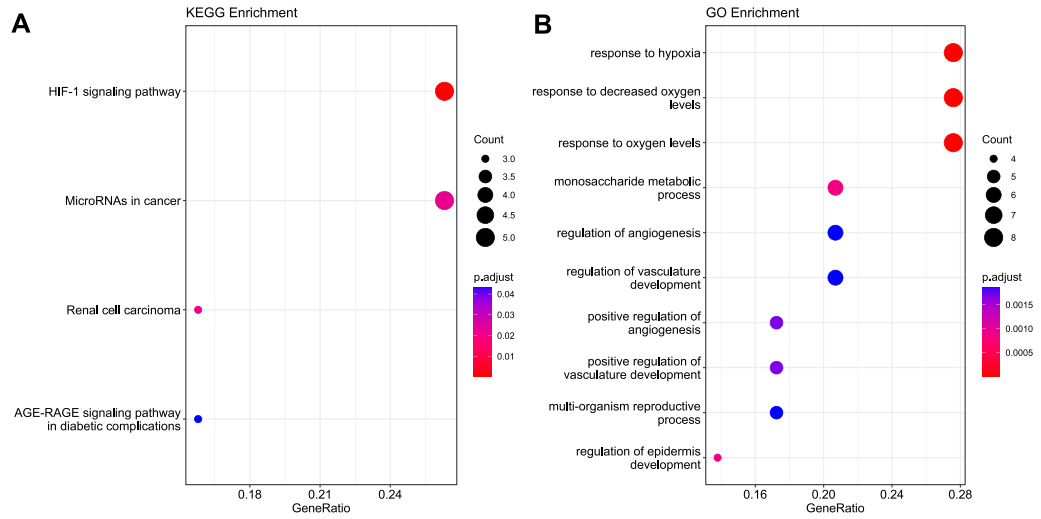


FIGURE 3
Functional enrichment analysis for the 30 validated genes. **(A)** KEGG pathway **(B)** Gene Ontology.

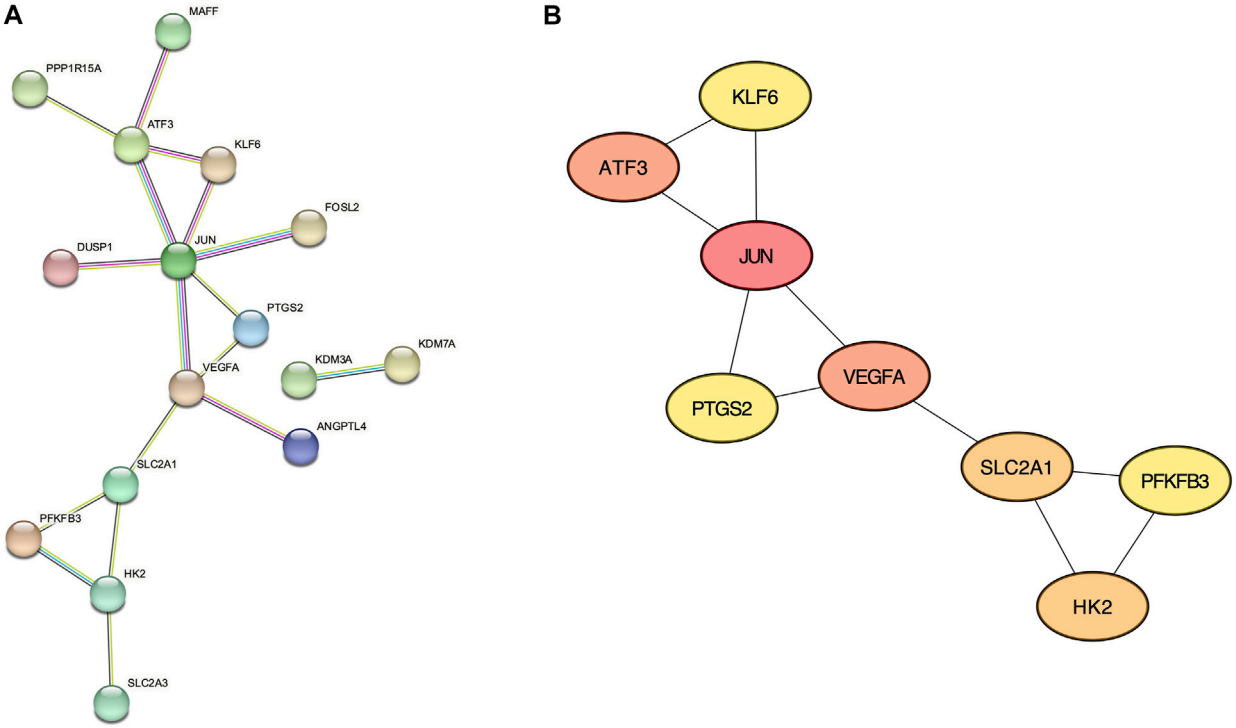


FIGURE 4
Protein-Protein interaction networks. **(A)** STRING analysis of validated genes. **(B)** CytoHubba with Maximal Clique Centrality analysis showing 8 hub genes.

Integrating network analysis with functional enrichment analyses

To understand the biological roles of these 30 validated genes, we performed GO and KEGG pathway analyses to identify the biological pathways that were significantly enriched (FDR ≤ 0.05) in these modules.

KEGG analyses revealed that the genes were highly enriched in the HIF-1 signaling pathway, microRNAs in cancer, renal cell carcinoma, and AGE-RAGE signaling pathway in diabetes complications, as shown in Figure 3A (Detailed information is provided in Supplementary Table S3). Gene Ontology analyses elucidated that these genes were enriched in response to hypoxia, monosaccharide metabolic process, angiogenesis regulation, vasculature development, reproductive process, and epidermis development, as shown in Figure 3B. Additional details are given in Supplementary Table S3.

Protein-protein interaction network analysis

From the 30 validated genes, we first performed the protein-protein interaction (PPI) analysis using the STRING database (Figure 4A). The resultant data is then imported to the Cytoscape plugin CytoHubba, and the top 8 genes with the highest Maximal Clique Centrality (MCC) score were considered hub genes: *JUN, ATF3, VEGFA, SLC2A1, HK2, PTGS2, PFKFB3, and KLF6* (Figure 4B and Supplementary Table S3).

Discussion

The interaction between host cells and Candida is central to the immunopathology of candidiasis in transplant patients; a comprehensive understanding of this synergy will identify new treatment strategies. Here, we investigate how human epithelial and endothelial cells communicate with different Candida species during infection. In this study, we constructed a weighted gene correlation network and performed differential gene expression analysis to identify genes that are important in host–candida interactions.

Comparative network analysis could rank genes for further investigation of their connectivity (Schadt et al., 2005). A distinct advantage of WGCNA is that it considers modules or gene clusters for constructing interactions, and the genes in a module are likely to be connected by the same regulatory pathways. Therefore, in this study, we aim to discover novel genes and molecular pathways in human-candida infection and to understand the regulation due to cell dynamics using the WGCNA and DESeq2 algorithms. Network depictions provided immediate insight into the relationships between the correlated modules. The construction of a gene coexpression network and differential gene expression analysis of the discovery and validation data set facilitated the identification of genes with similar biological functions by GO and KEGG analyses.

According to the results of functional enrichment analysis, the top 3 GO terms and topmost KEGG pathway were a response to hypoxia, response to decreased oxygen, response to oxygen levels (Figure 3B), and hypoxia-inducible factor 1 (HIF-1) signaling pathway (Figure 3A). HIF-1 is a transcription factor that functions as a master regulator of oxygen homeostasis. It has been shown that suppressing HIF-1 helps treat cancer and ischemia (Ziello et al., 2007). All organs during the process of transplantation undergo hypoxic and ischemic injury. Low oxygen levels trigger the colonization of candida infection in the human host, resulting in complications like allograft rejection in SOT patients (Akhtar et al., 2014). We identified eight hub genes using PPI network analysis. Four hub genes (*HK2, PFKFB3, SLC2A1, and VEGFA*) are involved in the HIF-1 signaling pathway. The hexokinase isoenzyme (*HK2*) elevates innate immunity in hepatocellular carcinoma (Perrin-Cocon et al., 2021). *HK2* and *PFKFB3* are involved in glycolysis which affects the immune response against fungal infection (Perrin-Cocon et al., 2021); specifically, after transplantation, the *PFKFB3* gene increase the risk of invasive pulmonary aspergillosis (Gonçalves et al., 2021). *Huang et al.* showed in their omics analysis that *SLC2A1* is involved in ischemic reperfusion injury in liver transplant patients and forms the core gene network (Huang et al., 2019a). Vascular Endothelial Growth Factor A (*VEGFA*) is associated with an increased risk of chronic kidney disease (Anderson et al., 2018) but induces vasculogenesis. Kidney vasculature comprises vascular smooth muscle and endothelial cells (Udan et al., 2012). One of the most challenging components to handle during a kidney transplant is through vasculogenesis and angiogenesis processes (Munro and Davies, 2018; Lebedenko and Banerjee, 2021). HIF-1 stimulates the *VEGFA* to maintain oxygen delivery and protect the kidney (Hunga et al., 2013).

The other top enriched KEGG pathways in our analysis were microRNAs in cancer (hsa05211) and renal cell carcinoma (hsa05206). MicroRNAs play a diverse role in cancer and infections (Yong and Dutta, 2009). Recent advances in microRNA therapeutics have shown the extensive use of microRNAs for cancer and infections (Rupaimoole and Slack, 2017). There has been increased support for microRNA therapeutics in solid organ transplantation, including kidney (Wilflingseder et al., 2015; Jin et al., 2017; Ledeganck et al., 2019), lung (Benazzo et al., 2022), and heart transplantation (Hamdorf et al., 2017). *Candida albicans* have been linked to cancerous processes by taking advantage of the compromised immune system (Ramirez-Garcia et al., 2016; Chung et al., 2017; Sultan et al., 2022). Our PPI network analysis identified four hub genes (*SLC2A1, VEGFA*, *JUN*, and *PTGS2*) enriched in the cancer-related pathways. *SLC2A1* belongs to a glucose transporter family and has been reported to be associated with

HCC (Kim et al., 2017). *SLC2A1* is also essential to IRI during liver transplantation (Huang et al., 2019b) and a diagnostic biomarker for colorectal cancer (CRC) (Liu et al., 2022). In CRC, the *SLC2A1* gene infiltrates the $CD4^+$ T cell, neutrophil, dendritic cells, and B cells (Liu et al., 2022). Candidiasis is one of the risk factors for Oral squamous cell carcinoma (OSCC). The transcriptomics data analysis revealed that *VEGFA* and *JUN* are highly regulated in OSCC invasion and metastasis (Vadovics et al., 2022). *JUN* is a member of the activator protein-1 family of oncogenic transcription factors, which is involved in various cancer-related and cell signaling pathways such as tumorigenesis, cell differentiation, and angiogenesis (Brennan et al., 2020). Post renal transplantation, the activation of c-JUN affects acute humoral rejection and acute T-cell-mediated rejection (Kobayashi et al., 2010). c-JUN is also associated with reduced graft function and plays an important role in renal pathophysiological events (Kobayashi et al., 2010). Prostaglandin E2 (PGE2) is an inflammatory mediator produced by the Prostaglandin-endoperoxide synthase (*PTGS2*) enzyme, and PGE2 promotes candida morphogenesis. In response to candida infection, PTGS2 activation promotes NF-kB and MAPK signaling pathways (Deva et al., 2003). In OSCC, *PTGS2* involves an inflammatory response to infection by promoting tumorigenesis (Cacina et al., 2018) and activating transcription factor 3 (*ATF3*), one of the 8 hub genes that regulate the *PTGS2* during acute inflammation (Hellmann et al., 2015) and helps in the homeostasis of the metabolism and immune system (Sha et al., 2017). Zhu et al. also showed that *ATF3* is one of the top hub genes in samples infected with 4 different candida species (Zhu et al., 2022). Using bioinformatics omics analysis, *ATF3* and Kruppel-like factor 6 (*KLF6,* hub gene) are shown to be the central players in ischemic reperfusion injury in liver transplant patients (Huang et al., 2019b). *KLF6* promotes inflammation and oxidative stress by regulating HIF-1 expression in macrophages (Kim et al., 2020).

Another enriched KEGG pathway was the AGE-RAGE signaling in diabetes complications (hsa04933). Endoplasmic reticulum stress due to AGE-RAGE plays an essential role in renal inflammation, diabetic nephropathy (Pathomthongtaweechai and Chutipongtanate, 2020) and early-stage renal disease (Meerwaldt et al., 2009). Advanced glycation end products (AGEs) may also play a role in the hardening of arteries after renal transplantation (Liu et al., 2015b). Our two hub genes, *JUN* and *VEGFA,* showed enrichment in the AGE-RAGE signaling pathway in diabetes complications. Poorly controlled diabetes increases the risk of fungal infections (Rodrigues et al., 2019). Some diabetes-related complications include cardiovascular disease, kidney disease, neuropathy, hearing loss, vision loss, Alzheimer's, liver disease, etc. (Deshpande et al., 2008; Prasad et al., 2016). *VEGFA* and *JUN* were identified as the central players in diabetic nephropathy (Oltean et al., 2015; Wang et al., 2021) and Alzheimer's disease (Zu et al., 2021) whereas, *VEGFA* was associated with diabetic retinopathy (Bucolo et al., 2021), cardiac autonomic neuropathy (Ravichandran et al., 2019), and non-alcoholic fatty liver disease-hepatocellular carcinoma (Shen et al., 2022). Each hub gene plays a vital and diverse role in the pathways and biological processes. Therefore, more research is warranted on the divergent roles of these genes' signaling and regulatory mechanisms during infection, cancer, and transplantation.

Limitations

WGCNA lacks resolution as it decomposes a group of genes into a single eigenvalue that may not correctly represent a single gene's expression profile or pathway changes. Further analysis may be needed to detect changes in the expression of individual processes. Another limitation of the study is the small sample size; therefore, we present this study as a proof of concept to be validated in a larger cohort. The current study used cell lines from epithelial and endothelial cells; thus, the identified gene markers should be validated from the peripheral blood transcriptome of candidiasis patients for non-invasive clinical relevance.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56093.

Author contributions

AM conceived the study design. SN developed the methodology and performed data analysis, and wrote the manuscript. SN and AM revised and edited the manuscript.

Acknowledgments

We would like to acknowledge Liu Yaoping for providing the data. We would also like to thank the office of scientific writing at the University of Tennessee Health Science Center for copy editing and proofreading the manuscript. We also thank the reviewers for their constructive feedback on improving the manuscript, including the experimental design and analysis suggestions. The publication costs for this article have been covered by the Center for Biomedical Informatics at the University of Tennessee Health Science Center.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.917636/full#supplementary-material

References

Akhtar, M. Z., Sutherland, A. I., Huang, H., Ploeg, R. J., and Pugh, C. W. (2014). The role of hypoxia-inducible factors in organ donation and transplantation: The current perspective and future opportunities. *Am. J. Transpl.* 14, 1481–1487. doi:10.1111/ajt.12737

Anderson, C. E., Hamm, L. L., Batuman, G., Kumbala, D. R., Chen, C. S., Kallu, S. G., et al. (2018). The association of angiogenic factors and chronic kidney disease. *BMC Nephrol.* 19, 117–118. doi:10.1186/s12882-018-0909-2

Arendrup, M. C., and Patterson, T. F. (2017). Multidrug-resistant candida: Epidemiology, molecular mechanisms, and treatment. *J. Infect. Dis.* 216, S445–S451. doi:10.1093/infdis/jix131

Barker, K. S., Park, H., Phan, Q. T., Xu, L., Homayouni, R., Rogers, P. D., et al. (2008). Transcriptome profile of the vascular endothelial cell response to Candida albicans. *J. Infect. Dis.* 198, 193–202. doi:10.1086/589516

Benazzo, A., Bozzini, S., Auner, S., Berezhinskiy, H. O., Watzenboeck, M. L., Schwarz, S., et al. (2022). Differential expression of circulating miRNAs after alemtuzumab induction therapy in lung transplantation. *Sci. Rep.* 121 12, 7072–7113. doi:10.1038/s41598-022-10866-w

Bozhilova, L. V., Whitmore, A. V., Wray, J., Reinert, G., and Deane, C. M. (2019). Measuring rank robustness in scored protein interaction networks. *BMC Bioinforma.* 20, 446. doi:10.1186/s12859-019-3036-6

Brennan, A., Leech, J. T., Kad, N. M., and Mason, J. M. (2020). Selective antagonism of cJun for cancer therapy. *J. Exp. Clin. Cancer Res.* 39, 184–216. doi:10.1186/s13046-020-01686-9

Brown, S. E., Schwartz, J. A., Robinson, C. K., O'Hanlon, D. E., Bradford, L. L., He, X., et al. (2019). The vaginal microbiota and behavioral factors associated with genital Candida albicans detection in reproductive-age women. *Sex. Transm. Dis.* 46, 753–758. doi:10.1097/OLQ.0000000000001066

Bucolo, C., Barbieri, A., Vigano, I., Marchesi, N., Bandello, F., Drago, F., et al. (2021). Short-and long-term expression of vegf: A temporal regulation of a key factor in diabetic retinopathy. *Front. Pharmacol.* 12, 2164. doi:10.3389/fphar.2021.707909

Cacina, C., Kaşarci, G., Bektaş, K., Unur, M., and Cakmakoglu, B. (2018). The COX2 genetic variants in oral squamous cell carcinoma in Turkish population. *Cell. Mol. Biol.* 64, 96–100. doi:10.14715/cmb/2018.64.14.16

Chen, R., Ge, T., Jiang, W., Huo, J., Chang, Q., Geng, J., et al. (2019). Identification of biomarkers correlated with hypertrophic cardiomyopathy with co-expression analysis. *J. Cell. Physiol.* 234, 21999–22008. doi:10.1002/jcp.28762

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8, S11–S17. doi:10.1186/1752-0509-8-S4-S11

Chung, L. M., Liang, J. A., Lin, C. L., Sun, L. M., and Kao, C. H. (2017). Cancer risk in patients with candidiasis: A nationwide population-based cohort study. *Oncotarget* 8, 63562–63573. doi:10.18632/oncotarget.18855

Cowen, L. E., Sanglard, D., Howard, S. J., Rogers, P. D., and Perlin, D. S. (2015). Mechanisms of antifungal drug resistance. *Cold Spring Harb. Perspect. Med.* 5, a019752. doi:10.1101/cshperspect.a019752

de Vries, D. H., Matzaraki, V., Bakker, O. B., Brugge, H., Westra, H. J., Netea, M. G., et al. (2020). Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response. *PLoS Pathog.* 16, e1008408. doi:10.1371/journal.ppat.1008408

Deshpande, A. D., Harris-Hayes, M., and Schootman, M. (2008). Epidemiology of diabetes and diabetes-related complications. *Phys. Ther.* 88, 1254–1264. doi:10.2522/ptj.20080020

Deva, R., Shankaranarayanan, P., Ciccoli, R., and Nigam, S. (2003). Candida albicans induces selectively transcriptional activation of cyclooxygenase-2 in HeLa cells: Pivotal roles of toll-like receptors, p38 mitogen-activated protein kinase, and NF-κB. *J. Immunol.* 171, 3047–3055. doi:10.4049/jimmunol.171.6.3047

Díez, A., Carrano, G., Bregon-Villahoz, M., Cuetara, M. S., Garcia-Ruiz, J. C., Fernandez-de-Larrinoa, I., et al. (2021). Biomarkers for the diagnosis of invasive candidiasis in immunocompetent and immunocompromised patients. *Diagn. Microbiol. Infect. Dis.* 101, 115509. doi:10.1016/j.diagmicrobio.2021.115509

Dix, A., Hunniger, K., Weber, M., Guthke, R., Kurzai, O., and Linde, J. (2015). Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. *Front. Microbiol.* 6, 171. doi:10.3389/fmicb.2015.00171

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy - analysis of affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi:10.1093/bioinformatics/btg405

Gonçalves, S. M., Antunes, D., Leite, L., Mercier, T., Horst, R. T., Vieira, J., et al. (2021). Genetic variation in PFKFB3 impairs antifungal immunometabolic responses and predisposes to invasive pulmonary aspergillosis. *MBio* 12, e0036921. doi:10.1128/mBio.00369-21

Hamam, J., Navellou, J. C., Bellanger, A. P., Bretagne, S., Winiszewski, H., Scherer, E., et al. (2021). New clinical algorithm including fungal biomarkers to better diagnose probable invasive pulmonary aspergillosis in ICU. *Ann. Intensive Care* 11, 41–49. doi:10.1186/s13613-021-00827-3

Hamdorf, M., Kawakita, S., and Everly, M. (2017). The potential of MicroRNAs as novel biomarkers for transplant rejection. *J. Immunol. Res.* 2017, 4072364. doi:10.1155/2017/4072364

Hellmann, J., Tang, Y., Zhang, M. J., Hai, T., Bhatnagar, A., Srivastava, S., et al. (2015). Atf3 negatively regulates Ptgs2/Cox2 expression during acute inflammation. *Prostagl. Other Lipid Mediat.* 0, 49–56. doi:10.1016/j.prostaglandins.2015.01.001

Horgan, J. M. (2012). Programming in R. *WIREs. Comp. Stat.* 4, 75–84. doi:10.1002/wics.183

Huang, S., Ju, W., Zhu, Z., Han, M., Sun, C., Tang, Y., et al. (2019). Comprehensive and combined omics analysis reveals factors of ischemia-reperfusion injury in liver transplantation. *Epigenomics* 11, 527–542. doi:10.2217/epi-2018-0189

Hunga, T. W., Liou, J. H., Yeh, K. T., Tsai, J. P., Wu, S. W., Tai, H. C., et al. (2013). Renal expression of hypoxia inducible factor-1α in patients with chronic kidney disease: A clinicopathologic study from nephrectomized kidneys. *Indian J. Med. Res.* 137, 102–110.

Huppler, A. R., Fisher, B. T., Lehrnbecher, T., Walsh, T. J., and Steinbach, W. J. (2017). Role of molecular biomarkers in the diagnosis of invasive fungal diseases in children. *J. Pediatr. Infect. Dis. Soc.* 6, S32–S44. doi:10.1093/jpids/pix054

Jin, P., Chen, H., Xie, J., Zhou, C., and Zhu, X. (2017). Essential role of microRNA-650 in the regulation of B-cell CLL/lymphoma 11B gene expression following transplantation: A novel mechanism behind the acute rejection of renal allografts. *Int. J. Mol. Med.* 40, 1840–1850. doi:10.3892/ijmm.2017.3194

Kadarmideen, H. N., Watson-Haigh, N. S., and Andronicos, N. M. (2011). Systems biology of ovine intestinal parasite resistance: Disease gene modules and biomarkers. *Mol. Biosyst.* 7, 235–246. doi:10.1039/c0mb00190b

Kim, G. D., Ng, H. P., Chan, E. R., and Mahabeleshwar, G. H. (2020). Kruppel-like factor 6 promotes macrophage inflammatory and hypoxia response. *FASEB J.* 34, 3209–3223. doi:10.1096/fj.201902221R

Kim, Y. H., Jeong, D. C., Pak, K., Han, M. E., Kim, J. Y., Liangwen, L., et al. (2017). SLC2A2 (GLUT2) as a novel prognostic factor for hepatocellular carcinoma. *Oncotarget* 8, 68381–68392. doi:10.18632/oncotarget.20266

Klepser, M. E. (2006). Candida resistance and its clinical relevance. *Pharmacotherapy* 26, 68S–75S. doi:10.1592/phco.26.6part2.68S

Kobayashi, A., Takahashi, T., Horita, S., Yamamoto, I., Yamamoto, H., Teraoka, S., et al. (2010). Activation of the transcription factor c-Jun in acute cellular and antibody-mediated rejection after kidney transplantation. *Hum. Pathol.* 41, 1682–1693. doi:10.1016/j.humpath.2010.04.016

Langfelder, P., and Horvath, S. W. G. C. N. A. (2008). Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Lebedenko, C. G., and Banerjee, I. A. (2021). Enhancing kidney vasculature in tissue engineering—current trends and approaches: A review. *Biomimetics* 6, 40. doi:10.3390/biomimetics6020040

Ledeganck, K. J., Gielis, E. M., Abramowicz, D., Stenvinkel, P., Shiels, P. G., and Van Craenenbroeck, A. H. (2019). MicroRNAs in AKI and kidney transplantation. *Clin. J. Am. Soc. Nephrol.* 14, 454–468. doi:10.2215/CJN.08020718

Li, B., Pu, K., and Wu, X. (2019). Identifying novel biomarkers in hepatocellular carcinoma by weighted gene co-expression network analysis. *J. Cell. Biochem.* 120, 11418–11431. doi:10.1002/jcb.28420

Li, C., and Xu (2019). Feature selection with the Fisher score followed by the Maximal Clique Centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma. *Sci. Rep.* 91 9, 17283–17311. doi:10.1038/s41598-019-53471-0

Liang, J. W., Fang, Z. Y., Huang, Y., Liuyang, Z. Y., Zhang, X. L., Wang, J. L., et al. (2018). Application of weighted gene Co-expression network analysis to explore the key genes in alzheimer's disease. *J. Alzheimer's Dis.* 65, 1353–1364. doi:10.3233/JAD-180400

Lin, S. M., Du, P., Huber, W., and Kibbe, W. A. (2008). Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* 36, e11. doi:10.1093/nar/gkm1075

Liu, X., Hu, A. X., Zhao, J. L., and Chen, F. L. (2017). Identification of key gene modules in human osteosarcoma by Co-expression analysis weighted gene Co-expression network analysis (WGCNA). *J. Cell. Biochem.* 118, 3953–3959. doi:10.1002/jcb.26050

Liu, X., Liu, K., Wang, Z., Liu, C., Han, Z., Tao, J., et al. (2015). Advanced glycation end products accelerate arteriosclerosis after renal transplantation through the AGE/RAGE/ILK pathway. *Exp. Mol. Pathol.* 99, 312–319. doi:10.1016/j.yexmp.2015.07.009

Liu, X. S., Yang, J. W., Zeng, J., Chen, X. Q., Gao, Y., Kui, X. Y., et al. (2022). SLC2A1 is a diagnostic biomarker involved in immune infiltration of colorectal cancer and associated with m6A modification and ceRNA. *Front. Cell. Dev. Biol.* 10, 853596. doi:10.3389/fcell.2022.853596

Liu, Y., Shetty, A. C., Schwartz, J. A., Bradford, L. L., Xu, W., Phan, Q. T., et al. (2015). New signaling pathways govern the host response to C. albicans infection in various niches. *Genome Res.* 125, 679–689. doi:10.1101/gr.187427.114

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, 550–621. doi:10.1186/s13059-014-0550-8

Maertens, A., Tran, V., Kleensang, A., and Hartung, T. (2018). Weighted gene correlation network analysis (WGCNA) reveals novel transcription factors associated with bisphenol A dose-response. *Front. Genet.* 9, 508. doi:10.3389/fgene.2018.00508

Meerwaldt, R., Zeebregts, C. J., Navis, G., Hillebrands, J. L., Lefrandt, J. D., and Smit, A. J. (2009). Accumulation of advanced glycation end products and chronic complications in ESRD treated by dialysis. *Am. J. Kidney Dis.* 53, 138–150. doi:10.1053/j.ajkd.2008.08.031

Mueller, A. J., Canty-Laird, E. G., Clegg, P. D., and Tew, S. R. (2017). Cross-species gene modules emerge from a systems biology approach to osteoarthritis. *npj Syst. Biol. Appl.* 3, 13–14. doi:10.1038/s41540-017-0014-3

Munro, D. A. D., and Davies, J. A. (2018). Vascularizing the kidney in the embryo and organoid: Questioning assumptions about renal vasculogenesis. *J. Am. Soc. Nephrol.* 29, 1593–1595. doi:10.1681/ASN.2018020179

Nishimoto, A. T., Sharma, C., and Rogers, P. D. (2020). Molecular and genetic basis of azole antifungal resistance in the opportunistic pathogenic fungus candida albicans. *J. Antimicrob. Chemother.* 75, 257–270. doi:10.1093/jac/dkz400

Oltean, S., Qiu, Y., Ferguson, J. K., Stevens, M., Neal, C., Russell, A., et al. (2015). Vascular endothelial growth factor-A165b is protective and restores endothelial glycocalyx in diabetic nephropathy. *J. Am. Soc. Nephrol.* 26, 1889–1904. doi:10.1681/ASN.2014040350

Pathomthongtaweechai, N., and Chutipongtanate, S. (2020). AGE/RAGE signaling-mediated endoplasmic reticulum stress and future prospects in non-coding RNA therapeutics for diabetic nephropathy. *Biomed. Pharmacother.* 131, 110655. doi:10.1016/j.biopha.2020.110655

Pendleton, K. M., Huffnagle, G. B., and Dickson, R. P. (2017). The significance of Candida in the human respiratory tract: Our evolving understanding. *Pathog. Dis.* 75, 29. doi:10.1093/femspd/ftx029

Perrin-Cocon, L., Vidalain, P. O., Jacquemin, C., Aublin-Gex, A., Olmstead, K., Panthu, B., et al. (2021). A hexokinase isoenzyme switch in human liver cancer cells promotes lipogenesis and enhances innate immunity. *Commun. Biol.* 41 4, 217–315. doi:10.1038/s42003-021-01749-3

Prasad, K., Dhar, I., Zhou, Q., Elmoselhi, H., Shoker, M., and Shoker, A. (2016). AGEs/sRAGE, a novel risk factor in the pathogenesis of end-stage renal disease. *Mol. Cell. Biochem.* 423, 105–114. doi:10.1007/s11010-016-2829-4

Puniya, B. L., Kulshreshtha, D., Verma, S. P., Kumar, S., and Ramachandran, S. (2013). Integrated gene co-expression network analysis in the growth phase of *Mycobacterium tuberculosis* reveals new potential drug targets. *Mol. Biosyst.* 9, 2798–2815. doi:10.1039/c3mb70278b

Ramirez-Garcia, A., Rementeria, A., Aguirre-Urizar, J. M., Moragues, M. D., Antoran, A., Pellon, A., et al. (2016). Candida albicans and cancer: Can this yeast induce cancer development or progression? *Crit. Rev. Microbiol.* 42, 181–193. doi:10.3109/1040841X.2014.913004

Ravichandran, S., Srivastav, S., Kamble, P. H., Chambial, S., Shukla, R., Sharma, P., et al. (2019). VEGF-A and cardiac autonomic function in newly diagnosed type 2 diabetes mellitus: A cross-sectional study at a tertiary care center. *J. Fam. Med. Prim. Care* 8, 3185–3190. doi:10.4103/jfmpc.jfmpc_537_19

Redding, S. W., Zellars, R. C., Kirkpatrick, W. R., McAtee, R. K., Caceres, M. A., Fothergill, A. W., et al. (1999). Epidemiology of oropharyngeal Candida colonization and infection in patients receiving radiation for head and neck cancer. *J. Clin. Microbiol.* 37, 3896–3900. doi:10.1128/JCM.37.12.3896-3900.1999

Revankar, S. G., Sanche, S. E., Dib, O. P., Caceres, M., and Patterson, T. F. (1998). Effect of highly active antiretroviral therapy on recurrent oropharyngeal candidiasis in HIV-infected patients. *Aids* 12, 2511–2513.

Rhodus, N. L., Bloomquist, C., Liljemark, W., and Bereuter, J. (1997). Prevalence, density, and manifestations of oral Candida albicans in patients with Sjogren's syndrome. *J. Otolaryngol.* 26, 300–305.

Rodrigues, C. F., Rodrigues, M. E., and Henriques, M. (2019). Candida sp. infections in patients with diabetes mellitus. *J. Clin. Med.* 8, E76. doi:10.3390/jcm8010076

Romo, J. A., Zhang, H., Cai, H., Kadosh, D., Koehler, J. R., Saville, S. P., et al. (2019). Global transcriptomic analysis of the Candida albicans response to treatment with a novel inhibitor of filamentation. *mSphere* 4, 006200-e719. doi:10.1128/mSphere.00620-19

Rupaimoole, R., and Slack, F. J. (2017). MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* 163 16, 203–222. doi:10.1038/nrd.2016.246

Sangeorzan, J. A., Bradley, S. F., He, X., Zarins, L. T., Ridenour, G. L., Tiballi, R. N., et al. (1994). Epidemiology of oral candidiasis in HIV-infected patients: Colonization, infection, treatment, and emergence of fluconazole resistance. *Am. J. Med.* 97, 339–346. doi:10.1016/0002-9343(94)90300-x

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37, 710–717. doi:10.1038/ng1589

Sha, H., Zhang, D., Zhang, Y., Wen, Y., and Wang, Y. (2017). ATF3 promotes migration and M1/M2 polarization of macrophages by activating tenascin-C via Wnt/β-catenin pathway. *Mol. Med. Rep.* 16, 3641–3647. doi:10.3892/mmr.2017.6992

Shen, H., Yu, H., Li, Q. Y., Wei, Y. T., Fu, J., Dong, H., et al. (2022). Hepatocyte-derived VEGFA accelerates the progression of non-alcoholic fatty liver disease to hepatocellular carcinoma via activating hepatic stellate cells. *Acta Pharmacol. Sin.* 43, 2917–2928. doi:10.1038/S41401-022-00907-5

Shoham, S., and Marr, K. A. (2012). Invasive fungal infections in solid organ transplant recipients. *Future Microbiol.* 7, 639–655. doi:10.2217/fmb.12.28

Sobel, J. D. (1985). Epidemiology and pathogenesis of recurrent vulvovaginal candidiasis. *Am. J. Obstet. Gynecol.* 152, 924–935. doi:10.1016/s0002-9378(85)80003-x

Sultan, A. S., Theofilou, V. I., Alfaifi, A., Montelongo-Jauregui, D., and Jabra-Rizk, M. A. (2022). Is Candida albicans an opportunistic oncogenic pathogen? *PLoS Pathog.* 18, e1010413. doi:10.1371/journal.ppat.1010413

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Front. Oncol.* 8, 374. doi:10.3389/fonc.2018.00374

Thomas, G., Bain, J. M., Budge, S., Brown, A. J. P., and Ames, R. M. (2020). Identifying Candida albicans gene networks involved in pathogenicity. *Front. Genet.* 11, 375–412. doi:10.3389/fgene.2020.00375

Udan, R. S., Culver, J. C., and Dickinson, M. E. (2012). Understanding vascular development: WIRE developmental biology. *Wiley Interdiscip. Rev. Dev. Biol.* 2, 327–346. doi:10.1002/wdev.91

Vadovics, M., Ho, J., Igaz, N., Alfoldi, R., Rakk, D., Veres, E., et al. (2022). Candida albicans enhances the progression of oral squamous cell carcinoma *in vitro* and *in vivo*. *MBio* 13, e0314421. doi:10.1128/mBio.03144-21

Wang, G., Zeng, L., Huang, Q., Lu, Z., Sui, R., Liu, D., et al. (2021). Exploring the molecular mechanism of liuwei dihuang pills for treating diabetic nephropathy by combined network pharmacology and molecular docking. *Evid. Based. Complement. Altern. Med.* 2021, 7262208. doi:10.1155/2021/7262208

Wilflingseder, J., Reindl-Schwaighofer, R., Sunzenauer, J., Kainz, A., Heinzel, A., Mayer, B., et al. (2015). MicroRNAs in kidney transplantation. *Nephrol. Dial. Transpl.* 30, 910–917. doi:10.1093/ndt/gfu280

Willis, A. M., Coulter, W. A., Fulton, C. R., Hayes, J. R., Bell, P. M., and Lamey, P. J. (1999). Oral candidal carriage and infection in insulin-treated diabetic patients. *Diabet. Med.* 16, 675–679. doi:10.1046/j.1464-5491.1999.00134.x

Wu, Y., Li, Y. H., Yu, S. B., Li, W. G., Liu, X. S., Zhao, L., et al. (2016). A genome-wide transcriptional analysis of yeast-hyphal transition in *Candida tropicalis* by RNA-Seq. *PLoS One* 11, e0166645. doi:10.1371/journal.pone.0166645

Xu, H., Fang, T., Omran, R. P., Whiteway, M., and Jiang, L. (2020). RNA sequencing reveals an additional Crz1-binding motif in promoters of its target genes in the human fungal pathogen Candida albicans. *Cell. Commun. Signal.* 18, 1. doi:10.1186/s12964-019-0473-9

Yin, K., Zhang, Y., Zhang, S., Bao, Y., Guo, J., Zhang, G., et al. (2019). Using weighted gene co-expression network analysis to identify key modules and hub genes in tongue squamous cell carcinoma. *Medicine* 98, e17100. doi:10.1097/MD.0000000000017100

Yong, S. L., and Dutta, A. (2009). MicroRNAs in cancer. *Annu. Rev. Pathol.* 4, 199–227. doi:10.1146/annurev.pathol.4.110807.092222

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

Zhang, G., Yao, X., Wang, C., Wang, D., and Wei, G. (2019). Transcriptome analysis reveals the mechanism underlying improved glutathione biosynthesis and secretion in Candida utilis during selenium enrichment. *J. Biotechnol.* 304, 89–96. doi:10.1016/j.jbiotec.2019.08.015

Zhu, G. D., Xie, L. M., Su, J. W., Cao, X. J., Yin, X., Li, Y. P., et al. (2022). Identification of differentially expressed genes and signaling pathways with Candida infection by bioinformatics analysis. *Eur. J. Med. Res.* 27, 43. doi:10.1186/s40001-022-00651-w

Ziello, J. E., Jovin, I. S., and Huang, Y. (2007). Hypoxia-inducible factor (HIF)-1 regulatory pathway and its potential for therapeutic intervention in malignancy and ischemia. *Yale J. Biol. Med.* 80, 51–60.

Zu, G., Sun, K., Zu, X., Han, T., and Huang, H. (2021). Mechanism of quercetin therapeutic targets for Alzheimer disease and type 2 diabetes mellitus. *Sci. Rep.* 111 11, 22959–23011. doi:10.1038/s41598-021-02248-5

Zuo, Z., Shen, J. X., Pan, Y., Pu, J., Li, Y. G., Shao, X. H., et al. (2018). Weighted gene correlation network analysis (WGCNA) detected loss of MAGI2 promotes chronic kidney disease (CKD) by podocyte damage. *Cell. Physiol. biochem.* 51, 244–261. doi:10.1159/000495205

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 16 January 2023
DOI 10.3389/fgene.2022.1084727

OPEN ACCESS

EDITED BY
Mallana Gowdra Mallikarjuna,
Indian Agricultural Research Institute (ICAR), India

REVIEWED BY
Weizhu Zeng,
Jiangnan University, China
Ankita Chatterjee,
National Chemical Laboratory (CSIR), India

*CORRESPONDENCE
Priya V. K,
priyavk85@gmail.com
Somdatta Sinha,
somdattasinha@iiserkol.ac.in

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

RECEIVED 30 October 2022
ACCEPTED 30 December 2022
PUBLISHED 16 January 2023

CITATION
V. K P and Sinha S (2023), A systems level approach to study metabolic networks in prokaryotes with the aromatic amino acid biosynthesis pathway.
Front. Genet. 13:1084727.
doi: 10.3389/fgene.2022.1084727

A systems level approach to study metabolic networks in prokaryotes with the aromatic amino acid biosynthesis pathway

Priya V. K[1]* and Somdatta Sinha[2]*

[1]National Institute of Technology Calicut, Kattangal, Kerala, India, [2]Indian Institute of Science Education and Research Kolkata, Mohanpur, West Bengal, India

Metabolism of an organism underlies its phenotype, which depends on many factors, such as the genetic makeup, habitat, and stresses to which it is exposed. This is particularly important for the prokaryotes, which undergo significant vertical and horizontal gene transfers. In this study we have used the energy-intensive Aromatic Amino Acid (Tryptophan, Tyrosine and Phenylalanine, TTP) biosynthesis pathway, in a large number of prokaryotes, as a model system to query the different levels of organization of metabolism in the whole intracellular biochemical network, and to understand how perturbations, such as mutations, affects the metabolic flux through the pathway - in isolation and in the context of other pathways connected to it. Using an agglomerative approach involving complex network analysis and Flux Balance Analyses (FBA), of the Tryptophan, Tyrosine and Phenylalanine and other pathways connected to it, we identify several novel results. Using the reaction network analysis and Flux Balance Analyses of the Tryptophan, Tyrosine and Phenylalanine and the genome-scale reconstructed metabolic pathways, many common hubs between the connected networks and the whole genome network are identified. The results show that the connected pathway network can act as a proxy for the whole genome network in Prokaryotes. This systems level analysis also points towards designing functional smaller synthetic pathways based on the reaction network and Flux Balance Analyses analysis.

KEYWORDS

metabolic pathways, aromatic amino acids biosynthesis, network analysis, flux balance analysis (FBA), systems biology

1 Introduction

Biochemical pathways in cells underlie cellular functions, and hence its phenotype. These are regulated by many direct and indirect, and hardwired and transient factors. Evolution of multi-step biochemical pathways in any species depends upon how natural selection shapes the evolution of a set of enzyme-coding genes catalysing the constituent chemical reactions, such that the required end-product is made (Flowers et al., 2007; Invergo et al., 2013). However, the genes, enzyme and pathways do not function independently. In each species, they exist in the context of a large biochemical network, consisting of other genes, enzymes and pathways interacting with each other, and with the intra- and extra-cellular environments. Hence in order to understand the interactions and effects in functionally related pathways, we need to study the properties of subsets of metabolic networks at different levels.

To study how pathways regulate their function with respect to each other, we chose the highly branched aromatic amino acid (Tryptophan-Tyrosine-Phenylalanine, TTP) biosynthesis pathway as an example. This pathway is responsible for the production of three aromatic amino

acids; Tryptophan, Tyrosine and Phenylalanine–all requiring high energy for their synthesis. The TTP pathway has been studied previously for its role in the production of secondary metabolites (Herrmann 1995; Herrmann and Weaver 1999), and its usage as target for several antibiotics, fungicides and herbicides (Roberts et al., 2002; Abell et al., 2005; Webby et al., 2005). The TTP pathway is present in most of the prokaryotes, but is lost in higher eukaryotes and mammals (Xie et al., 2003), thus requiring higher organisms to get some of these amino acids as food additives. Even in the TTP prototrophs, the evolutionary history of the pathway is convoluted due to instances of horizontal gene transfer and is characterized by many isozymes, bi-functional enzymes and gene fusions (Bentley and Haslam 1990; Xie et al., 2003; Richards et al., 2006; Priya et al., 2014).

Traditionally, specific pathways such as, the Tryptophan biosynthetic pathway, have been studied in depth both experimentally and theoretically using mathematical models (Yanofsky et al., 1987; Sinha 1988; Santillan and Mackey 2001; Castro-López et al., 2022). However, in the post-genomic era, most of the studies have focussed on network modelling and analysis of the whole cellular metabolism (Fairlamb 2002; Ma and Zeng 2003; Gerlee et al., 2009). In recent times, the principles of Systems Biology have been used extensively to study metabolic pathways at different scales (Nielsen 2017), and reconstruction of whole genome metabolic networks from their genome sequences has been an active area of study (Khodayari et al., 2016; Norsigian et al., 2018; Bagheri et al., 2019).

From the perspective of the intracellular biochemical network, the maze of neighbouring pathways, that are connected through sharing one or more metabolites, can influence the function and evolution of each other. Yet, study of pathways *in the context of each other* is rarely done across species. Hence in order to study the contextual influence of the inter-connected pathways, we use complex network analysis on the TTP pathway reactions network in 29 Bacteria and Archaea. Several FBA and network models have shown how various reactions are connected and used smaller subsystems to improve production or for finding new drug targets. But in these networks, the pathways present in one particular organism were studied, for example the network for disease associated pathway cluster for Huntington disease (Kakouri et al., 2019) or the network of interacting pathways to find drug targets (Raman et al., 2005; Chen et al., 2016). Our study is different from these since we are using data from 29 species of free-living Bacteria and Archaea from diverse environments and metabolic activities and we have formed a network of pathways that are connected to the TTP pathway that is common across the 29 species. This is a novel method to understand how the pathways are interconnected and function in context to each other. We have assessed the variations in the topological properties of the TTP reaction network nodes after adding the neighbouring pathways, in the combined reaction networks. Our results show the contextual variations of the topological properties of the TTP reaction network nodes in the combined network, and study their similarity across bacteria and archaea.

Network representation and analysis of metabolic pathways offers a convenient and useful mode for understanding the role of the connectivity patterns of the reaction nodes in interconnected pathways. However, the chemical reactions at each step decide the function of the pathway. Flux Balance Analysis (FBA), a constraint-based approach to model organisms based on mass-energy balance, and flux limitations (Kauffman et al., 2003) are used to understand how the reaction product flux functioned in the pathway. The FBA has been used previously for representing and modeling the growth of many organisms such as, *E. coli* (Edwards and Palsson 2000; Burgard and Maranas 2001), *L. lactis* (Flahaut et al., 2013), *S. coelicolor* A3(2) (Borodina et al., 2005), *G. oxydans* (Wu et al., 2014), *etc.* We used the FBA to study the effect of mutation or deletion of genes/reactions, present in the TTP pathway and other connected pathways - on the flux through the TTP pathway. This study yielded information on those reaction steps that have a direct effect on the production of aromatic amino acids, in the context of the larger reaction network. Comparing the network and FBA analysis results, we show that, at the systems level, the pathway activities are dependent on a smaller set of reactions that are important for its biochemical activities. This also indicates that a smaller reaction network of the important reactions and enzymes may be chemically engineered for a functional pathway instead of the existing whole metabolic pathway that has evolved through a step-by-step evolutionary historical contingency.

2 Results

***The TTP Pathway*:** A reconstructed common TTP pathway model for Bacteria and Archaea is shown in Figure 1. The pathway is divided into four sections (see Figure 1 legend) where the additional reactions specific to bacteria are shown in red and that for Archaea in blue boundaries at the top.

2.1 Network analysis

The directed reaction networks were constructed for the TTP pathway for 29 organisms (Supplementary Table S1), and their network properties such as Degree, Clustering Coefficient, Closeness Centrality and Betweenness Centrality were studied.

2.1.1 TTP pathway network and its network properties

The TTP reaction pathway is a linear network (Figure 2) with very few connections other than due to consecutive dependence. The Archaeal and Bacterial TTP networks are topologically similar. Figure 2 shows *E. coli* and *N. pharaonis* TTP networks as examples for the Bacterial and Archaeal organisms. The major difference between the two networks lies in the input region, since in many Archaea the DKFP pathway provides the precursors for the formation of 3-dehydroquinate, whereas in Bacteria it is from the Pentose phosphate pathway and Glycolysis (see Figure 1). The number of reaction nodes and edges varies among both bacterial and archaeal species. For example, in Bacteria, the reaction nodes vary between 18 (for *S. thermophilus*) and 25 (for *E. coli*), and the number of edges between 30 (*S. thermophilus*) and 50 (*E. coli*).

The difference in the number of nodes between the organisms is because there are multiple reactions that provide different paths for the production of the same metabolite. Due to this, the number of connected pathways also differs across the organisms under study. For further analysis, only those reactions and pathways are chosen that are common across all the 29 organisms (Supplementary Table S2 and S3).

The average Degree in these TTP networks is between 2.86 (*Synechocystis sp.* PCC 6803) and 4 (*E. coli*), which further shows

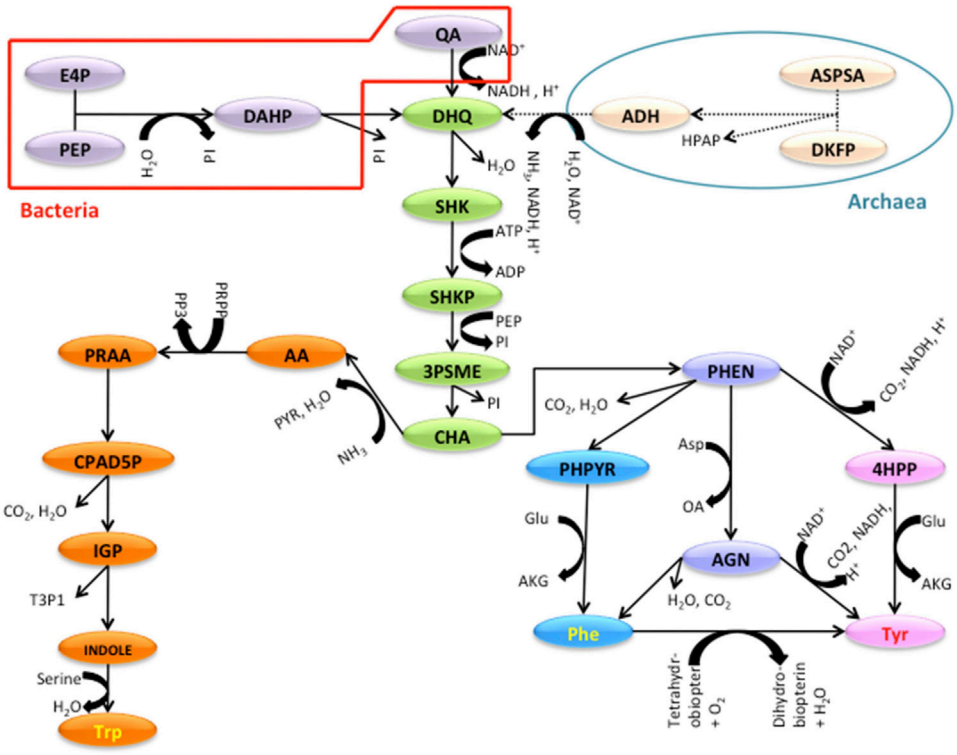


FIGURE 1
The TTP Pathway studied with the sections indicated. Red and blue shapes are specific to Bacterial and Archaeal TTP pathway. The Green substrates are part of the Shikimate section. Orange is for Tryptophan and Blue and Pink are for the Phenylalanine and Tyrosine parts. INPUT SECTION: In bacteria; E4P (D-Erythrose-4-phosphate), PEP (Phosphoenol pyruvate), DAHP (2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate), QA (Quinate), DKFP (6-deoxy-5-ketofructose-1-phosphate). In archaea; ASPSA (Aspartate semi aldehyde). SHIKIMATE SECTION: DHQ (3-dehydroquinate), SHK (Shikimate), SHKP (Shikimate 3-phosphate), 3PSME (5-O-(1-Carboxyvinyl)-3-phosphoshikimate), CHA (Chorismate). TRYPTOPHAN SECTION: AA (Anthranilate), PRAA (N-(5-Phospho-D-ribosyl)anthranilate), CPAD5P (1-(2-Carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate), IGP (Indoleglycerol phosphate), INDOLE, Trp (Tryptophan). PHENYLALANINE AND TYROSINE SECTION: PHEN (Prephenate), PHPYR (Phenylpyruvate), 4HPP (4-Hydroxyphenylpyruvate), AGN (L-Arogenate), Phe (Phenylalanine), Tyr (Tyrosine).

how sparsely connected the network is. Based on these properties, the Bacteria and Archaea networks do not differ much. Amongst the Bacteria, the Proteobacteria tend to have higher number of nodes and edges. The Gamma-proteobacteria, *E. coli* and *P. putida* has the highest number of nodes, edges and average degree for their TTP pathway network (Figure 3).

***Connected Pathway of TTP*:** A connected pathway is one in which at least one reaction of that pathway either produces or consumes a metabolite that is either consumed or produced by the TTP pathway. Even though there is a slight difference between the bacterial and the archaeal TTP pathway, the entire metabolic network of these organisms may differ greatly from each other. This may cause the pathways associated with the TTP pathway to differ between organisms. Therefore, only the reactions and associated pathways that are common among all the 29 organisms under study are discussed here (Supplementary Table S2 and S3).

2.1.2 Network properties of connected pathway networks

The average network properties of the connected networks, i.e., the TTP network combined with each of the connected networks (as given in Supplementary Tables S2 and S3), were calculated for all organisms. First the global properties of the connected pathway networks are given, and then local node-level properties are discussed.

Global properties of connected networks

Nodes: The number of nodes of the combined pathways are significantly different from their TTP network in all the Bacteria and Archaea (Wilcoxon test, p-value <.05) (Supplementary Figure S1). The highest number of nodes is in Microbial metabolism in "diverse environments" (map01120), Biosynthesis of Amino Acids (map01230), Purine metabolism (map00230) and Carbon metabolism (map01200). Except for the 2-Oxocarboxylic acid metabolism (map01210) and Methane metabolism networks (map00680) all the other connected networks of Archaea have lower number of nodes than its Bacterial counterpart. In Bacteria, the highest variation in the number of nodes is in map00330, map01120, map00230 and map00240.

Bacterial networks show larger variation (std dev range: 1.67–16.18) in node numbers than Archaeal networks (std dev range: 1.34–8.49), and the main contributor to this are the Proteobacteria. Except for TTP, and the other 7 out of 17 connected pathways (e.g., map00340, map01230, map00020, map01200, map00010 and map00260), the rest of the connected networks differ significantly between Bacteria and Archaea (Wilcoxon test, p-value <.05). Bacterial networks have significantly higher number of nodes compared to the Archaeal networks in few pathways, but in map01210 and map00680 they are significantly more in Archaea (Wilcoxon test, p-value <.05).

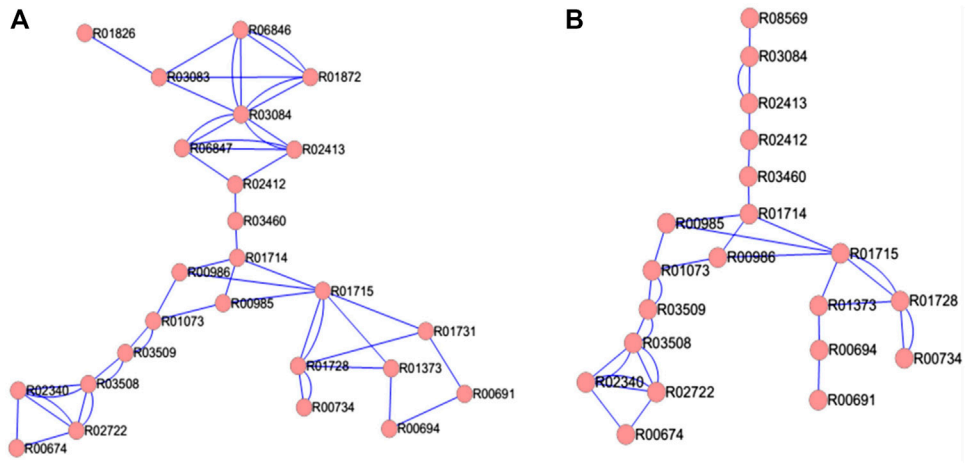


FIGURE 2
TTP network in **(A)** bacteria *E. coli* and **(B)** Archaea *N. pharaonis*. The double edges indicate reversible reactions.

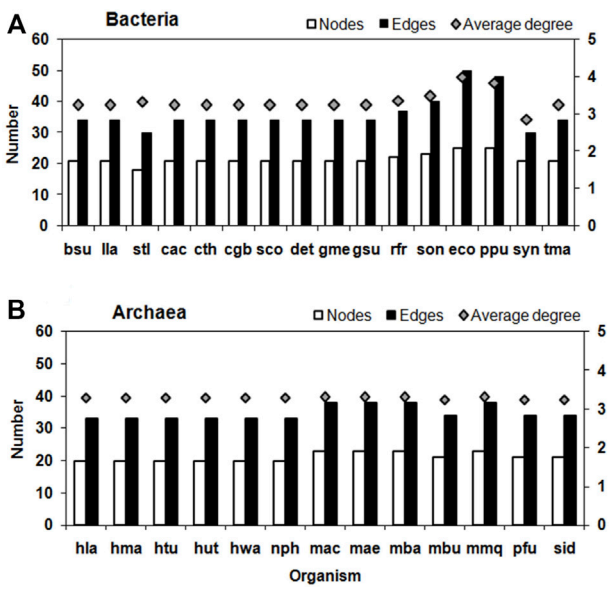


FIGURE 3
Number of Nodes, Edges (left *Y*-axis) and average degree (right *Y*-axis) of TTP pathway network in **(A)** Bacteria and **(B)** Archaea. See Supplementary Table S1 for the three lettered species names.

Edges: A similar distribution is seen in the edge numbers and degree in both Bacteria and Archaea (Supplementary Figure S2 and S3). In Archaea, for example, the Glycolysis pathway adds a higher number of edges than in Bacteria, suggesting larger number of connections between the nodes in Archaea. The addition of connected pathways significantly changes the degree in all the pathways, except map00330 and map00051 in Bacteria, and map00970 and map00051 in Archaea. Here also the variation is more in Bacteria than in Archaea. Contrary to all the other properties, the variation in the degree is slightly more in Archaea (std. dev. range: .03–.99) than in Bacteria (std. dev. range: .1–.87), and significant differences are observed in map00970, map01120, map01210, map00020, map00010, map00680, map00230, map00240 and map00030 between Bacteria and Archaea (Wilcoxon test, *p*-value <.05). Furthermore, addition of sparsely connected map00340, map00970 and map00270 decreases the average degree of the combined networks (Supplementary Figure S3).

Average Path Length: Addition of new nodes to the existing TTP pathway does not always increase the average path length of the network proportionately (Supplementary Figure S4), except for the addition of map00020 in Bacteria and map01210, map00020, map00010, map00260, map00270 and map00051 in Archaea. For all other pathways, the addition significantly changes the Average Path Length (APL) of the network (Wilcoxon test, *p*-value<.05). There are significant differences in the APL between Bacteria and Archaea in map01120, map01210, map01230, map00010, map00680, map00240 and map00030.

These results indicate that as the metabolic networks expand, due to addition of nodes, the network properties do not change proportionately–they depend on the connection point to the TTP pathway, and the topology of the added pathway. They also differ between and within Bacterial and Archaeal species for the same connected pathway even though the basic TTP pathway do not differ much between the two types of Prokaryotes.

Local properties of connected networks

The addition of the *connected* networks to TTP pathway network not only changes the global properties of the combined pathway networks, but also the properties of the individual TTP network nodes.

Degree: The comparison of ***Degree*** across the different connected networks show that there is no consistent difference between Bacterial and Archaeal networks. The 3 out of 15 common reactions showed no variations in degree, while 5 out of 15 have significantly different Degree in connected networks (>2 std dev) across different organisms in the connected pathways. Addition of certain pathways such as the map01230 introduces fairly large variations in the Degree for *E. coli*. (shown in Supplementary Figure S5), and *C. glutamicum*, *C. acetobutylicum*, *M. barkeri*, *R. perfringens* in the TTP reaction network nodes. The same in *H. turkmenica* show the least variability

TABLE 1 Changes in Degree and Clustering Coefficient in nodes.

	Degree	Clustering coefficient
No significant variation	R03508 R03509 R02340	R03460 R03508 R03509 R02340 R02722
Significant variation (Std. dev. >2)	R02722 R01073 R03460 R01714 R00674 R01073	R01073 R01373 R01714

in their degree across all pathways. This result, interestingly, clearly demonstrates that individual reactions change their connectivity pattern on addition of pathways, and this is not necessarily due to direct attachment of the connecting pathway to that node. It could also be due to changes in their biochemical interactions facilitated due to the new pathway topology in different organisms.

***Clustering Coefficient* (*CC*):** The CC of the TTP pathway reactions also change due to the addition of the connected networks (see Supplementary Figure S6). Although, out of 17 connected pathways, the CC of 4 remain the same, but 3 show significant differences (>2 std dev). These are for the addition of map01230, map00230, and map01120 as the addition of new nodes in these networks reduces the CC of these nodes. Bacteria and Archaea show similar variation in their CC. A summary of changes in Degree and Clustering Coefficient in TTP nodes are shown in Table 1.

***Closeness centrality*:** Addition of pathways tends to change the path length, which is reflected in the parameter Closeness centrality. The TTP pathway has a high Closeness centrality, and the addition of other pathways increase the number of Nodes and the Closeness centrality of the overall network. Analysis showed that most of the connected networks, with the exception of map00970, have a significantly different Closeness centrality when compared to the isolated TTP pathway. Addition of map01230 increases Closeness centrality while addition of map00250, map00330, map01120, map01210, map01200, map00260, map00230 - decreases it for the TTP nodes. The pathways, such as, map01120 and map01200 have varying results in different organisms due to the diverse environments in which these organisms survive. This increase and decrease in the network parameter Closeness centrality indicates that the local network properties of the TTP pathway reactions nodes can change in a non-consistent manner even when the network is expanding due to the addition of nodes (see Supplementary Figure S7).

***Betweenness centrality* (*BC*)**: BC of a node is an important property, as it signifies the central position of the node in the network in terms of transfer of information from all other nodes. There is a general decrease in this network parameter for most connected pathways across all TTP nodes. However, the addition of map01230 and map00970 also significantly alter the BC of the common reactions (z-score >3) across all organisms. R00674 show *almost no* variation in its BC among the combined pathways of different organisms, since it is at the terminal end of the network. The analysis of the change in BC in TTP nodes showed that, across organisms, most of the variation is observed in the map00030. The addition of this pathway to TTP changes the topology of the combined network in such a manner that it induces changes in the BC in several nodes. The reaction node R01073 in TTP pathway shows considerable increase in BC on addition of map00030 and map00340 due to the addition of pathways that are linear. BC of the terminal reactions, such as R00674 and R02722 in the TTP pathway, increases significantly due to the addition of the connected pathways which occur in very few cases. BC of a node being an important property in terms of transfer of information from all other nodes, our results show that only those pathways change the BC of the TTP nodes, which change the topology of the combined network based on where the added pathway is connected to the TTP network (Supplementary Figure S8).

TABLE 2 Average Betweenness and average Closeness values (for 29 organisms) for the nodes in the TTP pathway - in isolation and in the Combined Connected Network (CCN). The standard deviations are not shown as the values are very low.

	Betweenness		Closeness	
	TTP	CCN	TTP	CCN
R02412	.156	.007	.012	1.16 × 10−05
R03460	.183	.008	.014	1.16 × 10−05
R01373	.084	.007	.013	1.16 × 10−05
R01714	.205	.009	.016	1.17 × 10−05
R01715	.18	.017	.016	1.17 × 10−05
R00674	0	0	.008	1.17 × 10−05
R02722	.011	.047	.009	1.17 × 10−05
R02340	.05	.001	.009	1.17 × 10−05
R03508	.15	.045	.011	1.17 × 10−05
R03509	.178	.045	.012	1.17 × 10−05
R01073	.196	.065	.014	1.17 × 10−05
R00985	.102	.015	.016	1.17 × 10−05
R00986	.102	.015	.016	1.17 × 10−05
R03084	.085	.006	.009	1.16 × 10−05
R02413	.116	.006	.01	1.16 × 10−05

2.1.3 Combined Connected Network (CCN) of TTP

Till now the network properties of the TTP pathway network, in combination with each of the connected pathways (as in Supplementary Table S2 and S3), have been studied. The *Combined Connected Network* (CCN) is the combined network of the TTP pathway with all the connected pathways added together. It gives an integrated view of the TTP pathway embedded in the metabolic network of the 17 reaction pathways directly connected to it for each of the organisms under study. The question addressed here is how the network properties of the individual nodes of the TTP pathway network change in such a combined network, because of the

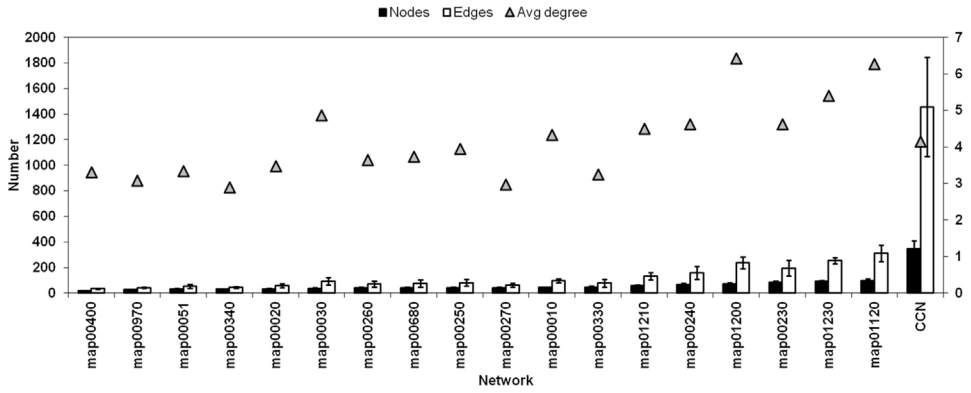


FIGURE 4
Network size (number of Nodes), Number of Edges, and the Average Degree of each connected networks and the CCN (Refer Supplementary Table S2 for pathway names).

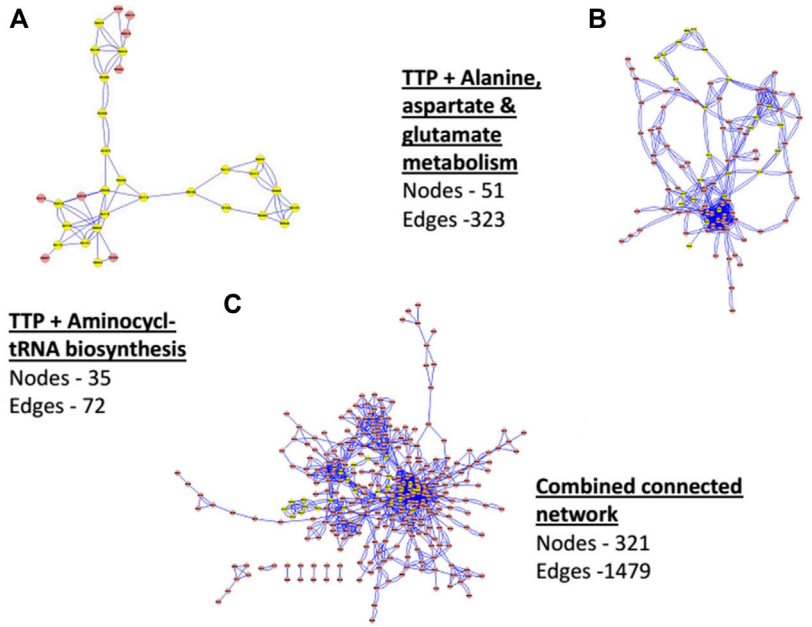


FIGURE 5
Network of TTP (with Yellow Nodes) with connected networks of **(A)** Aminoacyl tRNA biosynthesis pathway (map00970), **(B)** Alanine, Aspartate and Glutamate metabolism (map00250) and **(C)** Combined Connected Network (CCN).

change in the topology and connectivity patterns in the CCN. Combining the connected pathway networks to TTP added new nodes and edges to the TTP pathway. As will be shown below, some of these additions significantly change the topological properties of the TTP pathway reactions (Table 2). For example, the TTP reaction node R02722 is the only one that shows increase in Betweenness Centrality in the CCN. This is due to the addition of the highly interconnected *Glycine*, Serine and Threonine pathway in the CCN through that node. The addition of the highly interconnected pathways, such as the amino acid biosynthesis pathway, or addition of a few nodes, as in the case of map00970, could significantly alter the properties of the TTP nodes (Figure 4).

Figure 4 shows the comparison of a few network properties among each connected pathway in all organisms (see Supplementary Table S2 for pathway names) and the CCN. Network size (number of Nodes), total number of Edges, and the average Degree of each connected networks are shown along with that of the CCN. Figure 5 shows the topology of the TTP reaction network (Yellow nodes) when connected with A) Aminoacyl t-RNA biosynthesis pathway (map00970), B) Alanine, Aspartate and Glutamate metabolism (map00250), and C) Combined Connected Network (CCN). It is clear that increasing the number of Nodes does not necessarily increase the average Degree of the network (Figure 4; Figure 5).

TABLE 3 Percentage of Network hub reactions from CCN, which were shown to be essential in the FBA models. Organism specific: hubs of *E. coli* and *M. barkeri*; Common: common hubs of 29 Bacteria and Archaea).

Network	Organism	Degree hubs	Betweenness Centrality hubs	Closeness Centrality hubs
Organism specific	E. coli	0%	19.6%	8.26%
	M. barkeri	62.5%	72.72%	45.45%
Common	E. coli	16.7%	25.50%	24.4%
	M. barkeri	16.7%	41.18%	39%

For CCN, the average degree is 4.15 - quite low even though the average network size is large (321 Nodes). The number of Edges, though not additive, is also quite large (1479). This indicates, as is seen in Figures 5A–C also, that the CCN has a topology that is largely branched with many linear sections. The TTP pathway is largely linear and is a non-redundant network. Hence most of the nodes are equally important for the pathway to function, even though each of them has different network properties (as mentioned in the previous sections). Addition of other pathways can cause nodes to change their local network properties. Low Centrality measures point towards the fact that the CCN has a non-compact topology with large linear sections. This is due to the underlying chemical basis of the network, where substrate-product reactions are quite specific to their chemical nature, and the same chemical species cannot be obtained through different chemical reactions.

The "hubs" of the network parameters - Degree, Betweenness Centrality and Closeness Centrality - are reaction nodes in the network with the highest value of the respective parameters. We use a cut off for selecting Hubs as the "Nodes in top 20%" of each of the measures. The CCN have few Degree hubs, since these networks are characterized by a large number of Nodes with low degrees, and very small number of Nodes with high degrees, and non-redundant routes for metabolism. There are 66 reaction nodes in the CCN that are found as hubs common to all organisms. Most of them are either Betweenness Centrality or Closeness Centrality hubs (Supplementary Table S4). Since these hub reactions in the CCN are important for the network, these might also be important for the functioning of the TTP pathway in the integrated network. It is clear that many (9 of 15) of the TTP reactions have now increased their Betweenness Centrality and Closeness Centrality when in the context of other connected pathways. The nature of these networks is generally linear sequence of chemical reactions leading to formation of specific products. However, these specific reaction pathways interact to facilitate cross-talk to promote coordinated response of the cell. Therefore, increasing the centrality measures seems to be a functionally suitable strategy, since increasing degree may not be chemically possible. The changed network parameters of the TTP nodes in the combined network (CCN) points towards their role in changing/modifying their function when in context of other pathways. This can lead to change in their biochemical attributes (such as, reaction velocity, flux, regulation, *etc.*).

2.2 Flux balance analysis (FBA)

FBA is done in order to analyze the flux passing through the reaction steps of the TTP pathway during wild-type growth, and after perturbations (e.g., loss of reaction due to deletion mutation, or lowering of efficiency of the reaction), in order to understand the influence of different reactions on the working of the TTP pathway. The flux analysis (see Methods section), for the TTP pathway was done on the complete genome scale models of *E. coli* (Feist et al., 2007) and *M. barkeri* (Gonnerman et al., 2013). The genome scale *E. coli* model, considered here, consisted of a total of 2382 reactions, 1261 genes, and 1668 metabolites; and, the *M. barkeri* model consisted of 815 reactions, 750 genes, and 718 metabolites.

2.2.1 Flux analysis in TTP pathway

Production of aromatic amino acids (TTP) in the cell is a high energy consuming process (Akashi and Gojobori 2002). This energy cost is reflected in their low usage in the polypeptide chain, and in the flux passing through the TTP pathway in almost all the organisms. All the flux mentioned here on will be in $mmolgDW^{-1}h^{-1.}$ The Tryptophan section has the least amount of flux passing through it: .0418 for *E. coli*, and .0013 for *M. barkeri*. The Phenylalanine section (.1296 for *E. coli* and .0041 for *M. barkeri*), and Tyrosine section (.1018 for *E. coli* and .0035 for *M. barkeri*) have higher fluxes (Supplementary Figure S9 and S10). The list of reactions present in *E. coli* and *M. barkeri* is given in Supplementary Table S5.

Fluxes through the TTP pathway for *E. coli* and *M. Barkeri* are different

1) In both the organisms, fluxes through the Input and Shikimate section are higher than the rest of the sections, because the flux is undivided in these sections. At Chorismate synthase reaction (CHORS), the flux is distributed between the two branches depending on the coefficients of Trp and Phe-Tyr in the biomass equation. All the flux passes through TRPAS2 of the Trp section in *E. coli*. In *M. barkeri*, it takes the reaction TRPS1 to produce the same metabolite Tryptophan.
2) Compared to Bacteria *E. coli,* the Archaea *M. barkeri* has a lower flux. It may be noted that the growth rate for *E. coli* is higher than that of *M. barkeri*, which also shows up in the differences in the media and biomass equations of the two organisms. Out of 1339 unique reactions (as mentioned in section 2.2 of CCN of *E. coli*) present in the whole FBA, deletion of 175 reactions was found to be adversely affecting the production of the aromatic amino acids. We reduced the efficiency of the *E. coli* TTP pathway genes to find the effect of such changes in the production of the aromatic amino acids. A 100% reduction (deletion) of the TTP pathway genes is shown in (Supplementary Figure S11; Supplementary Table S8). Deletion of genes in the TTP pathway leads to no flux through any of the reactions except for TRPAS2, TRPS1, TRPS2 and TRPS3 which are alternate pathways to each other. If the bounds of the flux of the reactions are reduced to 90% of the flux one by one through the reactions, then a marked reduction is seen in the flux through the network (Supplementary Table S9). Suggesting that even though the amount of flux passing through the reactions are very low, they still play a major role in the biomass formation of the organisms. For example, constraining the flux through PHETA1 to −.117 (reversible reaction) leads to a reduction of the flux through the Input and Shikimate section to .247 (.274 in Wild type-WT) and .038 through the tryptophan section (.042 in WT) and .117 in Phenylalanine (.13 in WT) and .092 through Tyrosine (.102 in WT) (Supplementary Figure S12).

In *M. barkeri*, out of the 815 reactions present in the FBA model, the deletion of 250 reactions shows adverse effect on the production of TTP. Many of these pathways are common between *E. coli* and *M. barkeri*, but some of them are unique to either Bacteria or Archaea, as the metabolism of these two organisms are different–in some cases. For example, the pathway for Glycerophospholipid biosynthesis pathway influences TTP production in *E. coli*, while the Methanofuran B biosynthesis and Methanogenesis pathways influences TTP production in Archaea *M. barkeri*.

3) Reducing single gene efficiency does not significantly affect TTP production in *M. barkeri* because there are alternative reactions for some reactions, which provide other routes for producing the same metabolite. This indicates that the TTP pathway is more robust in this organism in terms of random gene/reaction deletions. Deletion of genes involved in all reactions, except ANS, ANS2, TRPS1, TRPS2, TRPS3, leads to no flux through the TTP pathway (Supplementary Table S10; Supplementary Figure S13). The reactions ANS has the alternate path ANS2 and TRSP1 has the alternate route formed by TRSP2 and TRSP3 because of which the flux flows through the pathway even in case of deletion of any one of them. Constraining the flux through the TTP reactions to 90% of the flux through those reactions has an effect on the growth rate and flux through the reactions (Supplementary Table S11). In the *E. coli* pathway, reduction in the efficiency of the input and shikimate pathway affects the flux, but not for the reactions ANS, ANS2, TRSP1, TRSP2 and TRSP3 due to the alternate routes, as previously mentioned. Decrease in the efficiency of reactions in the Phe and Tyr section also reduce the flux, out of which the reaction CHORM (chorismate mutase) affects the most, since the flux for the synthesis of Phe and Tyr pass through it. Reduction to 90% of the flux through the reaction has interesting results, for example, when the flux through CHORM is .0072 (.0077 in WT), the flux through the input and Shikimate section is .0084 (.009 in WT), through the Tryptophan section is .0012 (.0013 in WT) and through Phenylalanine is .0039 (.0042 in WT) and Tyrosine is .0033 (.0035 in WT) (Supplementary Table S11; Supplementary Figure S14).

2.3 Comparison of network analysis and FBA studies for the TTP pathway

Deletion of hubs can cause a network to lose its structural and functional integrity (Barabási and Oltvai, 2004). Our results (Supplementary Table S4) yielded TTP Network hubs (Nodes having high Degree, Betweenness Centrality, and Closeness Centrality). The reaction deletion studies using FBA analysis also provided a set of the reactions that, when deleted individually, affects the flux through the TTP pathway (Supplementary Table S7 and S12). These two results obtained using different theoretical approaches were compared with each other to find if the Network hubs (of high Degree, BC, and Closeness Centrality) and the essential genes (obtained from FBA reaction deletion analysis) overlap. Table 3 shows the percentage of Degree hubs, BC hubs and Closeness Centrality hubs that were identified using network analysis and also found to be essential reactions for TTP pathway using FBA. Organism specific reactions are those hubs that were identified from the CCN of either *E. coli* or *M. barkeri*. The *Common* hubs are the hubs that were identified to be common across all the 29 organisms that were used in the network analysis. The reactions that are common between Network hubs and the essential reactions from FBA mostly belong to Purine and Pyrimidine biosynthesis, Threonine and Lysine biosynthesis and TTP pathway.

The Network analysis of the CCN can predict some of the important nodes obtained from FBA analysis. It may be kept in mind that the CCN takes into account only 18 pathways and the reactions present in them, and gives equal weightage to all the reactions and connections. Whereas, in the genome scale FBA, the flux does not flow through all the reactions equally, and hence those reactions and the connections are not reflected in the essential reactions. This indicates that a reduced collection of connected networks can be used to find essential reactions. The list of common hubs across the 29 organisms can be used as a reference list for further studies for finding reactions essential for functioning of TTP pathway and to increase its productivity, since they provide similar result to organism-specific hubs. The list of Network hubs that were shown to be essential by the FBA analysis is given in Supplementary Table S6.

3 Discussion

The important role of "context" has been of long-standing empirical and theoretical interest in biological systems because of their multi-scale and interacting modular structures. Understanding context representations and its interaction with functional outcome in behaviour is an area of immense interest to both neurobiologists and in psychology (Rudy, 2009). In an interesting article, the multi-scale and modular structure of metabolic network was analysed to identify the context in which evolutionary processes may occur (Spirin V et al., 2006).

Studies involving molecular interactions of single genes or proteins in the context of their downstream partners and gene context-based modules have been done to evaluate their role in cellular response mechanisms in signalling, amino acids and carbohydrate metabolism pathways (Lan et al., 2013; Bhatt et al., 2018). We started with a general question; *do the topological features (as studied using network analysis) of a metabolic pathway vary when it is embedded in the larger network of other connected pathways, and does this variation affect the pathway function?* We approached to answer this query from a different perspective using two systems biology methods - topological properties (network analysis) and metabolic activity (Flux Balance Analysis) - of the aromatic amino acid biosynthesis (TTP) pathway in many species of Bacteria and Archaea. This pathway consists of quite high energy consuming reactions in the cell. It takes an equivalent of 52, 50 and 74.3 high-energy phosphate bonds for the production of Phenylalanine, Tyrosine and Tryptophan, respectively (Akashi and Gojobori 2002). This energy cost is thus reflected in their usage in the polypeptide chain, and in the metabolic flux passing through the TTP pathway.

The control of the production of aromatic amino acids is traditionally done by means of metabolic engineering in organisms such as *E. coli* and *C. glutamicum* (Katsumata and Ikeda 1993; Ikeda 2006). In these studies, systematic control of

genes in the TTP pathway (such as, *aroG, aroF, aroH, anthranilate synthase, pheA etc.*), which respond to the production of the end products, were mutated to increase the production of the aromatic amino acids. Here we have looked at the TTP pathway individually, as well as, when embedded at the larger metabolic network in Bacteria and Archaea. Such studies require various sources of genetic and biochemical information, such as, stoichiometry, structure of reaction pathways and alternative routes of reactions, along with genes and genomes of different organisms. The results presented highlight the fact that functioning of a biochemical reaction in the cell is intimately connected to its "context" (i.e., position of the pathway in the total biochemical network), and the topology of its connectivity to the larger set of reactions - both in the pathway and in the larger biochemical network.

Based on these analyses we are able to arrive at several conclusions. The Network analysis was undertaken to analyse the changes in network properties of TTP pathway reaction network in isolation and in combination with other pathways directly connected to it through sharing of metabolites as incoming or outgoing reactants. The TTP pathway, which is a predominantly linear and a sparse network, shows a low average degree in all organisms. The nodes in the centre of the network possess high Betweenness and high Closeness Centrality values, while the nodes at the extremities show the opposite characteristics. Out of the many pathways that are connected to the TTP pathway, the 17 pathways that were common among the 29 organisms were considered in this study. The network analysis with all connected pathways in all the organisms showed that - changes in the properties of the 15 TTP reaction network nodes not only depended on the topology of the added network, but also on the nodes to which the pathway was added. The Complete Combined Network (CCN), consisting of the TTP pathway and all the 17 connected networks, showed that the properties of the TTP nodes is not the same when considered in the context of the larger connected network. Nodes with low Degree, Betweenness Centrality or Closeness Centrality, either acquire more connections, or by virtue of the new connections that alter the resulting topology, change their network properties, and become hubs in the CCN. The different Degree, Betweenness Centrality and Closeness Centrality hubs were found for the CCN for all the organisms, and the common hubs were ascertained from them. Hence, analyzing pathways in isolation, and in combination with other networks, gives varying properties to the nodes in the network. How these changes in network topology and parameters of the TTP nodes influence the chemical activity leading to end product formations was analyzed using the Flux Balance Analysis.

The Flux Balance Analysis was done to study the flow of metabolites through the metabolic reaction network of the TTP pathway, and to compare it between Bacteria and Archaea, by taking *E. coli* and *M. barkeri* as representatives from the two phyla. The flux through TTP is very low in both the organisms with *M. barkeri* being lower than *E. coli*. *In silico* gene deletion studies of TTP pathway genes showed that fluxes in *M. barkeri* is more resistant to random attack than *E. coli*, due to the presence of isozymes. In both the organisms, the deletion or reduction of efficiency of the gene for Phenylalanine and Tyrosine production greatly affected the overall flux though the network. Deletion of reactions in the whole network showed that many pathways such as, Glycolysis, Histidine metabolism, etc, affect the production of these aromatic amino acids in both the groups of organisms. There are also differences in the pathways, affecting TTP between Bacteria and Archaea, due to their differences in metabolism, such as the Methanogenesis pathway.

A comparison between the network analysis and flux balance analysis of the isolated TTP and CCN of TTP pathways showed that many of the important reaction nodes or "hubs" (in terms of higher network parameters) in the TTP network were common with the essential reactions found by FBA. This points towards identifying a smaller set of reaction steps that can be used for experimental manipulation of the TTP pathway in the cell. This combined Network-FBA approach can be used to predict important reaction steps before attempting any engineering of any pathway for increase or suppression of functionality. Until now, whole genome metabolic networks have been studied by breaking them down into modules using network science (Alcalá-Corona, et al., 2021). This study endeavored to give an integrative view of pathway function and evolution across many prokaryotes, both at a single reaction pathway level, and also when embedded in the larger scheme of biochemical networks. Both the static network approach and the dynamic flux balance analysis offered different perspectives of the same pathway function by arriving at important reaction sets (hubs and essential reactions) that promises to have important applications. Thus, even though the proximate goal of this study (with the PPT pathway as an example) is to understand the contextual role of a specific pathway - in isolation and when embedded in the larger biochemical network of the cell - this approach to study biochemical pathways to understand their systemic properties in the context of biochemical functions inside the cell, may also offer better insight for identifying essential genes, reactions for drug targets, and mutations for improving pathway functions in any organism.

4 Materials and methods

4.1 Organisms under study

29 Archaeal and Bacterial species (Supplementary Table S1) were considered for the analysis, which consist of Proteobacteria, Halobacteria and Methanomicrobia. Details are given in Supplementary Information.

4.2 Division of the pathway

The TTP pathway was broken down into different levels; the lowest level being the individual reactions, thus at individual gene level. The next level was created by dividing the pathway into individual branches or sections that end with the production of important compounds, and the final level was the whole pathway. Figure 1 shows the schematic of a typical TTP pathway. For the

ease of understanding and analysis, the TTP pathway is divided into four sections. The first section is the **Input section**, where the genes for the enzymes that catalyze the reactions for the conversion of the initial precursors to 3-dehydroquinate is present. In bacteria, the pathway begins from Erythrose-4-phosphate and Phospho-enol-pyruvate. In many archaea, due to the absence of the oxidative Pentose Phosphate Pathway in several archaeal species (Soderberg 2005), the 3-dehydroquinate necessary for the initial steps of TTP production is produced by DKFP (Porat et al., 2006; Gulko et al., 2014). The second section is the **Shikimate section** of the pathway (Green substrates) which consists of five steps, in which dehydroquinate gets converted to chorismate. The third section is the **Tryptophan section** (Orange substrates), which converts Chorismate, the end product of Shikimate section to Tryptophan. The last section is the **Phenylalanine and Tyrosine section** (Purple and Pink substrates), which consist of genes for the enzymes that sequentially convert Chorismate to Phenylalanine and Tyrosine (Dosselaere and Vanderleyden 2001).

4.3 Network analysis

In this analysis, 29 organisms (Bacteria and Archaea) were selected for the study. The details of forming the reaction networks and the list of organisms is given in Supplementary Table S1. The networks were generated using in-house Perl programs. Network parameters such as Degree, Clustering Coefficient, Closeness centrality, Betweenness Centrality (Oldham et al., 2019) were calculated using the igraph package (Csardi and Nepusz 2006) of R (R Core Team 2014). Statistical analysis of the networks was carried out using R and in-house Perl programs.

4.4 Flux balance analysis

Flux balance analysis was conducted on *E coli* whole genome model (Feist et al., 2007), as a representative of Bacteria, and, the *M. barkeri* whole genome model (Gonnerman et al., 2013), as a representative of Archaea. The *E. coli* model (iAF1260) consists of 1261 metabolism associated genes, 2382 reactions, and 1668 metabolites. The *M. barkeri* model (iMG746) consists of 746 metabolism associated genes, 815 unique reactions and 718 metabolites. Both the models were simulated in minimal media. The FBA analysis was carried out using Cobrapy .26.0 (Ebrahim et al., 2013), Cobra package for MATLAB and calculations were carried out using in-house python and perl programming. All the data used in this study are available on request.

Data availability statement

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

Author contributions

SS designed and supervised the study. PVK performed the analyses. The manuscript was prepared by PVK and SS.

Funding

Council of Scientific and Industrial Research (CSIR), India for a fellowship to PVK.

Acknowledgments

SS thanks the Indian National Science Academy for the Honorary Senior Scientist award at IISER Kolkata. The authors thank the CSIR-Centre for Cellular and Molecular Biology, Hyderabad and IISER Mohali for providing the facilities where the study was initiated. Prof. Karthik Raman's guidance to PVK is gratefully acknowledged.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1084727/full#supplementary-material

References

Abell, C., Kerbarh, O., Payne, R. J., Sahr, T., and RebeilleF. (2005). Mechanistic and inhibition studies of chorismate-utilizing enzymes. *Biochem. Soc. Trans.* 33, 763–766. doi:10.1042/BST0330763

Akashi, H., and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* 99, 3695–3700. doi:10.1073/pnas.062526999

Alcalá-Corona, S. A., Sandoval-Motta, S., Espinal-Enríquez, J., and Hernández-Lemus, E. (2021). Modularity in biological networks. *Front. Genet.* 14 (12), 701331. doi:10.3389/fgene.2021.701331

Bagheri, M., Marashi, S-A., and Amoozegar, M. A. (2019). A genome-scale metabolic network reconstruction of extremely halophilic bacterium Salinibacter ruber. *PLoS One* 14, e0216336. doi:10.1371/journal.pone.0216336

Barabási, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5 (2), 101–13. doi:10.1038/nrg1272

Bentley, R., and Haslam, E. (1990). The shikimate pathway--a metabolic tree with many branches. *Crit. Rev. Biochem. Mol. Biol.* 25, 307–384. doi:10.3109/10409239009090615

Bhatt, V., Mohapatra, A., Anand, S., Bhusan, K. K., and Mande, S. S. (2018). FLIM-MAP: Gene context based Identification of functional modules in bacterial metabolic pathways. *Front. Microbiol.* 9, 2183. doi:10.3389/fmicb.2018.02183

Borodina, I., Krabben, P., and Nielsen, J. (2005). Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res.* 15, 820–829. doi:10.1101/gr.3364705

Burgard, A. P., and Maranas, C. D. (2001). Probing the performance Limits of the Escherichia coli metabolic network Subject to gene additions or deletions. *Biotechnol. Bioeng.* 74, 364–375. doi:10.1002/bit.1127

Castro-López, D. A., González de la Vara, L. E., Santillán, M., and Martínez-Antonio, A. (2022). A molecular dynamic model of tryptophan Overproduction in *Escherichia coli*. *Fermentation* 8, 560. doi:10.3390/fermentation8100560

Chen, D., Zhang, H., Lu, P., Liu, X., and Cao, H. (2016). Synergy evaluation by a pathway-pathway interaction network: A new way to predict drug combination. *Mol. Biosyst.* 12 (2), 614–623. doi:10.1039/c5mb00599j

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Syst.* 1695 (5), 1–9.

Dosselaere, F., and Vanderleyden, J. (2001). A metabolic node in action: Chorismate-utilizing enzymes in microorganisms. *Crit. Rev. Microbiol.* 27, 75–131. doi:10.1080/20014091096710

Ebrahim, A., Lerman, J. A., Palsson, B. Ø., and Hyduke, D. R. (2013). COBRApy: COnstraints-based reconstruction and analysis for Python. *BMC Syst. Biol.* 7, 74. doi:10.1186/1752-0509-7-74

Edwards, J. S., and Palsson, B. Ø. (2000) Metabolic flux balance analysis and the *in silico* analysis of Escherichia coli K-12 gene deletions. *BMC Bioinforma.* 1, 1. doi:10.1186/1471-2105-1-1

Fairlamb, A. H. (2002). Metabolic pathway analysis in trypanosomes and malaria parasites. *Philos. Trans. R. Soc. Lond B Biol. Sci.* 357, 101–107. doi:10.1098/rstb.2001.1040

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., et al. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3, 121. doi:10.1038/msb4100155

Flahaut, N. A. L., Wiersma, A., van de Bunt, B., Martens, D. E., Schaap, P. J., Sijtsma, L., et al. (2013). Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. *Appl. Microbiol. Biotechnol.* 97, 8729–8739. doi:10.1007/s00253-013-5140-2

Flowers, J. M., Sezgin, E., Kumagai, S., Duvernell, D. D., Matzkin, L. M., Schmidt, P. S., et al. (2007). Adaptive evolution of metabolic pathways in Drosophila. *Mol. Biol. Evol.* 24, 1347–1354. doi:10.1093/molbev/msm057

Gerlee, P., Lundh, T., Zhang, B., and Anderson, R. (2009). Gene divergence and pathway duplication in the metabolic network of yeast and digital organisms. *J. R. Soc. Interface* 6, 1233–1245. doi:10.1098/rsif.2008.0514

Gonnerman, M. C., Benedict, M. N., Feist, A. M., Metcalf, W. W., and Price, N. D. (2013). Genomically and biochemically accurate metabolic reconstruction of Methanosarcina barkeri Fusaro, iMG746. *Biotechnol. J.* 8, 1070–1079. doi:10.1002/biot.201200266

Gulko, M. K., Dyall-smith, M., Gonzalez, O., and Oesterhelt, D. (2014). How do Haloarchaea synthesize aromatic amino acids. *PLoS One* 9, e107475. doi:10.1371/journal.pone.0107475

Herrmann, K. M. (1995). The shikimate pathway: Early steps in the biosynthesis of aromatic compounds. *Plant Physiol.* 107, 907–919. doi:10.1105/tpc.7.7.907

Herrmann, K. M., and Weaver, L. M. (1999). The shikimate pathway. *Annu. Rev. plant Physiol. plant Mol. Biol.* 50, 473–503. doi:10.1146/annurev.arplant.50.1.473

Ikeda, M. (2006). Towards bacterial strains overproducing L -tryptophan and other aromatics by metabolic engineering. *Appl. Microbiol. Biotechnol.* 69, 615–626. doi:10.1007/s00253-005-0252-y

Invergo, B. M., Montanucci, L., Laayouni, H., and Bertranpetit, J. (2013). A system-level, molecular evolutionary analysis of mammalian phototransduction. *BMC Evol. Biol.* 13, 52. doi:10.1186/1471-2148-13-52

Kakouri, A. C., Christodoulou, C. C., Zachariou, M., Oulas, A., Minadakis, G., Demetriou, C. A., et al. (2019). Revealing clusters of connected pathways through Multisource data integration in Huntington's disease and Spastic Ataxia. *IEEE J. Biomed. Health Inf.* 23 (1), 26–37. doi:10.1109/JBHI.2018.2865569

Katsumata, R., and Ikeda, M. (1993). Hyperproduction of tryptophan in Corynebacterium glutamicum by pathway engineering. *Bio/Technology* 11, 921–925. doi:10.1038/nbt0893-921

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). Advances in flux balance analysis. *Curr. Opin. Biotechnol.* 14, 491–496. doi:10.1016/j.copbio.2003.08.001

Khodayari, A., and Maranas, C. D. (2016). A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat. Commun.* 7, 13806. doi:10.1038/ncomms13806

Lan, A., Ziv-Ukelson, M., and Yeger-Lotem, E. (2013). A context-sensitive framework for the analysis of human signalling pathways in molecular interaction networks. *Bioinformatics* 29 (13), 210–i216. doi:10.1093/bioinformatics/btt240

Ma, H., and Zeng, A. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 19, 270–277. doi:10.1093/bioinformatics/19.2.270

Nielsen, J. (2017). Systems biology of metabolism. *Annu. Rev. Biochem.* 86, 245–275. doi:10.1146/annurev-biochem-061516-044757

Norsigian, C. J., Kavvas, E., Seif, Y., Palsson, B. O., and Monk, J. M. (2018). iCN718, an Updated and improved genome-scale metabolic network reconstruction of Acinetobacter baumannii AYE. *Front. Genet.* 9, 121. doi:10.3389/fgene.2018.00121

Oldham, S., Fulcher, B., Parkes, L., Arnatkevic Iūtė, A., Suo, C., and Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. *PLoS One* 14 (7), e0220061. doi:10.1371/journal.pone.0220061

Porat, I., Sieprawska-lupa, M., Teng, Q., Bohanon, F. J., White, R. H., and Whitman, W. B. (2006). Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p -aminobenzoic acid in the archaeon Methanococcus maripaludis. *Mol. Microbiol.* 62, 1117–1131. doi:10.1111/j.1365-2958.2006.05426.x

Priya, V. K., Sarkar, S., and Sinha, S. (2014). Evolution of tryptophan biosynthetic pathway in microbial genomes: A comparative genetic study. *Syst. Synth. Biol.* 8, 59–72. doi:10.1007/s11693-013-9127-1

R Core Team (2014) R: A Language and environment for Statistical Computing, Version 2.6.2 (2008-02-08)

Raman, K., Rajagopalan, P., and Chandra, N. (2005). Flux balance analysis of mycolic acid pathway: Targets for anti-tubercular drugs. *PLoS Comput. Biol.* 1, e46. doi:10.1371/journal.pcbi.0010046

Richards, T. A., Dacks, J. B., Campbell, S. A., Blanchard, J. L., Foster, P. G., McLeod, R., et al. (2006). Evolutionary Origins of the eukaryotic shikimate pathway: Gene fusions, horizontal gene transfer, and Endosymbiotic Replacements. *Eukaryot. Cell* 5, 1517–1531. doi:10.1128/EC.00106-06

Roberts, C. W., Roberts, F., Lyons, R. E., Kirisits, M. J., Mui, E. J., Finnerty, J., et al. (2002). The shikimate pathway and its branches in Apicomplexan parasites. *J. Infect. Dis.* 185, 25–36. doi:10.1086/338004

Rudy, J. W. (2009). Context representations, context functions, and the parahippocampal-hippocampal system. *Learn Mem.* 16 (10), 573–585. doi:10.1101/lm.1494409

Santillan, M., and Mackey, M. C. (2001). Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data. *Proc. Natl. Acad. Sci.* 98, 1364–1369. doi:10.1073/pnas.98.4.136410.1371/journal.pone.0024704

Sinha, S. (1988). Theoretical study of tryptophan operon: Application in microbial technology. *Biotechnol. Bioeng.* 31, 117–124. doi:10.1002/bit.260310204

Soderberg, T. I. M. (2005). Biosynthesis of ribose-5-phosphate and erythrose-4-phosphate in archaea: A phylogenetic analysis of archaeal genomes. *Archaea* 1, 347–352. doi:10.1155/2005/314760

Spirin, V., Gelfand, M. S., Mironov, A. A., and Mirny, L. A. (2006). A metabolic network in the evolutionary context: Multiscale structure and modularity. *PNAS* 103 (23), 8774–8779. doi:10.1073/pnas.0510258103

Webby, C. J., Baker, H. M., Lott, J. S., Baker, E. N., and Parker, E. J. (2005). The Structure of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase from Mycobacterium tuberculosis reveals a common catalytic scaffold and ancestry for type I and type II enzymes. *J. Mol. Biol.* 354, 927–939. doi:10.1016/j.jmb.2005.09.093

Wu, X., Wang, X., and Lu, W. (2014). Genome-scale reconstruction of a metabolic network for Gluconobacter oxydans 621H. *Biosystems* 117, 10–14. doi:10.1016/j.biosystems.2014.01.001

Xie, G., Keyhani, N. O., Bonner, C. A., and Jensen, R. A. (2003). Ancient Origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol. Mol. Biol. Rev.* 67, 303–342. doi:10.1128/mmbr.67.3.303-342.2003

Yanofsky, C., Paluh, J. L., van Cleemput, M., and Horn, V. (1987). Fusion of trpB and trpA of Escherichia coli yields a partially active tryptophan synthetase polypeptide. *J. Biol. Chem.* 262, 11584–11590. doi:10.1016/s0021-9258(18)60848-8

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fgene.2023.1097381

OPEN ACCESS

EDITED BY
Mallana Gowdra Mallikarjuna,
Indian Agricultural Research Institute (ICAR), India

REVIEWED BY
Hou-Ling Wang,
Beijing Forestry University, China
Rajesh Kumar Singh,
Institute of Himalayan Bioresource Technology (CSIR), India

*CORRESPONDENCE
Tingting Zhang,
zting@shzu.edu.cn
Lei Ma,
malei1979@hotmail.com

RECEIVED 13 November 2022
ACCEPTED 06 April 2023
PUBLISHED 17 April 2023

CITATION
Zhang Z, Zhang T and Ma L (2023), Analysis of basic pentacysteine6 transcription factor involved in abiotic stress response in *Arabidopsis thaliana*. *Front. Genet.* 14:1097381. doi: 10.3389/fgene.2023.1097381

Analysis of basic pentacysteine6 transcription factor involved in abiotic stress response in *Arabidopsis thaliana*

Zhijun Zhang, Tingting Zhang* and Lei Ma*

College of Life Science, Shihezi University, Shihezi City, Xinjiang, China

Background: Abiotic stress is a significant environmental factor that limits plant growth. Plants have complex and diverse mechanisms for dealing with abiotic stress, and different response mechanisms are interconnected. Our research aims to find key transcription factors that can respond to multiple non -biological stress.

Methods: We used gene expression profile data of Arabidopsis in response to abiotic stress, constructed a weighted gene co-expression network, to obtain key modules in the network. The functions and pathways involved in these modules were further explored by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. Through the enrichment analysis of transcription factor, the transcription factor that plays an important regulatory role in the key module. Through gene difference expression analysis and building protein interaction networks, the important role of key transcription factors is verified.

Result: In weighted gene co-expression network, identified three gene modules that are primarily associated with cold stress, heat stress, and salt stress. Functional enrichment analysis indicated that the genes in these modules participate in biological processes such as protein binding, stress response, and others. Transcription factor enrichment analysis revealed that the transcription factor Basic Pentacysteine6 (BPC6) plays a crucial regulatory role in these three modules. The expression of the BPC6 gene is dramatically affected under a variety of abiotic stress treatments, according to an analysis of Arabidopsis gene expression data under abiotic stress treatments. Differential expression analysis showed that there were 57 differentially expressed genes in bpc4 bpc6 double mutant Arabidopsis relative to normal Arabidopsis samples, including 14 BPC6 target genes. Protein interaction network analysis indicated that the differentially expressed genes had strong interactions with BPC6 target genes within the key modules.

Conclusion: Our findings reveal that the BPC6 transcription factor plays a key regulatory function in Arabidopsis coping with a variety of abiotic stresses, which opens up new ideas and perspectives for us to understand the mechanism of plants coping with abiotic stresses.

KEYWORDS

arabidopsis, BPC, abiotic stress, transcription factor, enrichment of function

1 Introduction

Adverse environmental factors, such as abiotic stress, severely limit agricultural production, reduce crop yield and quality, and affect plant growth and development, thereby threatening food security. Extreme temperatures and soil salinity are common extreme environmental conditions in nature, and climate change further complicates these adverse factors (GONG et al., 2020). Therefore, it is crucial to understand the response mechanism of plants to abiotic stress.

Arabidopsis is widely used as a model organism for research in plant genetics, developmental biology and molecular biology. And it is an ideal experimental material for exploring the mechanisms of plant response to abiotic stresses (KILIAN et al., 2007). In previous studies, some important *Arabidopsis* genes and metabolic pathways have been shown that they are involved in the process to respond to abiotic stress with *Arabidopsis*. For example, the *C-repeat Binding Transcription Factor3* (*CBF3*) transcription factor plays a key role in the cold response pathway (HE et al., 2008). Cold-inducible RNA helicase Regulator of *CBF* gene expression1 (*RCF1*) regulates cold-responsive genes and enhances the cold tolerance of plants by clipping pre-mRNA (GUAN et al., 2013). Exogenous application of jasmonate significantly enhances *Arabidopsis* freezing tolerance (HU et al., 2013). Humic acid (HA) significantly induces *Heat Shock Protein-encoding* (*HSP*) genes, including *HSP101*, *HSP81.1*, *HSP26.5*, *HSP23.6*, and *HSP17.6A*, which promotes heat tolerance in *Arabidopsis* (CHA et al., 2020). The *Arabidopsis Temperature-Induced Lipocalins1* (*TIL1*) gene (*AT5G58070*) is an important component of thermotolerance (CHI et al., 2009). Sanguinarine affects heat tolerance in *Arabidopsis* by enhancing the expression of heat shock protein genes such as *HSP17. 6C-CI*, *HSP70*, and *HSP90.1* (HARA and KURITA, 2014). The *Arabidopsis* histone acetyltransferase *General Control Non-Derepressible5* (*GCN5*) is also an important component of *Arabidopsis* thermotolerance (HU et al., 2015). The *Arabidopsis* regulator *RCF2*, expressed by the *C-repeat Binding Factor* (*CBF*) gene, has been shown to be an integrator of hyperthermia signaling and a mechanism of *Heat Stress Transcription Factor* (*HSF*) and *HSP* activation (GUAN et al., 2014). Overexpression of *Arabidopsis Stress-Induced BTB Protein 1* (*SIBP1*) genes increases salt tolerance in transgenic *Arabidopsis* (WAN et al., 2019). MADS-box transcription factor *Agamous-Linke16* (*AGL16*) acts as a negative regulator in stress response in *Arabidopsis*. The absence of *AGL16* makes *Arabidopsis* resistant to salt stress (ZHAO et al., 2021). Different members of the *Phosphoglycerate Dehydrogenase* (*PGHD*) gene family have different effects on salt tolerance in *Arabidopsis*, and the response to salt stress depends on the specific gene (ROSA-TELLEZ et al., 2020). Many genes or pathways are also involved in the response to multiple abiotic stresses in *Arabidopsis*. Overexpression of *Cysteine2/Histidine2* (*C2H2*)-Type Zinc Finger of *Arabidopsis Thaliana6* (*ATZAT6*) in *Arabidopsis* can increase resistance to pathogen infection, salt, drought and freeze stress (SHI et al., 2014). DNA methylation is also an important mechanism to regulate abiotic stress resistance in plants (OGNEVA et al., 2019).

Despite the fact that these studies have discovered numerous genes and biological processes in response to abiotic stress, most of these studies have focused on the link between a single gene and a single abiotic stress scenario. However, in nature, stress conditions are frequently layered on a range of unfavorable environmental circumstances. Therefore, the biological processes by which plants respond to different abiotic stresses are not completely independent. Responses to various abiotic stresses are both independent and highly interrelated. Plants possess genes that can respond to several different abiotic stressors simultaneously. To improve plant yield and quality and expand agricultural production, research must be conducted on genes that can adapt to multiple abiotic stresses.

In this study, we retrieved expression data from *Arabidopsis* plants that were subjected to abiotic stress treatments to identify genes that respond to these stressors. We then used the Weighted Gene Co-expression Network Analysis (WGCNA) method to identify the gene modules that are mostly related to cold stress, heat stress, and salt stress. We used a comprehensive bioinformatics approach to analyze the molecular function, signaling pathway, and transcription factor enrichment results of the modules. Finally, we identified a transcription factor, *BPC6* (Basic Pentacysteine), that is highly related to these three stresses. The expression data of *BPC6* under abiotic stress showed that it is involved in *Arabidopsis* responding to various abiotic stresses. Our study will improve the understanding of plant abiotic stress response mechanisms and may play an important role in improving plant yield and quality and promoting agricultural production.

2 Result

2.1 Data pre-processing

The 18 gene expression profile data associated with abiotic stress in *Arabidopsis* were preprocessed, and only wild-type *Arabidopsis* samples from all datasets were retained, for a total of 97 samples. All data were normalized. Each dataset contained 20,642 genes. After removing batch effects and putting the 18 datasets together into a new matrix file, all 54 stress-related samples from the new matrix file were selected and control samples were removed. The results of an analysis of Median Absolute Deviation (MAD) are shown in Supplementary Table S1. For building the weighted gene co-expression network, the top 10,000 genes with the most variable expression levels were selected as input genes (mad $\geq$ 0.390897091). The clustering analysis showed that the 54 *Arabidopsis* samples were close to each other, with no significant outliers (Figure 1), and the overall effect was good.

2.2 Weighted gene co-expression network analysis

To build the scale-free network, we optimized the appropriate network weighting coefficient β. The β was calculated using the "pick Soft Threshold" function of the WGCNA package. When the threshold β was set to 3, the topology analysis showed that the scale-free topology fitting index (R2) was close to 90% (Supplementary Figure S1), indicates that the network was close to being a scale-free network. We established a co-expression network with a soft threshold β of 3. The genes with similar expression patterns were grouped into modules in the network, and a total of seven modules were identified (Figure 2). For

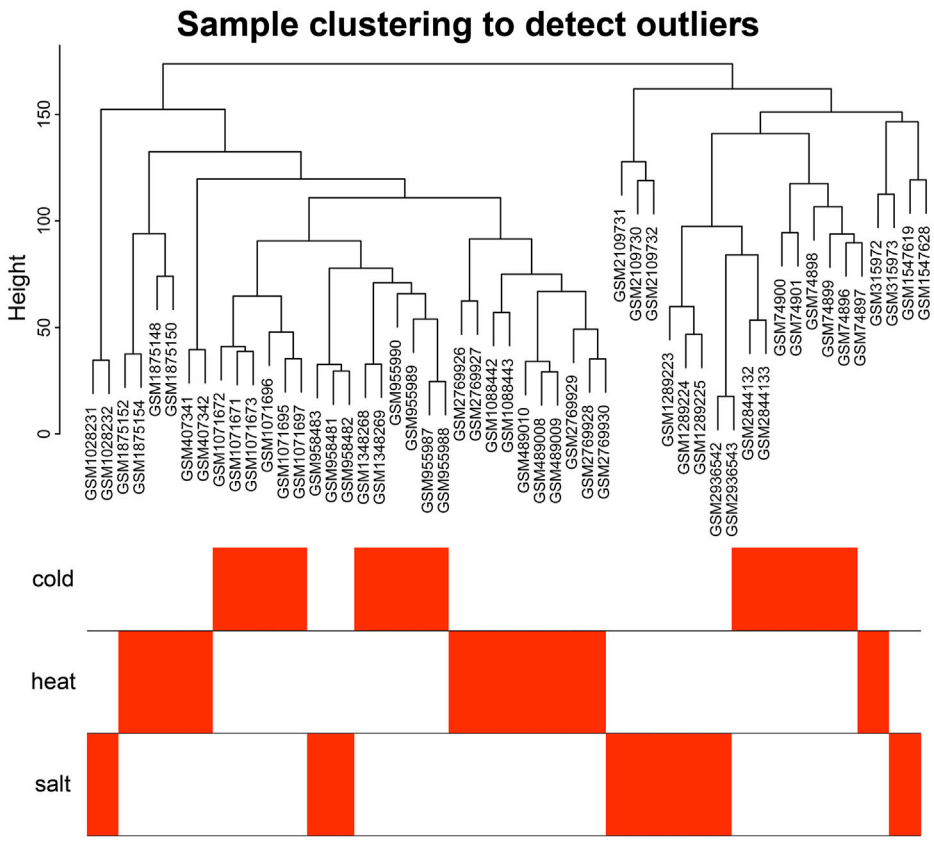


FIGURE 1
Sample cluster analysis diagram. Sample clustering analysis showed that no samples were outliers.

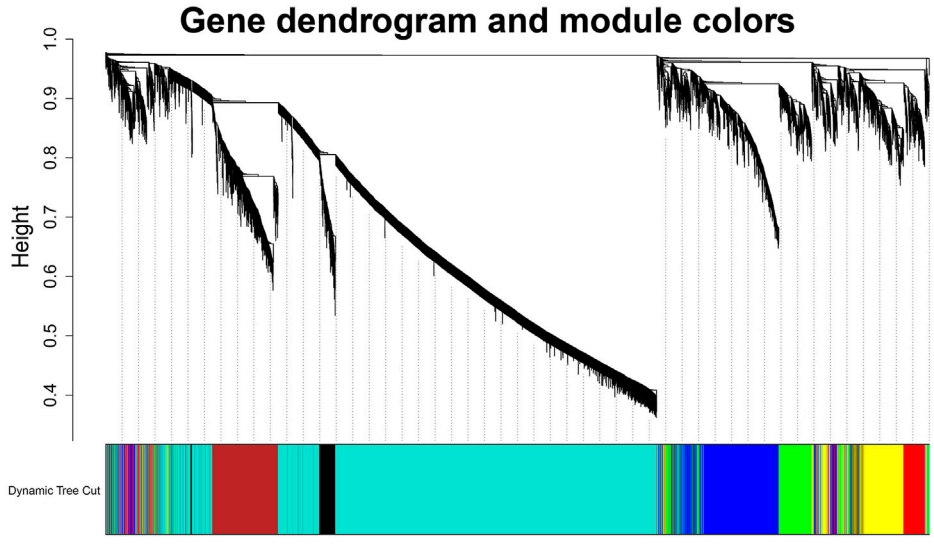


FIGURE 2
Identification diagram of gene co-expression module. Identification of gene co-expression modules via hierarchical average linkage clustering. The color row underneath the dendrogram shows the module assignment determined by the Dynamic Tree Cut.

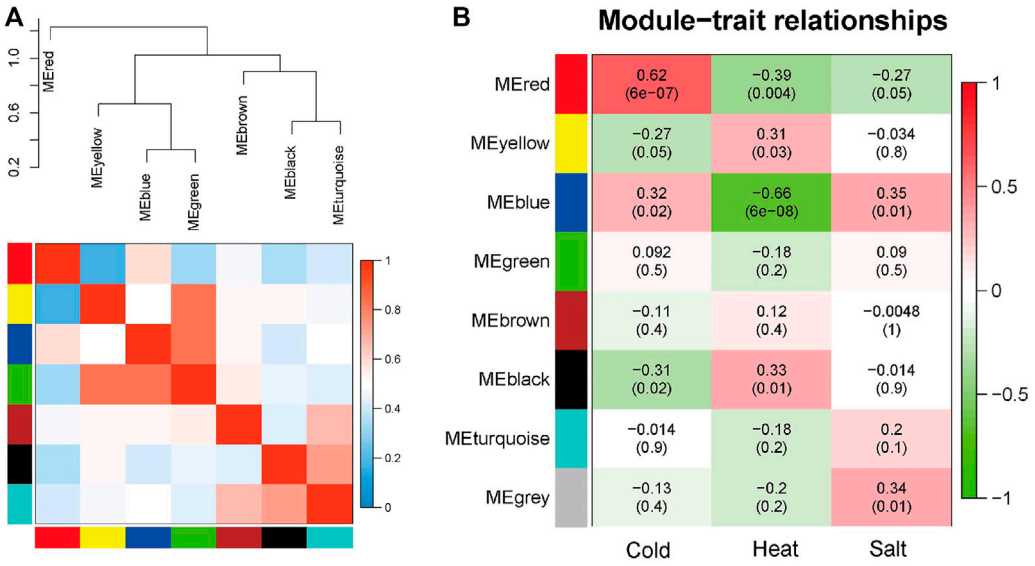


FIGURE 3
Hierarchical clustering dendrogram of module eigengenes and heatmap plot of the adjacencies in the eigengene network (labeled by their colors). **(A)** In the heatmap, the green color represents low adjacency (negative correlation), while a red represents high adjacency (positive correlation). **(B)** Correlation between sample grouping and gene modules. Each row of the table corresponds to a gene module, and each column corresponds to a group.

visualization, modules were named with colors: Black (258 genes), Blue (1,415 genes), Brown (882 genes), Green (654 genes), Red (571 genes), Turquoise (5,392 genes), and Yellow (827 genes).

The seven modules are primarily divided into three clusters (Figure 3A). Compared to the other modules (Figure 3B), the red module is most closely associated with cold stress, the black module is most closely associated with heat stress, and the blue module is most relevant to salt stress. The results demonstrate that the red (MM = 0.62, p = 6e-7), black (MM = 0.33, p = 0.01), and blue (MM = 0.35, p = 0.01) modules play crucial roles in the *Arabidopsis* response to abiotic stress. Therefore, these three modules are identified as the key modules.

2.3 Functional enrichment analysis of key modules

To better understand the biological functions of genes in key modules, the red, black, and blue modules were analyzed for GO function enrichment and KEGG pathway enrichment. The 93 GO terms are significantly enriched in the red module (Figure 4A). For biological processes, genes are mainly concentrated in the response to water shortage, abscisic acid and light stimulation, and signal transduction. For cellular components, genes are mainly enriched in membrane components, the plasma membrane, and cytoplasm. For molecular functions, genes are mainly enriched in protein binding.

The 24 GO terms are significantly enriched in the black module (Figure 4B). For biological processes, genes are mainly enriched in cell differentiation, root development, tissue development, defense responses, and plant epidermis development. For cellular components, genes were mainly enriched in the extracellular region. For molecular functions, genes were mainly enriched in protein transport, transcription factor activity, and sequence-specific DNA binding.

The 178 GO terms are significantly enriched in the blue module (Figure 4C). For biological processes, genes were mainly enriched in response to water shortage, defense response to bacteria, response to injury, and defense response to fungi. For cellular components, genes were mainly enriched in membrane components, the plasma membrane, extracellular region, and Golgi apparatus. For molecular functions, genes were mainly enriched in protein binding, transcription factor activity, sequence-specific DNA binding, and transcriptional regulatory region sequence specific DNA binding.

According to the KEGG pathway analysis, the red module is primarily involved in metabolic pathways (Figure 4D), the black module is mainly associated with phenylpropionic acid biosynthesis (Figure 4E), and the blue module is mainly involved in metabolic pathways and secondary metabolite synthesis (Figure 4F).

2.4 Transcription factor enrichment analysis of key modules

To further investigate the common biological mechanisms behind the three key modules responding to abiotic stress, transcription factor enrichment analysis of genes in the three key modules was performed using transcription factor enrichment in PlantTFDB (5.0). Figure 5A shows the enrichment results of the red module, Figure 5B shows the enrichment results of the black module, and Figure 5C shows the enrichment results of the blue module. Supplementary Tables S2–S4 show the regulatory relationship between genes and transcription factors in the red, black, and blue modules, respectively. The intersection of the three enrichment results (Figure 5D) showed that the *BPC6* transcription factor synthesized by the *AT5G42520* gene played a key regulatory role in all three modules simultaneously.

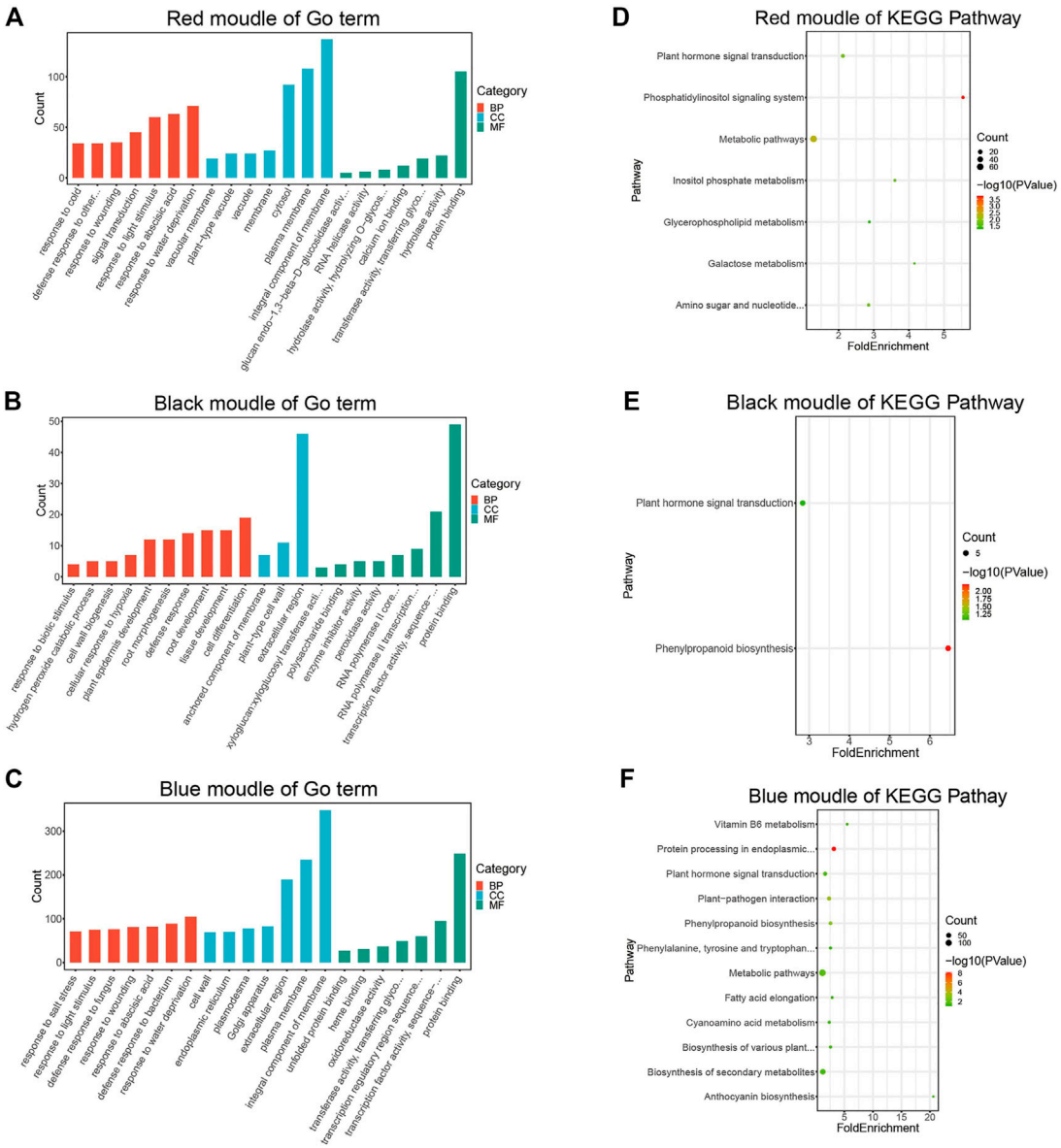


FIGURE 4
GO and KEGG enrichment analysis diagram. **(A)** GO Enrichment results of genes in the red module. **(B)** GO enrichment results of genes in the black module. **(C)** Enrichment results of genes in the blue module. **(D)** KEGG pathway enrichment results of genes in the red module. **(E)** KEGG pathway enrichment results of genes in the black module. **(F)** KEGG pathway enrichment results of genes in the blue module.

2.5 Expression of the *BPC6* gene in *Arabidopsis* under different abiotic stresses

To test whether the *BPC6* gene plays a key role in *Arabidopsis* responses to multiple abiotic stresses, we analyzed *Arabidopsis* expression profile data from the AtGenExpress project under different abiotic stresses and obtained the expression of the *BPC6* gene under different abiotic stresses (Figure 6). According to the expression profile data. The expression of the *BPC6* gene in *Arabidopsis* decreased significantly during the continuous cold stress period of 4°C. After 6 h of stress, the expression of the *BPC6* gene in plants began to increase significantly, and the change of expression in leaves was more significant. The expression level of the *BPC6* gene in the root continued to decrease initially, but after 6 h, the expression level began to increase and returned to the level before the stress.

During the 38°C/3 h heat stress period, the expression level of the *Arabidopsis BPC6* gene slowly decreased. After the stress was stopped, the expression level first increased, then decreased, then increased again, and then decreased once more. At 24 h, the expression level of the *BPC6* gene in the leaves was still low, while the expression level in roots returned to the level before the heat stress.

During the 150 mM/L NaCl salt stress, the expression of the *BPC6* gene in *Arabidopsis* firstly decreased within half an hour of salt stress, then increased within 1 h, then decreased within 6 h, and then increased within 12 h. After 24 h, the expression of the *BPC6* gene in leaves recovered to the pre-stress level, but the expression in roots was

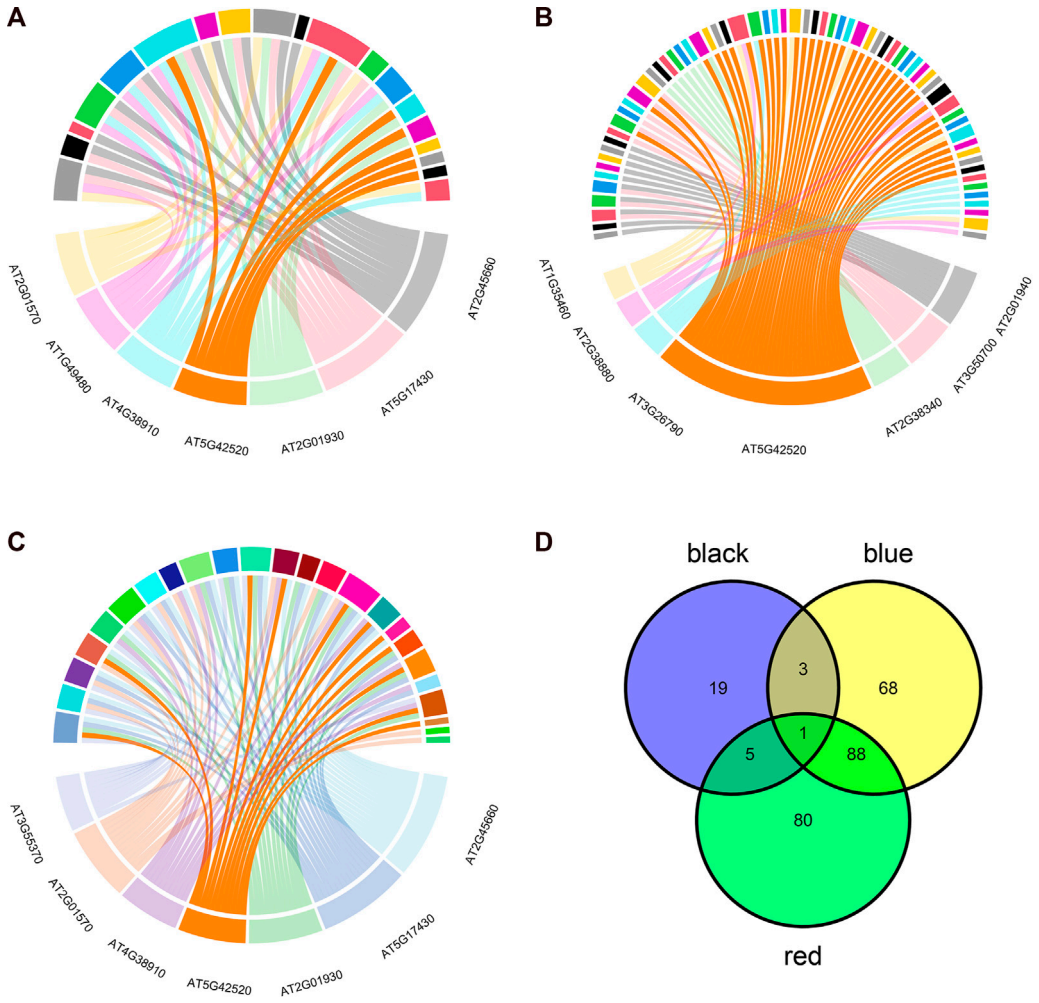


FIGURE 5
Transcription factor enrichment analysis results. **(A)** Transcription factor enrichment results of genes in the red module. **(B)** Transcription factor enrichment results of genes in the black module. **(C)** Transcription factor enrichment results of genes in the blue module. **(D)** Transcription factor enrichment results are intersected.

significantly lower than that before stress. These results indicated that the *Arabidopsis BPC6* gene is involved in the *Arabidopsis* response to various abiotic stresses, which confirms our data analysis results.

2.6 Effect of BPC mutant

After comparing *bpc4 bpc6* double mutant *Arabidopsis* with normal samples, a total of 57 genes were differentially expressed (DEG), with 27 genes being downregulated and 30 genes being upregulated (refer to Supplementary Figure S2, Supplementary Table S7). Out of the 30 upregulated genes, five were identified as *BPC6* target genes according to the plantregmap database, accounting for 20% of all upregulated genes. Similarly, eight of the 27 downregulated genes were *BPC6* target genes, accounting for 29.6% of all downregulated genes (Figure 7B). However, the expression of genes regulating *BPC6* remained unchanged (Figure 7A). These findings suggest that a substantial proportion of the DEGs were target genes regulated by *BPC6*, underscoring the critical role of the *BPC* gene in modulating the expression of these genes.

To ensure the accuracy of our findings, we analyzed the Protein-Protein Interaction (PPI) network from both a global and regional perspective. The results of the PPI network analysis of DEGs and target genes controlled by *BPC6* in key modules revealed a significant overall interaction link between these genes (Figure 8A). The PPI network has 48 DEGs, accounting for 84% of the total differential genes. The PPI network contains 157 of the 188 *BPC6* target genes in the red module, accounting for 84% of the total. The PPI network contains 59 of the 80 *BPC6* target genes in the black module, accounting for 74% of the total. The PPI network contains 337 of the 431 *BPC6* target genes in the blue module, accounting for 78% of the total.

PPI network analysis was performed on DEGs and the different key modules. The PPI network analysis results of the red module and 57 DEGs showed that 115 (64%) of the 180 target genes had obvious interactions with 42 (74%) DEGs (Figure 8B). The PPI network analysis results of the black module and

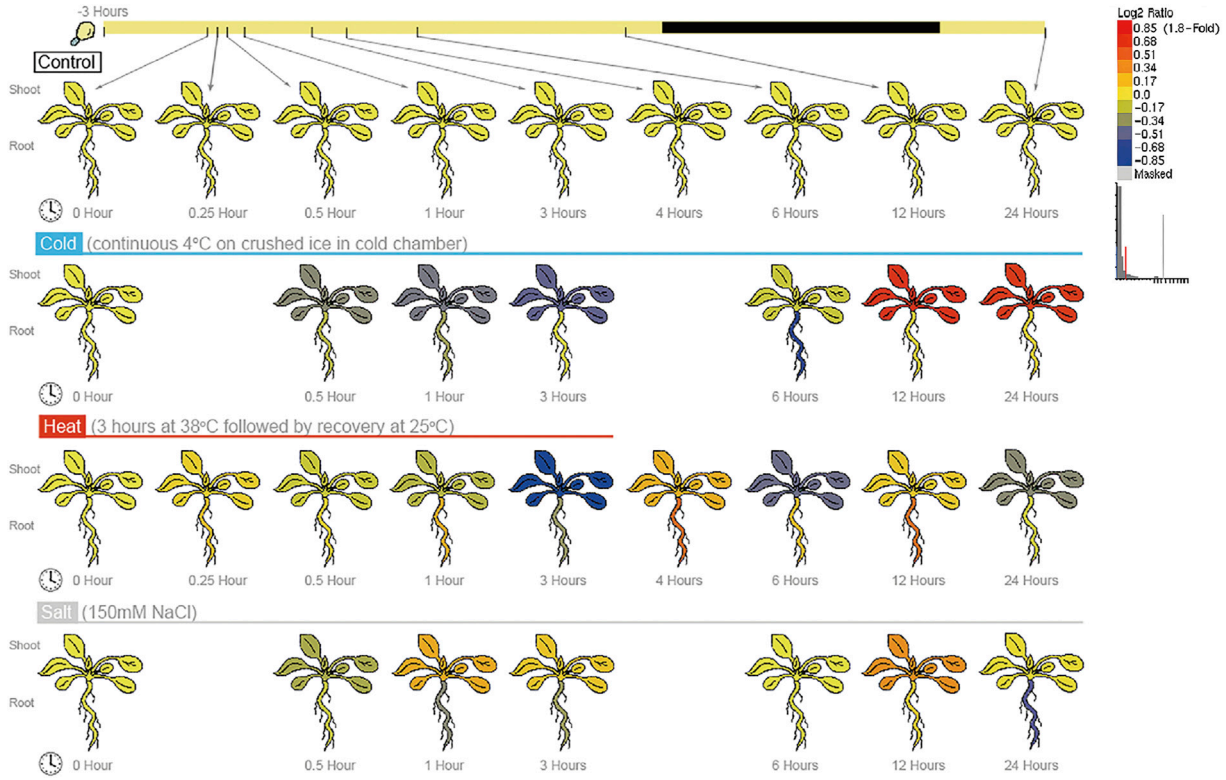


FIGURE 6
Expression of BPC6 gene under different abiotic stresses.

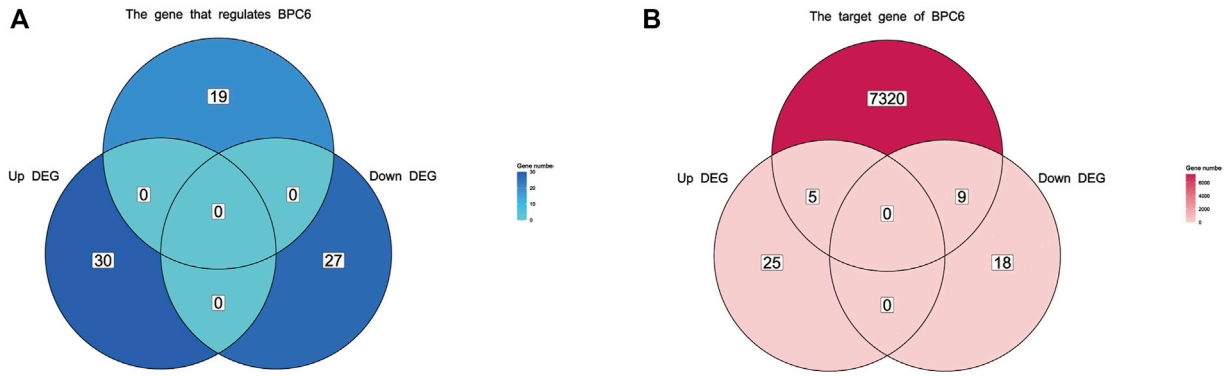


FIGURE 7
Venn diagram of differentially expressed genes and BPC upstream regulatory genes and downstream target genes. **(A)** Venn diagram of differentially expressed genes and upstream regulatory genes of BPC. **(B)** Venn diagram of differentially expressed genes and target genes regulated by BPC.

57 DEGs showed that there were obvious interactions between 42 (53%) of the 80 target genes and 38 (67%) DEGs (Figure 8C). The PPI network analysis results of the blue module and 57 DEGs showed that 313 (73%) of the 431 target genes had obvious interactions with 46 (81%) DEGs (Figure 8D). For different key modules, most of the target genes were regulated by *BPC6* and most of the DEGs had obvious interactions.

3 Discussion

The ever-increasing global population and the hard-to-increase arable land have aggravated the negative impact on human survival. The best solution to this problem is to increase crop yield per unit area. However, abiotic stress has a strong negative impact on plant growth and crop yield. Abiotic stress factors such as extreme

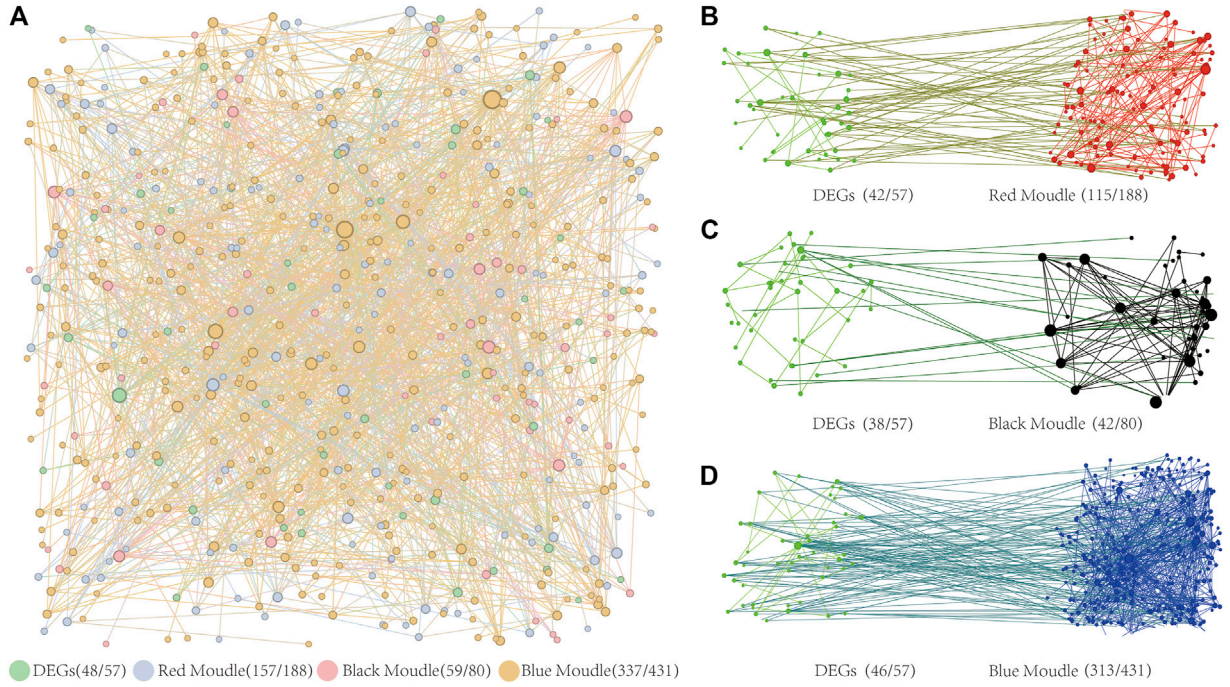


FIGURE 8
PPI Network Diagram. **(A)** PPI network diagram of three key modules and differentially expressed genes. **(B)** The PPI network diagram of target genes regulated by BPC6 and differentially expressed genes in the red module. **(C)** The PPI network diagram of target genes regulated by BPC6 and differentially expressed genes in the black module. **(D)** The PPI network diagram of target genes regulated by BPC6 and differentially expressed genes in the blue module.

temperatures and soil salinization seriously affect crop production every year, making it significant to improve plant tolerance to abiotic stress. As a model plant, *Arabidopsis* exhibits strong adaptability to environmental stress and is widely used to study various abiotic stress response mechanisms. Most of the previous work has focused on studying *Arabidopsis* response mechanisms to a single stress. However, many extreme climatic conditions occur simultaneously in nature, and the mechanisms by which plants responding to different stresses are not independent of each other. At present, the shared response mechanisms of plants to cope with multiple abiotic stresses are unclear. The *BPC6* transcription factor is an important regulatory transcription factor, and studying its mechanism of participation in coping with abiotic stress is significant.

In this study, we constructed a weighted gene co-expression network using *Arabidopsis* gene expression data and identified three modules that were most associated with cold, heat, and salt stress. The GO enrichment analysis showed that the blue module was mainly involved in the response to water shortage, and had a superior response to bacteria and fungi. The red module was mainly involved in the response to water shortage, abscisic acid, and so on. The black module was mainly involved in cell differentiation, plant development, protein transport, and transcription factor activity. The KEGG pathway enrichment analysis showed that the blue and red modules were mainly involved in the metabolic pathway, while the black module was mainly involved in the phenyl-propionic acid synthesis pathway. The three modules were then enriched for transcription factors, and the results showed that most of the genes in the three modules were simultaneously regulated by the *BPC6* transcription factor. The expression of the *BPC6* gene in *Arabidopsis* was analyzed under different abiotic stresses, and the results showed that the expression of the *BPC6* gene changed significantly under different abiotic stresses.

The DEGs between the *bpc4 bpc6* double mutant *Arabidopsis* samples and normal samples were analyzed. The differential expression analysis shows that compared with normal samples, the *bpc4 bpc6* double mutant *Arabidopsis* must have a total of 57 DEGs. Sequence analysis showed that the *BPC* gene family has a total of 7 genes in *Arabidopsis*, which are divided into three classes: class I proteins *BPC1* (*AT2G01930*), *BPC2* (*AT1G14685*) and *BPC3* (*AT1G68120*); class II proteins *BPC4* (*AT2G21240*), *BPC5* (*AT4G38910*) and *BPC6* (*AT5G42520*); and class III protein *BPC7* (*AT2G35550*). They are all ubiquitously expressed transcriptional activators and repressors, except for *BPC5*, which is considered a pseudogene (MEISTER et al., 2004). There is functional overlap between different classes. Single gene mutations do not produce visible phenotypic effects, and severe morphological phenotypes occur only in higher-order mutants between class I and class II members. Therefore, to study the function of the *BPC6* gene through gene mutant, it is necessary to knock out all *BPC* genes that are similar to the *BPC6* gene function. The control group samples and gene mutant samples used in sequencing are cultivated in normal environments without coercion. Therefore, the only reason for generating differential genes is the mutant of the *BPC4* and *BPC6* genes. The number

of different genes is very small. It may be due to the overlapping function of other unintended *BPC* genes with *BPC6*, which caused the physiological biochemical activity of gene knocking *Arabidopsis* not be significantly affected. There are 13 genes among the DEGs that are direct or indirect targets of the *BPC6* transcription factor, accounting for 23% of the total number of different genes. The PPI network analysis of the *BPC6* target genes and DEGs in the key modules can be seen that most of the differences can have a strong interaction with most of the differential genes in the key module in the key module. This proves that the analysis results of the weighted gene co-expression network.

Cytokinin plays an important role in plant growth and development and also participates in the response process of plants to non-biological stress. Research indicates that cytokinins can regulate ion channels, antioxidant enzyme activity, protect chlorophyll and cell membrane stability, and modulate the balance of hormones in plants. The promotes the growth and differentiation of roots, thus increasing plant adaptability to abiotic stress (Sabagh et al., 2021). In addition, cytokinins can also regulate plant abiotic stress responses by interacting with other signaling molecules such as ABA, SA, and ROS (GUPTA and HUANG, 2014; GAO et al., 2019). The type-B *Arabidopsis* response regulator (ARR) transcription factors have emerged as primary targets of cytokinin signaling and are required for essentially all cytokinin-mediated changes in gene expression. By cooperating with other transcription factors, ARR can affect the process and effect of cytokinin in plants (ARGUESO et al., 2010). *BPC* transcription factors are a potential set of coregulators regulating cytokinin responses. Disruption of multiple *BPC* genes in *Arabidopsis* thaliana reduces its sensitivity to cytokinin. Further, a significant number of *BPC6* regulated genes are also direct targets of the type-B ARRs (SHANKS et al., 2018). Therefore, cytokinin is likely to be a key substance involved in *Arabidopsis*'s response to abiotic stress by the *BPC6* transcription factor.

The *BPC* transcription factor family plays a crucial role in regulating gene expression in plants. These proteins are located in the nucleus and regulates the transcription process by specifically binding to the GA dinucleotide repeat sequence of the gene. *BPC* proteins were first discovered in barley in 2003 (SANTI et al., 2003), and subsequently in *Arabidopsis* in 2004 (MEISTER et al., 2004). *BPC* genes have a broad expression pattern in *Arabidopsis*, more than 3,000 *Arabidopsis* genes contain at least one GA-rich segment in their regulatory region. BPC transcription factors are essential for normal plant growth and development. The *Arabidopsis BPC1* transcription factor has been shown to bind to a GA-rich consensus sequence in the Seedstick (*STK*) promoter *in vitro*, and this binding induces conformational changes. Vivo *BPCs* also bind to the consensus boxes, and when these were mutated, expression from the *STK* promoter was derepressed, resulting in ectopic expression in the inflorescence. GA consensus sequences in the *STK* promoter to which *BPCs* bind are essential for the recruitment of the corepressor complex to this promoter (SIMONINI et al., 2012). *Shootmeristemless* (*STM*) and *Brevipedicellus/Knat1* (*BP*) genes are both direct targets of *BPCs*, and *BPC* transcription factors also play an important role in the fine regulation of cytokinin content in meristem (SIMONINI and KATER, 2014). *BPC6* can interact with two *Arabidopsis Polycomb-Repressive Complexes* (*PRC1.PRC2*) to affect the expression of a large number of genes (HECKER et al., 2015). *BPCs* can bind to the promoter of transcription factors *Abscisic Acid Insensitive4* (*AAI4*), inhibit the expression of *ABI4* in roots, and promote lateral root (LR) development in *Arabidopsis* (MU et al., 2017). *BPCs* also significantly affect the function of cytokinins in *Arabidopsis*, and disruption of multiple *BPCs* in *Arabidopsis* results in reduced sensitivity to cytokinins (SHANKS et al., 2018). *BPCs* may also promote *Arabidopsis* ovule and seed development by limiting the transcription of *Fusca3* (*FUS3*) (ROSCOE et al., 2019; CARELLA, 2020). Class I *BPC* works by directly binding to the GA/CT cis-element in *FUS3* and limiting its expression (Wu et al., 2020).

Previous studies have demonstrated that the *BPC* transcription factor family plays an important role in regulating plant growth and development. However, from the perspective of abiotic stress, our study expounds a brand-new research result, that is, the *BPC6* transcription factor is involved in the process of plants responding to various abiotic stresses. Compared with previous studies, our advantages lie in the large sample size, abundant data, novel research angles, and diverse research methods. We illustrate new findings with existing data.

Base on further discussion of the results of this study, the following points can be paid attention to: First of all, the specificity and regulatory mechanism of *BPC6* transcription factor in various abiotic stresses can be further explored. Additionally, the interactions and regulatory networks between *BPC6* and other regulators can be studied to gain a deeper understanding of its role in plant abiotic stress responses.

Second, this study found that the blue and red modules were mainly involved in the metabolic pathway, and the black module was mainly involved in the phenylpropionic acid synthesis pathway. This suggests that metabolic and synthetic pathways have important roles in plant responses to abiotic stresses. Therefore, future studies can further focus on the key genes and regulatory mechanisms in these pathways to better understand the physiological and metabolic regulation of plants under abiotic stress.

In addition, the results of this study indicated that different modules have different response characteristics to abiotic stresses. For example, blue modules are mainly involved in the response to water deprivation, while red modules are mainly involved in the response to abscisic acid. This suggests that plants may require different adaptive mechanisms for optimal growth and survival in response to different abiotic stresses. Therefore, future studies could delve deeper into these adaptive mechanisms and response traits to guide plant breeding and planting practices to improve plant adaptability and stress resistance.

Finally, the results of this study demonstrate that key genes and regulatory mechanisms in plants under abiotic stress can be effectively identified using the WGCNA approach. Therefore, future research can apply the WGCNA method to more plant species and different types of abiotic stress to establish a more comprehensive and accurate plant abiotic stress response network, and provide a more scientific basis for plant breeding and cultivation. In addition, by combining other bioinformatics methods, such as gene expression profiling and functional annotation, deeper information and mechanisms of the abiotic stress response network can be further explored. At the same time, since *Arabidopsis* is a model organism, our research results can also guide the study of other plants, which is of great significance for agricultural production and food security.

4 Conclusion

The *BPC* transcription factor family is very important in plants and can regulate various plant growth and development processes. From the perspective of abiotic stress, this study explored the role of the *BPC6* transcription factor in *Arabidopsis* response to abiotic stress. It confirmed that *Arabidopsis BPC6* transcription factor can participate in coping with various abiotic stresses by regulating the expression of many genes. Analysis of *Arabidopsis* gene expression data validated this result. This study proves that the biological processes of *Arabidopsis* in response to different abiotic stresses are not isolated, but have commonality at the level of transcription factors. This work provides new ideas and perspectives for the study of plant responses to abiotic stress.

5 Materials and methods

5.1 Data acquisition

In order to study the mechanism of *Arabidopsis* response to abiotic stress, we searched the GEO (https://www.ncbi.nlm.nih.gov/geo/) database using "*Arabidopsis*" as the keyword. To query gene expression profiles associated with abiotic stress in *Arabidopsis*, we downloaded 18 groups of gene expression profiles related to abiotic stress. These included 6 groups related to salt stress, 6 groups related to heat stress, and 6 groups related to cold stress. We only retained wild-type *Arabidopsis* expression data in the gene expression profiles, resulting in a total of 97 samples. The detailed information of all gene expression profiles is shown in Supplementary Table S8. We also searched the GEO database using "*Arabidopsis*" and "*BPC*" as keywords and obtained gene expression data of *Arabidopsis* thaliana with *BPC* gene mutant (GSE68437). This data set contains eight samples, of which two *bpc4 bpc6* double mutant *Arabidopsis* samples and two control samples were retained. All data used whole plants as material to be sequenced.

5.2 Data pre-processing

Gene expression profiles were downloaded in TXT format from the GEO database. The R software package was used to process the matrix files and filter out low-quality data. The probe ID was converted to a gene symbol, invalid expression data were deleted, and the expression data of duplicate gene symbols were averaged. The expression profiles without log2 transformation were log2 transformed using R language. We used the combat package to remove batch effects from all expression profiles, and merged them into a matrix file. The expression data from all stress-treated *Arabidopsis* samples in the matrix file were merged into a new matrix file. Subsequently, we performed WGCNA using the new matrix files containing only stress-treated *Arabidopsis* samples. The GSE68437 dataset was used for gene differential expression analysis.

5.3 Weighted gene co-expression network analysis

In statistics, the median absolute deviation (MAD) is a robust measure of sample bias on univariate numerical data. At the same time, it can also represent the population parameters estimated by the MAD of the sample. We used the MAD algorithm to select the expression data of the top 10,000 genes as input data for WGCNA.

WGCNA is regarded as a methodology to reconstruct a free-scale gene co-expression network and concurrently identify modules consisting of highly correlated genes to appraise connectivity between external clinical traits and the module. Eigengene is used for summarizing relationships among internal gene membership. In this study, we applied the one-step network construction and module detection function of the WGCNA package (https://horvath.genetics.ucla.edu/html/Co-expressionNetwork/Rpackages/WGCNA/Tutorials/) in R to handle the analysis of the expression profiles of *Arabidopsis*, which contained 20 cold-treated samples, 18 heat-treated samples, and 15 NaCl-treated samples. We correlated gene clusters with each other and external sample features. The weighted adjacency matrix was calculated to represent the connection strength of each pair of genes. According to the scale-free topology network, the soft thresholding power was set to 4. Then, a hierarchical clustering dendrogram composed of rich branches was established. The dynamic tree-cutting method was used to complete module identification, the minimum size of the gene dendrogram is 25, and the grouping information of samples is made by setting the value of 1 under stress and 0 under no stress as the grouping standard. Finally, modules were associated with groups using module-group associations based on Module Membership (MM) and Gene Salience (GS).

5.4 Identification of key modules

We evaluate the relationship between module and sample grouping by using the correlation between module eigengenes and sample grouping. When dealing with sample features, statistical significance measures between module feature genes and features can be defined. For example, using correlation values or *p*-values, modules with high feature significance values are considered to be associated with sample grouping.

5.5 Functional and pathway enrichment analysis of key modules

Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of genes in key modules were performed using the online DAVID (https://david.ncifcrf.gov/). The gene list of key modules was uploaded to the DAVID database to obtain the GO enrichment and KEGG pathway enrichment results. Results with $p < 0.05$ were considered significant, and the obtained enrichment analysis results were visualized using the ggplot2 package.

5.6 Transcription factor enrichment analysis

Transcription factor enrichment analysis of genes in key modules was performed using the online plantTFDB database (http://planttfdb.gao-lab.org/). The gene list of key modules was uploaded to the plantTFDB database, and the enrichment results of transcription factors of key modules were obtained. R language was used for subsequent analysis of transcription factor enrichment results.

5.7 Analysis of key gene expression

Arabidopsis eFP Browsers (http://bar.utoronto.ca/#GeneExpressionAndProteinTools) from the AtGenExpress project were used to analyze the expression profiles of *Arabidopsis* genes under different abiotic stresses, using the *Arabidopsis* eFP in the BAR database (KILIAN et al., 2007).

5.8 Gene expression data validation

To verify our data analysis results, we used the plantregmap database (http://plantregmap.gao-lab.org/) to obtain all target genes regulated by *BPC6* and all genes that regulate *BPC6* in *Arabidopsis*. We also used the limma package (https://bioconductor.org/packages/release/bioc/html/limma.html) to analyze the gene differential expression of the which two *bpc4 bpc6* double mutant *Arabidopsis* samples and normal samples in the GSE68437 dataset ($|\log_2^{FC}| > 2$, adj.$p < 0.05$). Additionally, we utilized the STRING (https://string-db.org/) to perform PPI networks analysis on DEGs and key genes within the three modules, and then used Gephi V0.10.1 to visualize the PPI network. Finally, we took the intersection of *BPC6*-related genes and DEGs to verify our data analysis results.

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

ZZ and LM contributed to the conception and design. LM and TZ provided administrative support. ZZ contributed to data collection, analysis and interpretation. ZZ, LM, and TZ contributed to manuscript writing and provided final approval of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The research was supported by the National Natural Science Foundation of China (32060300, 32060145, and 31860308), International Science and Technology Cooperation Project of Bingtuan (2020BC002), the China Scholarship Council and the Science Foundation of Shihezi University (RCZK201953). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1097381/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Soft threshold analysis result.

SUPPLEMENTARY FIGURE S2
Volcano plot of differentially expressed genes.

SUPPLEMENTARY TABLE S1
Median Absolute Deviation (MAD) analysis results.

SUPPLEMENTARY TABLE S2
Regulatory relationship between genes and transcription factors in red module.

SUPPLEMENTARY TABLE S3
Regulatory relationship between genes and transcription factors in black module.

SUPPLEMENTARY TABLE S4
Regulatory relationship between genes and transcription factors in blue module.

SUPPLEMENTARY TABLE S5
The gene list and regulatory relationship that can regulate BPC6 in Arabidopsis.

SUPPLEMENTARY TABLE S6
The target gene and regulatory relationship of BPC6 in Arabidopsis.

SUPPLEMENTARY TABLE S7
Results of gene differential expression analysis.

SUPPLEMENTARY TABLE S8
Detailed information about GEO data.

References

Argueso, C. T., Raines, T., and Kieber, J. J. (2010). Cytokinin signaling and transcriptional networks. *Curr. Opin. plant Biol.* 13 (5), 533–539. doi:10.1016/j.pbi.2010.08.006

Carella, P. (2020). Stop the FUSS: BPCs restrict *FUSCA3* transcription to promote ovule and seed development. *Plant Cell* 32 (6), 1779–1780. doi:10.1105/tpc.20.00295

Cha, J. Y., Kang, S. H., Ali, I., Lee, S. C., Ji, M. G., Jeong, S. Y., et al. (2020). Humic acid enhances heat stress tolerance via transcriptional activation of Heat-Shock Proteins in Arabidopsis. *Sci. Rep.* 10 (1), 15042. doi:10.1038/s41598-020-71701-8

Chi, W. T., Fung, R. W., Liu, H. C., Hsu, C. C., and Charng, Y. Y. (2009). Temperature-induced lipocalin is required for basal and acquired thermotolerance in Arabidopsis. *Plant Cell Environ.* 32 (7), 917–927. doi:10.1111/j.1365-3040.2009.01972.x

Gao, S., Xiao, Y., Xu, F., Gao, X., Cao, S., Zhang, F., et al. (2019). Cytokinin-dependent regulatory module underlies the maintenance of zinc nutrition in rice. *New Phytol.* 224 (1), 202–215. doi:10.1111/nph.15962

Gong, Z. Z., Xiong, L. M., Shi, H. Z., Yang, S., Herrera-Estrella, L. R., Xu, G., et al. (2020). Plant abiotic stress response and nutrient use efficiency. *Sci. China Life Sci.* 63 (5), 635–674. doi:10.1007/s11427-020-1683-x

Guan, Q., Wu, J., Zhang, Y., Jiang, C., Liu, R., Chai, C., et al. (2013). A DEAD box RNA helicase is critical for pre-mRNA splicing, cold-responsive gene regulation, and cold tolerance in Arabidopsis. *Plant Cell* 25 (1), 342–356. doi:10.1105/tpc.112.108340

Guan, Q., Yue, X., Zeng, H., and Zhu, J. (2014). The protein phosphatase RCF2 and its interacting partner NAC019 are critical for heat stress-responsive gene regulation and thermotolerance in Arabidopsis. *Plant Cell* 26 (1), 438–453. doi:10.1105/tpc.113.118927

Gupta, B., and Huang, B. (2014). Mechanism of salinity tolerance in plants: Physiological, biochemical, and molecular characterization [J]. *Int. J. Genomics* 2014 (1), 701596.

Hara, M., and Kurita, I. (2014). The natural alkaloid sanguinarine promotes the expression of heat shock protein genes in Arabidopsis. *Acta Physiol. Plant.* 36 (12), 3337–3343. doi:10.1007/s11738-014-1681-y

He, F., Jiuqing, K. A. N. G., Zhou, X., Su, Z., Qu, L., and Gu, H. (2008). Variation at the transcriptional level among Chinese natural populations of *Arabidopsis thaliana* in response to cold stress. *Chin. Sci. Bull.* 19, 2989–2999. doi:10.1007/s11434-008-0403-5

Hecker, A., Brand, L. H., Peter, S., Simoncello, N., Kilian, J., Harter, K., et al. (2015). The Arabidopsis GAGA-binding factor BASIC PENTACYSTEINE6 recruits the POLYCOMB-REPRESSIVE COMPLEX1 component LIKE HETEROCHROMATIN PROTEIN1 to GAGA DNA motifs. *Plant Physiol.* 168 (3), 1013–1024. doi:10.1104/pp.15.00409

Hu, Y., Jiang, L., Wang, F., and Yu, D. (2013). Jasmonate regulates the inducer of cbf expression-C-repeat binding factor/DRE binding factor1 cascade and freezing tolerance in Arabidopsis. *Plant Cell* 25 (8), 2907–2924. doi:10.1105/tpc.113.112631

Hu, Z., Song, N., Zheng, M., Liu, X., Liu, Z., Xing, J., et al. (2015). Histone acetyltransferase GCN5 is essential for heat stress-responsive gene activation and thermotolerance in Arabidopsis. *Plant J.* 84 (6), 1178–1191. doi:10.1111/tpj.13076

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., et al. (2007). The AtGenExpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J.* 50 (2), 347–363. doi:10.1111/j.1365-313X.2007.03052.x

Meister, R. J., Williams, L. A., Monfared, M. M., Gallagher, T. L., Kraft, E. A., Nelson, C. G., et al. (2004). Definition and interactions of a positive regulatory element of the Arabidopsis INNER NO OUTER promoter. *Plant J.* 37 (3), 426–438. doi:10.1046/j.1365-313x.2003.01971.x

Mu, Y., Zou, M., Sun, X., He, B., Xu, X., Liu, Y., et al. (2017). BASIC PENTACYSTEINE proteins repress ABSCISIC ACID INSENSITIVE4 expression via direct recruitment of the polycomb-repressive complex 2 in Arabidopsis root development. *Plant Cell Physiol.* 58 (3), 607–621. doi:10.1093/pcp/pcx006

Ogneva, Z. V., Suprun, A. R., Dubrovina, A. S., and Kiselev, K. V. (2019). Effect of 5-azacytidine induced DNA demethylation on abiotic stress tolerance in *Arabidopsis thaliana*. *Plant Prot. Sci.* 55 (2), 73–80. doi:10.17221/94/2018-pps

Rosa-Tellez, S., Anoman, A. D., Alcantara-Enguidanos, A., Garza-Aguirre, R. A., Alseekh, S., and Ros, R. (2020). PGDH family genes differentially affect Arabidopsis tolerance to salt stress. *Plant Sci.* 290, 110284. doi:10.1016/j.plantsci.2019.110284

Roscoe, T. J., Vaissayre, V., Paszkiewicz, G., Clavijo, F., Kelemen, Z., Michaud, C., et al. (2019). Regulation of FUSCA3 expression during seed development in Arabidopsis. *Plant Cell Physiol.* 60 (2), 476–487. doi:10.1093/pcp/pcy224

Sabagh, A. E., Islam,, M. S., Skalicky, M., Raza, M. A., Singh, K., Hossain, M. A., et al. (2021). Salinity stress in wheat (*Triticum aestivum* L.) in the changing climate: Adaptation and management strategies [J]. *Front. Agron.* doi:10.3389/fagro.2021.661932

Santi, L., Wang, Y. M., Stile, M. R., Berendzen, K., Wanke, D., Roig, C., et al. (2003). The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. *Plant J.* 34 (6), 813–826. doi:10.1046/j.1365-313x.2003.01767.x

Shanks, C. M., Hecker, A., Cheng, C. Y., Brand, L., Collani, S., Schmid, M., et al. (2018). Role of BASIC PENTACYSTEINE transcription factors in a subset of cytokinin signaling responses. *Plant J.* 95 (3), 458–473. doi:10.1111/tpj.13962

Shi, H., Wang, X., Ye, T., Chen, F., Deng, J., Yang, P., et al. (2014). The cysteine2/histidine2-type transcription factor zinc finger of arabidopsis THALIANA6 modulates biotic and abiotic stress responses by activating salicylic acid-related genes and C-REPEAT-BINDING factor genes in Arabidopsis. *Plant Physiol.* 165 (3), 1367–1379. doi:10.1104/pp.114.242404

Simonini, S., and Kater, M. M. (2014). Class I BASIC PENTACYSTEINE factors regulate HOMEOBOX genes involved in meristem size maintenance. *J. Exp. Bot.* 65 (6), 1455–1465. doi:10.1093/jxb/eru003

Simonini, S., Roig-Villanova, I., Gregis, V., Colombo, B., Colombo, L., and Kater, M. M. (2012). Basic pentacysteine proteins mediate MADS domain complex binding to the DNA for tissue-specific expression of target genes in Arabidopsis. *Plant Cell* 24 (10), 4163–4172. doi:10.1105/tpc.112.103952

Wan, X., Peng, L., Xiong, J., Li, X., Wang, J., Li, X., et al. (2019). AtSIBP1, a novel BTB domain-containing protein, positively regulates salt signaling in *Arabidopsis thaliana*. *J. Plants (Basel)* 8 (12), 573. doi:10.3390/plants8120573

Wu, J., Mohamed, D., Dowhanik, S., Petrella, R., Gregis, V., and Li, J., et al. (2020). Spatiotemporal restriction of FUSCA3 expression by class I BPCs promotes ovule development and coordinates embryo and endosperm growth [J]. *Plant Cell* 32 (6), 1886–1887. doi:10.1111/head.13881

Zhao, P-X., Zhang, J., and Chen, S-Y. (2021). New Phytologist.Arabidopsis MADS-box factor AGL16 is a negative regulator of plant response to salt stress by down regulating salt-responsive genes [J]

frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 02 May 2023
DOI 10.3389/fgene.2023.1151962

OPEN ACCESS

EDITED BY
Mallana Gowdra Mallikarjuna,
Indian Agricultural Research Institute (ICAR), India

REVIEWED BY
Enrique Hernandez-Lemus,
National Institute of Genomic Medicine (INMEGEN), Mexico
Xiujun Zhang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Ying Xu,
xuy9@sustech.edu.cn
Zhenyu Huang,
zhenyuh19@mails.jlu.edu.cn

RECEIVED 27 January 2023
ACCEPTED 11 April 2023
PUBLISHED 02 May 2023

CITATION
Xiao G, Guan R, Cao Y, Huang Z and Xu Y (2023), KISL: knowledge-injected semi-supervised learning for biological co-expression network modules.
Front. Genet. 14:1151962.
doi: 10.3389/fgene.2023.1151962

KISL: knowledge-injected semi-supervised learning for biological co-expression network modules

Gangyi Xiao[1], Renchu Guan[1], Yangkun Cao[2], Zhenyu Huang[1]* and Ying Xu[3]*

[1]College of Computer Science and Technology, Jilin University, Changchun, China, [2]School of Artificial Intelligence Jilin University, Changchun, China, [3]School of Medicine, Southern University of Science and Technology, Shenzhen, Guangdong, China

The exploration of important biomarkers associated with cancer development is crucial for diagnosing cancer, designing therapeutic interventions, and predicting prognoses. The analysis of gene co-expression provides a systemic perspective on gene networks and can be a valuable tool for mining biomarkers. The main objective of co-expression network analysis is to discover highly synergistic sets of genes, and the most widely used method is weighted gene co-expression network analysis (WGCNA). With the Pearson correlation coefficient, WGCNA measures gene correlation, and uses hierarchical clustering to identify gene modules. The Pearson correlation coefficient reflects only the linear dependence between variables, and the main drawback of hierarchical clustering is that once two objects are clustered together, the process cannot be reversed. Hence, readjusting inappropriate cluster divisions is not possible. Existing co-expression network analysis methods rely on unsupervised methods that do not utilize prior biological knowledge for module delineation. Here we present a method for identification of outstanding modules in a co-expression network using a knowledge-injected semi-supervised learning approach (KISL), which utilizes apriori biological knowledge and a semi-supervised clustering method to address the issue existing in the current GCN-based clustering methods. To measure the linear and non-linear dependence between genes, we introduce a distance correlation due to the complexity of the gene-gene relationship. Eight RNA-seq datasets of cancer samples are used to validate its effectiveness. In all eight datasets, the KISL algorithm outperformed WGCNA when comparing the silhouette coefficient, Calinski-Harabasz index and Davies-Bouldin index evaluation metrics. According to the results, KISL clusters had better cluster evaluation values and better gene module aggregation. Enrichment analysis of the recognition modules demonstrated their effectiveness in discovering modular structures in biological co-expression networks. In addition, as a general method, KISL can be applied to various co-expression network analyses based on similarity metrics. Source codes for the KISL and the related scripts are available online at https://github.com/Mowonhoo/KISL.git.

KEYWORDS

biological co-expression network, factor analysis, semi-supervised learning algorithm, network modules identification, feature selection

1 Introduction

To study the functions of genes at a system level, a key is to understand how genes work together. A basic assumption is that co-expressed genes tend to work in the same subsystem. Co-expression networks (GCN) (Yip and Horvath, 2007a) are commonly used to describe such subsystems based on statistical correlations among the expressions of the relevant genes. Typically, each node in such an undirected network represents a distinct gene and a weighted edge between two nodes denotes the two genes with correlated expressions while the edge weight represents the correlation level.

One goal when studying such a network is to discover densely connected subnetworks, also referred to as functional modules or clusters, as co-expressed genes tend to be transcriptionally coregulated. WGCNA (Zhang and Horvath, 2005) is a most widely used software for GCN construction, and can be used to identify modules of highly co-expressed genes. Briefly, WGCNA constructs a weighted co-expression network based on the Pearson correlation coefficients among provided gene expressions; uses a topological overlap structure measure (TOM) (Ravasz et al., 2002) of nodes to identify modules; and utilizes eigengene and intramodule hub genes to summarize such modules (Langfelder and Horvath, 2008). WGCNA identifies gene modules by using hierarchical clustering, giving rise to a tree-like structure. The advantage of the hierarchical clustering method is its simplicity, but the process for generating a hierarchical clustering tree is irreversible.

Multiple developments have been made aiming to improve the TOM measure. Among them, Li et al. proposed a bottom-up multi-node topological overlap measure (MTOM) that selects nodes with the highest neighborhood size to form modules based on multiple nodes. (Yip and Horvath, 2007b) developed a generalized topological overlap measure, called GTOM. Compared to TOM that considers only the nodes directly adjacent to the target gene pair, GTOM considers neighboring nodes that are within K steps away from the target gene pair, where K is a parameter to be selected by the user. Thus, GTOM is more sensitive to higher-order connections. Hou et al. (2021) introduced the K-means method to WGCNA to add additional steps to improve the module-identification results of WGCNA. A few other algorithms have been deployed to analyze gene co-expression networks, such as the flow simulation-based module discovery method (MCL) (Hwang et al., 2006), the graph partitioning-based method (Qcut) (Ruan and Zhang, 2008), and the density model-based method (MCODE) (Bader and Hogue, 2003).

One common issue with all these methods is: they use only unsupervised methods for clustering or module identification, but do not make effective use of prior biological knowledge. In addition, WGCNA uses hierarchical clustering to identify gene modules. One drawback of hierarchical clustering is that once two objects are clustered together, the process cannot be reversed. Therefore, regrouping of inappropriately clustered items is not doable. Analyses of the improved methods of WGCNA for refining its module identification results shows that the methods could not solve the problem of generating an unreasonable number of clusters. The purpose of this paper is to develop an effective method for module identification in a co-expression network to improve the of these two issues in existing methods.

Here we present a method for identification of outstanding modules in a co-expression network using a knowledge-injected semi-supervised learning approach (KISL), which utilizes *apriori* biological knowledge and a semi-supervised clustering (Basu et al., 2004) method to address the issue existing in the current GCN-based clustering methods. A comparative analysis of our algorithm with the WGCNA method on eight human cancer datasets has revealed the effectiveness of our algorithm in discovering modular structures in co-expression networks, paving the way for more accurate and useful GCN analysis.

2 Methods

2.1 WGCNA and KISL algorithms

We sought to identify modules consisting of highly functionally related genes. The structure of our algorithm is shown in Figure 1, consisting of three main stages. The first stage covers data preprocessing, variance analysis and feature selection to generate a gene expression profile matrix. The second stage is to construct clustering constraints by using factor analysis, to perform Gene Ontology (GO) enrichment analysis, and to perform factor analysis based on gene expression profiles for the set of genes covered by enriched GO/BP pathways. The result is a factor-loading matrix. The factor coefficients are binarized through thresholding, a subset of genes affected by a single factor is screened to form the "must-link" gene clusters, and all gene clusters from the pathway screening together form the *apriori* constraints for module identification in the co-expression network. The third stage is to construct the GCN and then use a semi-supervised algorithm in combination with the *apriori* constraints for identification of the GCN functional modules.

The inputs to the semi-supervised algorithm are the GCN network, the *apriori* constraints and the number of clusters k (the value of k is set according to the learning curve by the user given a value interval for k). The main purpose of the algorithm is to calculate the connectivity of genes to the module mean vector in each module and to assign genes to the modules that are most highly connected to them. Here, the mean vector μ_j of module j is defined as in Eq. (1).

$$\mu_j = \frac{1}{|C_j|}\sum_{x \epsilon C_j} x_i \tag{1}$$

where x_i is the expression profile of gene i, C_j is the set of all genes in module j, and $|C_j|$ denotes the number of genes in module j. We calculate the distance $d_{ij} = \|x_i - \mu_j\|_2$ between the sample x_i and each mean vector μ_j ($1 \le j \le k$). We count count_j ($j = 1, 2, \ldots, k$) of other samples in the constraint set containing sample x_i in each clustering cluster. The distance $d_{ij} = d_{ij} + \text{count}_j$ between sample x_i and module j is adjusted according to the constraint. For each gene i we set its module label to the label of the mean vector that minimizes d_{ij}. We then recalculate the mean vector of genes in each module and repeat the previous steps until no cluster assignment changes or the preset maximum number of iterations is reached. Additionally, tool KISL includes several additional functions designed to aid the user in visualizing input data and results. These functions rely on basic plotting functions provided in python and the R packages

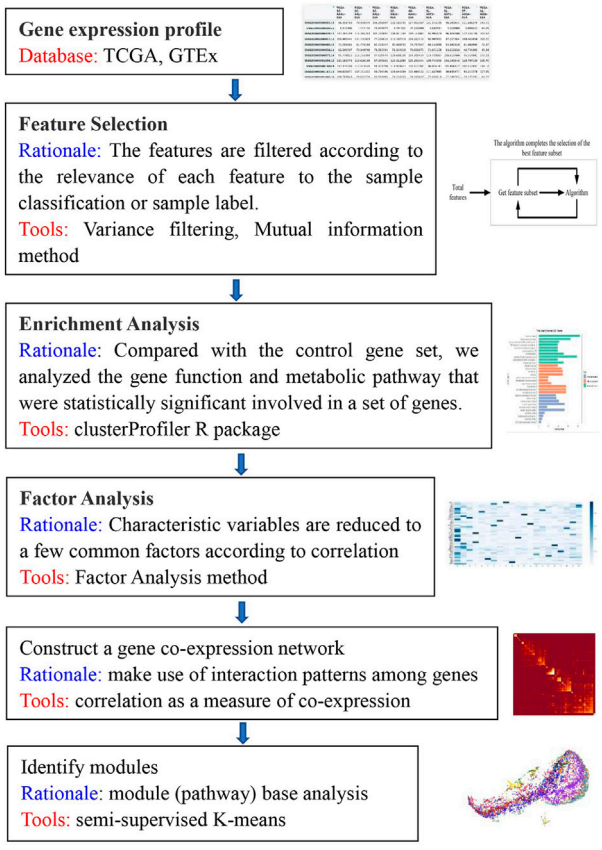


FIGURE 1
Overview of the algorithm. This flowchart briefly describes the main. steps of the KISL algorithm.

WGCNA (Langfelder and Horvath, 2008). The code of the KISL algorithm are available online at https://github.com/Mowonhoo/KISL.git.

2.2 Construction of the gene co-expression network

Measuring the co-expression relationship between genes is a key issue in the construction of gene co-expression networks. However, commonly used correlation measures, including linear (e.g., Pearson correlation) and monotonic (e.g., Spearman correlation) dependence measures, are not sufficient to observe the nature of real biological systems. Szekely et al. (Székely et al., 2007; Székely and Rizzo, 2009) proposed distance correlation for both linear and non-linear dependencies. Distance correlation reveals more about the complex biological relationships between gene profiles than other correlation metrics, which helps to provide more meaningful modules in the analysis of gene co-expression networks. However, the time complexity associated with computing the distance is high and requires more computational resources (Hou et al., 2022). However, for biological analysis we seek higher reliability and completeness of information mining, therefore, in this study, we use distance correlation to measure the relationship between genes. To optimize the time spent by the algorithm, the features can be optionally downscaled by using the principal component analysis (PCA) method before calculating the correlation coefficients between genes, and feature retention is filtered by setting a threshold based on the PCA variance interpretation rate.

The distance correlation coefficient can reveal an arbitrary relationship between the variables. When the Pearson correlation coefficient is 0, we cannot determine whether the two variables are independent, but if the distance correlation coefficient is 0, then we can conclude that the two variables are independent of each other (Pearson and Galton, 1895; Székely et al., 2007; Székely and Rizzo, 2009). The distance correlation coefficient of two variables u and v is denoted as $\hat{d}\mathrm{corr}(u,v)$. When $\hat{d}\mathrm{corr}(u,v)=0$, the two variables are independent of each other. The larger $\hat{d}\mathrm{corr}(u,v)$ is, the stronger the correlation between u and v. Let the random sample of the overall (u, v) be $\{(u,v), i=1,2,\ldots,n\}$ and Szekely et al. (Székely et al., 2007; Székely and Rizzo, 2009) defined the sample estimate of the distance correlation coefficient between two random variables u and v as Eq. 2.

$$\hat{d}\mathrm{corr}(u,v)=\frac{\hat{d}cov(u,v)}{\sqrt{\hat{d}cov(u,u)\hat{d}cov(v,v)}} \tag{2}$$

where $\hat{d}cov^2(u,v)=\widehat{S}_1+\widehat{S}_2-2\widehat{S}_3$, $\widehat{S}_1$, $\widehat{S}_2$ and $\widehat{S}_3$ are shown in Eqs 3, 4, 5, respectively.

$$\widehat{S}_1=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|u_i-u_j\|_{d_u}\|v_i-v_j\|_{d_v} \tag{3}$$

$$\widehat{S}_2=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|u_i-u_j\|_{d_u}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|v_i-v_j\|_{d_v} \tag{4}$$

$$\widehat{S}_3=\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|u_i-u_l\|_{d_u}\|v_i-v_l\|_{d_v} \tag{5}$$

Similarly, $\hat{d}cov(u,u)$ and $\hat{d}cov(v,v)$ can be calculated.

The gene adjacency matrix is obtained by power-lawing the gene correlation matrix with a "soft" threshold power, and then the TOM of the adjacency network is calculated to construct the gene co-expression network. The construction of gene co-expression networks based on the TOM metric has been shown to have better results than direct module identification based on the adjacency graph (Langfelder et al., 2008).

We have kept the Pearson correlation coefficient for measuring the interrelationship between genes among the optional parameters of the functional function used to construct the co-expression network in order to increase the applicability and scalability of our algorithm and to meet the various needs of users. We have also given the mutual information method (MI) as an optional parameter, so that users can choose the parameters according to their needs. A MI measures the entropy of gene interactions to evaluate their relationship. In comparing linear and non-linear methods for measuring gene dependence, Zhang et al. found that the mutual information method combined linear and non-linear interactions has some advantages over linear or non-linear methods (Jiang and Zhang, 2022). Moreover, the MI between two variables is symmetric, which means that MI-based methods infer undirected interactions (Jia and Zhang, 2022). Additionally, we simulated and generated 10 pairs of high-dimensional variables with different dependencies, and then used them to measure the relationship between these variable pairs in order to compare the

characteristics of distance correlation, mutual information, and Pearson correlation coefficient to capture the complex relationship between variables. Calculations are performed using Python packages sklearn (Pedregosa et al., 2011), dcor (Ramos-Carreño and Torrecilla, 2022), and scipy (Virtanen et al., 2020). The supplementary Material 6 (Supplementary Figure S1) contains the pertinent results.

2.3 Topological characteristics of GCN

Network topology analysis is an important tool for understanding network characteristics at the system level. Network centrality analysis and global network topology analysis are two levels used to analyze the network from the system level. A key concept in network analysis is node connectivity (centrality). A central node (called a hub) is a node that is densely connected to other nodes. Co-expression networks have global topological properties of scale-free distributions, functional modular networks, and small-world properties. For weighted networks, Zhang and Horvath et al. (Zhang and Horvath, 2005) also defined the corresponding connectivity, intramodule connectivity metric and generalized scale-free topology for weighted networks.

1) Connectivity in weighted networks

The connectivity metric based on the weighted adjacency network is defined as Eq. 6.

$$W_i = \sum_{j=1}^{n} w_{ij} \tag{6}$$

where w_{ij} is the adjacency between two nodes i and j. Thus, if a node has high adjacency with many other nodes, then it has high connectivity W_i based on the weighted adjacency network.

A network connectivity metric is defined for a specific module's genes (intramodule connectivity). The intramodule connectivity (unweighted network node connectivity also commonly referred to as "degree") of gene i within module q is calculated as in Eq. 7.

$$within(k_i^{(q)}) = \sum_j w_{ij} \ (j = 1, 2, \ldots, n(q)) \tag{7}$$

where $n(q)$ denotes the number of genes within module q.

2) Module density

The dense connectivity property between genes within module q can be measured by the average neighboring degree of module genes, defined as the module density, as shown in Eq. 8.

$$\text{Density}\left(A^{(q)}\right) = \frac{\sum_i \sum_{j \neq i} w_{ij}^{(q)}}{n^{(q)}\left(n^{(q)} - 1\right)} \tag{8}$$

where $A^{(q)}$ denotes the $n^{(q)} \times n^{(q)}$ adjacency matrix corresponding to the subnetwork formed by the genes of module q.

3) Generalized scale-free topology

The frequency distribution $p(k)$ of node connectivity in a gene neighborhood network follows the power law $p(k) \sim k^{-\gamma}$. where k is the node connectivity (Langfelder et al., 2008). The square of the correlation between $\log_{10} p(k)$ and $\log_{10} k$ can be used to measure the degree to which the network satisfies the scale-free topology, i.e., the model fit index R^2 for a linear model regressing $\log_{10} p(k)$ on $\log_{10} k$. If the R^2 value is close to 1, there is a linear relationship between $\log_{10} p(k)$ and $\log_{10} k$.

2.4 Construction methods for *a priori* constraints

Thanks to the results of work in related fields of research it has been possible to obtain many biological explanations of the relationships between genes. The Gene Ontology (GO) database is one of the common gene annotation systems used in bioinformatics research, and it defines a structured standard biological model that allows the description of gene and protein functions in various organisms in terms of cellular components, biological processes and molecular functions.

The enrichment analysis enables the annotation and classification of genes to obtain a subset of genes grouped according to different gene functions, and the annotated results can be transformed to constitute *a priori* constraints for module identification algorithms to improve the modular biological interpretation of functional module identification of co-expression networks. We introduced factor analysis (Swisher et al., 2004; Ferrando, 2021), a statistical method for extracting common factors from groups of variables, to construct intergenic correlation constraints. The British psychologist C.E. Spearman first proposed it. Factor analysis can identify the common influences embedded in multiple variables. By grouping variables of the same nature into a common factor, the number of variables can be reduced. as shown in Eq. 9 below.

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + \epsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + \epsilon_2 \\ \cdots \\ X_P = a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + \epsilon_p \end{cases} \tag{9}$$

where F denotes the common factor, X denotes the original variable, and ϵ denotes the part of the original variable that cannot be represented by the common factor. The number of original variables is generally satisfied as greater than or equal to the number of factors (i.e., $m \leq p$). The factors F are independent of each other and have a variance of 1. The correlation between the common factor and ϵ is 0 and the correlation between ϵ is 0.

Before performing factor analysis, the Kaiser-Meyer-Olkin test (KMO test) and Bartlett's test of sphericity were performed on the features to determine whether the gene expression profile was suitable for factor analysis. Then, by calculating the eigenvalues of the gene correlation matrix and ranking them, the common factors with eigenvalues greater than 1 were extracted according to Kaiser's principle, and the cumulative total variance contribution rate was ensured to be greater than 0.85 according to the variance contribution rate accumulation principle. This process ensures that the extracted common factors cover enough information contained in the original gene expression profile and better replace the original gene characteristics. The factor loading coefficients are then derived and transformed by orthogonal rotation of the loading coefficients to obtain the factor loading matrix and then to analyze the characteristics of the factor coefficients for each gene. The factor loading coefficient matrix is then binarized to filter out the subset of

genes that depend on a certain common factor in the same pathway, and these genes are only highest correlated with this main factor. The constrained gene set is obtained by performing factor analysis on all GO terms enriched in the gene expression profile and then by merging the subsets with common overlapping genes.

2.5 Clustering evaluation metrics

The silhouette coefficient (RousseeuwSilhouettes, 1987), the Calinski-Harabasz index (Caliński and Harabasz, 1974) and the Davies-Bouldin index (Davies and Bouldin, 1979) are common and valid internal measures to evaluate the validity of clustering. The silhouette coefficient is a measure of how similar an observation is to its own cluster compared to other clusters, and it takes values from −1 to 1. A value of 1 indicates that the clusters are far from each other and clearly distinguished, a value of 0 indicates that the distance between clusters is non-significant, and a value of −1 indicates that the clusters are incorrectly assigned. The Calinski-Harabasz index is also known as the variance ratio criterion. For cluster q, the Calinski-Harabasz index is given by the ratio of the between-cluster dispersion mean to the within-cluster dispersion, and a higher Calinski-Harabaz index indicates better clustering. The physical meaning of the Davies-Bouldin index is the ratio of the sum of the mean sample distance (i.e., intracluster sample distance) of each cluster to the distance between the centroids of the two clusters (i.e., intercluster sample distance); given two clusters, the smaller the value is, the better.

2.6 Gene function annotation tools

The database for annotation, visualization and integrated discovery (DAVID) provides researchers with a comprehensive set of functional annotation tools to understand the biological significance behind large lists of genes (Huang et al., 2009). DAVID integrates biological data and analysis tools to provide systematic, integrated biofunctional annotation information for large-scale gene and protein lists to help users extract biological information. Here, we used the rich scores from the DAVID functional annotation clustering tool—the geometric mean (logarithmic scale) of the p values of the members of the corresponding annotation clusters for ranking their biological significance. The clusterProfiler R package was used to obtain the Gene Ontology terms of all differentially expressed genes (Yu et al., 2012).

2.7 Datasets

The tumor sample dataset used in this experiment was obtained from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) database, including BLCA (bladder urothelial carcinoma), BRCA (breast invasive carcinoma), COAD (colon adenocarcinoma), KIRC (kidney renal clear cell carcinoma), LUAD (lung adenocarcinoma), LUSC (lung squamous cell carcinoma), PAAD (pancreatic adenocarcinoma) and STAD (stomach adenocarcinoma) RNA-Seq data for eight tumors, and normal samples for each tumor were obtained from the Genotype-Tissue Expression (GTEx) database. The GTEx project aims to establish a repository of samples and data for studying the relationships between genetic variants, gene expression and other molecular phenotypes in a wide range of human tissues (GTEx Consortium, 2013; GTEx Consortium, 2015). First, the eight cancer datasets obtained from TCGA and GTEx databases were analyzed for differences by using the R package DESeq2 (Gentleman et al., 2004; Love et al., 2014). We set the screening criteria for differential genes as padj < 0.05, $|\log_2 \mathrm{Fold}Change| > 1$, followed by variance filtering to screen out genes with variance less than or equal to 0, i.e., consistent expression activity on all samples. The selection of features is then done using the mutual information method. The sample type is the phenotype (clinical trait) that we employ for gene screening. After feature selection filtering, the final retained samples and gene counts are provided in Supplementary Material 1 (Supplementary Table S1). Source codes for the KISL and the related scripts are available online at https://github.com/Mowonhoo/KISL.git. The datasets from Gene expression RNA-seq were performed using TCGA: https://www.cancer.gov/tcga.

3 Results and DISCUSSION

3.1 Effect of distance correlation on various datasets

(Székely and Rizzo, 2009) verified that the value of the distance correlation is always smaller than the absolute value of the Pearson correlation for bivariate normal data. Therefore, if the distance correlation coefficient between two random variables is greater than the Pearson correlation coefficient then a complex relationship exists between them - non-binary normal data and non-linear nonmonotonic relationship. In general, correlation values greater than 0.8 are described as strong correlation, while values less than 0.5 are described as weak correlation (Castro Sotos et al., 2009). To measure the proportion of complex relationships in the dataset, we selected gene pairs with distance correlation coefficients greater than 0.5 from eight datasets. Next, we analyzed the distribution of Pearson correlation coefficients for the retained gene pairs. In the PAAD dataset, 70.88% of the gene pairs had Pearson correlation coefficients less than 0.5 (Figure 2G). In addition, the ratios in the LUSC dataset (Figure 2F), LUAD dataset (Figure 2E) and STAD dataset (Figure 2H) were 66.37%, 61.04% and 50.62%, respectively, as shown in Supplementary Material 2 (Supplementary Table S2). Both our algorithm and the standard WGCNA method use a ‘soft’ threshold power in the construction of the GCN, which amplifies the difference between strong and weak correlations. When using Pearson correlation coefficients, gene pairs with complex relationships have small correlation coefficient values, and the presence of the soft threshold further leads to a smaller weight of the two genes and increases the error, making the clustering results inaccurate.

It has been reported that biological networks show scale-free topology (STF) (Langfelder and Horvath, 2008; Barabási et al.,

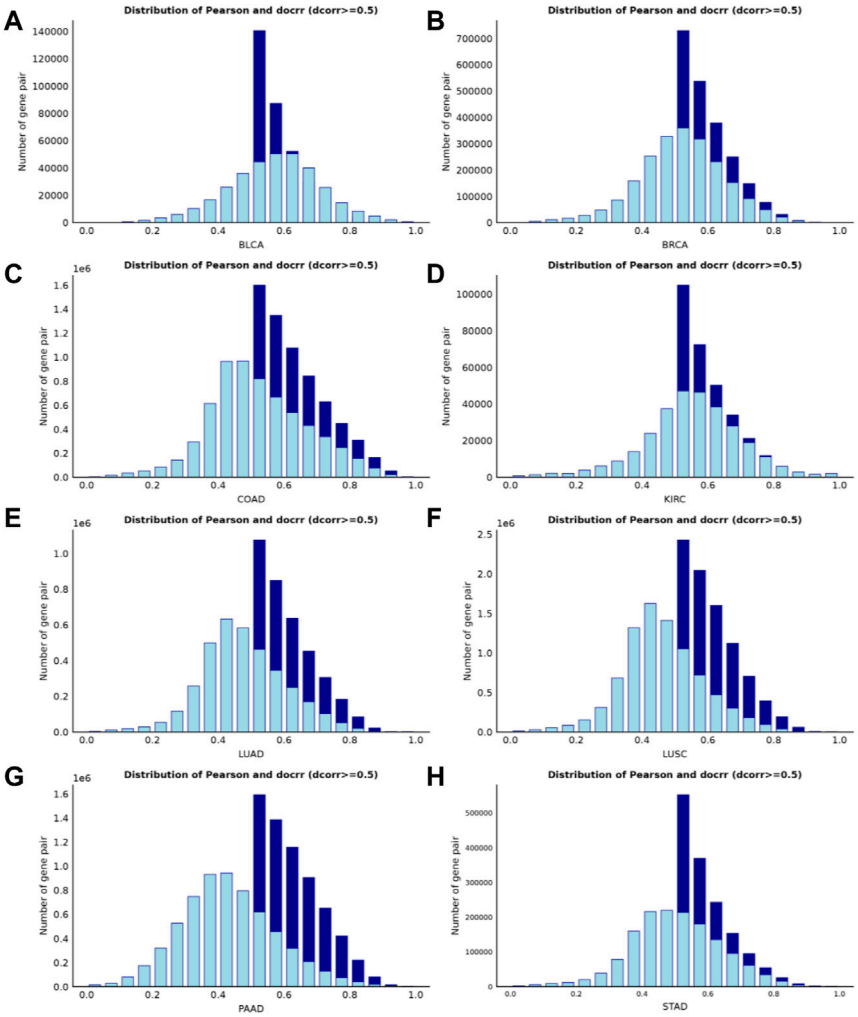


FIGURE 2
Histogram of correlation coefficients for interactions with high distance correlation scores (>0.5). The bright blue borders in each panel represents the Pearson correlations, and the dark blue borders represents the distance correlations. When using the criterion that the Pearson correlation coefficient must be greater than 0.5, more than 50% of the complex correlated data information on four of the datasets (Figures 2E–H) would be lost.

2011). It is important in SFT networks to identify the dominant hub nodes because they usually have significant influence on the network. In the case of biological networks it may mean that the genes, proteins or metabolites represented by these nodes are biologically important (Albert, 2005; Andrecut et al., 2008; Nafis et al., 2015; Atiia et al., 2020). Therefore, we investigate the SFT of the two correlation coefficients for the eight datasets. The closer the SFT fit index is to 1, the better. In Figure 3 the left panel shows the histogram of network connectivity and the right panel shows the logarithmic plot of the corresponding histogram. The approximate linear relationship (high R^2 values) indicates the approximate scale-free topology. We find that for eight datasets, both Pearson correlation coefficients and distance correlation coefficients achieve SFT when a suitable "soft" threshold power is chosen to define the adjacency matrix, and in five of them (Figures 3A–E), distance correlation shows an advantage in the scale-free fit index.

3.2 Constructing clustering constraints

The KMO test and Bartlett's test of Sphericity were used to determine whether a gene expression profile was suitable for factor analysis before all GO terms enriched in the gene expression profile were subjected to factor analysis. In this paper, the number of contained genes is greater than 5, the threshold value set by KMO test is greater than 0.6, and the *p*-value of Bartlett's test of sphericity is set to less than 0.05 (*p*-value is less than the significance level value of 0.05, indicating a high correlation between genes in the expression profile data) of GO term for factor analysis to construct constrained gene sets. From Figure 4, we can see that the percentage of GO terms enriched in each gene expression profile data that were evaluated to be suitable for factor analysis ranged from approximately 40%–72%, which indicates that we can effectively extract *a priori* biological knowledge by introducing factor analysis methods. The factor loading matrix is binarized by setting an appropriate factor screening threshold (we

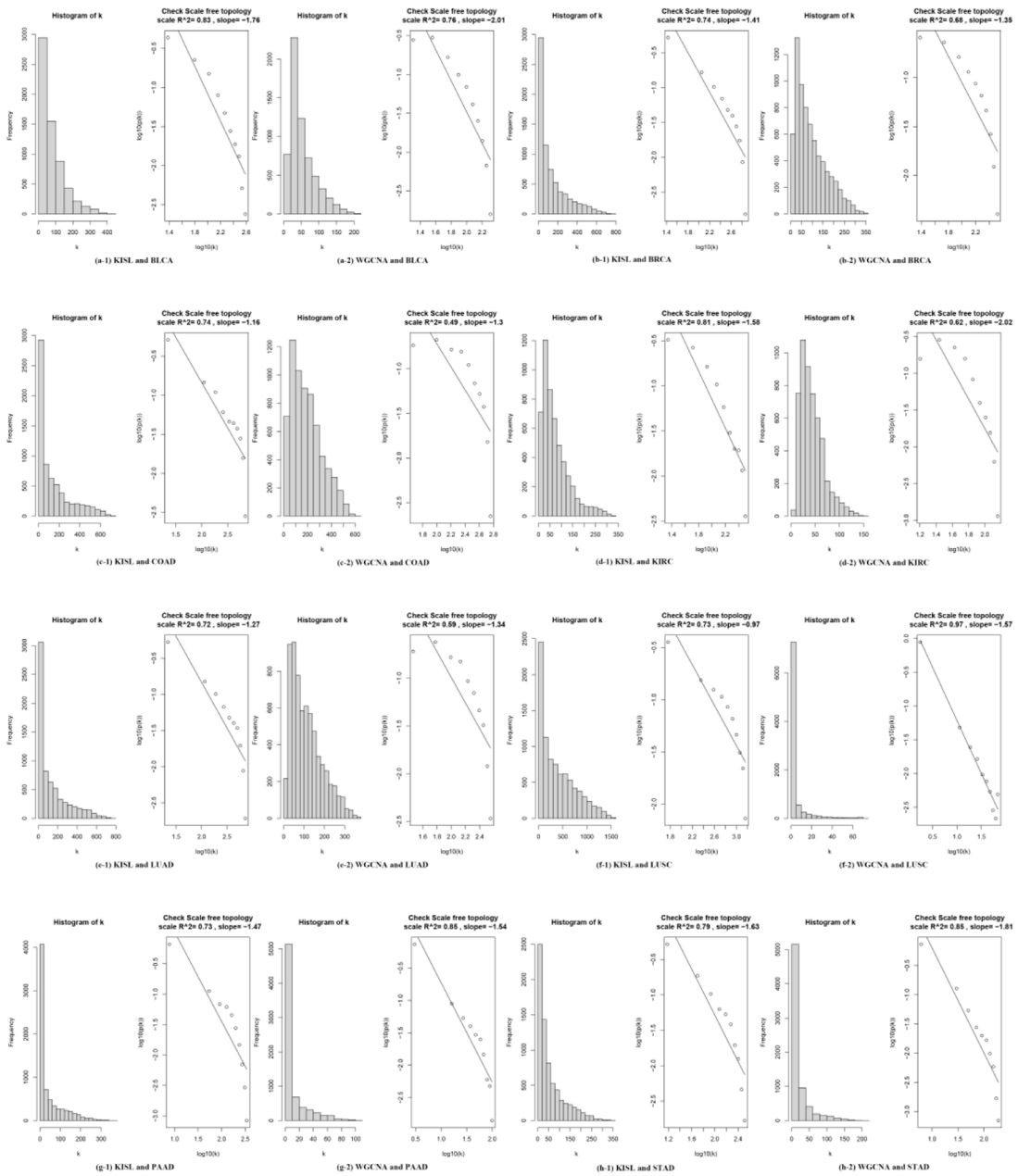


FIGURE 3
shows the scale-free topological properties of the co-expression network. The left panel shows the histogram of the network connectivity, and the right panel shows the logarithmic plot of the corresponding histogram. The approximate linear relationship (high R2 values) represents the approximate scale-free topology. The scale-free topology is at least approximately satisfied when a suitable "soft" threshold is chosen to define the adjacency matrix for the eight selected real datasets.

set the threshold to 0.2, then each gene factor coefficient greater than 0.2 is set to 1, and less than that is set to 0). Finally, the set of constrained genes that significantly depend on a single common factor in the same pathway is obtained from the binarized factor loading matrix. All subsets of genes in all GO terms that depend on a single principal factor are filtered out, and the subsets with common overlapping genes are merged to obtain the constrained gene set. According to the clustering constraint construction process described above, the final constrained gene sets based on *a priori* biological knowledge are obtained on each dataset, and the constrained gene sets are summarized as shown in (Supplementary Table S3).

3.3 Evaluation based on internal metrics of clustering algorithms

In this section, we use the silhouette coefficient, the Calinski-Harabasz index and the Davies-Bouldin index to evaluate the quality of the WGCNA and KISL clustering results. As shown in Figure 5, the

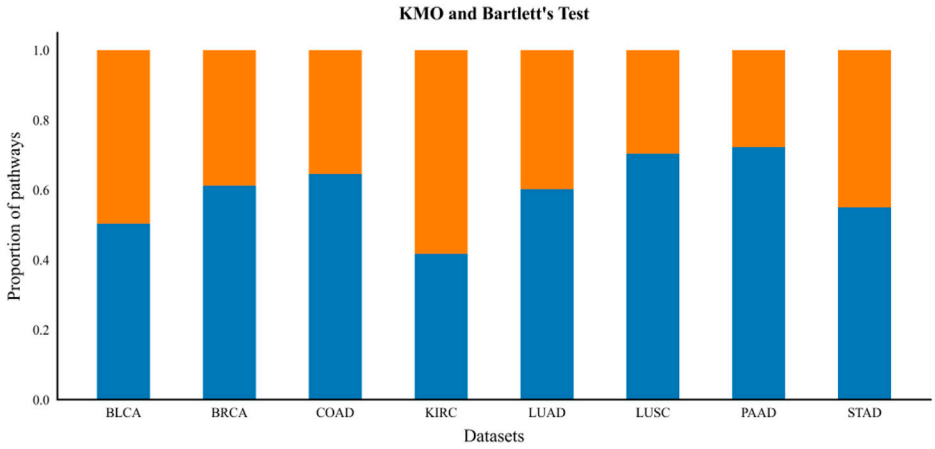


FIGURE 4
KMO and Bartlett's test. The blue bars below the figure indicate the proportion of gene expression profiles of GO Term suitable for factor analysis after the KMO test and Bartlett's test of sphericity.

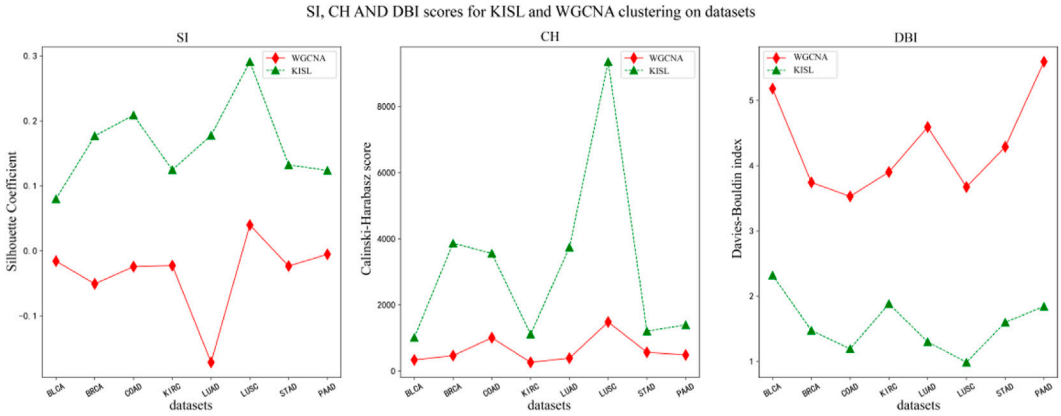


FIGURE 5
Silhouette coefficient, Calinski–Harabasz score and Davies–Bouldin index for the WGCNA and KISL algorithms. The evaluation value obtained by the KISL algorithm was the best in all the datasets.

KISL algorithm obtained the highest silhouette coefficient and Calinski-Harabasz index evaluation values in all eight datasets, while obtaining the lowest Davies-Bouldin index evaluation value. Taking the silhouette coefficient evaluation metric as an example, three of the datasets, COAD, LUAD, and LUSC, obtained a boost of more than 0.3 on the dataset, and two datasets, BRCA and STAD, obtained a boost of more than 0.15 with the smallest evaluation value on the BLCA dataset but also slightly improved. It is also important to note that the silhouette coefficient value obtained by the base method is negative on most of the datasets, especially on the LUAD dataset, where it is the worst and even reaches −0.17, which means that many sample points are assigned to the wrong cluster. Our algorithm also obtained the best evaluation values for both the Calinski-Harabasz index and Davies-Bouldin index evaluation metrics. The clusters obtained by KISL have better clustering evaluation values and better aggregation of the obtained gene modules. The details of the three evaluation values of the clusters are shown in (Supplementary Table S4). In Figure 6, we plot the results of the silhouette coefficient analysis for the KISL algorithm (the left side) and the Pearson-based WGCNA (the right side) corresponding to the eight datasets. The closer the silhouette coefficient to 1, the better the clustering result. The evaluation value obtained by the KISL algorithm was the highest in all the datasets.

3.4 Analysis of the nature of the recognition module

The module significance measure was defined as the average gene significance of all genes in the module. We used absolute values to define the relevance-based gene module significance metric. The results of the significance of each module identified on the eight datasets are shown in Figure 7. We use a gene module significance of 0.4 (the red dashed line) as the threshold, and we find that our algorithm obtains more high gene significance

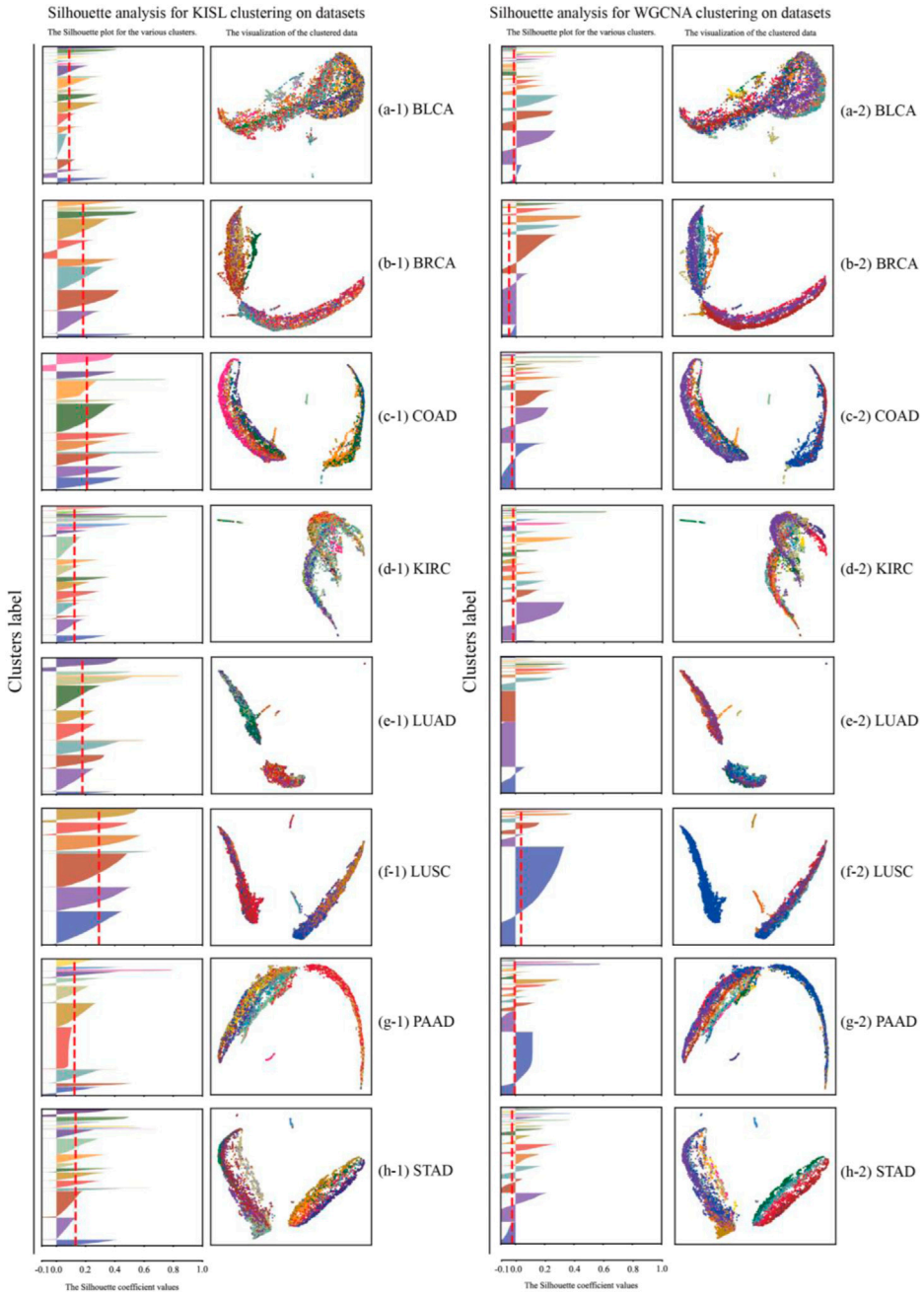


FIGURE 6
Silhouette coefficient analysis for the WGCNA and KISL algorithms. The left panel shows the results of silhouette coefficient analysis of the clusters obtained by the KISL. The right panel shows the results obtained by the base method WGCNA on the corresponding dataset. The evaluation value (the red dashed line) obtained by the KISL algorithm was the highest in all the datasets. In each panel, the left part represents the silhouette coefficient value of each sample, the *y*-axis represents the sample sequence, and the *x*-axis represents the silhouette coefficient size. UMAP visualization results are displayed on the right side of each panel.

modules on five of the datasets BLCA, BRCA, KIRC, LUAD, and STAD (Figures 7A, B, D, E, H), while on the other three datasets our method obtains the same number of high significance modules as the base method.

A network connectivity metric is defined for module-specific genes (intramodule connectivity). The intramodular connectivity of genes within a module is calculated, and the dense connectivity property between genes within a module is measured by the average adjacency of the module genes, defined as the module density. Figure 8 shows the comparison between the density of each module obtained by the KISL algorithm and the base method, where a larger average module density is obtained on seven of

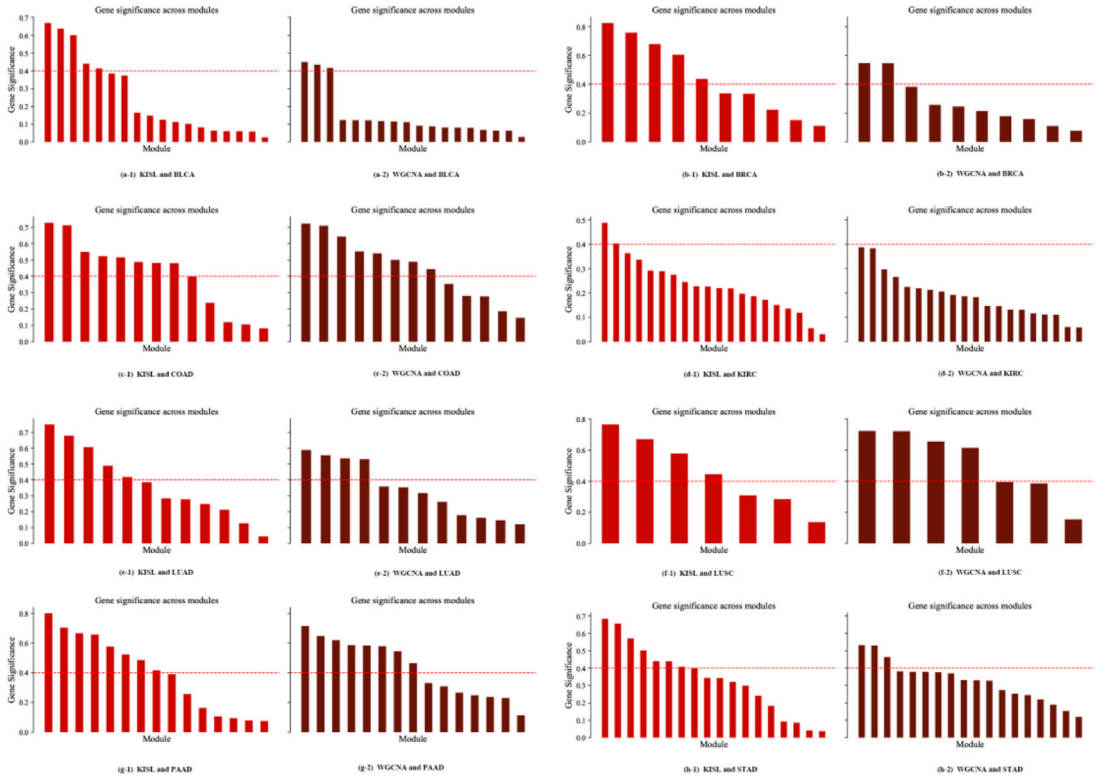


FIGURE 7
Module significance metric. The gene module significance threshold is set to 0.4 (the red dashed line), and our algorithm obtains more high gene significance modules on five of the datasets, BLCA, BRCA, KIRC, LUAD and STAD (Figures 7A,B,D,E,H), while on the other three datasets our method obtains the same number of high significance modules as the base method.

the datasets and a larger number of modules with greater density are possessed. In addition, the top 3 modules with the highest module density on all eight datasets are found by our algorithm.

3.5 Comparison of module gene enrichment analysis

Co-expressed genes often act synergistically and participate in the same biological processes (van Dam et al., 2012). Therefore, algorithms that identify modules that are highly enriched for specific gene classes are more reasonable (Rau et al., 2013). To compare the average enrichment scores and stability of the algorithms, we use the recommended parameters of the WGCNA package for module identification, and to keep the number of modules identified by the two algorithms equal, the number of modules obtained by the WGCNA method is used to initialize the K values of our algorithms.

In the current analysis, we obtained the enrichment scores of each cluster in the functional annotation clustering of DAVID. The higher the enrichment score, the lower the *p*-value and therefore the more significant the enrichment. The module enrichment score is an important indicator to evaluate the rationality of a module. We discuss the average enrichment scores of modules from gene co-expression networks constructed by two different algorithms to measure the degree of enrichment of co-expression networks. As shown in Table 1, the modules from KISL have higher DAVID average enrichment scores in the six data sets, indicating that the division of their modules is more reasonable. Higher DAVID enrichment scores for each module can be viewed in (Supplementary Table S5), where the modules identified by KISL have the highest top 3 enrichment scores in the five datasets, and the top 3 modules have one or two enrichment scores in the other three datasets.

To verify whether the identification modules obtained by KISL are biologically meaningful, the highly enriched (Top 5) biological process (BP) terms of the network modules in GO terms were summarized for the LUSC sample, as shown in Table 2. Overall, the enrichment of GO terms shows the biological significance of the modules obtained by KISL.

4 Conclusion

Co-expression analysis is useful for exploring patterns of gene networks, identifying gene functional modules, and mining cancer-associated markers at the system level. By using the enriched information of the current sample as a constraint, we aim to perform semi-supervised clustering. Other clustering methods only take into account the algorithm parameters, not the sample itself. Therefore, we propose the KISL method to try to improve these methods. KISL algorithm measures linear and non-linear dependencies between genes by using distance correlation, which is appropriate for the complexity of the relationship between genes. In cases where outliers significantly influence the correlation coefficient value, distance correlation is a better alternative because it is distribution-free and

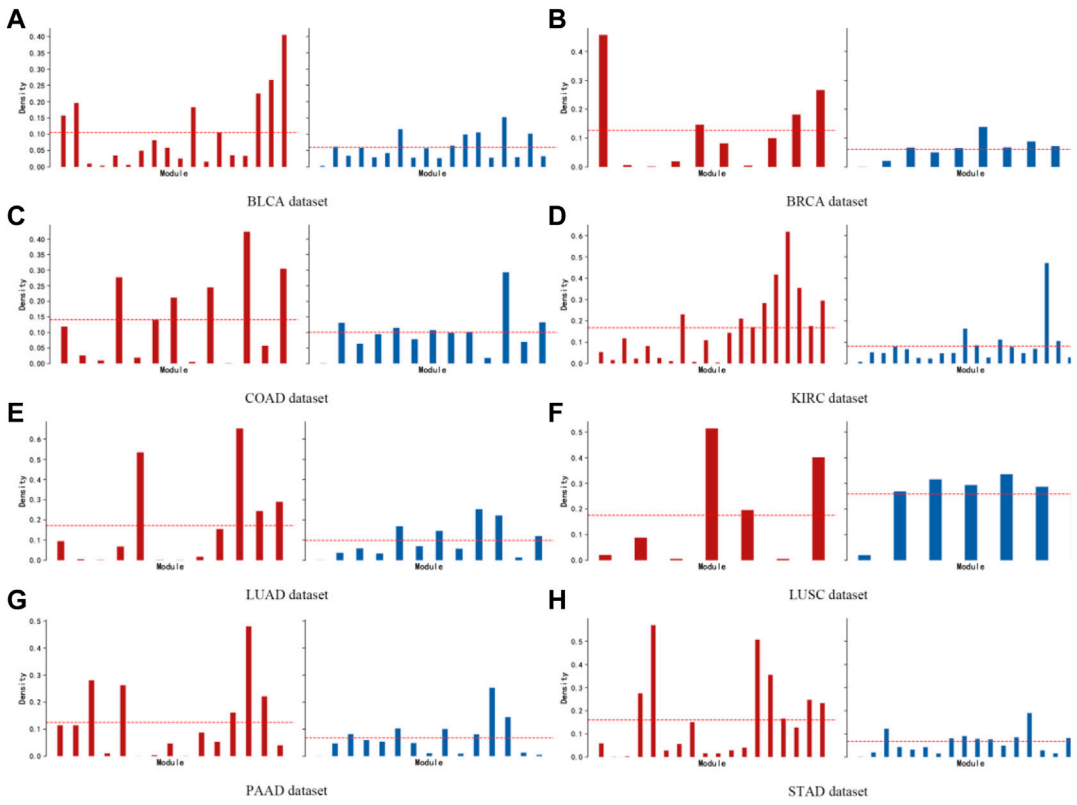


FIGURE 8
Module density. The brown bar plot indicates the module density obtained by the KISL algorithm on each dataset, and the blue bar plot indicates the results of the widely used method WGCNA. Our method obtains a larger average module density on seven of the datasets and has a larger number of modules with a larger density. In addition, the top 3 modules with the highest module density on all eight datasets are obtained by our algorithm.

TABLE 1 Average DAVID enrichment score for each dataset.

	BLCA	BRCA	COAD	KIRC	LUAD	LUSC	STAD	PAAD
WGCNA	5.20	**9.53**	3.88	4.29	6.25	2.82	3.64	**5.98**
KISL	**6.11**	8.44	**4.12**	**5.64**	**7.36**	**6.55**	**4.63**	5.58

The bold words in Table 1 indicate the maximum value of the column, and the KISL algorithm obtains the maximum value on most data sets.

TABLE 2 GOTERM BP on LUSC dataset.

Module	GOTERM BP
module0	O-glycan processing; innate immune response in mucosa; antibacterial humoral response; antimicrobial humoral immune response mediated by antimicrobial peptide; protein O-linked glycosylation
module1	DNA replication; DNA unwinding involved in DNA replication; spliceosomal snRNP assembly; mitochondrial translation; DNA-dependent DNA replication
module2	epithelial cell differentiation; epidermis development; intermediate filament organization; immunoglobulin production; keratinization
module3	cilium movement; flagellated sperm motility; microtubule-based movement; cilium assembly; outer dynein arm assembly
module4	cell division; chromosome segregation; mitotic spindle assembly checkpoint; mitotic cell cycle; mitotic spindle organization
module5	immunoglobulin production; immune response; positive regulation of B-cell activation; phagocytosis, recognition; phagocytosis, engulfment
module6	signal transduction; vasculogenesis; angiogenesis; positive regulation of angiogenesis; cell adhesion

better suited to complex relationships. Moreover, using biological knowledge based on GO terms to construct clustering constraints, a semi-supervised method is used to identify network modules, which can more effectively partition the network.

After comparing the silhouette coefficient, the Calinski-Harabasz index and the Davies-Bouldin index evaluation metric values of the modules identified by KISL with the widely used WGCNA, our algorithm obtained the best performance on eight real-world cancer sample datasets. The clustering produced by the method in this paper has a better clustering evaluation value, and the obtained gene modules have better aggregation. Based on enrichment analysis, the identified modules were effective in discovering modular structures in biological co-expression networks. The KISL method is a general method for analyzing biological co-expression networks based on similarity metrics.

In addition, we plan to incorporate more useful biological knowledge in the future, such as protein–protein interaction networks and gene regulatory networks, which could allow us to better identify co-expressed gene modules. Genomics and transcriptomics are increasingly being applied to aid in clinical diagnosis and prognosis; thus, in addition to discussing module identification in co-expression network analysis, it is also important to develop effective methods for comparative network analysis. As part of our future research, we plan to explore how co-expression networks can be compared. It is our future goal to examine comparative methods of co-expression networks.

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

GX designed the research and implemented Knowledge injected semi-supervised learning algorithms and wrote the manuscript. ZH and YX contributed scientific background, biological interpretation, and consistency of the result of the experiment. YC conducted data preprocessing. RG provided the background knowledge of models, helped to focus the relevance of the contribution and aided in revising the manuscript.

Funding

Our work is supported by the National Key Research and Development Program of China No. 2021YFF1201200, the National Natural Science Foundation of China No. 62172187 and No. 61972174, Liaoning Provincial Archives Science and Technology Project (Grant No. 2021-X-012 and Grant No. 2022-X-017), and Guangdong Universities' Innovation Team Project (No. 2021KCXTD015) and Guangdong Key Disciplines Project (No. 2021ZDJS138).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1151962/full#supplementary-material

References

Albert, R. (2005). Scale-free networks in cell biology. *J. Cell Sci.* 118, 4947–4957. doi:10.1242/jcs.02714

Andrecut, m., kauffman, s. A., and madni, a. M. (2008). Evidence of scale-free topology in gene regulatory network of human tissues. *Int. J. Mod. Phys. C* 19, 283–290. doi:10.1142/s0129183108012091

Atiia, A. A., Hopper, C., Inoue, K., Vidal, S., and Waldispühl, J. (2020). Computational intractability law molds the topology of biological networks. *Appl. Netw. Sci.* 5, 34–22. doi:10.1007/s41109-020-00268-0

Bader, G. D., and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinforma.* 4, 2. doi:10.1186/1471-2105-4-2

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi:10.1038/nrg2918

Basu, S., Banerjee, A., and Mooney, R. J. (2004). "Active semi-supervision for pairwise constrained clustering," in Proceedings of the 2004 SIAM International Conference on data mining *333–344* (Society for Industrial and Applied Mathematics). doi:10.1137/1.9781611972740.31

Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* 3 (1), 1–27. doi:10.1080/03610927408827101

Castro Sotos, A. E., Vanhoof, S., Van Den Noortgate, W., and Onghena, P. (2009). The transitivity misconception of PEARSON'S correlation coefficient. *Stat. Educ. Res. J.* 8, 33–55. doi:10.52041/serj.v8i2.394

Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. *PAMI-* 1, 224–227. doi:10.1109/tpami.1979.4766909

Ferrando, P. J. (2021). Seven decades of factor analysis: From yela to the present day. *Psicothema* 33, 378–385. doi:10.7334/psicothema2021.24

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80. doi:10.1186/gb-2004-5-10-r80

GTEx Consortium (2015). Human genomics. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348, 648–660. doi:10.1126/science.1262110

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi:10.1038/ng.2653

Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., et al. (2022). Distance correlation application to gene co-expression network analysis. *BMC Bioinforma.* 23, 81. doi:10.1186/s12859-022-04609-x

Hou, J., Ye, X., Li, C., and Wang, Y. K. (2021). K-module algorithm: An additional step to improve the clustering results of WGCNA Co-expression networks. *Genes* 12, 87. doi:10.3390/genes12010087

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

Hwang, W., Cho, Y.-R., Zhang, A., and Ramanathan, M. (2006). A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol. Biol. Amb.* 1, 24. doi:10.1186/1748-7188-1-24

Jia, Z., and Zhang, X. (2022). Accurate determination of causalities in gene regulatory networks by dissecting downstream target genes. *Front. Genet.* 13, 923339. doi:10.3389/fgene.2022.923339

Jiang, X., and Zhang, X. (2022). Rsnet: Inferring gene regulatory networks by a redundancy silencing and network enhancement technique. *BMC Bioinforma.* 23, 165. doi:10.1186/s12859-022-04696-w

Langfelder, P., and Horvath, S. (2008). Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics* 24, 719–720. doi:10.1093/bioinformatics/btm563

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, 550. doi:10.1186/s13059-014-0550-8

Nafis, S., Kalaiarasan, P., Brojen Singh, R. K., Husain, M., and Bamezai, R. N. K. (2015). Apoptosis regulatory protein-protein interaction demonstrates hierarchical scale-free fractal network. *Brief. Bioinform.* 16, 675–699. doi:10.1093/bib/bbu036

Pearson, K., and Galton, F. V. I. I. (1895). Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58, 240–242.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Weiss, R., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Ramos-Carreño, C., and Torrecilla, J. L. (2022). dcor: distance correlation and energy statistics in Python. *Orig. Softw. Publ.* 22, 101326. doi:10.5281/zenodo.7484447

Rau, C. D., Wisniewski, N., Orozco, L. D., Bennett, B., Weiss, J., and Lusis, A. J. (2013). Maximal information component analysis: A novel non-linear network analysis method. *Front. Genet.* 4, 28. doi:10.3389/fgene.2013.00028

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi:10.1126/science.1073374

RousseeuwSilhouettes, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Ruan, J., and Zhang, W. (2008). Identifying network communities with a high resolution. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 77, 016104. doi:10.1103/PhysRevE.77.016104

Swisher, L. L., Beckstead, J. W., and Bebeau, M. J. (2004). Factor analysis as a tool for survey analysis using a professional role orientation inventory as an example. *Phys. Ther.* 84, 784–799. doi:10.1093/ptj/84.9.784

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.* 35, 2769–2794. doi:10.1214/009053607000000505

Székely, G. J., and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* 3, 1236–1265. doi:10.1214/09-aoas312

van Dam, S., Cordeiro, R., Craig, T., van Dam, J., Wood, S. H., and de Magalhães, J. P. (2012). GeneFriends: An online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* 13, 535. doi:10.1186/1471-2164-13-535

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2

Yip, A. M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinforma.* 8, 22. doi:10.1186/1471-2105-8-22

Yip, A. M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinforma.* 8, 22. doi:10.1186/1471-2105-8-22

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

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, 17. doi:10.2202/1544-6115.1128

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 May 2023
DOI 10.3389/fgene.2023.1141011

OPEN ACCESS

EDITED BY
Rinku Sharma,
Brigham and Women's Hospital,
United States

REVIEWED BY
Jing Li,
Fujian Medical University, China
Eliana Saul Furquim Werneck Abdelhay,
National Cancer Institute (INCA), Brazil

*CORRESPONDENCE
Jesús Espinal-Enríquez,
jespinal@inmegen.gob.mx

RECEIVED 09 January 2023
ACCEPTED 25 April 2023
PUBLISHED 09 May 2023

CITATION
Hernández-Gómez C, Hernández-Lemus E and Espinal-Enríquez J (2023), CNVs in 8q24.3 do not influence gene co-expression in breast cancer subtypes. *Front. Genet.* 14:1141011. doi: 10.3389/fgene.2023.1141011

CNVs in 8q24.3 do not influence gene co-expression in breast cancer subtypes

Candelario Hernández-Gómez[1,2], Enrique Hernández-Lemus[1,2] and Jesús Espinal-Enríquez[1,2]*

[1]Computational Genomics Division, National Institute of Genomic Medicine, México City, Mexico, [2]Center for Complexity Sciences, Universidad Nacional Autónoma de México, México City, Mexico

Gene co-expression networks are a useful tool in the study of interactions that have allowed the visualization and quantification of diverse phenomena, including the loss of co-expression over long distances in cancerous samples. This characteristic, which could be considered fundamental to cancer, has been widely reported in various types of tumors. Since copy number variations (CNVs) have previously been identified as causing multiple genetic diseases, and gene expression is linked to them, they have often been mentioned as a probable cause of loss of co-expression in cancerous networks. In order to carry out a comparative study of the validity of this statement, we took 477 protein-coding genes from chromosome 8, and the CNVs of 101 genes, also protein-coding, belonging to the 8q24.3 region, a cytoband that is particularly active in the appearance of breast cancer. We created CNVS-conditioned co-expression networks of each of the 101 genes in the 8q24.3 region using conditional mutual information. The study was carried out using the four molecular subtypes of breast cancer (Luminal A, Luminal B, Her2, and Basal), as well as a case corresponding to healthy samples. We observed that in all cancer cases, the measurement of the Kolmogorov-Smirnov statistic shows that there are no significant differences between one and other values of the CNVs for any case. Furthermore, the co-expression interactions are stronger in all cancer subtypes than in the control networks. However, the control network presents a homogeneously distributed set of co-expression interactions, while for cancer networks, the highest interactions are more confined to specific cytobands, in particular 8q24.3 and 8p21.3. With this approach, we demonstrate that despite copy number alterations in the 8q24 region being a common trait in breast cancer, the loss of long-distance co-expression in breast cancer is not determined by CNVs.

KEYWORDS

gene co-expression networks, breast cancer subtypes, copy number variations, conditional mutual information, luminal breast cancer, HER2+ breast cancer, basal breast cancer

Introduction

Regulation of gene expression involves several processes by which the information contained in the genome is transformed into proteins. These processes within eukaryotic cell include signaling, chromatin remodeling, covalent histone modification, and transcription initiation, among others. Impairing of those processes are fundamental for the development

of cancer, promoting tumor growth, cell proliferation, angiogenesis, or evasion of the immune response (Bian et al., 2022).

According to the World Health Organization, in 2020 around 685,000 people died from breast cancer (World Health Organization, 2020). It is the fifth cause of death from cancer. However, it is the first place in new diagnoses, with 2.26 million. This apparent contradiction between cases of death and those detected for breast cancer is largely explained by early detection (more than 90% of breast tumors are detected without metastases) and the relative good knowledge of the disease. Although breast cancer patients have a survival rate greater than 80% 5 years after diagnosis, this depends on the subtype.

The most commonly used classification system for breast cancer is the PAM50 gene expression signature, which divides breast cancer into four main subtypes: Luminal A, Luminal B, HER2+ and Basal-like subtypes Perou et al. (2000). The molecular classification of breast cancer arises from advances in genomic sequencing, taking into account the gene expression signature of 50 genes relevant for the disease. This classification has allowed a better evaluation, diagnosis and treatment of breast cancer.

The luminal A subtype is the most diagnosed breast cancer (between 40% and 50% of all cases) and is also the one with the best prognosis, with hormone receptor suppressors being a good therapy. The luminal B subtype is less common (between 20% and 30% of cases) but more aggressive, although with a good response to chemotherapy. About 15% of breast cancers have an overexpression of the Her2 gene and this makes them particularly aggressive. The worst prognosis is for patients with triple-negative cancer, which represent about 15% of diagnosed cases Tsang and Tse (2020).

Copy number variations (CNVs), refer to genomic changes that involve deletions or duplications of large DNA segments ranging in size from 1 KB to several megabases. Typically, a person has two copies of each gene inherited from their parents, but there are naturally occurring variations to this number. These genetic variants can include deletions, duplications, or insertions in the paternal or maternal chromosomes, or both, and are present in healthy individuals. In a more standardized definition, CNVs are stretches of DNA larger than 1 kb that display copy number differences in the normal or reference population (Scherer et al., 2007).

In cancer, CNVs can have a significant impact on gene expression and contribute to the development and progression of the disease, for instance, in oncogene amplification (Gajria and Chandarlapaty, 2011; Swain et al., 2023), tumor suppression gene deletion (Ried et al., 2019; Gupta et al., 2021), genomic instability (Duijf et al., 2019; Kalimutho et al., 2019; Neuse et al., 2020), or even drug resistance (Lim and Ma, 2019; Pös et al., 2021).

The 8q24 genomic region is a specific location on the long (q) arm of human chromosome 8. Amplifications and deletions of this region are involved in the development of certain types of cancer, such as prostate (Gu et al., 2020; Wilson and Kanhere, 2021), colon (Killian et al., 2006; Anauate et al., 2019; Nait Slimane et al., 2020), or bladder cancer (Kiltie, 2010). Research has identified several genetic variations within the 8q24 region that are associated with an increased risk of developing these cancers. In particular, the 8q24.3 region has previously been identified as one with significant activity in various types of cancer (Mahmood et al., 2014; Brusselaers et al., 2019; Ambele et al., 2020; Zheng et al., 2021), including breast cancer (Dorantes-Gilardi et al., 2021).

To analyze next-generation sequence data, contemporary biology often uses correlation networks to integrate the multiple sources of data. One of the most commonly implemented tools are the Gene co-expression networks (GCNs). GCNs are mathematical constructions based on the patterns of statistical correlation between genes across different phenotypes. These networks can help identify functionally related genes and pathways and provide insights into the underlying mechanisms of complex biological processes, such as cancer.

Previous studies found that the gene co-expression networks of cancerous samples differ significantly from those of healthy samples (Rai et al., 2017; Dorantes-Gilardi et al., 2021; Dorantes-Gilardi et al., 2020). In adjacent-to-tumor breast tissue, gene co-expression networks show a higher connection between genes from different chromosome, indicating coordination and cooperation between genes. However, this co-expression is dramatically lost in cancer GCNs, both when all subtypes are analyzed together (Espinal-Enríquez et al., 2017) and for subtype-specific GCNs (Alcalá-Corona et al., 2017; García-Cortés et al., 2020; González-Espinoza et al., 2021). The genes in cancerous samples tend to co-express mainly with their nearest neighbors and lose co-expression relations with medium and long distance genes. This phenomenon has been observed in lung cancer (Andonegui-Elguera et al., 2021), clear cell renal carcinoma (Zamora-Fuentes et al., 2020; Zamora-Fuentes et al., 2022), as well as other thirteen types of cancer (Garcia-Cortes et al., 2022).

The cause of the loss of co-expression in cancerous sample networks is still unknown. However, a general alteration in the transcriptional regulatory program could be underlying this effect. Therefore, assessing the influence that CNVs may exert on gene co-expression networks results appealing. In a previous work (Hernández-Gómez et al., 2022), we demonstrated that in Luminal B breast cancer molecular subtype, the copy number alterations of chromosome 8 influences marginally the gene co-expression landscape. Notwithstanding, the intrinsic heterogeneity of breast cancer molecular subtype could be differentially affected by CNVs, and concomitantly, the associated co-expression network.

Taking into account the previous studies that found differences in gene co-expression networks between cancerous and healthy samples, in this work, we proposed to analyze the influence of CNVs of the 8q24.3 region in the gene co-expression networks for each breast cancer molecular subtype. We analyzed the topological influence, the association of CNVs with network hubs, and the role of such hubs in a subtype-specific fashion.

Materials and methods

Cancer and healthy samples were obtained from The Cancer Genome Atlas Consortium (TCGA) and preprocessed according to (Espinal-Enríquez et al., 2017). All samples were classified according to (Dorantes-Gilardi et al., 2021), resulting in 210 samples for Luminal A, 189 samples for Luminal B, 101 samples for HER2+, 215 samples for Basal, and 113 samples for normal adjacent-to-tumor tissues. The expression of 477 genes coding for proteins on chromosome 8 and the CNVs of 101 genes in the 8q24.3 region were analyzed for each sample.

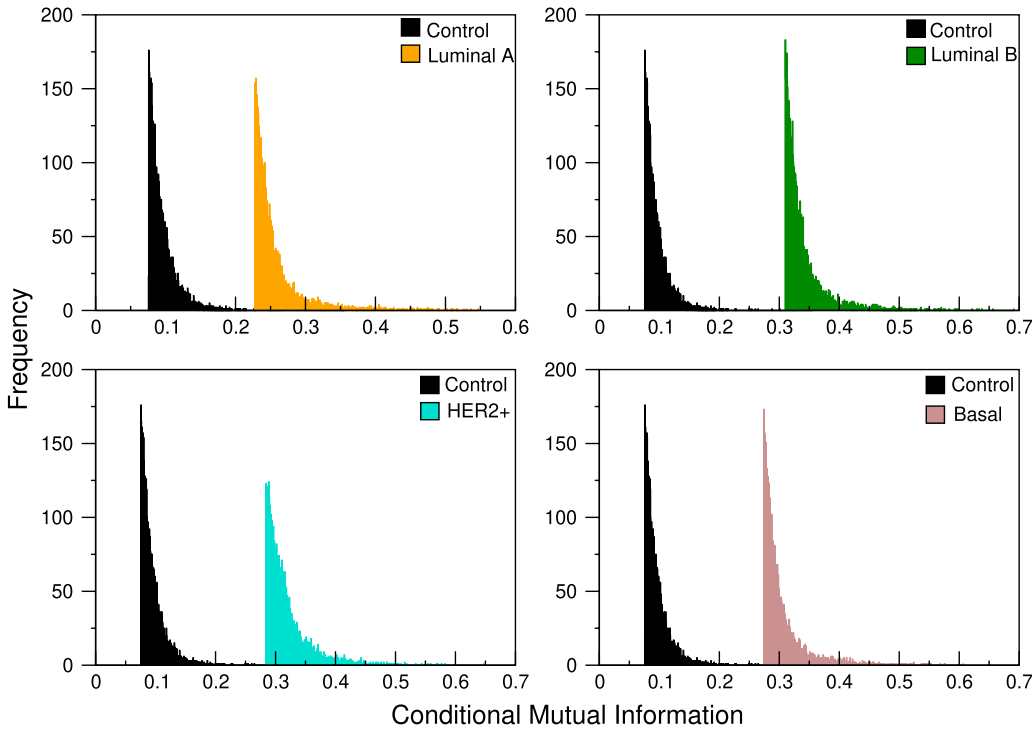


FIGURE 1
CMI comparison between control network and the four breast cancer subtypes. The abrupt cut in the left tail is due to the fact that we take, for each analysis, a pre-specified number of links, always staying with those that have the highest conditional mutual information values. In the distributions shown here, the cut selects only the first 3,500 links.

We used the copy number alteration observed in chromosome 8 for the five phenotypes. A total of 101 CNVs were obtained using ascat data. For each of these CNV values, we constructed a CNV-specific gene co-expression network. To infer the conditional mutual information (CMI) for all phenotypes, we calculated as in (Hernández-Gómez et al., 2022), taking into account the co-expression between genes depending on the CNV values of chromosome 8 to observe the effect of variations in copy number on the co-expression of the entire genome. In this way, we obtained one network per CNV value, each of which can be considered a layer of a multi-CNVs co-expression network.

CMI calculations are thus the core of our analytic approach. In brief, CMI reflects the degree to which a random variable (here, the expression level g_i of a given gene i) is statistically dependent on another random variable (the expression level g_j of gene j) given a third random variable (the copy number landscape in the given genomic region k, CNV_k) potentially affecting the relationship between g_i and g_j.

$CMI(g_i, g_j|CNV_k)$ thus reflects the amount of information we have about the expression of gene i given our knowledge of the expression of gene j in the presence of copy number alterations in the region k. For the present case, $CMI(g_i, g_j|CNV_k)$ answers the following question: is the copy number landscape in the regions changing the way two genes are locally co-expressed or not?

To provide a statistically meaningful response to this question, it is necessary, nor only to provide systematic calculations of $CMI(g_i, g_j|CNV_k)$ for all the considered genes i and j and all the regions k, but also to perform rigorous hypothesis testing. To do this, we have resorted to the quite general and non-parametric, Kolmogorov-Smirnov test; since no assumptions need to be made in the nature of the probability distributions for gene expression nor copy number variants.

Hence, after constructing the networks, we calculated the Kolmogorov-Smirnov statistic to quantify differences between CMI layers. Once the CMI networks were constructed, we compared the number of intra-cytoband, inter-cytoband, and inter-arm *cis*-gene pairs for all chromosomes in the five phenotypes. We also evaluated the variations of these numbers depending on the CMI cutoff values and observed whether the intra-cytoband, inter-cytoband, and inter-arm numbers changed in accordance with the cutoff values.

Finally, we analyzed the most relevant genes in terms of their topological properties. We identified those genes that are both relevant for the structure and relevant for the proper function of a given phenotype.

Conditional mutual information

Mutual information $I(X; Y)$ is a measure of the mutual dependence between two random variables. It quantifies the amount of information that one random variable contains about the other. In other words, it measures the amount of reduction in uncertainty about one random variable given knowledge of the other random variable. Conditional mutual information, $I(X; Y|Z)$, is the value of the mutual information between two random variables X and Y given (i.e., conditional to) the value of a third random value Z.

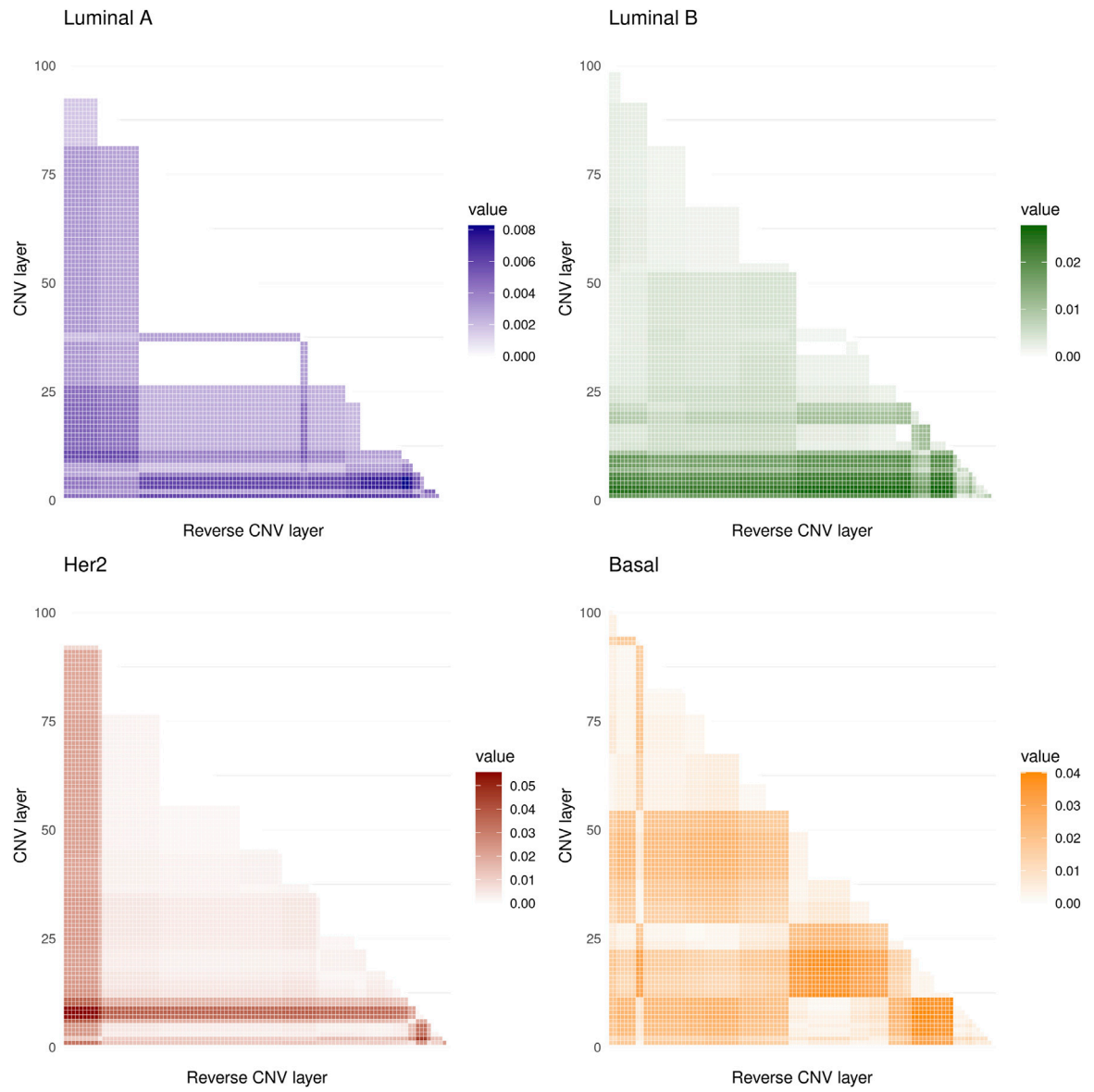


FIGURE 2
The heat maps show the values of the D statistic between the different distributions of the CMI values for the four molecular subtypes analyzed. 5,050 comparisons were made in all cases. The lowest values of D were obtained for the Luminal A subtype while the highest occurred in the Her2 subtype.

Conditional mutual information measures the amount of reduction in uncertainty about one random variable given knowledge of the other random variable, but only in the context of a specific value of the third random variable.

$$I(X;Y|Z) = \sum_{z\epsilon z} \sum_{y\epsilon Y} \sum_{x\epsilon X} p(x,y,z) \log\left(\frac{p_z(z)p(x,y,z)}{p_{x,z}(x,z)\,p_{y,z}(y,z)}\right) \quad (1)$$

Where $p(x, y, z)$ is the joint probability of X, Y and Z, $p(x, y)$ is the joint probability of X and Y and so on. It is worth noticing that conditional mutual information can only provide information about the dependence between the random variables, and cannot provide information about the causality between them.

Conditional mutual information calculations in this work were made with the infotheo library of the R programming language (Meyer and Meyer, 2009).

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test is a statistical test used to determine whether a sample of data comes from a known distribution. It is a non-parametric test, meaning that it makes no assumptions about the form of the distribution of the data. The test compares the empirical cumulative distribution function of the

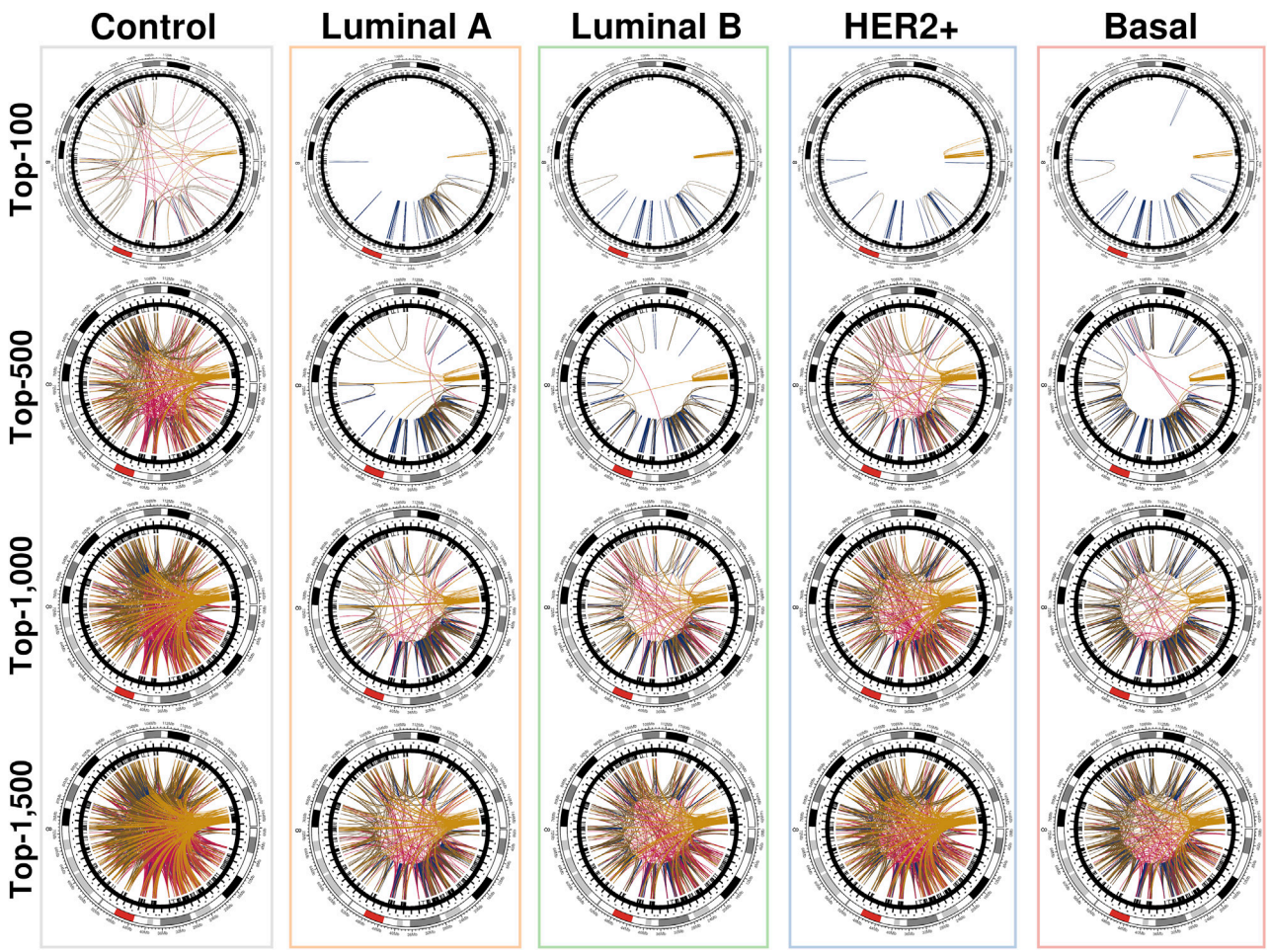


FIGURE 3
Top interactions of the five CMI networks at different cut-offs (100, 500, 1,000, and 1,500 edges). At first, intra-cytoband interactions dominate, mainly in q24.3, p21.3, p11.21 and p11.23; afterwards, inter-cytoband interactions (particularly in p-arm) grow, and finally, inter-arm edges arise. Red arcs at the external circle represents the centromere of Chr8. It can be clearly appreciated that for the normal tissue network, the distribution of interactions is remarkably more homogeneous than any cancer network, where interactions are preferentially located to neighboring regions. Circos plots were made with the R programming language package *circlize* (Cui et al., 2016).

sample data to the cumulative distribution function of the known distribution, and quantifies the (maximal) difference between the two. If the difference is large enough, the null hypothesis that the sample data comes from the known distribution is rejected.

The KS test hence compares the cumulative distributions F_1 and F_2 of two probability functions f_1 and f_1 by quantifying the K-S statistic, defined as

$$D_{n,m} = sup_x |F_{1,n}(x) - F_{2,m}(x)| \quad (2)$$

The null hypothesis is rejected (at significance level α), whenever:

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n \cdot m}} \quad (3)$$

where $c(\alpha) = \sqrt{-ln(\frac{\alpha}{2}) \cdot \frac{1}{2}}$

In the present context the KS test is appropriate since the sample sizes are sufficiently large and the CMI distributions can be safely assumed to be continuous. All tests were done using the ks.test library of the R programming language.

Results and discussion

Here we report the main results of analyzing the conditional mutual information distributions associating the pairwise co-expression of genes conditional on the copy number landscape of the respective regions. These are data-based probabilistic tools to assess to what extent gene co-expression is affected by the underlying CNV structure in the same samples.

Copy number alterations in 8q24.3 do not influence gene co-expression in breast cancer

Each of the 101 genes of the 8q24.3 region for which the CNVs were used as the conditional variable in Eq. 1 producing 101 different distributions of CMI values, whose typical profile can be seen in Figure 1. Since the distributions suggest that the differences between them are minimal for all subtypes, we

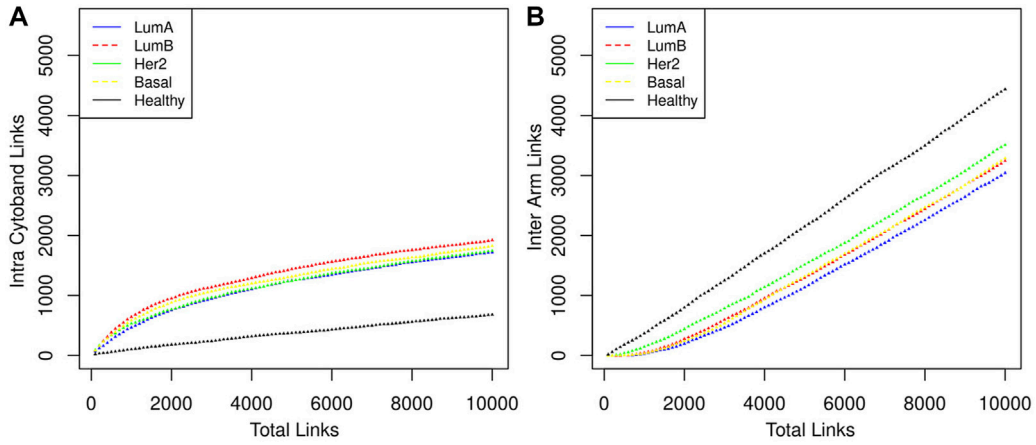


FIGURE 4
The distribution of the links in three categories is shown. **(A)** Intra cytoband. Co-expression with nearest neighbors is something that genes do in both healthy and cancerous phenotypes, although this tendency is markedly greater in the latter case. **(B)** Inter arm. In this category the behavior between the healthy phenotype and the cancerous ones is very marked, indeed Luminal B and Basal cases overlap throughout almost the entire range. It should be noted that the *order of appearance* of the links is determined by the magnitude of the CMI.

performed the K-S test in each case. For each subtype we take each of the 101 distributions and compare them with the remaining 100; since $D_{n,m} = D_{m,n}$, we have $101 \times 100/2 = 5{,}050$ comparisons for each subtype. This is, we tested the null hypothesis that the distributions of CMI values are the same for each layer.

The values obtained are shown in Figure 2 where it can be seen that the maximum values of the D statistic for any subtype are low, the larger values are for the Her2 subtype and are approximately equal to 0.05, and the minima correspond to Luminal A, with values of around 0.008. Based on these values, we conclude that CNVs within the 8q24.3 region do not significantly affect the expression of genes located on chromosome 8.

It is worth noting that there is no significant variation between conditional layers for any of the phenotypes analyzed here. With this, we show that copy number alterations are not a significant factor altering the gene co-expression landscape in breast cancer or in healthy tissue in this region of chromosome 8.

Copy number variation is indeed one of several aspects that can influence (either individually or on a cooperative fashion) gene expression and co-expression patterns. Since we have been studying local gene co-expression phenomena and intuitively, one expects that the influence of CNVs on gene expression will also be predominantly local, we decided to perform a comprehensive analysis looking at all the pairwise co-expression relationships within chromosome 8, conditional on the full CNV variant landscape of the 8q24.3 region.

Intra-chromosomal co-expression analysis

Intra-chromosomal gene co-expression refers to the simultaneous expression of genes that are located on the same chromosome. This means that they are physically close in the DNA sequence.

Intra-chromosomal gene co-expression can occur for different reasons. For example, genes that are physically close to one another on a chromosome may be regulated by the same regulatory elements, such as enhancers or promoters. This can lead to the coordinated expression of these genes.

The following results aim to present a broader view of this phenomenon in the context of breast cancer molecular subtypes.

The networks shown in Figure 3 were constructed using the first distribution and are representative of the behavior of all conditional layers. There, circos plots of gene co-expression interactions in chromosome 8 for the top-100, 500, 1,000 and 1,500 highest CMI values are depicted for all phenotypes.

We can notice that Figure 3 is better understood when compared to Figure 4, which shows the cumulative growth of intra-cytoband and inter-arm links. By observing the growth line corresponding to the network of healthy samples as a reference, it can be seen how each subtype differs from the healthy reference network in terms of the growth of intra-cytoband and inter-arm interactions. Firstly, there is a lineal growth of intra-cytoband and inter-arm interactions in the healthy case, which is not the case of any breast cancer subtype. Additionally, all subtypes behave similarly in both panels, but with small differences. In Figure 4A, all breast cancer co-expression networks have a fast growth of intra-cytoband links, which is inversely proportional to the slow increase in the inter-arm edges.

In (García-Cortés et al., 2020) we demonstrated that the loss of inter-chromosomal interactions in breast cancer is evident in all phenotypes. Furthermore, the intensity of this loss is in agreement with the *malignancy* of the subtype: the most remarkable difference with respect to the healthy tissue network was observed in the Basal subtype, followed by HER2, then Luminal B, and finally, the most similar behavior to the control phenotype was observed in Luminal A.

Despite the Basal subtype being the most aggressive and the one with the worst prognosis, in the particular case of Chr8 intra-chromosomal edges, the most different behavior compared with the healthy case is observed in the Luminal B network (red lines in Figure 4).

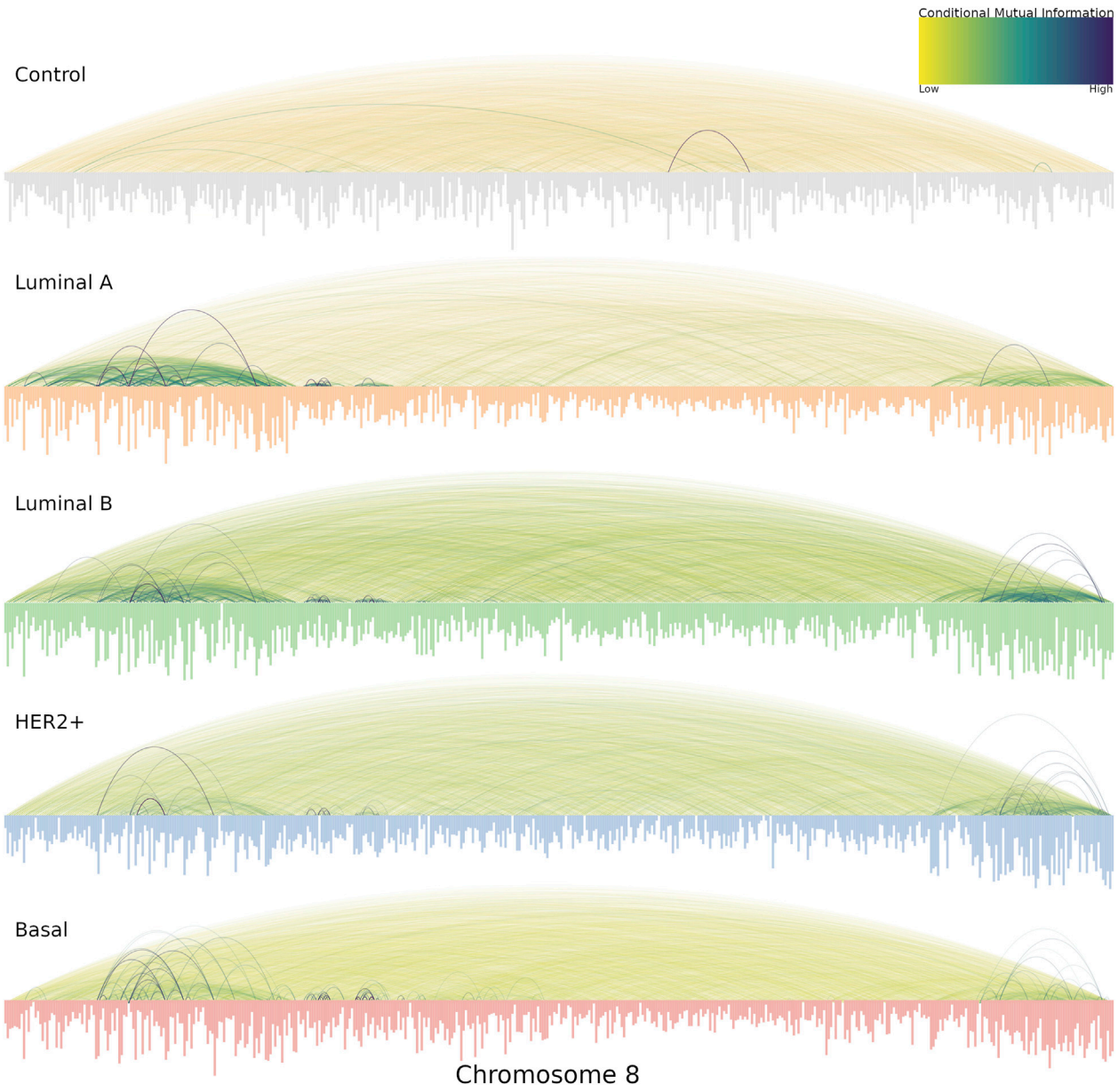


FIGURE 5
Chromosome 8 co-expression interactions for the five phenotypes are shown, with the genes placed according to its gene start position. The size of the genes is proportional to its degree. The color of co-expression interactions is related to the CMI values. Notice that for all cancer networks, the strongest interactions occur in the extreme places, in particular, 8p21 and 8q24.3.

In Figure 4 the total number of links ranges from 1 to (the top) 10,000, which gives a good sample of the behavior we want to illustrate. Intra-cytoband links grow a lot in the first few hundred larger CMI values in cancers but then they tend to saturate (see change of curvature in the plots). Eventually, all possible links will be formed and the curves will reach their maximum value. Finally, all link saturation lines are well above/below (panels A and B in Figure 4, respectively) the behavior of the healthy phenotype, thus showing the deficit of links at long distances as previously reported.

Another relevant aspect that we noticed in the cancer chr8 gene co-expression networks is the location of highest co-expression values. This can be appreciated in Figure 5. In the case of the healthy network (labeled *Control* at the top) the vast majority of interactions present similar co-expression values, that is the reason for which several edges in the network present similar color. On the other hand, in the case of all breast cancer subtypes, highly dense regions of strong co-expression values are evident. Importantly, in all cancer cases, the q24.3 region contains a hotspot of strong interactions. Importantly, in all cancer cases, a hotspot of strong interactions is present in the q24.3 region. On the one hand, luminal networks present a large region from p23.3 to p11.23, while HER2 and Basal subtypes present a much more localized p-arm hotspot at p21.1-p21.3

Additionally, in the cancer networks, the degree of nodes is clearly higher than in their healthy counterparts. That is represented by the length of bars depicted below for each network of Figure 5.

Interestingly, the most connected genes in all cancer networks are located at 8p21.3, except in the HER2 network, where the most connected genes belong to the 8q24.3 region. Conversely, the healthy network's most connected genes belong to q13 and q22 regions.

We have previously reported the appearance of a highly connected region located at 8q24.3 in breast cancer subtypes Dorantes-Gilardi et al. (2021). There, we showed that 8q24.3 is the only region in the entire genome in which all breast cancer subtypes present the same set of highly co-expressed interactions (top-100,00). That results is one of the motivations of this work, being focused in Chromosome 8. Here, we demonstrate that 8q24.3 is still important in terms of co-expression in breast cancer subtypes, but also 8p21 emerges as a relevant region.

In the case of HER2 network, it is worth noting that the HER2-enriched subtype was indeed named so, because of the amplification of a specific part of chromosome 17. In this work, we observe that 8q24.3 and 8p21.3 regions are also important but they do not depend on the copy number alterations, such as the case of amplification of 17q12 region, which is also related with global genomic instability (Ellsworth et al., 2008).

We want to stress that all of these results were obtained with TCGA-derived data. Further research must include other datasets in order to corroborate that these results are consistent independently of the data source.

Conclusions and perspectives

The main conclusions of this work can be recapitulated in the form of a summary of findings, as follows:

1. Copy number alterations in the 8q24.3 region do not significantly affect gene co-expression in chromosome 8. Therefore, the loss of long-distance co-expression must be triggered by a different mechanism.
2. Basal and Luminal B breast cancer subtypes have the most remarkable loss of long-distance co-expression in this region.
3. HER2+ subtype has a worse prognosis than Luminal B, however, Luminal B behaves more differently from the healthy tissue. Perhaps, Luminal B has another mechanism involved in the co-expression program and the observed behavior in the chromosome 8 co-expression network is a manifestation of that.

For our dataset, CNVs does not influence gene co-expression networks in breast cancer in this region. However, copy number alterations are known to affect gene expression at different levels. The loss of long distance co-expression is strongly maintained in all cancer phenotypes, but in a different intensity.

The analysis performed here has been implemented for breast cancer molecular subtypes. Another classification approaches such as the TNM system, which is based on the tumor progression, should be incorporated to broaden the implications of copy number alterations in terms of their role on tumor progression. Further research in this line must be addressed to evaluate other aspects of CNVs in breast cancer.

Finally, this kind of analyses using different omic-approaches will definitively enhance our perspective and understanding of complex diseases such as breast cancer. We can envision to perform similar analysis at a whole genome scale in the future, though this endeavor will imply a high computational burden due to combinatorial effects. However, it is necessary to determine whether or not the copy number alterations observed in cancer are associated with the appearance of the phenomenon of loss of long-distance co-expression.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

CH-G organized data, developed code, performed calculations, analyzed data, drafted the manuscript; EH-L designed the methodological approach, analyzed data, discussed results, co-supervised the project. JE-E devised and coordinated the project, contributed to the methodological strategy, analyzed data and integrated the biological results, discussed results, co-supervised the project; All authors read and approved the final manuscript.

Funding

This work was supported by the Consejo Nacional de Ciencia y Tecnología (Estancia Posdoctoral de Incidencia, CHG), and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México.

Acknowledgments

Authors want to thank Diana García-Cortés and José M. Zamora-Fuentes for productive discussions.

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

References

Alcalá-Corona, S. A., de Anda-Jáuregui, G., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network modularity in breast cancer molecular subtypes. *Front. physiology* 8, 915. doi:10.3389/fphys.2017.00915

Ambele, M. A., Van Zyl, A., Pepper, M. S., Van Heerden, M. B., and Van Heerden, W. F. (2020). Amplification of 3q26 2, 5q14 3, 8q24 3, 8q22 3, and 14q32 33 are possible common genetic alterations in oral cancer patients. *Front. Oncol.* 10, 683. doi:10.3389/fonc.2020.00683

Anauate, A. C., Leal, M. F., Wisnieski, F., Santos, L. C., Gigek, C. O., Chen, E. S., et al. (2019). Analysis of 8q24 21 mirna cluster expression and copy number variation in gastric cancer. *Future Med. Chem.* 11, 947–958. doi:10.4155/fmc-2018-0477

Andonegui-Elguera, S. D., Zamora-Fuentes, J. M., Espinal-Enríquez, J., and Hernández-Lemus, E. (2021). Loss of long distance co-expression in lung cancer. *Front. Genet.* 12, 625741. doi:10.3389/fgene.2021.625741

Bian, X., Jiang, H., Meng, Y., Li, Y.-p., Fang, J., and Lu, Z. (2022). Regulation of gene expression by glycolytic and gluconeogenic enzymes. *Trends Cell Biol.* 32, 786–799. doi:10.1016/j.tcb.2022.02.003

Brusselaers, N., Ekwall, K., and Durand-Dubief, M. (2019). Copy number of 8q24. 3 drives hsf1 expression and patient outcome in cancer: An individual patient data meta-analysis. *Hum. genomics* 13, 54–12. doi:10.1186/s40246-019-0241-3

Cui, Y., Chen, X., Luo, H., Fan, Z., Luo, J., He, S., et al. (2016). Biocircos. Js: An interactive circos javascript library for biological data visualization on web applications. *Bioinformatics* 32, 1740–1742. doi:10.1093/bioinformatics/btw041

Dorantes-Gilardi, R., García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2021). k-core genes underpin structural features of breast cancer. *Sci. Rep.* 11, 16284–16317. doi:10.1038/s41598-021-95313-y

Dorantes-Gilardi, R., García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2020). Multilayer approach reveals organizational principles disrupted in breast cancer co-expression networks. *Appl. Netw. Sci.* 5, 47–23. doi:10.1007/s41109-020-00291-1

Duijf, P. H., Nanayakkara, D., Nones, K., Srihari, S., Kalimutho, M., and Khanna, K. K. (2019). Mechanisms of genomic instability in breast cancer. *Trends Mol. Med.* 25, 595–611. doi:10.1016/j.molmed.2019.04.004

Ellsworth, R. E., Ellsworth, D. L., Patney, H. L., Deyarmin, B., Love, B., Hooke, J. A., et al. (2008). Amplification of her2 is a marker for global genomic instability. *BMC cancer* 8, 297–299. doi:10.1186/1471-2407-8-297

Espinal-Enríquez, J., Fresno, C., Anda-Jáuregui, G., and Hernández-Lemus, E. (2017). Rna-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep.* 7, 1760–1819. doi:10.1038/s41598-017-01314-1

Gajria, D., and Chandarlapaty, S. (2011). Her2-amplified breast cancer: Mechanisms of trastuzumab resistance and novel targeted therapies. *Expert Rev. anticancer Ther.* 11, 263–275. doi:10.1586/era.10.226

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernández-Lemus, E., and Espinal-Enríquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol.* 10, 1232. doi:10.3389/fonc.2020.01232

Garcia-Cortes, D., Hernandez-Lemus, E., and Espinal-Enriquez, J. (2022). Loss of long-range co-expression is a common trait in cancer. bioRxiv.

González-Espinoza, A., Zamora-Fuentes, J., Hernández-Lemus, E., and Espinal-Enríquez, J. (2021). Gene co-expression in breast cancer: A matter of distance. *Front. Oncol.* 11, 726493. doi:10.3389/fonc.2021.726493

Gu, Y., Lin, X., Kapoor, A., Chow, M. J., Jiang, Y., Zhao, K., et al. (2020). The oncogenic potential of the centromeric border protein fam84b of the 8q24 21 gene desert. *Genes* 11, 312. doi:10.3390/genes11030312

Gupta, S., Sukov, W. R., Vanderbilt, C. M., Shen, W., Herrera-Hernandez, L., Lohse, C. M., et al. (2021). A contemporary guide to chromosomal copy number profiling in the diagnosis of renal cell carcinoma. *Urologic Oncol. Seminars Orig. Investigations* 40, 512. doi:10.1016/j.urolonc.2021.04.042

Hernández-Gómez, C., Hernández-Lemus, E., and Espinal-Enríquez, J. (2022). The role of copy number variants in gene co-expression patterns for luminal b breast tumors. *Front. Genet.* 13, 806607. doi:10.3389/fgene.2022.806607

Kalimutho, M., Nones, K., Srihari, S., Duijf, P. H., Waddell, N., and Khanna, K. K. (2019). Patterns of genomic instability in breast cancer. *Trends Pharmacol. Sci.* 40, 198–211. doi:10.1016/j.tips.2019.01.005

Killian, A., Sarafan-Vasseur, N., Sesboüé, R., Le Pessot, F., Blanchard, F., Lamy, A., et al. (2006). Contribution of the bop1 gene, located on 8q24, to colorectal tumorigenesis. *Genes, Chromosomes Cancer* 45, 874–881. doi:10.1002/gcc.20351

Kiltie, A. E. (2010). Common predisposition alleles for moderately common cancers: Bladder cancer. *Curr. Opin. Genet. Dev.* 20, 218–224. doi:10.1016/j.gde.2010.01.002

Lim, Z.-F., and Ma, P. C. (2019). Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *J. Hematol. Oncol.* 12, 134–218. doi:10.1186/s13045-019-0818-2

Mahmood, S. F., Gruel, N., Chapeaublanc, E., Lescure, A., Jones, T., Reyal, F., et al. (2014). A sirna screen identifies rad21, eif3h, chrac1 and tanc2 as driver genes within the 8q23, 8q24. 3 and 17q23 amplicons in breast cancer with effects on cell growth, survival and transformation. *Carcinogenesis* 35, 670–682. doi:10.1093/carcin/bgt351

Meyer, P. E., and Meyer, M. P. E. (2009). Package 'infotheo'. R Packag. version 1

Nait Slimane, S., Marcel, V., Fenouil, T., Catez, F., Saurin, J.-C., Bouvet, P., et al. (2020). Ribosome biogenesis alterations in colorectal cancer. *Cells* 9, 2361. doi:10.3390/cells9112361

Neuse, C. J., Lomas, O. C., Schliemann, C., Shen, Y. J., Manier, S., Bustoros, M., et al. (2020). Genome instability in multiple myeloma. *Leukemia* 34, 2887–2897. doi:10.1038/s41375-020-0921-y

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *nature* 406, 747–752. doi:10.1038/35021093

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., et al. (2021). Dna copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* 44, 548–559. doi:10.1016/j.bj.2021.02.003

Rai, A., Pradhan, P., Nagraj, J., Lohitesh, K., Chowdhury, R., and Jalan, S. (2017). Understanding cancer complexome using networks, spectral graph theory and multilayer framework. *Sci. Rep.* 7, 1–16. doi:10.1038/srep41676

Ried, T., Meijer, G. A., Harrison, D. J., Grech, G., Franch-Expósito, S., Briffa, R., et al. (2019). The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. *Mol. aspects Med.* 69, 48–61. doi:10.1016/j.mam.2019.07.007

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., et al. (2007). Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* 39, S7–S15. doi:10.1038/ng2093

Swain, S. M., Shastry, M., and Hamilton, E. (2023). Targeting her2-positive breast cancer: Advances and future directions. *Nat. Rev. Drug Discov.* 22, 101–126. doi:10.1038/s41573-022-00579-0

Tsang, J., and Tse, G. M. (2020). Molecular classification of breast cancer. *Adv. anatomic pathology* 27, 27–35. doi:10.1097/PAP.0000000000000232

Wilson, C., and Kanhere, A. (2021). 8q24 21 locus: A paradigm to link non-coding rnas, genome polymorphisms and cancer. *Int. J. Mol. Sci.* 22, 1094. doi:10.3390/ijms22031094

World Health Organization (2020). Cancer Today international agency for research on cancer. Available at: https://gco.iarc.fr/today/home. [Accessed: 2022-11-17].

Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. (2020). Gene expression and co-expression networks are strongly altered through stages in clear cell renal carcinoma. *Front. Genet.* 11, 578679. doi:10.3389/fgene.2020.578679

Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. (2022). Oncogenic role of mir-217 during clear cell renal carcinoma progression. *Front. Oncol.* 12, 934711. doi:10.3389/fonc.2022.934711

Zheng, Y., Lei, T., Jin, G., Guo, H., Zhang, N., Chai, J., et al. (2021). Lncpsca in the 8q24. 3 risk locus drives gastric cancer through destabilizing ddx5. *EMBO Rep.* 22, e52707. doi:10.15252/embr.202152707

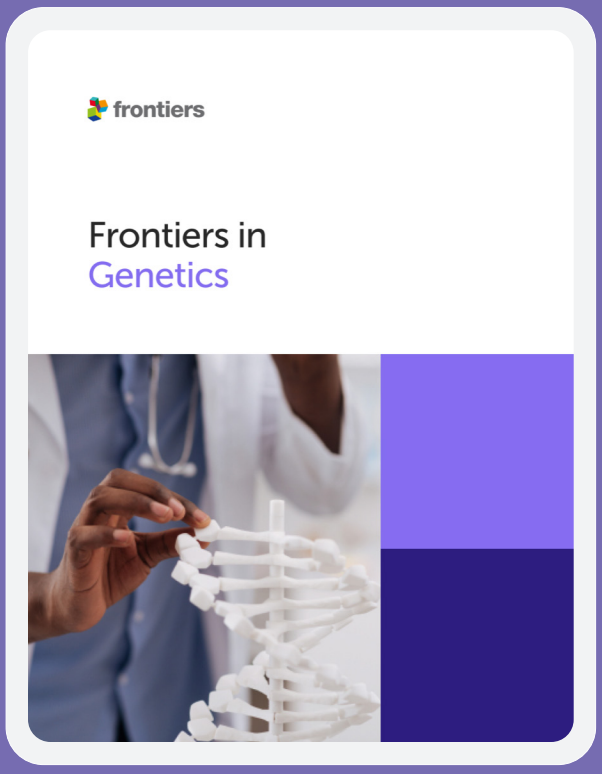
Frontiers in Genetics

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