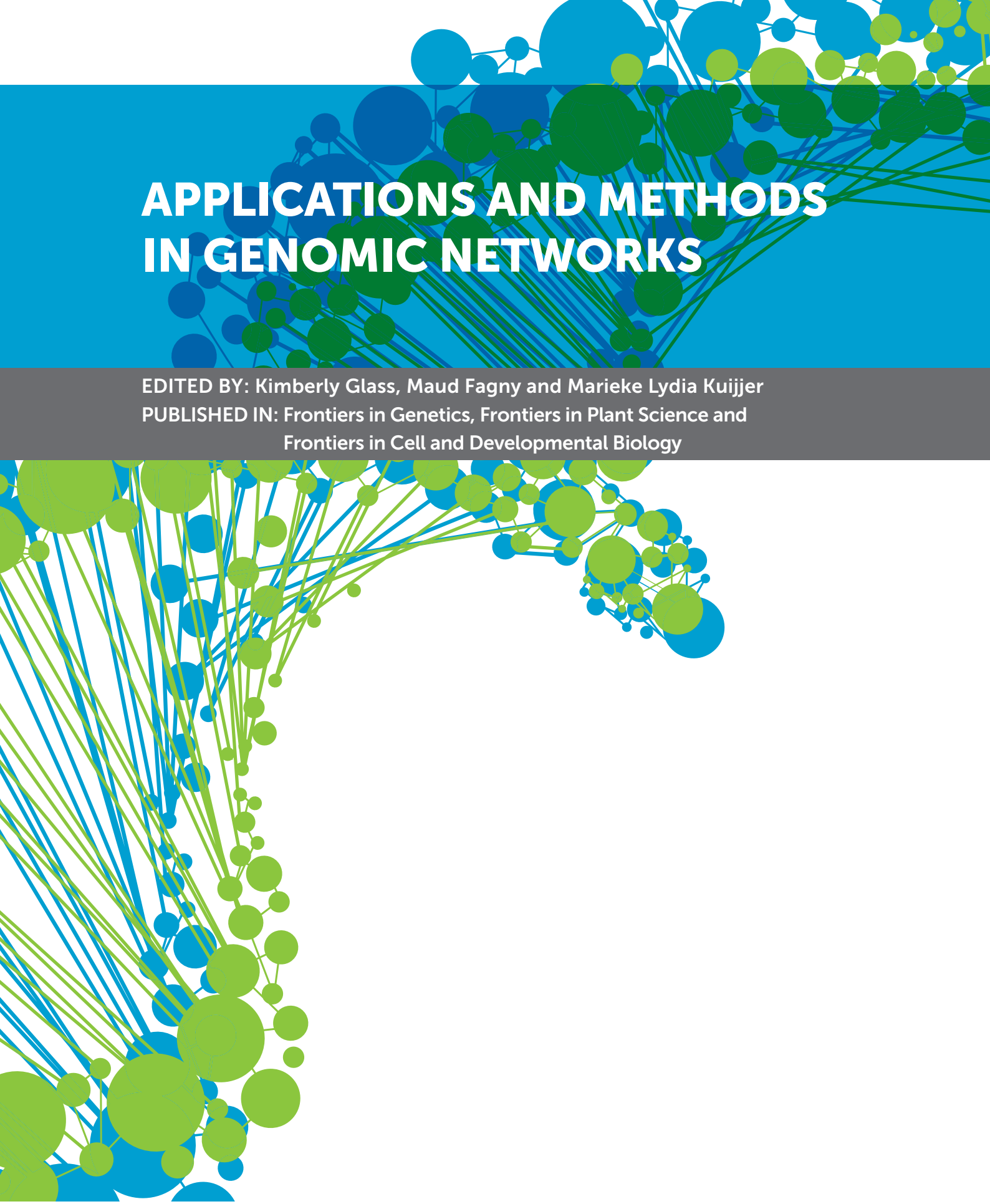

APPLICATIONS AND METHODS IN GENOMIC NETWORKS

EDITED BY: Kimberly Glass, Maud Fagny and Marieke Lydia Kuijjer

PUBLISHED IN: Frontiers in Genetics, Frontiers in Plant Science and Frontiers in Cell and Developmental Biology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-482-2
DOI 10.3389/978-2-88976-482-2

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

APPLICATIONS AND METHODS IN GENOMIC NETWORKS

Topic Editors:
Kimberly Glass, Brigham and Women's Hospital, Harvard Medical School, United States
Maud Fagny, Institut National de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), France
Marieke Lydia Kuijjer, University of Oslo, Norway

Citation: Glass, K., Fagny, M., Kuijjer, M. L., eds. (2022). Applications and Methods in Genomic Networks. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-482-2

Table of Contents

05 ***Editorial: Applications and Methods in Genomic Networks***
Maud Fagny, Kimberly Glass and Marieke L. Kuijjer

08 ***GelFAP: Gene Functional Analysis Platform for* Gastrodia elata**
Jiaotong Yang, Qiaoqiao Xiao, Jiao Xu, Lingling Da, Lanping Guo, Luqi Huang, Yue Liu, Wenying Xu, Zhen Su, Shiping Yang, Qi Pan, Weike Jiang and Tao Zhou

19 ***Gene Expression and Co-expression Networks are Strongly Altered Through Stages in Clear Cell Renal Carcinoma***
Jose María Zamora-Fuentes, Enrique Hernández-Lemus and Jesús Espinal-Enríquez

35 ***Inference of Genetic Networks From Time-Series and Static Gene Expression Data: Combining a Random-Forest-Based Inference Method With Feature Selection Methods***
Shuhei Kimura, Ryo Fukutomi, Masato Tokuhisa and Mariko Okada

46 ***Generating Ensembles of Gene Regulatory Networks to Assess Robustness of Disease Modules***
James T. Lim, Chen Chen, Adam D. Grant and Megha Padi

64 ***Abiotic Stress-Responsive miRNA and Transcription Factor-Mediated Gene Regulatory Network in* Oryza sativa*: Construction and Structural Measure Study***
Rinku Sharma, Shashankaditya Upadhyay, Sudeepto Bhattacharya and Ashutosh Singh

77 ***Gene-Microbiome Co-expression Networks in Colon Cancer***
Irving Uriarte-Navarrete, Enrique Hernández-Lemus and Guillermo de Anda-Jáuregui

93 ***CoExp: A Web Tool for the Exploitation of Co-expression Networks***
Sonia García-Ruiz, Ana L. Gil-Martínez, Alejandro Cisterna, Federico Jurado-Ruiz, Regina H. Reynolds, NABEC (North America Brain Expression Consortium), Mark R. Cookson, John Hardy, Mina Ryten and Juan A. Botía

106 ***Integrated Protein–Protein Interaction and Weighted Gene Co-expression Network Analysis Uncover Three Key Genes in Hepatoblastoma***
Linlin Tian, Tong Chen, Jiaju Lu, Jianguo Yan, Yuting Zhang, Peifang Qin, Sentai Ding and Yali Zhou

122 ***Loss of Long Distance Co-Expression in Lung Cancer***
Sergio Daniel Andonegui-Elguera, José María Zamora-Fuentes, Jesús Espinal-Enríquez and Enrique Hernández-Lemus

134 ***An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation***
Soledad Ochoa, Guillermo de Anda-Jáuregui and Enrique Hernández-Lemus

147 ***Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms***
Yoonjee Kang, Denis Thieffry and Laura Cantini

157 ***System-Level Analysis of Alzheimer's Disease Prioritizes Candidate Genes for Neurodegeneration***
Jeffrey L. Brabec, Montana Kay Lara, Anna L. Tyler and J. Matthew Mahoney for the Alzheimer's Disease Neuroimaging Initiative

175 ***Luminal A Breast Cancer Co-expression Network: Structural and Functional Alterations***
Diana García-Cortés, Enrique Hernández-Lemus and Jesús Espinal-Enríquez

191 ***Community Detection in Large-Scale Bipartite Biological Networks***
Genís Calderer and Marieke L. Kuijjer

200 ***Gene Targeting in Disease Networks***
Deborah Weighill, Marouen Ben Guebila, Kimberly Glass, John Platig, Jen Jen Yeh and John Quackenbush

207 ***Comparing Statistical Tests for Differential Network Analysis of Gene Modules***
Jaron Arbet, Yaxu Zhuang, Elizabeth Litkowski, Laura Saba and Katerina Kechris

221 ***Filtering of Data-Driven Gene Regulatory Networks Using* Drosophila melanogaster *as a Case Study***
Yesid Cuesta-Astroz, Guilherme Gischkow Rucatti, Leandro Murgas, Carol D. SanMartín, Mario Sanhueza and Alberto J. M. Martin

frontiers | Frontiers in Genetics

EDITORIAL
published: 08 June 2022
doi: 10.3389/fgene.2022.936015

Editorial: Applications and Methods in Genomic Networks

Maud Fagny[1,2]*, *Kimberly Glass*[3,4,5]* and *Marieke L. Kuijjer*[6,7,8]*

[1]EcoAnthropology Lab, UMR 7206 CNRS/MNHN/Universite Paris Diderot, Muséum National d'Histoire Naturelle, Paris, France, [2]Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE — Le Moulon, Gif-sur-Yvette, France, [3]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, United States, [4]Harvard Medical School, Boston, MA, United States, [5]Harvard Chan School of Public Health, Boston, MA, United States, [6]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, Oslo, Norway, [7]Department of Pathology, Leiden University Medical Center, Leiden, Netherlands, [8]Leiden Center for Computational Oncology, Leiden University Medical Center, Leiden, Netherlands

Keywords: network inference, network modeling, biological network analysis, genomic networks, network applications

Editorial on the Research Topic

Applications and Methods in Genomic Networks

High-throughput technologies are generating large quantities of data. These data provide a snapshot of the molecular environment and can include transcriptomic, epigenomic, and genomic information. Network approaches are a powerful way to model the biological processes measured by these data. Over the past decade, network inference and reconstruction algorithms have been developed and applied in a variety of organisms and tissues to model interactions between genes and gene products in the cell. Network approaches hold great promise in facilitating our understanding of biological processes, as well as their relationship to health and disease. However, there are many challenges that impede translating 'omics data into meaningful networks, and in leveraging networks effectively to gain new insights into biological mechanisms and/or impact patient outcomes. Networks derived from 'omics data are often very large and therefore difficult to model, analyze, and interpret. The Research Topic on "Applications and Methods in Genomic Networks" covers several areas—from discussions about how to handle data prior to network modeling, to the presentation of innovative and novel methods for biological network inference and analysis, to how to make the results available to and usable by the genomic network community, to applications illustrating the impact of network approaches in a wide range of research fields in biology and medicine.

First, this collection contains articles tackling a wide variety of issues related to genomic network inference and analysis. Cuesta-Astroz et al. propose an approach to improve data filtering, reduce noise, and increase signal in biological networks. An important challenge in the field of network biology is to develop inference methods that retrieve actual regulatory relationships while limiting the number of false positives, and that are not overly sensitive to noise. Random forest-based methods are efficient at detecting true regulatory relationships, but create a high proportion of false positives and thus, pruning networks built with these approaches is necessary to avoid spurious regulatory relationships. To solve this issue, Kimura et al. built a pipeline combining an efficient random forest-based network inference method with a series of feature selection methods, which significantly improved the quality of the inferred network. Network inference methods should also lead to consistent results across datasets obtained from the same biological condition. This is particularly important for data-driven approaches applied to single-cell datasets, which are known to have a high level of inherent

OPEN ACCESS

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

***Correspondence:**
Maud Fagny
maud.fagny@inrae.fr
Kimberly Glass
rekrg@channing.harvard.edu
Marieke L. Kuijjer
marieke.kuijjer@ncmm.uio.no

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

Received: *04 May 2022*
Accepted: *20 May 2022*
Published: *08 June 2022*

Citation:
Fagny M, Glass K and Kuijjer ML (2022) Editorial: Applications and Methods in Genomic Networks. Front. Genet. 13:936015. doi: 10.3389/fgene.2022.936015

noise. Kang et al. propose a blueprint to benchmark network inference approaches in this context, that includes most genes present in the network, taking into account both their presence in the network and the weight of their relationships, and assesses the biological soundness of the inferred networks by comparing them to gold-standard regulatory relationships extracted from public databases.

Once researchers have properly filtered their data, inferred the networks, and filtered out spurious connections, the networks are ready for analysis. A common way of making sense of genomic networks, which often include thousands of nodes and many more edges, is to look for modules or communities, *i.e.* groups of nodes that are enriched for links to each other relative to other parts of the network. Identifying network communities, and comparing them across conditions, are two ways of identifying condition-specific regulatory relationships and extracting new biological knowledge from a network. However, finding modules within a network is an NP-hard problem, meaning that existing approaches for large networks approximate the best community structure. This raises the question of the robustness of the network structure detected and its biological interpretability. Several papers in this collection address this issue from different angles. A mini-review by Calderer and Kuijjer compares different algorithms to infer modules from bipartite networks and proposes different scores aiming at assessing the quality of each method. Three other articles focus on the comparison of networks between conditions. In a perspective piece, Weighill et al., highlight the promise of using a weighted gene degree, or "gene targeting score," toglobally compare networks inferred from data representing different conditions, in order to identify key regulatory processes in disease. Lim et al. developed Constrained Random Alteration of Network Edges (*CRANE*), a new algorithm to identify robust disease-related regulatory modules. Finally, Arbet et al. share a new algorithm aimed at identifying differentially co-expressed modules and propose an R implementation, *discoMod*. This tool tests whether connections between co-expressed genes differ between conditions, and allows the user to assess how regulatory relationships within a module vary between conditions.

Finally, two papers tackled an important issue in the genomic network field: how to disseminate genomic network results and make them usable by the broader scientific community. Yang et al. built a web platform that hosts co-expression network results from *Gastrodia elata,* an important herb in traditional Chinese medicine. The platform gives access to a series of tools that facilitate result-browsing and allows the user to perform functional analysis of genes. Garcia-Ruiz et al. propose CoExp, a web platform that allows researchers to manipulate, compare, and analyze 109 co-expression networks. Importantly, the types of web tools presented here, based on open data and widely used programming languages and softwares, can be emulated and applied to a wide range of topics and organisms.

Rapidly developing research on how to best infer the genomic networks has led to the publication of algorithms and software that are crucial tools in systems biology. Many of these tools focus on unraveling the biological networks involved in regulating gene expression at the level of a cell, tissue, or organism. The application of these tools is leading to crucial discoveries in fields as diverse as Alzheimer's disease (Brabec et al.), the control of mitochondrial gene expression in *D. melanogaster* (Cuesta-Astroz et al.), the response to abiotic stress in rice (Sharma et al.), and cancer.

In this collection, five articles from the Computational Genomics Division of the National Institute of Genomic Medicine in Mexico City use mutual information approaches to explore gene co-expression networks associated with diverse cancer stages. Zamora-Fuentes et al. analyzed both gene expression and co-expression modeled on data obtained from different stages of clear cell renal carcinoma, and found substantial differences in network topology across cancer stages, with a loss of interchromosomal (*trans*) interactions compared to control networks. A similar observation was made in lung cancer by (Andonegui-Elguera et al.). Guarcia-Cortés et al. further analyzed differences in inter- and intrachromosomal (*cis*) interactions in the luminal A subtype of breast cancer. They found that *cis*-communities were enriched in copy number deletions, representing a potential mechanism of strengthened *cis*-co-expression and loss of *trans*-co-expression in cancer. Ochoa et al. also focused on breast cancer, modeling multilayer networks based on various types of omics data and identifying potential regulatory patterns of breast cancer subtype expression. Finally, through combining gene-microbiome networks with co-expression networks in colon cancer Uriart-Navarrete et al. characterized discriminating features between early and late stage cancer.

In summary, this Research Topic presents a wide variety of novel methods for network pre-processing, modeling, benchmarking, and comparison, as well as applications of network analysis to integrate different data layers, study the control of gene expression in model organisms, as well as investigate altered associations and network properties in response to environmental triggers and disease. We believe that, together, these articles form a strong basis for discussions and future projects supporting novel method development in genomic network science, as well as future applications of large-scale network modeling in biology.

AUTHOR CONTRIBUTIONS

MF, KG, and MLK wrote the manuscript together.

FUNDING

MF was supported by the Marie Sklodowska-Curie grant PATTERNS (845083) and the INRAE WIREMAIZE project.

KG is supported by R01HL155749 from the National Heart, Lung, and Blood Institute within the National Institutes of Health. MLK is supported by the Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (187615), the Norwegian Cancer Society (214871), and the Norwegian Research Council (313932).

Conflict of Interest: The author declares 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 Plant Science

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fpls.2020.563237

GelFAP: Gene Functional Analysis Platform for *Gastrodia elata*

Jiaotong Yang[1]†, Qiaoqiao Xiao[1]†, Jiao Xu[1], Lingling Da[2], Lanping Guo[3], Luqi Huang[3], Yue Liu[4], Wenying Xu[2], Zhen Su[2], Shiping Yang[2], Qi Pan[1], Weike Jiang[1] and Tao Zhou[1]**

[1] Source Institute for Chinese and Ethnic Materia Medica, Guizhou University of Traditional Chinese Medicine, Guiyang, China, [2] College of Biological Sciences, China Agricultural University, Beijing, China, [3] National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [4] College of Horticulture, Qingdao Agricultural University, Qingdao, China

Gastrodia elata, also named Tianma, is a valuable traditional Chinese herbal medicine. It has numerous important pharmacological roles such as in sedation and lowering blood pressure and as anticonvulsant and anti-aging, and it also has effects on the immune and cardiovascular systems. The whole genome sequencing of *G. elata* has been completed in recent years, which provides a strong support for the construction of the *G. elata* gene functional analysis platform. Therefore, in our research, we collected and processed 39 transcriptome data of *G. elata* and constructed the *G. elata* gene co-expression networks, then we identified functional modules by the weighted correlation network analysis (WGCNA) package. Furthermore, gene families of *G. elata* were identified by tools including HMMER, iTAK, PfamScan, and InParanoid. Finally, we constructed a gene functional analysis platform for *G. elata*[1]. In our platform, we introduced functional analysis tools such as BLAST, gene set enrichment analysis (GSEA), and *cis*-elements (motif) enrichment analysis tool. In addition, we analyzed the co-expression relationship of genes which might participate in the biosynthesis of gastrodin and predicted 19 mannose-binding lectin antifungal proteins of *G. elata*. We also introduced the usage of the *G. elata* gene function analysis platform (GelFAP) by analyzing *CYP51G1* and *GFAP4* genes. Our platform GelFAP may help researchers to explore the gene function of *G. elata* and make novel discoveries about key genes involved in the biological processes of gastrodin.

Keywords: *Gastrodia elata*, co-expression network, functional module, gene functional analysis platform, functional enrichment analysis

OPEN ACCESS

Edited by:
Maud Fagny,
UMR 7206 Eco-Anthropologie et Ethnobiologie (EAE), France

Reviewed by:
Jinpeng Wang,
Institute of Botany, Chinese Academy of Sciences, China
Won Kyong Cho,
Seoul National University, South Korea
Etienne Delannoy,
UMR 9213 Institut des Sciences des Plantes de Paris Saclay (IPS2), France

***Correspondence:**
Tao Zhou
taozhou88@163.com
Jiaotong Yang
y_jiaotong@163.com

† These authors have contributed equally to this work

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

Received: *18 May 2020*
Accepted: *16 September 2020*
Published: *22 October 2020*

Citation:
Yang J, Xiao Q, Xu J, Da L, Guo L, Huang L, Liu Y, Xu W, Su Z, Yang S, Pan Q, Jiang W and Zhou T (2020) GelFAP: Gene Functional Analysis Platform for Gastrodia elata. Front. Plant Sci. 11:563237. doi: 10.3389/fpls.2020.563237

INTRODUCTION

Gastrodia elata, a kind of perennial herb of Orchidaceae, is one of the traditional Chinese herbal medicines. The growth cycle of *G. elata* is generally about 3 years, including the development stages of the seed, protocorm, juvenile tuber, immature tuber, mature tuber, and scape (Yuan et al., 2018). *G. elata* is a typical heterotrophic plant, which has a symbiotic relationship with at least two fungi during its life cycle. One is *Mycena* that offers nutrition for the seed germination of *G. elata*, and the other is *Armillaria mellea* that offers nutrition and energy for the vegetative propagation corms of *G. elata* development into tubers (Xu, 1981, 1989). The mannose-binding

[1] http://www.gzybioinformatics.cn/Gel

lectin antifungal proteins of *G. elata* (GAFPs) play important roles in its growth during *G. elata* and *A. mellea*, establishing a stable symbiotic association (Yuan et al., 2018). *G. elata* has important functions such as in sedation and lowering blood pressure and as anticonvulsant and anti-aging, and it also has effects on the immune and cardiovascular systems. Its pharmacological action makes it widely used in clinical settings (Shan et al., 2016). As an important medicinal plant, *G. elata* has many active chemical ingredients, such as gastrodins, 4-hydroxybenzyl alcohols, vanillyl alcohols, vanillins, polysaccharides, sterols, and organic acids (Shan et al., 2016). Among them, gastrodin is one of the important components for its beneficial effects. Gastrodin biosynthesis pathway from toluene to 4-hydroxytoluene can be catalyzed by monooxygenase of cytochrome P450 (CYP450) (Carmona et al., 2009), and then CYP450 further catalyzes the oxidation of 4-hydroxytoluene to p-hydroxybenzyl alcohol; finally, glycogenase is synthesized through glycosyltransferase (UGT) (Tsai et al., 2016). Therefore, exploring the function of genes that can catalyze the synthesis of gastrodin from the CYP450 and UGT gene family will help to explore the molecular mechanism of gastrodin biosynthesis.

The development of high-throughput sequencing technology has greatly enriched the research methods in the field of life sciences, and it not only improves the efficiency of scientific research but also promotes the development of basic research. In the past decade, whole genome sequencing had been completed in typical model plants and crops, and many species even owned their gene function analysis platforms, which were established by the integration of multiple omics data. Reiser et al. (2017) had established the Arabidopsis Information Resource (TAIR) platform, which covered detailed functional annotation information of each gene and various auxiliary analysis tools, thereby greatly improving research efficiency in scientific fields. Tian et al. (2018) had also built a gene function analysis platform MCENet, which contained a large number of *Zea mays* gene co-expression networks constructed by transcriptomic data and gene function analysis tools, so as to study gene function and synergy between different genes. Recently, Wang et al. (2020) analyzed the genomics data of 13 species in 9 genera of Malvaceae, such as genome-wide association analysis site (GWAS) information and single nucleotide mutation site (SNP) information, as well as a total of 374 sets of transcriptomic and proteomic data, and established a functional genomic hub for Malvaceae plants, which provided a powerful online analysis tool for scientists to carry out mallow family gene function analysis. Therefore, it is necessary to develop a gene function analysis platform for *G. elata* by integrating various annotations, which may contribute to deeper gene function analysis and mining.

The whole genome sequencing of *G. elata* was completed in 2018 (Yuan et al., 2018), making a certain accumulation in transcriptome data of *G. elata*. We collected the transcriptome data of 39 samples, and of these samples, 27 were from the Sequence Read Archive (SRA) in the National Center for Biotechnology Information (NCBI) and 12 were generated by our group. In order to use these data adequately and effectively, we constructed the co-expression network of *G. elata* and identified its functional modules to predict gene function. Furthermore, we constructed a *G. elata* gene function analysis platform (GelFAP) with analysis tools, such as BLAST, GSEA, and *cis*-element enrichment analysis tools, which will help to further explore the novel functions of genes in *G. elata*.

MATERIALS AND METHODS

RNA-Seq Data Processing

The quality control of *G. elata* transcriptome data was performed by FastQC software (version 0.11.2). After removing the unqualified transcriptome data samples, we used TopHat (version 2.1.0) (Trapnell et al., 2009) to map the clean reads to the reference genome and calculated the fragments per kilobase of exon model per million reads mapped (FPKM) values by the Cufflinks software (version 2.2.1) (Trapnell et al., 2010).

Co-expression Network Construction

Here, the Pearson correlation coefficient (PCC) algorithm was used to construct the gene co-expression networks of *G. elata*. We firstly calculated the correlation between different genes according to the expression values of genes in all 37 samples. Genes with high correlation had similar expression patterns in different samples, which could be considered as gene pairs with co-expression relationship. Then, we calculated the network density and the scale-free topology fitting index R^2 based on the PCC changes and selected the appropriate PCC to construct the gene co-expression network based on the maximizing scale-free topology fitting index R^2 and relative small network density. Correlation can be evaluated by PCC, and the formula is as follows:

$$\mathrm{PCC}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

PCC_{xy} is the Pearson correlation coefficient between gene x and gene y, n represents the total number of samples, x_i represents the FPKM values of gene x in the i sample, y_i represents the FPKM value of gene y in sample i, $\overline{x}$ represents the average value of gene x in n samples, and $\overline{y}$ is the average value of gene y in n samples.

Gene Set Enrichment Analysis

Gene set enrichment analysis was used as a method for annotating gene sets by calculating the degree of overlap between a specific gene set and various clearly defined gene sets and then defining an enriched gene set by the hypergeometric test, Fisher's exact test, or χ^2 test. Multiple test correction methods for GSEA, including Yekutieli, Bonferroni, Hochberg, Hochberg, Hommel, and Holm, could be used to reduce the false positive rate of GSEA analysis. These methods could perform enrichment analysis on gene ontology (GO) annotations, Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations, and Pfam domain of specific gene sets (Yi et al., 2013). The

hypergeometric test was set as a default method for users to perform gene set enrichment analysis. The formula is as follows:

$$P = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}$$

N represents the number of genes in *G. elata*, K represents the number of genes in an annotated gene set a, n represents the number of genes submitted by the user, and k represents the overlapped number of genes submitted by the user and the same genes in gene set a.

Enrichment Analysis of *Cis*-Elements (Motifs)

For the genes which needed to be analyzed, we used the following steps to calculate the Z score and P value of each motif. Firstly, we scanned the promoter region (1k, 2k, or 3k from annotated genes based on the gene structure "gff" file) of each gene that was submitted by the user and obtained the number of matches for each motif. Secondly, we selected genes to form a gene list from *G. elata* genome for 1,000 times randomly, and the number of genes was equal to the number of users who have submitted. Thirdly, we scanned the 3-kb promoter region of each gene list and calculated the average number of each motif. Finally, we calculated the Z score and P value of each motif based on the following formula. If the P value was less than 0.05, it meant that the motif was significantly enriched.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$P \text{ value} = 1 - P \text{ norm}\left(\overline{X}, \mu, \frac{\sigma}{\sqrt{n}}\right)$$

Module Identification and Annotation

We used the weighted gene correlation network analysis (WGCNA) package (Langfelder and Horvath, 2008) of R language to identify the functional modules. The process mainly included four steps. Firstly, we defined the gene co-expression correlated matrix, which weighted the correlation between genes, and determined the software threshold β based on the maximizing scale-free topology fitting index (R^2). Secondly, the blockwiseModules function was used to construct a scale-free network, and then module partition analysis was executed to identify functional modules. Thirdly, modules were defined by the dynamic tree cutting algorithm. Lastly, modules with high similarity were merged to get the final modules. Through this package, we identified the functional modules of *G. elata* co-expression network and further annotated their functions *via* gene set enrichment analysis.

Orthologous Protein Prediction and Protein–Protein Interaction Network Construction

InParanoid (Sonnhammer and Ostlund, 2015) was a software developed by Perl script for constructing orthologous groups, and its normal operation could not do without the BLAST software. We used InParanoid software (Sonnhammer and Ostlund, 2015) to predict orthologous relationship between rice/maize and *G. elata* with a cutoff over 60% bootstrap. We then mapped the protein–protein interaction (PPI) network of maize and rice to *G. elata* to construct *G. elata* PPI networks.

Gene Family Classification

We used the localized iTAK software to predict the transcription factors and transcription regulators of *G. elata* with default parameters, and the operation command was "iTAK.pl+protein_sequence." We downloaded the hidden Markov model file of the conserved domain of ubiquitin proteases from the Ubiquitin and Ubiquitin-like Conjugation Database (UUCD) (Gao et al., 2013) and used the HMMER software to predict the ubiquitin proteases of *G. elata*. The *e*-value parameter used in this calculation process was derived from the threshold recommended by the UUCD (Gao et al., 2013). In order to predict EAR motif-containing proteins and CYP450 proteins, we first collected 20,542 EAR motif-containing proteins and 19,221 CYP450 protein sequences from the PlantEAR (Yang et al., 2018) and CYP450 databases (Nelson, 2009), respectively. Then, we predicted the orthologous relationship between collected proteins and *G. elata* proteins by InParanoid (bootstrap >60%) and further defined the EAR motif-containing proteins and CYP450 proteins based on the orthologous relationship.

Search and Visualization Platform Construction

GelFAP was constructed based on CentOS Linux, Apache server, MySQL database, and PHP language. The software used for network visualization in the platform was a JavaScript package Cytoscape.js with open resources (Franz et al., 2016).

PLATFORM CONTENTS

Data Resources and Functional Annotation

Gastrodia elata genomic data, including 3,779 scaffold sequences, gene location files, gene sequences, 18,969 transcript sequences, and 18,969 protein sequences, was derived from the National Genomics Data Center (NGDC) (Accession number: GWHAAEX00000000) of China produced by the National Resource Center for Chinese Materia Medica of China Academy of Chinese Medical (Yuan et al., 2018). The gene functions of *G. elata* were annotated by comparing nucleic acids or protein sequences with various functional annotation databases, including nr, KOG, TAIR (Reiser et al., 2017), COG, Swiss-Prot, and TrEMBL (**Figure 1A**). In addition, 27 transcriptome data samples were obtained from the SRA in NCBI (Accession

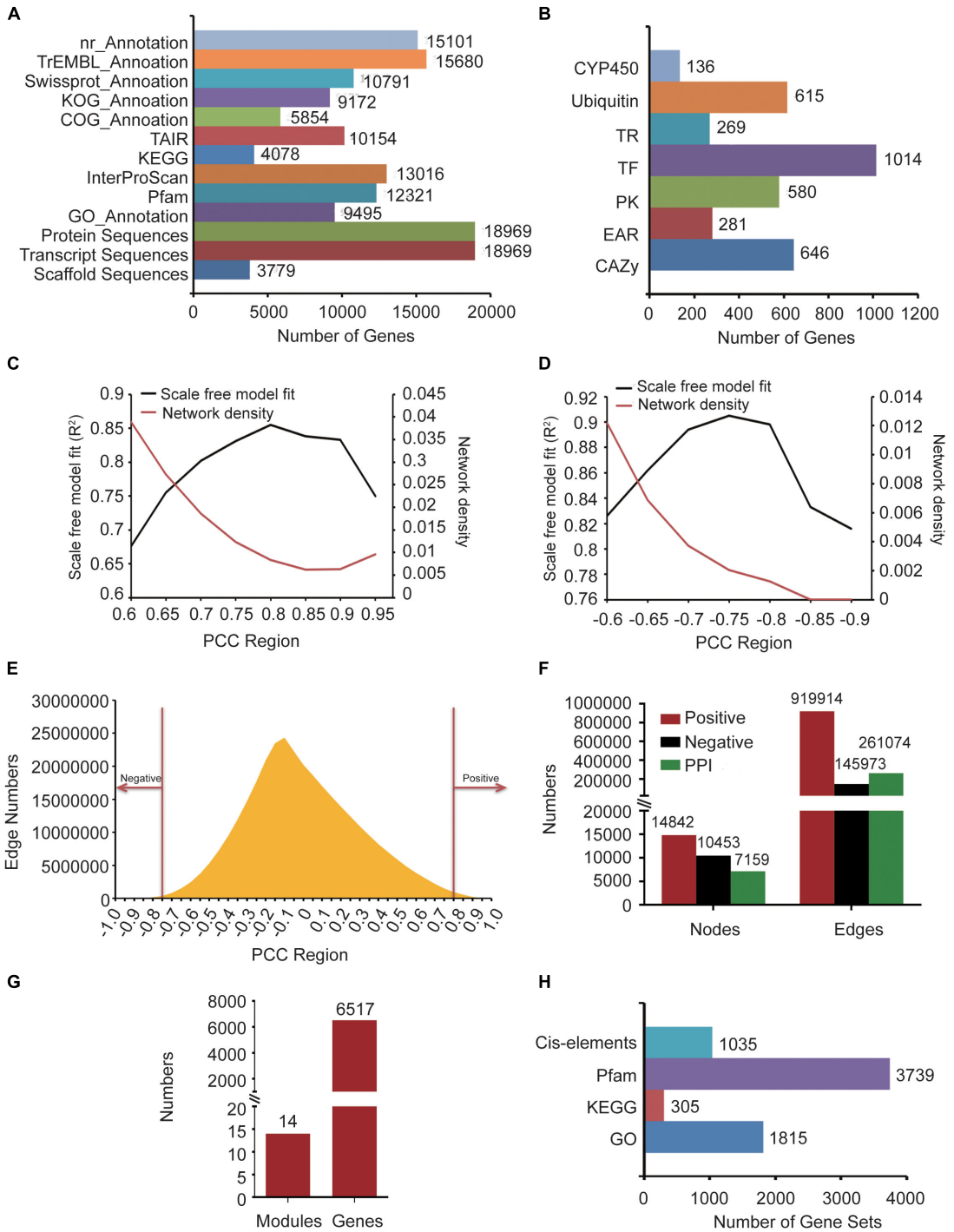

FIGURE 1 | The information about *G. elata* gene function analysis platform. **(A)** Gene function annotation information. **(B)** Gene family classification information. **(C)** Network density and scale-free model fitting (R^2) of the positive co-expression network based on changing Pearson correlation coefficient (PCC) cutoffs. **(D)** Network density and scale-free model fitting (R^2) of the negative co-expression network based on changing PCC cutoffs. **(E)** Distribution diagram of the relationship between PCC and the number of edges. **(F)** Statistics of nodes and edges in the positive co-expression network, negative co-expression network, and PPI network. **(G)** Predicted gene functional modules and involved genes. **(H)** The background gene sets of the GSEA and motif enrichment analysis tools.

number: SRP064423, SRP108465 and SRP118053) and 12 samples were produced by our group. We used the InterProScan (Jones et al., 2014) software to obtain GO terms of 9,495 genes and InterProScan domain annotations of 13,016 genes. The GO annotations were obtained from Gene Ontology Consortium (Gene Ontology Consortium, 2015). Pfam domain annotation

information of 12,321 genes was predicted by the local PfamScan tool (El-Gebali et al., 2019). KEGG orthology annotation information of 4,078 genes was predicted by GhostKOALA (Kanehisa et al., 2016), which was supported by the KEGG website. Finally, the orthologous relationship between *G. elata* and *Arabidopsis thaliana* was analyzed by the InParanoid tool, and *Arabidopsis thaliana* annotation information of 10,154 genes in *G. elata* was obtained (**Figure 1A**).

Gene Family Identification

Pfam is a protein family database, which contained multiple sequence alignment results and hidden Markov model (HMM) profiles of conserved regions from many gene families (El-Gebali et al., 2019). HMMER is a homolog searching tool based on HMM profiles (Potter et al., 2018). The gene families could be identified by combining Pfam with HMMER. Several platforms could also be used to identify gene families; for example, the analysis tools provided by the iTAK website were used for the identification of transcriptional regulators and protein kinases (Zheng et al., 2016), and HMM profiles offered by the UUCD database were used to identify members of the ubiquitin protease family (Gao et al., 2013). In addition, gene families could also be predicted by the orthologous relationship between different species.

To identify the CYP450 gene family numbers, 20,657 CYP450 protein sequences were downloaded from the CYP450 website (Nelson, 2009). Then, we constructed a library according to the downloaded CYP450 protein sequences and aligned the *G. elata* protein sequences with this library. From the results, we obtained 1,455 protein sequences whose *e*-value was less than 1e-5. Among them, 136 protein sequences with the CYP450 domain (PF00067.21) were identified as candidate members of the CYP450 family by HMMER. We used the iTAK software to identify the transcription factors, transcription regulators, and protein kinases of *G. elata* and obtained 1,014 transcription factors, 269 transcription regulators, and 580 protein kinases. We also used UUCD's HMM profile to predict the ubiquitin proteases of G. elata, and 615 ubiquitin proteases were identified. To identify the carbohydrate-active enzymes (CAZy), we downloaded the genes of A. thaliana CAZy gene family from the CAZy database (Lombard et al., 2014), matched the CAZy gene family to *G. elata* according to their orthologous relationship, and predicted 646 CAZy genes of *G. elata* (**Figure 1B**). We also collected the EAR motif-containing proteins of 71 plants from the PlantEAR platform (Yang et al., 2018) and identified 281 EAR motif-containing proteins in *G. elata* according to their orthologous relationship (**Figure 1B**).

Network Construction and Functional Module Identification

Co-expression Network

After removing the non-compliant transcriptome data samples by FastQC tools, we obtained 39 *G. elata* transcriptome data samples, including RNA-seq samples of SRP108465, SRP064423, SRP279888, and SRP118053 in SRA (**Supplementary Table S1**). The reads of RNA-seq samples were mapped to the *G. elata* genome and detailed alignment information was obtained by TopHat (**Supplementary Table S1**). In addition, the FPKM expression values of genes in each sample were obtained by computation using the Cufflinks software. Then, we calculated the PCC value between every two genes in different samples by WGCNA package of R language. Biological networks are usually scale-free networks and the network density is relatively low. Based on this principle, we analyzed PCC value over 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 and 0.95 to evaluate the scale-free model fitting index R^2 and network density of the positive co-expression network. PCC $>$ 0.8 had the largest scale-free model fitting index (R^2) and the network density was relatively low (**Figure 1C**). We also chose the PCC threshold of the negative co-expression network based on the same method (**Figure 1D**). Finally, we chose PCC $>$ 0.8 and PCC $<$ −0.75 to determine the positive co-expression network and the negative co-expression network, respectively, (**Figure 1E**). We obtained a positive co-expression network with 14,842 nodes and 919,914 edges and a negative co-expression network with 10,453 nodes and 145,973 edges (**Figure 1F**).

Protein–Protein Interaction Network

The PPI network of maize and rice had been constructed in recent years (Zhu et al., 2016; Liu et al., 2017). So, we constructed the *G. elata* PPI network by predicting the orthologous relationship between maize and *G. elata* and mapped the maize PPI network to *G. elata*. By the same method, we also mapped the rice PPI network to *G. elata*. Finally, we obtained a PPI network with 7,159 nodes and 261,074 edges (**Figure 1F**).

Functional Module Identification

The co-expression network we constructed covered 14,842 genes, so we used the WGCNA to divide these genes into modules. WGCNA is a method used to construct a gene co-expression network based on gene expression profiles. By evaluating the relationship between soft threshold and scale-free model fitting index, we chose 7 as the soft threshold (**Supplementary Figure S1A**). Similarly, the relationship between soft threshold and mean connectivity showed that a soft threshold of 7 had a lower mean connectivity (**Supplementary Figure S1B**). Finally, we merged the modules after performing the dynamic tree cutting algorithm and then further identified gene functional modules based on the similarity between modules (**Supplementary Figure S1C**). We obtained 14 functional modules with 6,517 genes (**Figure 1G**).

Functional Enrichment Analysis Tools

We annotated *G. elata* genes by gene sets of 1,815 GO annotations, 305 KEGG orthology and 3,739 Pfam (**Figure 1H**). Then, we constructed the GSEA online tool by the algorithm described in the "Materials and Methods" section.

Motifs are short and conserved sequences of the gene promoter region. It could be recognized by various transcription factors and participated in the regulation of gene expression. We also collected 1,035 motifs from the PlantEAR (Yang et al., 2018) and ccNET platforms (**Figure 1H**; You et al., 2017). Using the motif analysis algorithm in the "Materials and Methods" section,

we constructed an online motif enrichment analysis tool, which could perform motif analysis for the gene of *G. elata*.

The Structure of GelFAP

Based on the constructed gene co-expression networks, gene family classification, and functional analysis tools, the *G. elata* gene function analysis platform was constructed. The platform contained six main sections, namely Home, Browse, Gene family, Tools, KEGG, and Download and Help (**Figure 2**). Among them, there were network search and module search secondary menu functions under the network. The Tools section contained four secondary menus – Search, BLAST analysis, GSEA analysis, and cis-element analysis. The Gene family section contained CYP450, transcription factors, protein kinases, ubiquitin proteases, carbohydrate-active enzyme families, and EAR motif-containing proteins. The Pathway section contained pathways predicted by GhostKOALA (Kanehisa et al., 2016). In addition, the platform also provided the Download and Help page to assistant users to obtain data sources and help. The construction of the platform may contribute to the functional analysis of *G. elata* genes.

APPLICATION

Analysis of Putative Gastrodin Biosynthesis-Related Genes

The gastrodin biosynthesis may be regulated by CYP450, UGT, PAL, C4H, 4-HBS, and ADH family genes (Bai et al., 2016; Tsai et al., 2016). As shown in **Supplementary Figure S2**, many genes in this pathway had an obvious co-expression relationship. The PAL gene had a co-expression relationship with the C4H, CYP450, and ADH genes, and the CYP450 gene also had a co-expression relationship with UTG and ADH. Therefore, there may be an important synergistic relationship between them and they further participated in the regulation of gastrodin biosynthesis (**Supplementary Figure S2**).

A previous study had indicated that the CYP51G1 gene may be involved in the biosynthesis of gastrodin (Tsai et al., 2016), so we guessed that the function of this gene may be regulated by transcription factors that targeted on its upstream. We used the motif enrichment analysis tool to predict the transcription factors that might target on the CYP51G1 gene promoter region and found that multiple transcription factors were significantly enriched, including DRE1, MADS, and HD-zip transcription factors (**Supplementary Figure S3**). Therefore, these transcription factors may be the most probable genes that participated in the biosynthesis of gastrodin by regulating the CYP51G1 gene.

We selected the top 300 genes co-expressed with Arabidopsis CYP51G1 from the ATTED-II database (Obayashi et al., 2018) and compared them with the top 300 co-expressed genes of *G. elata* CYP51G1 (**Supplementary Figure S4**). The results demonstrated that there were 19 pairs of orthologous relationship. It had been reported that many genes of Arabidopsis had different functions (**Supplementary Figure S4**). For example, CPI1 (AT5G50375) was related to plant defense response (Cao et al., 2020), mMDH1 (AT1G53240) may be related to plant response to low temperature (Nakaminami et al., 2014), PGD1 (AT1G64190) could regulate the growth of Arabidopsis (Lim et al., 2009), TBL35 (AT5G01620) was related to xylan acetylation and growth (Yuan et al., 2016), and ARA12 (AT5G67360) was

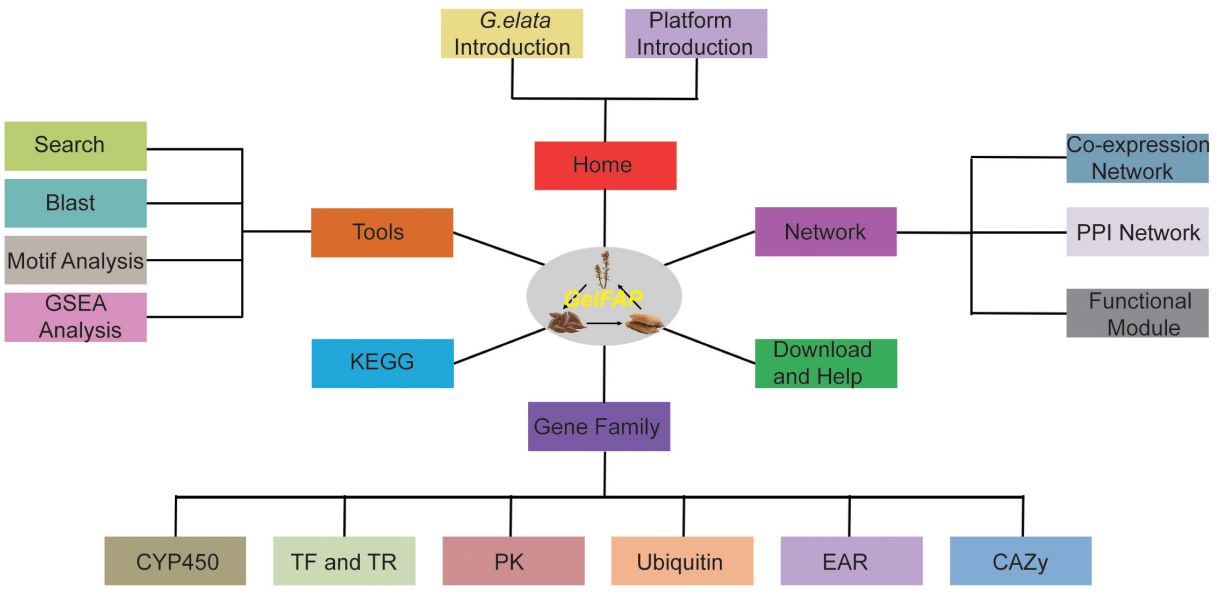

FIGURE 2 | The framework structure of *G. elata* gene function analysis platform. GelFAP included six main sections. Home section was the introduction of *G. elata* and the platform. The network section contained the co-expression network, PPI network, and function modules. The gene family section contained CYP450 family genes, transcription factors, transcription regulators, protein kinases, ubiquitin proteasomes, CAZy genes, and EAR motif-containing proteins. Tools section included Search, BLAST, motif analysis, and GSEA toolkit. KEGG, Download and Help became section separately.

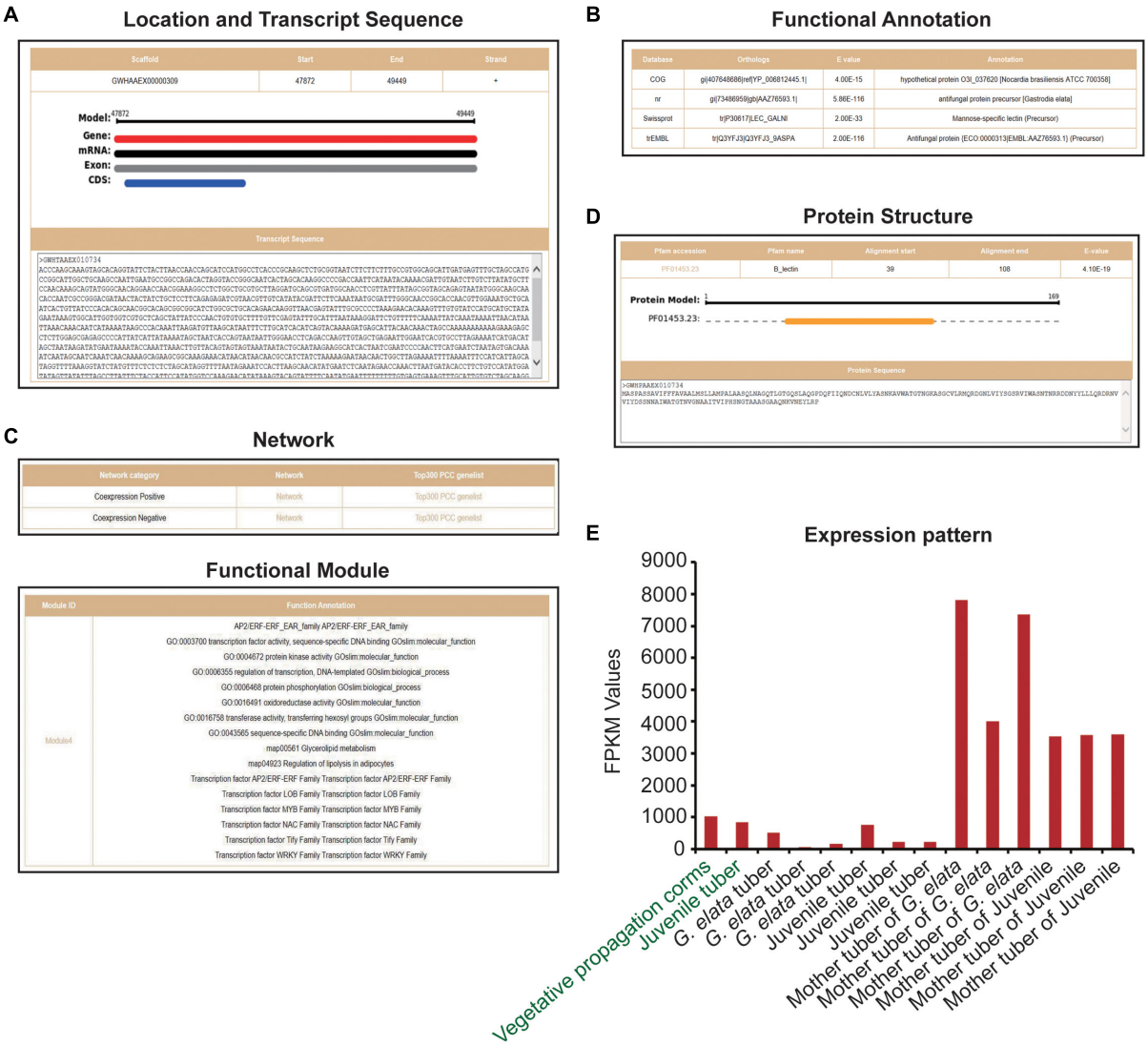

FIGURE 3 | Gene details of *GAFP4*. **(A)** Location and transcript sequences. **(B)** Functional annotations. **(C)** Network and functional modules. **(D)** Protein structure. **(E)** Expression pattern of *GAFP4* gene; green represents transcriptome samples of SRP064423, and black represents transcriptome samples produced by us.

related to release mucilage of seed coat (Rautengarten et al., 2008). Therefore, these reported genes in Arabidopsis may help to predict the function of *G. elata* CYP51G1 gene.

GAFP Identification and Functional Analysis

We obtained 12 *G. elata* mannose-binding lectin antifungal protein (GAFP) sequences from previous researches (Wang et al., 2016; Wang et al., 2019) and GenBank. By comparing these sequences with *G. elata* protein sequences, we obtained 23 protein sequences (*e*-value < 1e-3) and further identified them by the protein domain B_lection (PF01453.23). Finally, 19 *G. elata* proteins were identified as mannose-binding lectin antifungal proteins (**Supplementary Table S2**). The heterologous expression of *G. elata* GAFP4 gene (GWHGAAEX010734) in Arabidopsis thaliana could increase the resistance against *Botrytis cinerea*, and the heterologous expression of GAFP4 in cotton could also increase the resistance against Verticillium wilt (Wang et al., 2016; Wang et al., 2019). Here, we took *G. elata* antifungal protein GAFP4 as an example to analyze its functions by GelFAP. We searched the gene details and obtained the structure information and transcript sequences (**Figure 3A**), annotation information (**Figure 3B**), networks and functional modules (**Figure 3C**), protein structure and sequences (**Figure 3D**), and expression values (**Figure 3E**). We found that this gene had only one exon and CDS, and gene length was 1,557bp (**Figure 3A**). In addition, the functional annotation information indicated that this gene was annotated as an antifungal protein in the nr and TrEMBL databases (**Figure 3B**). Protein structure and sequence information suggested that this protein had a B_lectin domain. Related researches showed that the protein with B_lectin domains

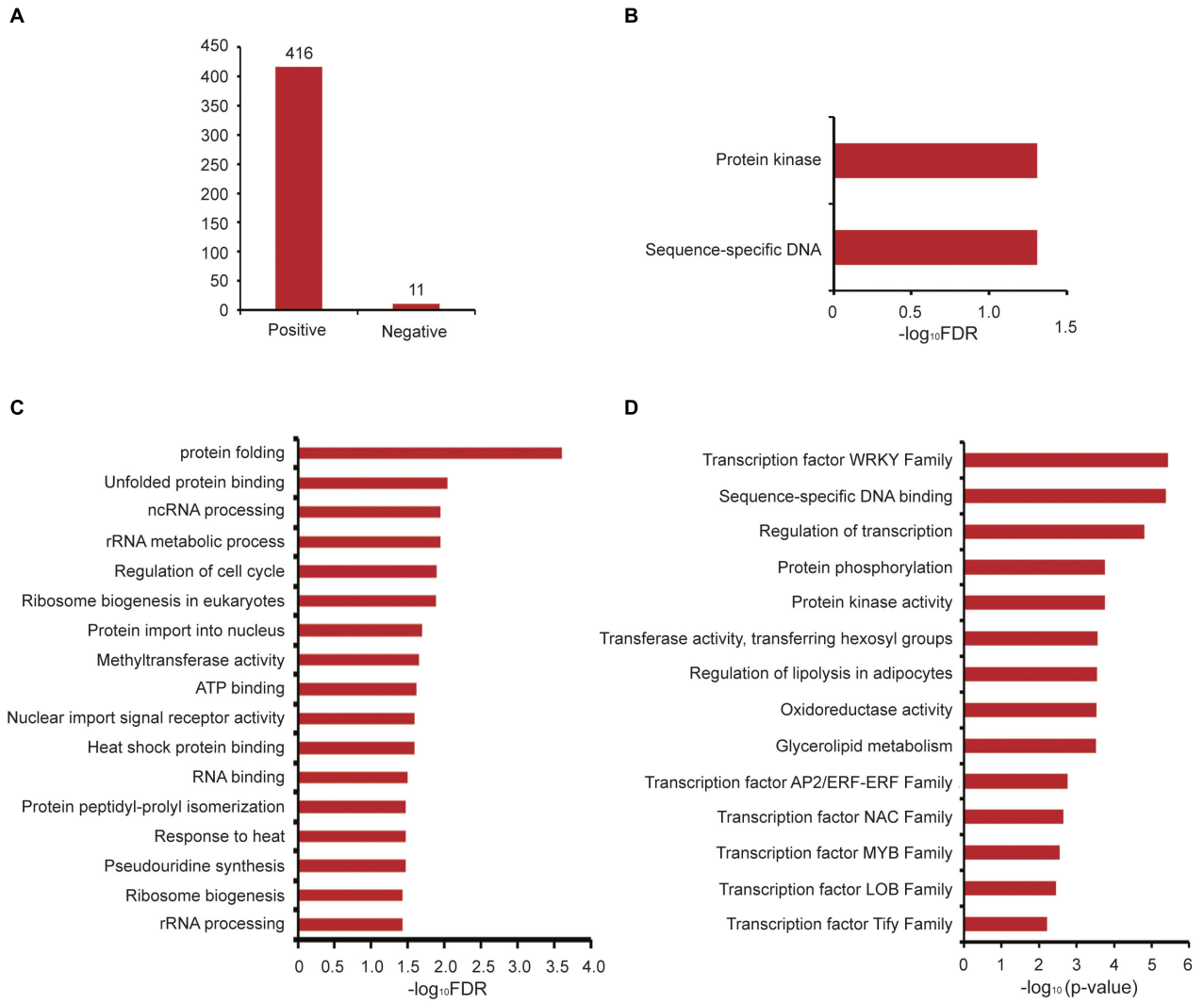

FIGURE 4 | Co-expression and functional module analysis of *GAFP4*. **(A)** The number of positive and negative co-expressed genes of *GAFP4* gene. **(B)** GSEA of *GAFP4* positive co-expressed genes. **(C)** GSEA of *GAFP4* negative co-expressed genes. **(D)** GSEA annotation for the module that contained *GAFP4*.

had antibacterial and antiviral functions (Cox et al., 2006; Sun et al., 2016; Wang et al., 2016; Wang et al., 2019; Yin et al., 2019; **Figure 3D**). Therefore, this domain may be an important structure for GAFP4 to perform its function.

Next, we analyzed GAFP4 gene function by the co-expression network, and the search results showed that GAFP4 had a positive co-expression relationship with 416 genes and a negative co-expression relationship with 11 genes (**Figure 4A** and **Supplementary Table S3**). GSEA of GAFP4 positive co-expressed genes revealed that this gene might have functions of protein kinase activity and sequence-specific DNA binding (Fisher's exact test, FDR $<$ 0.05) (**Figure 4B**). Therefore, GAFP4 may be co-expressed with several transcription factors (TFs) to perform its DNA-binding function. GSEA of GAFP4 negative co-expressed genes revealed its possible function in heat response, protein folding, methyltransferase, regulation of cell cycle, and so on (Fisher's exact test, FDR $<$ 0.05) (**Figure 4C**). In addition, we obtained a function module that contained GAFP4 (**Supplementary Table S4**). GSEA of this module showed significant enriched transcription factor family members, including ERF, MYB, and WRKY families. Moreover, protein kinase activity, transferase activity, oxidoreductase activity, and glycerolipid metabolism were also enriched in the module (Fisher's exact test, P value $<$ 0.05) (**Figure 4D**). When plants were infected by bacteria or viruses, the plant transcription factor families ERF (Wang et al., 2018; Zhu et al., 2019), MYB (Ibraheem et al., 2015; Shan et al., 2016), WRKY (Chen et al., 2013; Peng et al., 2016; Wang et al., 2017; Gao et al., 2018; Liu et al., 2018; Li et al., 2020), and protein kinase (Kim and Hwang, 2011; Shen et al., 2012) showed response functions. Therefore, GAFP4 may be co-expressed with many antibacterial TFs or form functional modules with TFs to further play its role in antibacterial defense response. Therefore, by analysis of GAFP4 gene in GelFAP, we found that it might have antibacterial effect functions. At

present, its antibacterial function has been verified in cotton and Arabidopsis (Wang et al., 2016; Wang et al., 2019), and many other functions still need to be explored in the future.

DISCUSSION

Gastrodia elata is a valuable traditional Chinese herbal medicine and has numerous important pharmacological roles. The whole genome sequencing of *G. elata* has been completed in recent years and its transcriptome data also has a certain accumulation (Tsai et al., 2016; Yuan et al., 2018). In this study, we firstly used the genome and transcriptomes of *G. elata* to construct *G. elata* gene co-expression networks and functional modules and provided related gene function analysis and annotation tools, including the BLAST search tool, GSEA tool, and motif enrichment analysis tool. The gene co-expression networks were of great significance for exploring gene functions, such as comparing networks between orthologous gene pairs in model specie and *G. elata*, which could provide more information for gene function researches. Similarly, gene function enrichment analysis tools also played important roles in *G. elata* gene functional researches. For example, gene enrichment analysis tools could analyze possible downstream functions of differentially expressed genes in the transcriptome. Finally, the gene families such as CYP450, transcription factors, protein kinases, ubiquitin proteases, and carbohydrate-active enzymes were classified and predicted, and the results were integrated into the *G. elata* gene functional analysis platform. Therefore, our platform can provide more data sources and analysis methods for researchers to study the gene function of *G. elata*, which may improve the efficiency of the research for *G. elata* genes.

Gastrodia elata established a symbiotic relationship with Armillaria during the growth process, and it was reported that GAFPs played important roles in establishing this relationship, but which kind of GAFPs were not mentioned. We identified 19 *G. elata* GAFPs based on the sequence information provided by the platform, and provided candidate genes for the follow-up research on the establishment of symbiotic relationship. Furthermore, we took the *GAFP4* gene as an example to introduce the application method of the platform. The analysis results indicated that *GAFP4* might be involved in various regulatory processes including antibacterial, and it had also been reported to have the function of antibacterial (Wang et al., 2016; Wang et al., 2019). Therefore, the platform we built has a certain feasibility and practicality.

The *G. elata* gene function analysis platform is established by us for the first time. Users can submit their interesting genes to the platform and then obtain information of various existing and processed annotations. However, there is still much room for improvement in the accumulation of omics data. In the future, we will continue to update and maintain the *G. elata* gene function analysis platform, such as collecting and integrating more transcriptome, proteome, metabolome data, etc. We expect that this platform will contribute to the study of molecular mechanisms in the process of gastrodin biosynthesis, and further help to solve the problems about variety and quality improvement of *G. elata*.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP064423, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP108465, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP118053, and https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP279888.

AUTHOR CONTRIBUTIONS

JY designed this study. JY and QX constructed the platform and completed the draft. LD, LG, and JX produced and processed the transcriptome. ZS, WX, and YL participated in the construction of this platform. SY and QP participated in the revision of the manuscript. TZ, WJ, and LH directed this work and provided financial support.

FUNDING

This work was supported by the ability establishment of sustainable use for valuable Chinese Medicine Resources (Grant No. 2060302), the High-level Innovative Talents of Guizhou Province of China (Qian Ke He Platform and Talent [2018]5638), Guizhou Education Department Innovation Group Major Research Projects (Qian Jiao He KY Zi [2018]022), Ph.D. Startup Foundation of Guizhou University of Traditional Chinese Medicine [2019]141 and [2020]32, the Science and Technology Project in Guizhou Province of China (Qian Ke He Platform and Talent [2019]5611), and National Natural Science Foundation of China [81960694].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.563237/full#supplementary-material

REFERENCES

Bai, Y., Yin, H., Bi, H., Zhuang, Y., Liu, T., and Ma, Y. (2016). De novo biosynthesis of Gastrodin in *Escherichia coli*. *Metab. Eng.* 35, 138–147. doi: 10.1016/j.ymben.2016.01.002

Cao, Y., He, Q., Qi, Z., Zhang, Y., Lu, L., Xue, J., et al. (2020). Dynamics and endocytosis of Flot1 in *Arabidopsis* require CPI1 function. *Int. J. Mol. Sci.* 21:1552. doi: 10.3390/ijms21051552

Carmona, M., Zamarro, M. T., Blazquez, B., Durante-Rodriguez, G., Juarez, J. F., Valderrama, J. A., et al. (2009). Anaerobic catabolism of aromatic

compounds: a genetic and genomic view. *Microbiol. Mol. Biol. Rev.* 73, 71–133. doi: 10.1128/MMBR.00021-08

Chen, L., Zhang, L., Li, D., Wang, F., and Yu, D. (2013). WRKY8 transcription factor functions in the TMV-cg defense response by mediating both abscisic acid and ethylene signaling in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1963–E1971. doi: 10.1073/pnas.1221347110

Cox, K. D., Layne, D. R., Scorza, R., and Schnabel, G. (2006). Gastrodia antifungal protein from the orchid Gastrodia elata confers disease resistance to root pathogens in transgenic tobacco. *Planta* 224, 1373–1383. doi: 10.1007/s00425-006-0322-0

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32, 309–311. doi: 10.1093/bioinformatics/btv557

Gao, J., Bi, W., Li, H., Wu, J., Yu, X., Liu, D., et al. (2018). WRKY transcription factors associated with NPR1-mediated acquired resistance in barley are potential resources to improve wheat resistance to *Puccinia triticina*. *Front. Plant Sci.* 9:1486. doi: 10.3389/fpls.2018.01486

Gao, T. S., Liu, Z. X., Wang, Y. B., Cheng, H., Yang, Q., Guo, A. Y., et al. (2013). UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. *Nucleic Acids Res.* 41, D445–D451. doi: 10.1093/nar/gks1103

Gene Ontology Consortium (2015). Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi: 10.1093/nar/gku1179

Ibraheem, F., Gaffoor, I., Tan, Q., Shyu, C. R., and Chopra, S. (2015). A sorghum MYB transcription factor induces 3-deoxyanthocyanidins and enhances resistance against leaf blights in maize. *Molecules* 20, 2388–2404. doi: 10.3390/molecules20022388

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W. Z., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Kim, D. S., and Hwang, B. K. (2011). The pepper receptor-like cytoplasmic protein kinase CaPIK1 is involved in plant signaling of defense and cell-death responses. *Plant J.* 66, 642–655. doi: 10.1111/j.1365-313X.2011.04525.x

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, H., Wu, J., Shang, X., Geng, M., Gao, J., Zhao, S., et al. (2020). WRKY transcription factors shared by BTH-induced resistance and NPR1-mediated acquired resistance improve broad-spectrum disease resistance in wheat. *Mol. Plant Microbe Interact.* 33, 433–443. doi: 10.1094/MPMI-09-19-0257-R

Lim, H., Cho, M. H., Jeon, J. S., Bhoo, S. H., Kwon, Y. K., and Hahn, T. R. (2009). Altered expression of pyrophosphate: fructose-6-phosphate 1-phosphotransferase affects the growth of transgenic *Arabidopsis* plants. *Mol. Cells* 27, 641–649. doi: 10.1007/s10059-009-0085-0

Liu, Q., Li, X., Yan, S., Yu, T., Yang, J., Dong, J., et al. (2018). OsWRKY67 positively regulates blast and bacteria blight resistance by direct activation of PR genes in rice. *BMC Plant Biol.* 18:257. doi: 10.1186/s12870-018-1479-y

Liu, S., Liu, Y., Zhao, J., Cai, S., Qian, H., Zuo, K., et al. (2017). A computational interactome for prioritizing genes associated with complex agronomic traits in rice (Oryza sativa). *Plant J.* 90, 177–188. doi: 10.1111/tpj.13475

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Nakaminami, K., Matsui, A., Nakagami, H., Minami, A., Nomura, Y., Tanaka, M., et al. (2014). Analysis of differential expression patterns of mRNA and protein during cold-acclimation and de-acclimation in *Arabidopsis*. *Mol. Cell Proteomics* 13, 3602–3611. doi: 10.1074/mcp.M114.039081

Nelson, D. R. (2009). The cytochrome p450 homepage. *Hum. Genomics* 4, 59–65. doi: 10.1186/1479-7364-4-1-59

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: a plant Coexpression database Based on Investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* 59:440. doi: 10.1093/pcp/pcx209

Peng, X., Wang, H., Jang, J. C., Xiao, T., He, H., Jiang, D., et al. (2016). OsWRKY80-OsWRKY4 module as a positive regulatory circuit in rice resistance against *Rhizoctonia solani*. *Rice (NY)* 9:63. doi: 10.1186/s12284-016-0137-y

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

Rautengarten, C., Usadel, B., Neumetzler, L., Hartmann, J., Bussis, D., and Altmann, T. (2008). A subtilisin-like serine protease essential for mucilage release from *Arabidopsis* seed coats. *Plant J.* 54, 466–480. doi: 10.1111/j.1365-313X.2008.03437.x

Reiser, L., Subramaniam, S., Li, D., and Huala, E. (2017). Using the *Arabidopsis* information resource (TAIR) to find information about *Arabidopsis* genes. *Curr. Protoc. Bioinformatics* 60, 1.11.1–1.11.45. doi: 10.1002/cpbi.36

Shan, T., Rong, W., Xu, H., Du, L., Liu, X., and Zhang, Z. (2016). The wheat R2R3-MYB transcription factor TaRIM1 participates in resistance response against the pathogen Rhizoctonia cerealis infection through regulating defense genes. *Sci. Rep.* 6:28777. doi: 10.1038/srep28777

Shen, Q., Bao, M., and Zhou, X. (2012). A plant kinase plays roles in defense response against geminivirus by phosphorylation of a viral pathogenesis protein. *Plant Signal. Behav.* 7, 888–892. doi: 10.4161/psb.20646

Sonnhammer, E. L., and Ostlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* 43, D234–D239. doi: 10.1093/nar/gku1203

Sun, Y. Y., Liu, L., Li, J., and Sun, L. (2016). Three novel B-type mannose-specific lectins of *Cynoglossus semilaevis* possess varied antibacterial activities against Gram-negative and Gram-positive bacteria. *Dev. Comp. Immunol.* 55, 194–202. doi: 10.1016/j.dci.2015.10.003

Tian, T., You, Q., Yan, H., Xu, W., and Su, Z. (2018). MCENet: a database for maize conditional co-expression network and network characterization collaborated with multi-dimensional omics levels. *J. Genet. Genomics* 45, 351–360. doi: 10.1016/j.jgg.2018.05.007

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tsai, C. C., Wu, K. M., Chiang, T. Y., Huang, C. Y., Chou, C. H., Li, S. J., et al. (2016). Comparative transcriptome analysis of *Gastrodia elata* (Orchidaceae) in response to fungus symbiosis to identify gastrodin biosynthesis-related genes. *BMC Genomics* 17:212. doi: 10.1186/s12864-016-2508-6

Wang, D., Fan, W., Guo, X., Wu, K., Zhou, S., Chen, Z., et al. (2020). MaGenDB: a functional genomics hub for Malvaceae plants. *Nucleic Acids Res.* 48, D1076–D1084. doi: 10.1093/nar/gkz953

Wang, J., Tao, F., Tian, W., Guo, Z., Chen, X., Xu, X., et al. (2017). The wheat WRKY transcription factors TaWRKY49 and TaWRKY62 confer differential high-temperature seedling-plant resistance to *Puccinia striiformis* f. sp. tritici. *PLoS One* 12:e0181963. doi: 10.1371/journal.pone.0181963

Wang, M., Zhu, Y., Han, R., Yin, W., Guo, C., Li, Z., et al. (2018). Expression of *Vitis amurensis* VaERF20 in *Arabidopsis thaliana* improves resistance to *Botrytis cinerea* and *Pseudomonas* syringae pv. Tomato DC3000. *Int. J. Mol. Sci.* 19:696. doi: 10.3390/ijms19030696

Wang, Y., Liang, C., Wu, S., Jian, G., Zhang, X., Zhang, H., et al. (2019). Vascular-specific expression of Gastrodia antifungal protein gene significantly enhanced cotton *Verticillium wilt* resistance. *Plant Biotechnol. J.* 18, 1498–1500. doi: 10.1111/pbi.13308

Wang, Y., Liang, C., Wu, S., Zhang, X., Tang, J., Jian, G., et al. (2016). Significant improvement of cotton *Verticillium wilt* resistance by manipulating the expression of gastrodia antifungal proteins. *Mol. Plant* 9, 1436–1439. doi: 10.1016/j.molp.2016.06.013

Xu, J. T. (1981). [A brief report on the nutrition sources of seed germination of *Gastrodia elata* (author's transl)]. *Zhong Yao Tong Bao* 6:2.

Xu, J. T. (1989). [Studies on the life cycle of Gastrodia elata]. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* 11, 237–241.

Yang, J., Liu, Y., Yan, H., Tian, T., You, Q., Zhang, L., et al. (2018). PlantEAR: functional analysis platform for plant EAR motif-containing proteins. *Front. Genet.* 9:590. doi: 10.3389/fgene.2018.00590

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–W103. doi: 10.1093/nar/gkt281

Yin, X., Mu, L., Li, Y., Wu, L., Yang, Y., Bian, X., et al. (2019). Identification and characterization of a B-type mannose-binding lectin from *Nile tilapia* (*Oreochromis nilot*icus) in response to bacterial infection. *Fish Shellfish Immunol.* 84, 91–99. doi: 10.1016/j.fsi.2018.09.072

You, Q., Xu, W., Zhang, K., Zhang, L., Yi, X., Yao, D., et al. (2017). ccNET: database of co-expression networks with functional modules for diploid and polyploid Gossypium. *Nucleic Acids Res.* 45, 5625–5626. doi: 10.1093/nar/gkw1342

Yuan, Y., Jin, X., Liu, J., Zhao, X., Zhou, J., Wang, X., et al. (2018). The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* 9:1615. doi: 10.1038/s41467-018-03423-5

Yuan, Y., Teng, Q., Zhong, R., and Ye, Z. H. (2016). Roles of *Arabidopsis* TBL34 and TBL35 in xylan acetylation and plant growth. *Plant Sci.* 243, 120–130. doi: 10.1016/j.plantsci.2015.12.007

Zheng, Y., Jiao, C., Sun, H. H., Rosli, H. G., Pombo, M. A., Zhang, P. F., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhu, G., Wu, A., Xu, X. J., Xiao, P. P., Lu, L., Liu, J., et al. (2016). PPIM: a protein-protein interaction database for maize. *Plant Physiol.* 170, 618–626. doi: 10.1104/pp.15.01821

Zhu, Y., Li, Y., Zhang, S., Zhang, X., Yao, J., Luo, Q., et al. (2019). Genome-wide identification and expression analysis reveal the potential function of ethylene responsive factor gene family in response to *Botrytis cinerea* infection and ovule development in grapes (*Vitis vinifera* L.). *Plant Biol. (Stuttg)* 21, 571–584. doi: 10.1111/plb.12943

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 November 2020
doi: 10.3389/fgene.2020.578679

Gene Expression and Co-expression Networks Are Strongly Altered Through Stages in Clear Cell Renal Carcinoma

Jose María Zamora-Fuentes[1], *Enrique Hernández-Lemus*[1,2] *and Jesús Espinal-Enríquez*[1,2]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de Mexico, Mexico City, Mexico

Clear cell renal carcinoma (ccRC) is a highly heterogeneous and progressively malignant disease. Analyzing ccRC progression in terms of modifications at the molecular and genetic level may help us to develop a broader understanding of its patho-physiology and may give us a glimpse toward improved therapeutics. In this work, by using TCGA data, we studied the molecular progression of the four main ccRC stages (i, ii, iii, iv) in two different yet complementary approaches: (a) gene expression and (b) gene co-expression. For (a) we analyzed the differential gene expression between each stage and the control non-cancer group. We compared the progression molecular signature between stages, and observed those genes that change their expression patterns through progression stages. For (b) we constructed and analyzed co-expression networks for the four ccRC progression stages, as well as for the control phenotype, to observe whether and how the co-expression landscape changes with progression. We separated genomic interactions into intra-chromosome (*cis*-) and inter-chromosome (*trans*-). Finally, we intersected those networks and performed functional enrichment analysis. All calculations were made over different network sizes, from the top 100 edges to top 1,000,000. We show that differential expression is quite similar between ccRC progression stages. However, interestingly, two genes, namely SLC6A19 and PLG show a significant progressive decrease in their expression according to ccRC stage, meanwhile two other genes, SAA2-SAA4 and CXCL13 show progressive increase. Despite the high similarity between gene expression profiles, all networks are substantially different between them in terms of their topological features. Control network has a larger proportion of *trans*- interactions, meanwhile for any stage, the amount of *cis*-interactions is higher, independent of the network cut-off. The majority of interactions in any network are phenotype-specific. Only 189 interactions are shared between the five networks, and 533 edges are ccRC-specific, independent of the stage. The small resulting connected components in both cases are formed by genes with the same differential expression trend, and are associated with important biological processes, such as cell cycle or immune system, suggesting that activity of these categories follows the differential expression trend. With this approach we have shown that, even if the

OPEN ACCESS

Edited by:
Kimberly Glass, Brigham and Women's Hospital and Harvard Medical School, United States

Reviewed by:
Jie Zhang, Indiana University, United States
Ali Salehzadeh-Yazdi, University of Rostock, Germany

***Correspondence:**
Jesús Espinal-Enríquez
jespinal@inmegen.gob.mx

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

Received: *30 June 2020*
Accepted: *18 September 2020*
Published: *03 November 2020*

Citation:
Zamora-Fuentes JM, 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

expression program is similar during ccRC progression, the co-expression programs strongly differ. More research is needed to understand the delicate interplay between expression and co-expression, but this is a first approach to enclose both approaches in an integrative view aimed at a deeper understanding in gene regulation in tumor evolution.

Keywords: clear cell renal carcinoma, gene co-expression networks, SLC6A19 progressive underexpression, PLG progressive underexpression, cancer progression stages, SAA2-SAA4 progressive overexpression, CXCL13 progressive overexpression, loss of long-range co-expression

1. INTRODUCTION

The term *renal cell cancer* refers to a heterogeneous group of cancers derived from renal tubular cells. In the last years, pathology-based and basic cancer research programmes have characterized different renal tumor entities (Moch, 2013). Renal cell carcinoma is a group of malignancies arising from the epithelium of the renal tubules (Moch, 2013). Renal cancer may be seen as several histologically defined cancers. Those present different genetic drivers, epigenetic marks, clinical courses, and also therapeutic responses (Ricketts et al., 2018).

Histologically, renal cancer has been divided into three major subtypes, clear cells, papillary renal cell carcinoma, and chromophobe renal cell carcinoma (Moch et al., 2016). Clear cell renal cell carcinoma (ccRC) is the most common subtype (≈75%); papillary renal cell carcinoma (PRCC) accounts for 15–20% and is subdivided into types 1 and 2; and chromophobe renal cell carcinoma (ChRCC) represents ≈5% of renal cell carcinomas (Jaffe et al., 2001; Moch et al., 2016).

Molecular and genomics characterization of these tumors have been conducted elsewhere. For instance, the Cancer Genome Atlas Consortium (TCGA, The Cancer Genome Atlas Research Network, 2013, 2016) has provided the most common deregulated processes in kidney cancer in general (The Cancer Genome Atlas Research Network, 2013), as well as in ccRC in particular (The Cancer Genome Atlas Research Network, 2016). Events such as Krebbs cycle downregulation, upregulation of pentose phosphate pathway genes or important genomics rearrangements in TERT region have been observed as recurrent deregulated processes.

Inside the ccRC subtype, particular subgroups have been identified. Such subgroups have been related to epigenetic modifications, somatic mutations, or genomic rearrangements within the TERT promoter region (Ricketts et al., 2018). Proteins associated with Warburg effect, as well as molecular predictors of late stage (Neely et al., 2016), have also been associated to ccRC. Several references regarding mutations of von Hippel-Landau (VHL) tumor suppressor gene have also been reported (Kaelin, 2004; Cowey and Rathmell, 2009; Arjumand and Sultana, 2012).

Regarding epigenetic modifications, comprehensive revisions have reported an increasing number of them (see Jung et al., 2009; Redova et al., 2011; Li et al., 2015). For instance, for ccRC, miR-99a, miR-106a, miR-125b, miR-144, miR-203, miR-378, or mir-28-5p have shown a dual behavior, oncogenic and oncosuppressive (Wang et al., 2016; Braga et al., 2019). Genes such as the aforementioned VHL, or RASSF1A, CDH1, and APAF1 have been found to be susceptible to hypermethylation (Dmitriev et al., 2014; Braga et al., 2015).

Despite all those advances in characterizing molecular features of renal cancer, histo-pathological aspects still contain crucial information for accurate clinical interventions. In those terms, progression stages (according to the Gold standard reference in cancer staging, Edge et al., 2010) provide us important elements to have, in combination with molecular characteristics, a broader and more integrative point of view regarding renal cancer. Hence, understanding progression in terms of molecular and genetic factors could help us to understand the disease with higher accuracy.

In this work, we used information from molecular and histo-pathological factors to unveil specific characteristics that change during progression stages. To this end, we focused on the molecular progression of clear cell Renal carcinoma (ccRC) by two different yet complementary approaches: (a) gene expression and (b) gene co-expression. For (a) we analyzed the differential expression of all genes at the four progression stages vs. the control non-cancer group, and between stages, to observe the gene expression pattern for each progression stage. We compared the progression signature between stages, and observed whether or not a set of genes change their expression patterns through progression stages.

For (b), we constructed and analyzed co-expression networks for the four ccRC progression stages, as well as for the control phenotype and compared between them, in order to have a quantitative indicator to distinguish and observe whether or not the co-expression landscape changes progressively.

In previous works from our group, we observed abrupt changes in the way that genes co-express: for instance, we have documented a substantial decrease of inter-chromosome (*trans*-) gene-gene interactions in breast cancer (Espinal-Enriquez et al., 2017; Dorantes-Gilardi et al., 2020; García-Cortés et al., 2020). We decided to separate gene-gene interactions into intra-chromosome (*cis*-) and inter-chromosome (*trans*-). We performed functional enrichment analyses for each whole-network, and also by communities inside networks, by assuming that network structure may guard functional features of an oncogenic phenotype (Alcalá-Corona et al., 2016, 2017, 2018; Hernández-Lemus et al., 2019).

We wanted to quantify similarities and differences between consecutive progression stages, since with this information one may isolate those features that are conserved or change between one stage to the following. To do so, we obtained the network

Abbreviations: ccRC, Clear Cell Renal Carcinoma; Log_2FC, Log_2 Fold Change; MI, Mutual Information; TCGA, The Cancer Genome Atlas.

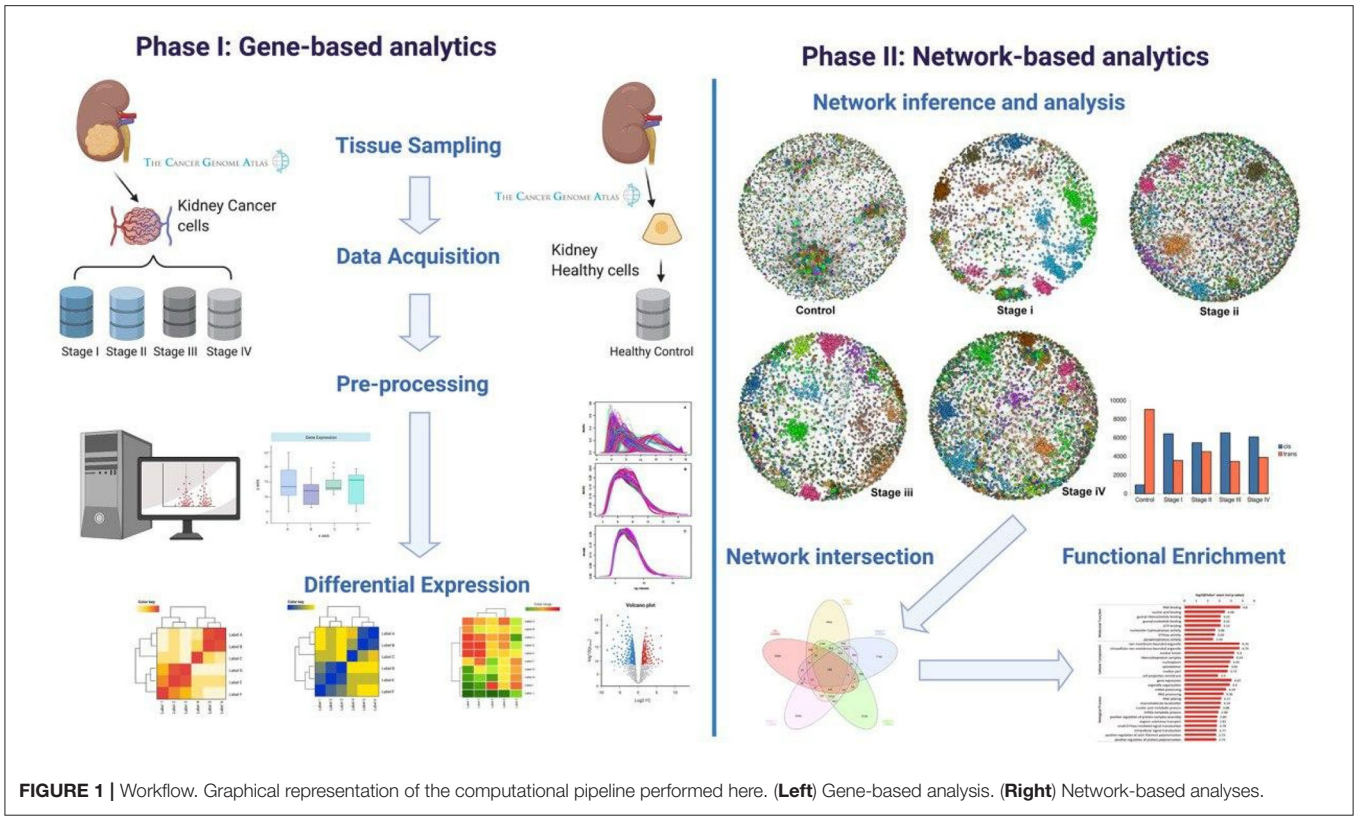

FIGURE 1 | Workflow. Graphical representation of the computational pipeline performed here. (**Left**) Gene-based analysis. (**Right**) Network-based analyses.

intersections and differences between consecutive progression phenotypes, starting with Stage I vs. Control network, Stage II vs. Stage I, etc. In a complementary task, we intersected the five networks (the four stages and control) to observe which genes and interactions are conserved throughout all phenotypes. Additionally, we intersected the four progression stage networks to observe those interactions that appear in cancer but are not present in a healthy phenotype. The resulting networks were then analyzed via over-representation analysis. We observed those processes involved in the resulting networks and also the respective differential expression patterns.

2. MATERIALS AND METHODS

A graphical representation of our methodology can be found in **Figure 1**. Our workflow can be broadly divided into two main branches: gene-based and network-based analyses. These in turn, can be divided into four main steps: (1) Data acquisition, (2) Pre-processing, (3) High-level processing, and (4) Functional enrichment.

2.1. Data Acquisition

We obtained the complete dataset from GDC clear cell renal carcinoma repository (https://portal.gdc.cancer.gov/repository). For this purpose, we developed a set of scripts that uses as input the TCGA project transcriptomic data and metadata (in this case, ccRC). The scripts collect all transcriptome profiling samples, as well as clinical data available for the same samples. The RNA-seq transcriptomic profiles were pruned, keeping those genes with valid numeric values and its associated ENSEMBL ID.

TABLE 1 | RNA-Seq data from ccRC patients per progression stage.

Tissue	Control	Stage I	Stage II	Stage III	Stage IV
ccRC	72	272	59	123	82

Tumor samples were separated into stages according to the *tumor_stage* variable, provided by TCGA for each clinical file. In the case that *tumor_stage* value was *not reported*, we decided to discard that sample.

We used RNA-Seq level 3 gene expression files from The Cancer Genome Atlas from 608 ccRC samples. We divided these patients by cancer progression stage, as well as control non-tumor tissue. Number of cases for each stage is shown in **Table 1**.

2.2. Data Pre Processing

We carried out a data pre-processing pipeline in three phases. (1) pre-normalization quality control, (2) batch and bias corrections (normalization) and (3) post-normalization quality control. Data pre-processing was conducted as previously (Drago-García et al., 2017; Espinal-Enriquez et al., 2017; de Anda-Jáuregui et al., 2019b,c; García-Cortés et al., 2020; Serrano-Carbajal et al., 2020). Briefly, we assessed (a) biotype abundances, to assure that samples contained protein coding genes. (b) gene counts expression boxplots were also evaluated per biotype to

confirm that the highest median expression corresponded to protein coding genes. (c) Finally, we evaluated the number of detected genes per sample, by using saturation plots. These steps were performed with standard R package *NOISeq* (Tarazona et al., 2011). Normalization method for correct Length bias (full) and GC content (full) was Within-lane. Additionally we applied a "TMM" normalization to eliminate RNA composition biases between libraries and prepare data to find Differentially Expressed Genes. Risso et al. (2011). PCAs and plots are shown in **Supplementary Material 1**. Genes were filtered by mean expression values ($mean > 10$). Normalization to correct batch effect was performed by using ARSyN (Nueda et al., 2012) implemented in NOISeq package. Scripts to perform pre-processing analysis can also be found at https://github.com/josemaz/kidney-stages.

2.3. Differential Expression

Differential gene expression analysis was performed to compare gene expression between each ccRC stage vs. control. This analysis was performed via empirical Bayes moderation of the standard errors using *edgeR* package (Robinson et al., 2010). To consider a gene as differentially expressed, we considered a Log_2Fold Change ($|LFC| > 2.0$) cut-off.

2.3.1. Statistical Significance and Multiple Hypothesis Testing

To account for multiple comparisons of gene profiles, we implemented Benjamini & Hochberg False Discovery Rate correction calculations. The FDR-adjusted *p*-value cut-off was set to be 0.05 for each comparison.

We also performed a multi-group comparison based on Likelihood ratio test (LRT) method to obtain all group contrasts (Love et al., 2014). With this method, implemented in the DEseq2 R package, we used the deviation of each group in the calculation of *p*-values for every contrast. We filtered genes with a corrected *p*-value less than 0.05 and log-fold change $-0.5 > |LFC| > 0.5$ for each contrast. This last, searching for differentially expressed genes, not only between genes of cancer stages and control samples, but also between stages.

Since ccRC data is separated into stages, we observed those genes that change in agreement with the stages, i.e., differential expression increases or decreases progressively with stages. To determine the significance of those differences, we performed a Wilcoxon signed rank test between individual gene expression at different stages.

2.4. Network Analysis

We used the mutual information (MI) statistical dependence measure to quantify co-expression between genes. We used the MI implementation on the ARACNe algorithm (Margolin et al., 2006), as previously described (Alcalá-Corona et al., 2017, 2018; Espinal-Enriquez et al., 2017; de Anda-Jáuregui et al., 2019c; García-Cortés et al., 2020), to determine all gene-gene interactions in the genome for the four ccRC stages and for control networks. With this procedure we inferred five networks, one for each stage and one for the control phenotype.

2.4.1. Network Interactions Assessment

In order to have those interactions with a higher relevance (as given by their mutual information values) for each phenotype, and in view of the so-called network sparsification problem, (determination of the number of *significant* edges that represent better the network structure consistent with the data), we decided to perform network cut-offs spanning over several scales well above and well beyond our working thresholds to account for possible size-effects. The cut-off thresholds range from the top 100 interactions, to the top 1,000,000 interactions, i.e., five orders of magnitude in network size. We performed those cut-offs to assess whether the effects under study, such as in the *cis-* rates was indeed due to network size.

Network visualizations were performed using *Cytoscape V 3.8.1* (Shannon et al., 2003), as well as the *iGraph* Python library (Csardi and Nepusz, 2006).

Since a relevant question underlies on whether in these networks, the effect of loss of *trans-* co-expression was also lost as in breast cancer (Espinal-Enriquez et al., 2017; de Anda-Jáuregui et al., 2019a,b,c; Dorantes-Gilardi et al., 2020; García-Cortés et al., 2020), we separated co-expression interactions into *cis-* (intra-chromosome), and *trans-* (inter-chromosome). We observed the *cis-/trans-* ratio for each phenotype.

2.5. Stages Intersections

One of the most important issues that might be addressed with a dataset such as the one we have, by means of the methodology exposed here, is how the co-expression landscape is modified throughout cancer progression. Derived from the latter, we compared the differences and intersections between the control network, and each progression stage. First, we observed the differences between network interactions, i.e., those gene-gene interactions that are not shared between phenotypes. Concomitantly, we observe those genetic interactions shared between control network and any ccRC progression stage.

Additionally, a question derived from the latter, is which interactions are conserved between all phenotypes, and also important, between cancer stages only. For that purpose, we performed a multi-group intersection to obtain the sub-network integrated by those links shared by all phenotypes, and also the ccRC-only sub-network.

2.6. Functional Enrichment

Functional enrichment analysis was performed using the *g:profiler* (Raudvere et al., 2019) API for Python. *g:Profiler* uses the hypergeometric test to measure the significance of a functional term in the input gene list (Reimand et al., 2007, 2011, 2016). Multiple testing corrections were performed by the g:SCS algorithm as implemented in *g:Profiler* with significance level $a = 0.05$; and a False Discovery Rate of 0.05.

It is worth noticing that in order to consider the network structure in the functional enrichment, the g:SCS algorithm was implemented over network communities, and not over the whole networks. For community detection in networks we performed the *Infomap* algorithm (Rosvall and Bergstrom, 2008),

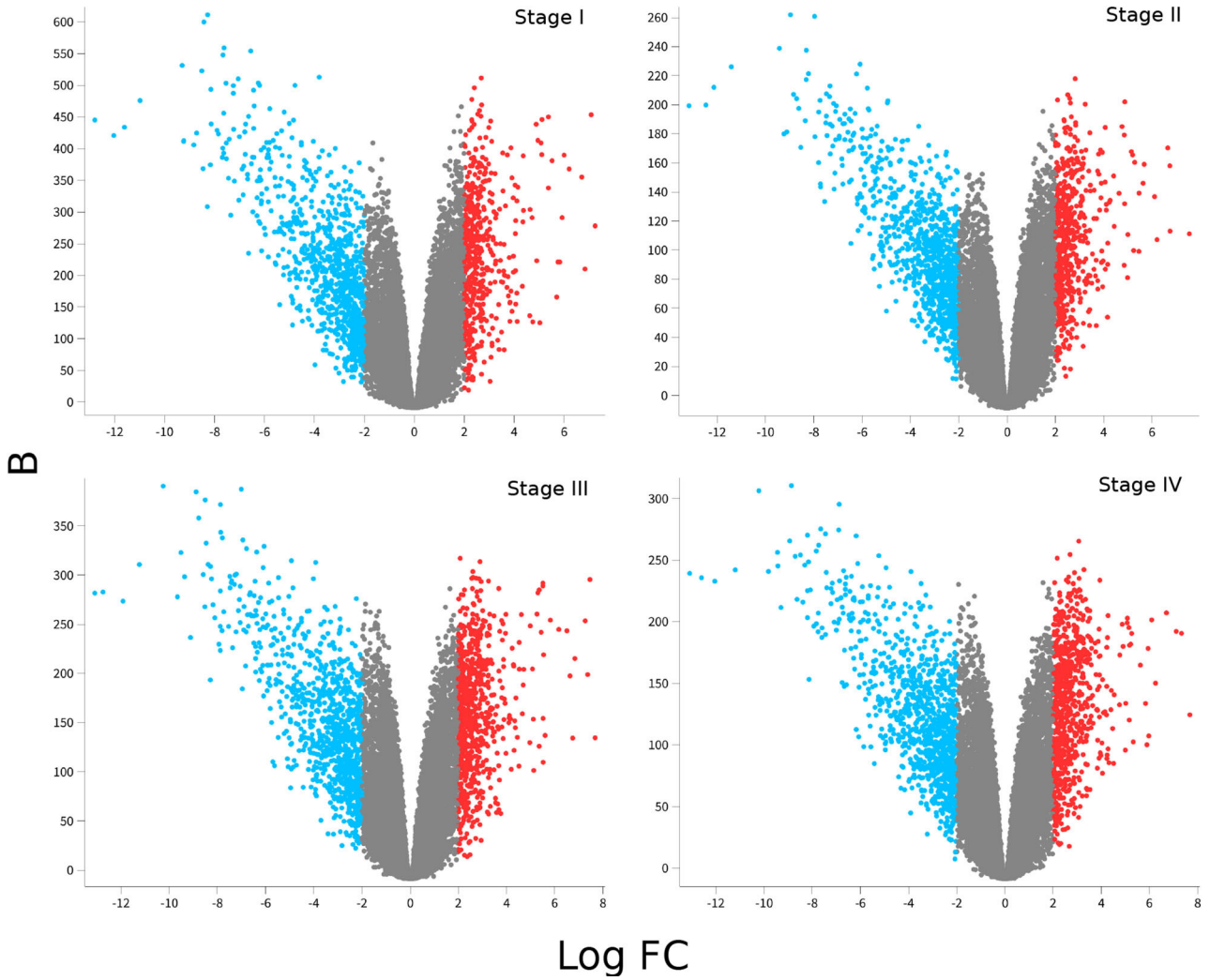

FIGURE 2 | Differential gene expression for each ccRC stage. In these volcano-plots, the differential expression between each stage vs. control samples is depicted. Red dots represent overexpressed genes, meanwhile underexpressed ones are in blue. Notice that underexpressed genes are more broadly distributed than overexpressed ones, and Log_2FC values are similar in the four figures; however, B-statistics change depending on the ccRC stage.

as implemented in Alcalá-Corona et al. (2016), Alcalá-Corona et al. (2017), and Alcalá-Corona et al. (2018).

In order to provide a clear and easy-to-follow manner to reproduce the results reported here, the five expression matrices, and all code for developing this work are provided in https://github.com/josemaz/kidney-stages. In this repository it can be found the code to reproduce all results, since the data download until functional enrichment.

3. RESULTS AND DISCUSSION

3.1. Differential Expression Is Similar Between ccRC Stages

After *low-level processing* of the four tumor stage data and control samples, we performed differential expression analysis for each stage compared with control samples (**Supplementary Material 2**).

Figure 2 shows volcano plots for differentially expressed genes in the four stages. Large similarity in the distribution of genes and range of values for the four stages is visible. The rank of differentially expressed genes is also similar. **Table 2** shows the Spearman's correlation of ranks between the four stages. As it can be observed, Spearman's $\rho_{corr} > 0.948$ in all cases, evidencing the similitude between differentially expressed gene ranks.

3.1.1. SLC6A19 and PLG Genes Show Progressively Decreasing Expression

Despite the fact that the four volcano plots are similar, and Spearman's correlation between all stages is high, some genes appear to be expressed according to tumor progression stages, such as the case of genes observed in **Figure 3**. Interestingly,

SLC6A19 and PLG, both show a remarkable decrease in their expression during progression stages (**Figure 3**).

3.1.2. SAAC2-SAAC4 and CXCL13 Genes Show Progressively Increasing Expression

Now regarding the overexpression of genes during ccRC progression, we found that only two genes, namely SAAC2-SAAC4 and CXCL13 genes, are overexpressed according to tumor progression stages, as it can be observed at the left side of **Figure 3**. It is worth to note that in the four cases, those genes are differentially expressed between control and any stage, but also between consecutive stages. This result may have clinical relevance since these protein-coding genes may be used as biomarkers of clear cell renal carcinoma progression.

Furthermore, we conducted a multi-group differential expression analysis, to observe whether or not said difference in gene expression also appeared between stages. In all cases, these genes are differentially expressed. However, between stage III and IV, the Log_2FC was set to 0.5. This means that the expression values of the four genes is different but not as largely different as in the previous stages. This could be due to the clinical and histo-pathological features that both stages may share.

To the best of our knowledge, the SLC6A19 gene has not been previously reported as importantly underexpressed in renal cancer, however, in the Human Protein Atlas, SLC6A19 underexpression has been reported as a biomarker for renal cancer (https://www.proteinatlas.org/ENSG00000174358-SLC6A19/pathology). SLC6A19 is highly expressed in kidney tissue (Fagerberg et al., 2014). Hence, its underexpression may bring relevant functional consequences.

TABLE 2 | Spearman correlation between rank of differentially expressed genes for all stages.

ccRC stage	Stage I	Stage II	Stage III	Stage IV
Stage I	1	0.995	0.974	0.948
Stage II	0.995	1	0.995	0.958
Stage III	0.974	0.994	1	0.997
Stage IV	0.948	0.958	0.997	1

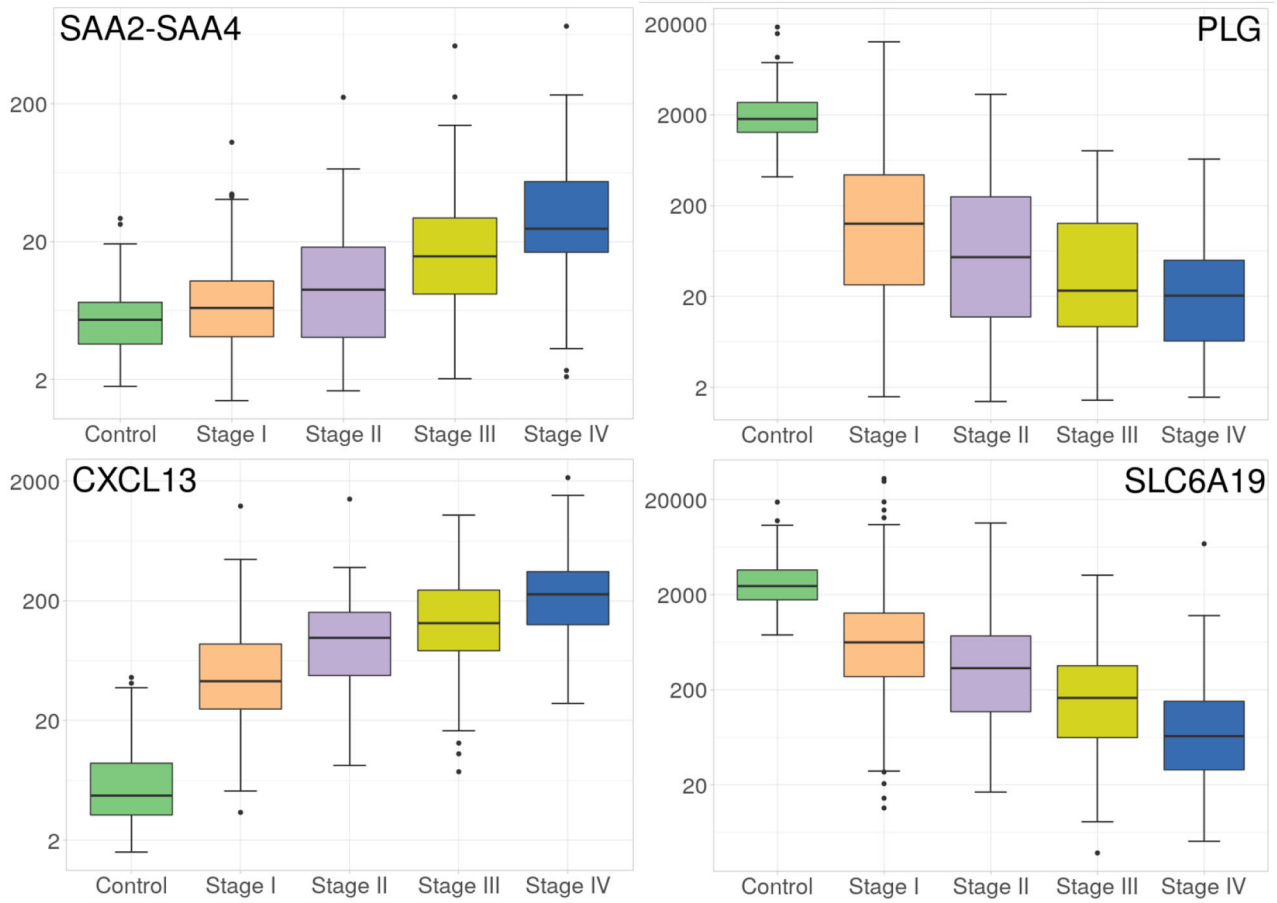

FIGURE 3 | Progressive increase and decrease in expression of four genes at the different ccRC stages. These barplots show the average expression of SAAC2-SAAC4 and CXCL13 genes **(left)**, and SLC6A19 and PLG genes **(right)**. Different colors represent the progression stages. Notice that the Y axis (gene expression) is, in all cases, depicted in log scale.

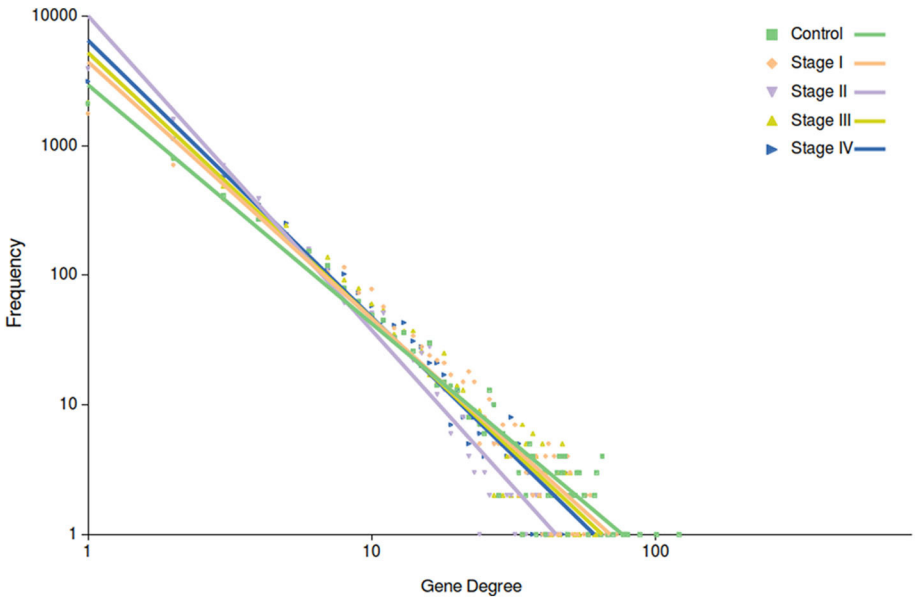

FIGURE 4 | Degree distribution of the five networks. In this plot, points correspond to the degree distribution for each phenotype. Color code is the same than **Figure 3**. Curve fitting ($y = ax^b$) to each degree distribution is also depicted. Notice that control network distribution slope (light green) is the lowest one.

PLG gene also presents a remarkable decrease throughout stages advance (**Figure 3**). Previously, PLG has been reported has decreased and a possible biomarker for renal carcinoma (Luo et al., 2018; Zhang et al., 2020).

In the case of CXCL13 overexpression, recently (Jiao et al., 2020), it has been found to be related to tumor-infiltrating immune cells, as well as bad prognosis in ccRC. In our case, we not only found the gene overexpresssed, but also progressively increased through the four stages.

Regarding SAA2-SAA4 gene, its overexpression has been observed as *unfavorable* in renal cancer, but at the same time *favorable* in breast cancer (https://www.proteinatlas.org/ENSG00000255071-SAA2-SAA4/pathology). SAA2-SAA4 is a naturally-occurred fusion between two serum amyloid genes (A2 and A4). SAA2-SAA4 overexpression has also been associated in metastatic brain tumor derived from papillary thyroid carcinoma (Schulten et al., 2016). It also has been associated with liver metastasis from colorectal tumor (Sayagués et al., 2016). The fact that SAA2-SAA4 overexpression has been associated with metastasis from neighboring primary tumor is matter of further research. However, it is worth mentioning that expression of this gene is progressively increased through ccRC progression stages.

To our knowledge, this is the first time that expression of SAA2-SAA4, CXCL13, PLG, and SLC6A19 have been reported to be differentially expressed through progression stages in clear cell renal carcinoma, showing a possible novel line of research related with ccRC genomic progressive alterations.

3.2. Control Network Is Topologically Different to Any Tumor Network

We found that all networks are substantially different among them, but the control one presents a more striking difference in terms of its topological features. Control network has a larger proportion of *trans*- interactions, whereas for any cancer stage the amount of intra-chromosome (*cis*-) interactions are more abundant than *trans*- ones.

TABLE 3 | Parameters of the non-linear curve fitting in all networks for the top 10,000 interactions.

Parameter	Control	Stage I	Stage II	Stage III	Stage IV
a	2941.8	4430.5	10047	5215.8	6490.8
b	−1.842	−1.982	−2.4266	−2.052	−2.137
Correlation	0.992	0.981	0.973	0.98	0.987
R-square	0.935	0.931	0.953	0.939	0.959

Among the most important network parameters to examine is the degree distribution $p(k)$. It is well-known that the k vs. $p(k)$ plot and its parameters for curve fitting may reflect several properties related to the system itself. In the case of top 10,000 edges cut-off, we may observe that in all cases the distribution is well-fitted to a power law distribution ($y = ax^b$). The differences are observed in **Figure 4**, in the different slopes of the curve fittings, as well as at the parameter level. The slope of control network degree distribution (light green) is the lowest one (−1.842), compared to the ccRC stages. **Table 3** contains the parameters for the non-linear curve fitting of the five networks. The latter may describe that *long-range communication* is a feature in a healthy phenotype.

3.3. Statistical Networks Differences

3.3.1. There Is a Preferential *cis*- Co-expression in ccRC Networks

Giant connected components of each network are depicted in **Figure 5**. Genes are colored according the chromosome each gene belongs to. In the control network, genes co-express with

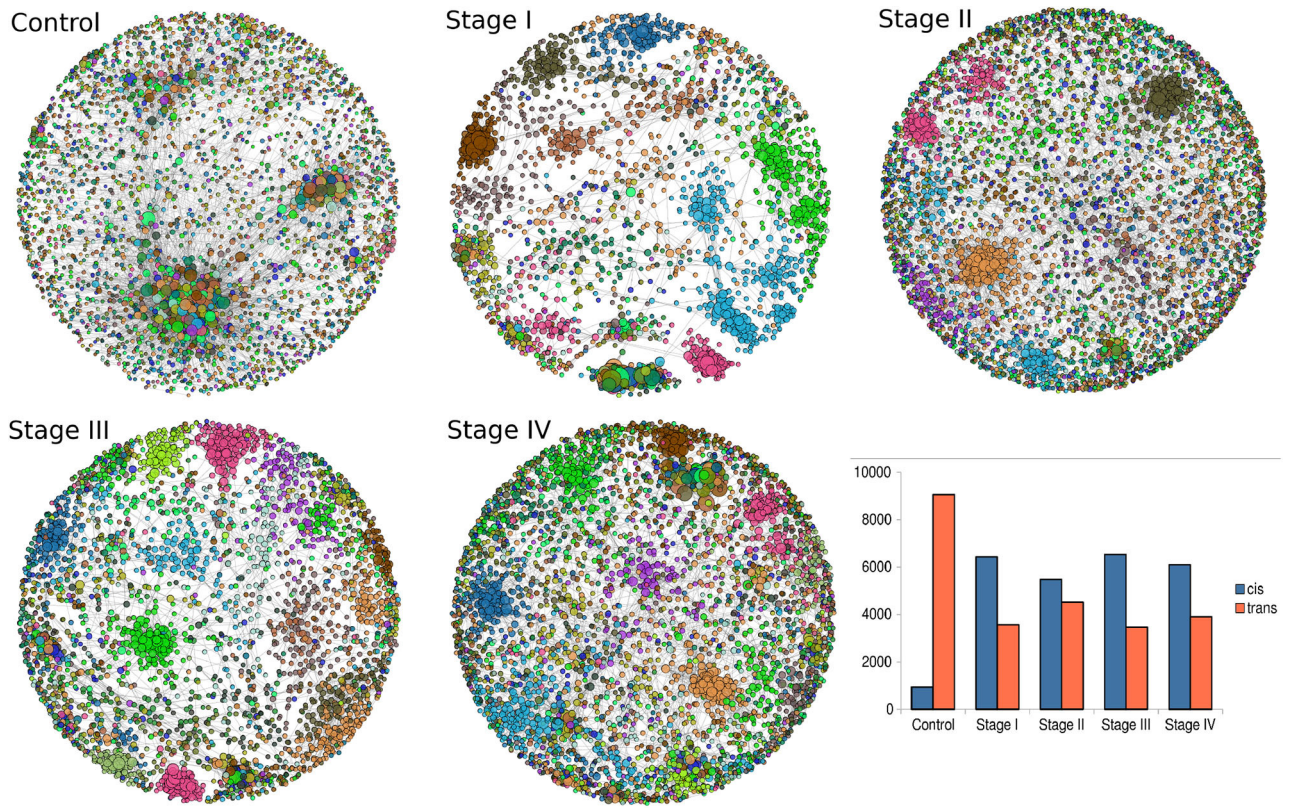

FIGURE 5 | Network topologies of ccRC per stage. Top to bottom figures correspond to the largest connected component of control, stage I, stage II, stage III, and Stage IV, respectively. Color of nodes correspond to the chromosome to which each gene belongs to. The bar-chart represents the proportion of *cis*- (blue) and *trans*- (orange) interactions.

genes from any chromosome, with a high prevalence of *trans*- interactions. Conversely, for the ccRC stages, in all cases there is preferential *cis*- co-expression. This is also reflected in the bar-charts at the bottom right part of **Figure 5**. Orange bars represent the number of *trans*- interactions, meanwhile *cis*- links are represented by blue bars.

3.3.2. *cis*-/*trans*- Ratios Do Not Re-trace Progression Stages

In previous works from our group (García-Cortés et al., 2020), we have shown that *cis*-/*trans*- ratio increased with severity of breast cancer subtypes, being Luminal A, Luminal B, HER2+ and Basal the order of *cis*-/*trans*- ratios. There, we also shown that breast control network is the only graph that contains more *trans*- interactions than *cis*- ones.

Intuitively, one may expect (based on our previous experience with breast cancer) a progressive decrease in the number of *trans*- interactions, starting from the largest number in control network, decreasing throughout ccRC stages. However, this is not the case, as it can be also appreciated from the bar-charts, as well as from networks. The ccRC network with less *trans*- co-expression links is stage III, followed by stage I, stage IV, and finally stage II. However, the difference between control and any stage is also evident.

3.3.3. Chromosome-Specific *cis*- Rates Are Different Between Phenotypes

Once the proportion of global *cis*-/*trans*- interactions were obtained, isolated chromosome *cis*- rates were calculated. We defined the *cis*- *rate* as the number of *cis*- edges divided by the total number of edges in each network. As it can be observed in the barplot of **Figure 6**, for the control network, all chromosomes but ChrY have a $cis - rate < 1$, but in the case of Chr Y, all phenotypes have a *cis*- rate > 1. In general, stage III network has the highest *cis*- rates at the chromosome level.

3.4. Topological Differences Do Not Follow Progression Stages

As a first approach, network cut-off was set to top-10,000 edges, ranked by MI values. Each network contains a different number of genes. Since these networks are obtained from gene expression of kidney tissue, one may naively expect similarities in terms of genes and even interactions. Additionally, given that the networks under study were separated into progression stages, it also would be expected that consecutive stages were more similar between them than with the rest of networks.

In **Figure 8**, we show the number of shared interactions between phenotypes, as well as their differences. As expected, control network is the most different in terms of number

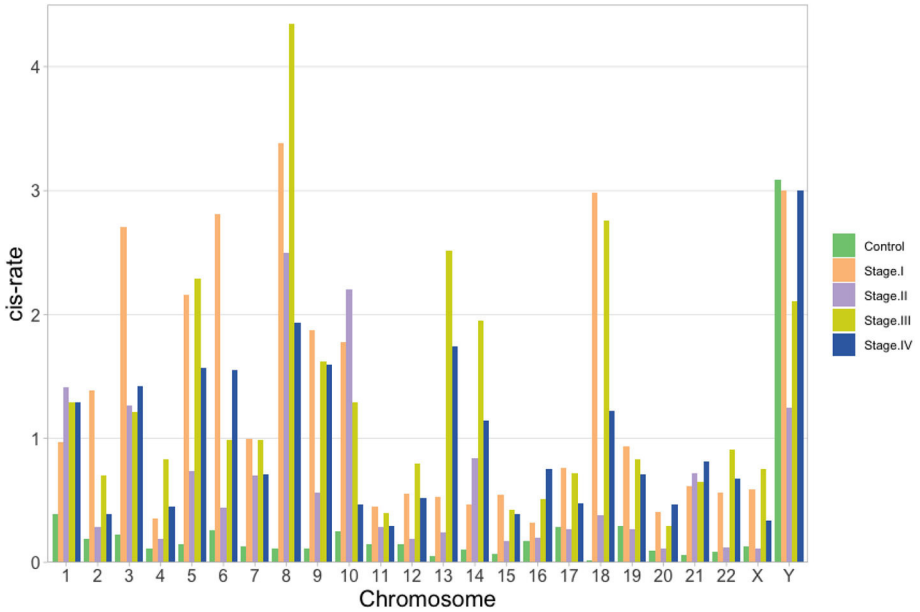

FIGURE 6 | *cis-* rate (*cis edges/# of genes*) per chromosome for the five networks: green, orange, violet, yellow and blue for control, stage I, stage II, stage III, and stage IV, respectively. In all cases but for ChrY, the ratio is lower than 1 for the control network.

of shared links with the ccRC networks. The percentage of divergence is 94% in the more similar case (stage I).

ccRC networks also differ vastly between them, more than 60% difference in any case. Stage II network is the most different, in terms of number of shared edges. Conversely, stage I and stage III networks are the more similar pair, even stage I and stage IV keep more shared edges between them (74%) than with stage II.

The latter results is surprising, taking into account the high similitude in terms of differential gene expression in the four phenotypes (**Table 2**). Network topologies and the concomitant co-expression programs do not coincide with the gene expression signatures of ccRC progression stages.

Additionally, the small number of shared genomic interactions between control and ccRC networks also reflects, a radical rearrangement of the transcriptional program between health and disease.

Biologically, the decrease in network commonalities between phenotypes is a clear indicative that each one of the ccRC stages behave differently. This could be important, since each network maps a specific snapshot of the co-expression landscape at different moments of the carcinogenic process. Analysis of those unique co-expression features could help in the understanding of the cancer progression process.

3.4.1. Most Interactions Are Phenotype-Specific

In **Figure 7**, the intersection of co-expression interactions for the five phenotype networks is depicted. As it can be observed, the largest number of links belongs to the non-shared sets for the five networks. This indicates that, independently of the phenotype, networks are structurally different. As in the previous figure, the largest difference occurs in the control network (9,295 unique edges). Five thirty-three edges are shared between the four ccRC phenotypes. This is the set of co-expression interactions that appear at any stage of clear cell renal carcinoma.

3.5. Network Topologies at Different MI Cut-Offs

Since cut-off election is still a non-closed problem in network science (the so-called network sparsification problem), we decided to cover a wide range of cut-offs to assess the observed result in the previous sections. We pruned the original networks (16,000 genes, 130 millions of edges) into small mutual-information-ranked sets, from Top-100 to Top-1,000,000 edges, i.e., covering five orders of magnitude. See **Supplementary Material 3**.

3.5.1. Proportion of Networks Intersection Decrease With Network Sizes

In **Figure 8** one can appreciate that the proportion of intersections between all phenotypes (control and ccRC), as well as in ccRC-only networks, is maintained in a wide range of network cut-offs. It can be clearly appreciated the consecutive decrease of the proportion of shared links according to networks growing in size.

3.5.2. Chromosomal Connectivity Differences Between Control and Cancer Networks Are Independent of the MI Cut-Off

Regarding the *cis-* and *trans-* difference between control and cancer networks, in **Figure 9** we may observe that *trans-* interactions in control are always higher than any tumor ccRC network, despite the MI cut-off value.

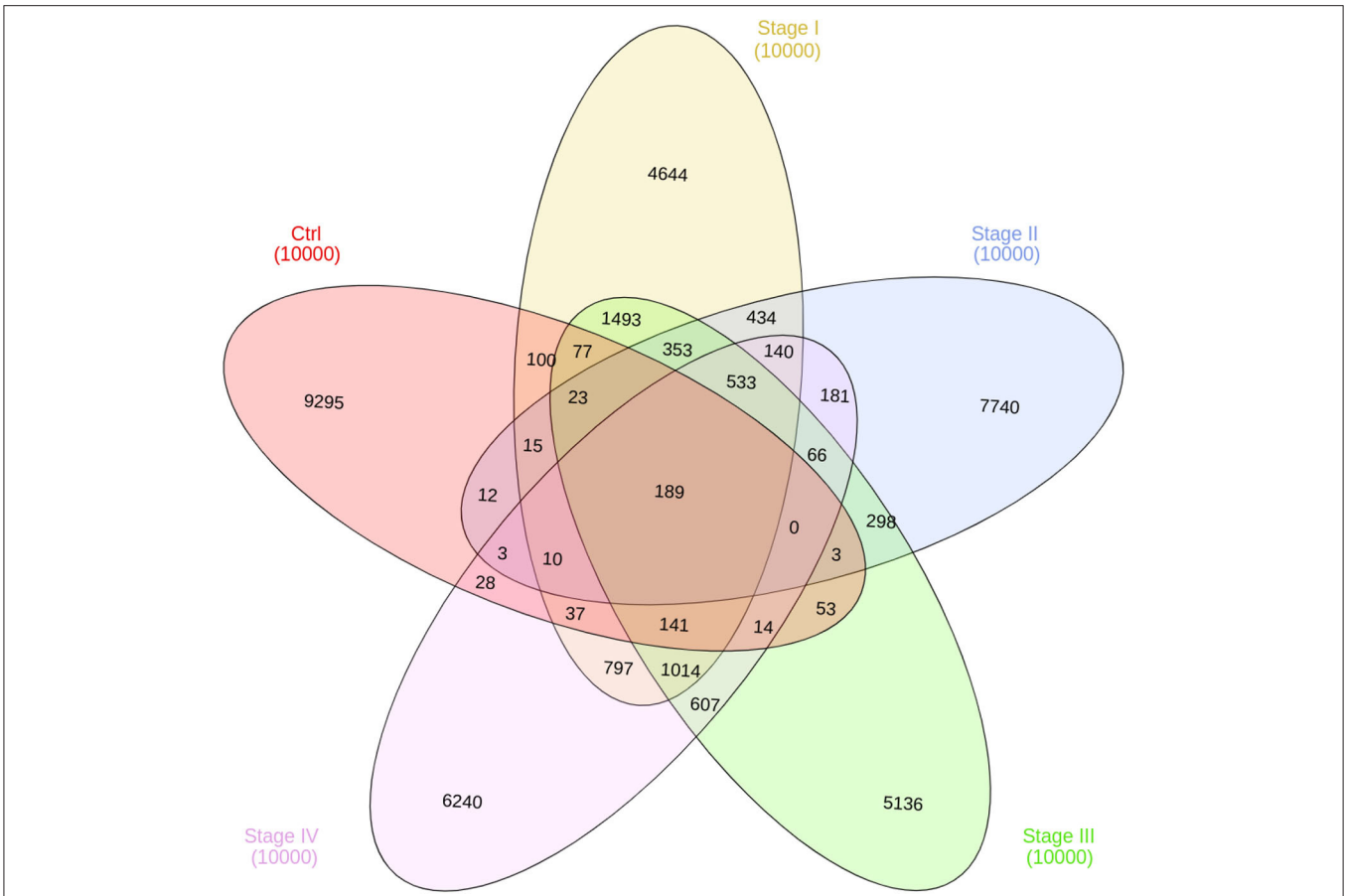

FIGURE 7 | Edge intersection of all networks. Venn diagram shows, in each set, the number of edges per phenotype. The number reflect the shared genes between networks, as well as network-specific interactions. Notice that out of 10,000 interactions, only 189 edges are shared between the five networks.

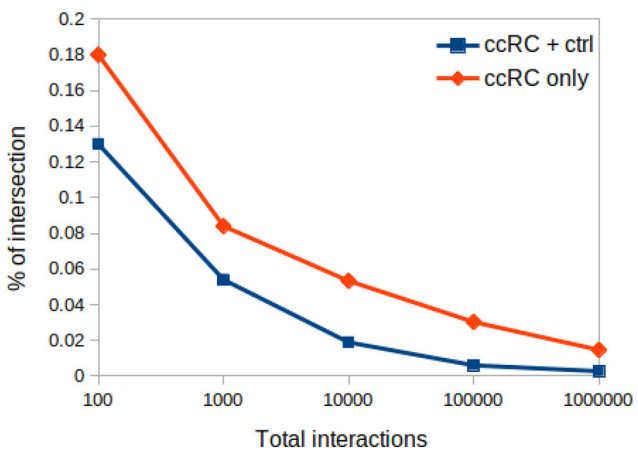

FIGURE 8 | Proportion of networks intersection at different network cut-offs. In this plot, proportion of network intersection between the four ccRC stages (orange diamonds), and those with control network (blue squares) is depicted. X-axis represent different network cut-off values.

It can be also appreciated that *trans-* interactions tend to converge according to the size increase. This result is expected since the more edges appear in the network, the more *cis-* edges have been "loaded" to prior cut-offs. This results also coincide with a recent finding in breast cancer networks, where consecutive non-overlapping layers of 100,000 edges (ranked top-to-bottom MI) contain more *cis-* interactions in top layers, and decreasing as they get close to the noise layer (Dorantes-Gilardi et al., 2020).

3.5.3. Cancer Networks Present a Shift in the Order of cis-Rate in a Small Range of Interactions

In **Figure 9** can also be observed that from the beginning range (100) to approximately 3,000 edges, the rank of *trans-* interactions is stage $I \rightarrow III \rightarrow IV \rightarrow II$. However, in the range 3,000-to-10,000 edges this rank in ccRC networks changes from $I \rightarrow III \rightarrow IV \rightarrow II$ to $II \rightarrow IV \rightarrow III \rightarrow I$. That acquired order is preserved until the already commented convergence at 1,000,000 edges.

As previously mentioned, the rank of *cis-*/*trans-* proportion does not follow progression of ccRC at any cut-off value. Hence we may conclude that differences in intra/inter-chromosomal network interactions are not a very informative parameter to evaluate progression in ccRC. Further investigation on the aforementioned shift is needed to have a more complete idea of the phenomenon.

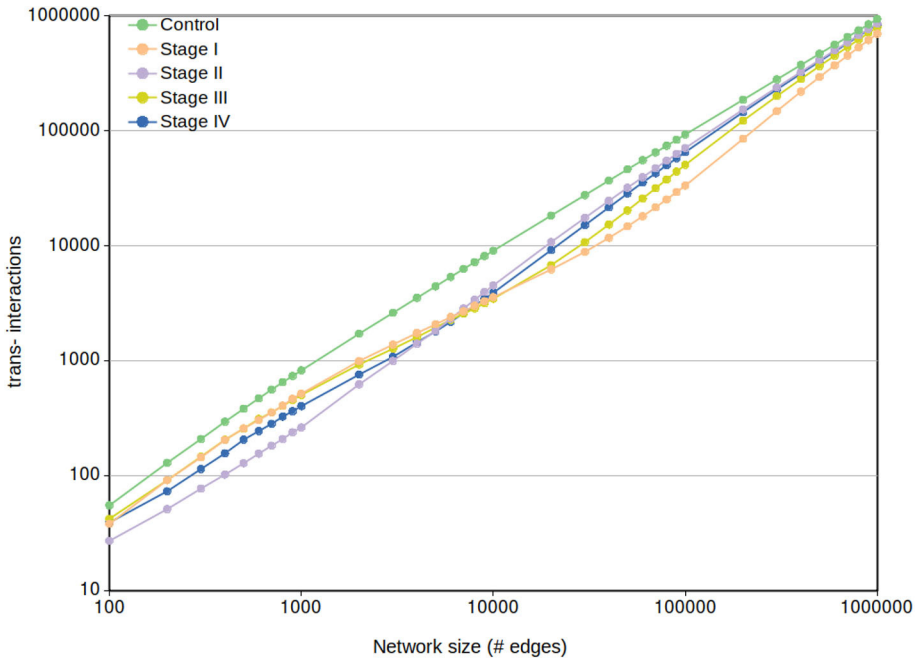

FIGURE 9 | Network *trans*- interactions at different cut-offs. In this plot, X-axis represents the cut-off network value for the five different networks (control and the four ccRC stages). Y-axis shows the number of inter-chromosomal interactions per each network cut-off. To note that the control network *trans*- edges are larger than any ccRC progression stage at any cut-off network value.

3.6. 189 Biologically Relevant Edges Are Shared in the Five Phenotypes

189 co-expression interactions are shared between the five networks. Those interactions are depicted in **Figure 10**. The resulting network is composed of 230 genes and 189 edges. Genes are colored according to their differential gene expression.

Interestingly, network components of this common sub-network are mostly clustered according to the differential expression trend: there are clusters composed by over-expressed genes only, as well as under-expressed-only ones. It is worth mentioning that the Spearman's correlation between the rank of differentially expressed genes is higher than 0.95 for any stage (**Table 2**).

Additionally, the small connected components are enriched for particular and specific biological processes. For example, the first component, which contain genes such as KIF20A, KIF18B, or UBE2C, is enriched for apical and tight junction assembly. This is a highly overexpressed component, which indicates that for any stage, tight junction and apical junction assembly are exacerbated processes. Conversely, the third component, with genes such as EGR2, EGR3, ATF, or FOSB is completely underexpressed, and it is enriched for immune response-related processes, which could mean that the immune response is depleted at any stage of ccRC.

3.7. Enriched Categories Are Independent of the Cut-Off-Value

Figure 11 shows the enriched categories obtained by intersecting the four progression stages (and excluding control interactions). Analog to **Figure 10**, in this case (533 edges) we have genes colored by their differential expression values, meanwhile enriched categories are painted by different colors depending on the component to which those processes belong. It is worth to mention that this figure only includes processes with a $p - value < 10^{-10}$. The complete list of enriched processes for both cases, all phenotypes and ccRC-only, in the five cut-off network values, is included in **Supplementary Material 4**. Additionally, network visualizations of enriched processes in the ccRC-only network intersection for 100,000 and 1,000,000 interactions are also included in **Supplementary Material 4**.

Another shared feature of this figure with **Figure 10** is that gene clusters with the same trend of differential expression have enriched categories. Among the top-enriched categories we may find cell-cycle-related (yellow), IFN-γ-related (dark blue), and T-cell-related (green) processes.

Since in this work one of the most relevant questions we made was related to the network structure at different progression stages in clear cell renal carcinoma, we calculated the over-expression analysis over network communities by means of the *infomap* algorithm (Rosvall and Bergstrom, 2008). We performed the enrichment analysis over separated sets of genes according to the community to which genes belong.

Given the fact that large networks often contains more communities than small networks, we performed the enrichment analysis for different cut-off values of network intersections. Independent of the network cut-off, intersections of ccRC-only networks always present this set of enriched categories, associated with cell-cycle, immune system, tridimensional structure of DNA and chromatin, or Transcription regulation (**Supplementary Material 4**).

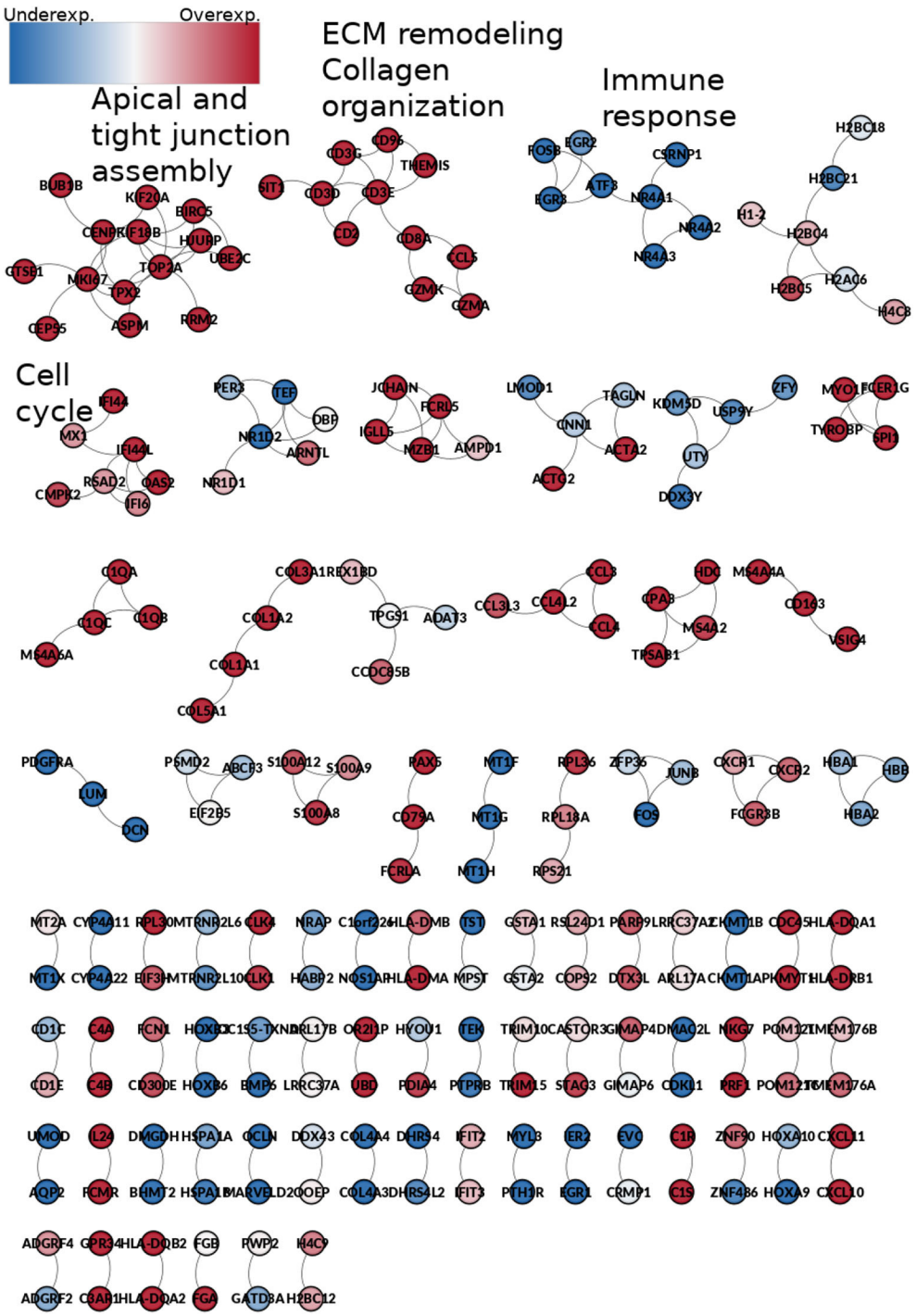

FIGURE 10 | Network from shared interactions between the five phenotypes. The resulting network is composed of 189 edges and 230 genes. Those are colored according to the differential expression compared with the control group. Notice that network smaller components have a similar expression pattern. Some components are enriched to specific GO categories, meaning that those processes are increased or decreased during the whole process of ccRC progression.

4. CONCLUDING REMARKS

Clear cell renal carcinoma is a complex disease. It involves several layers of complexity. It must be dissected to have a comprehensive landscape allowing for a better understanding of its progression.

In previous work we observed an important increment in *cis*- ratio in breast cancer molecular subtypes, according to the

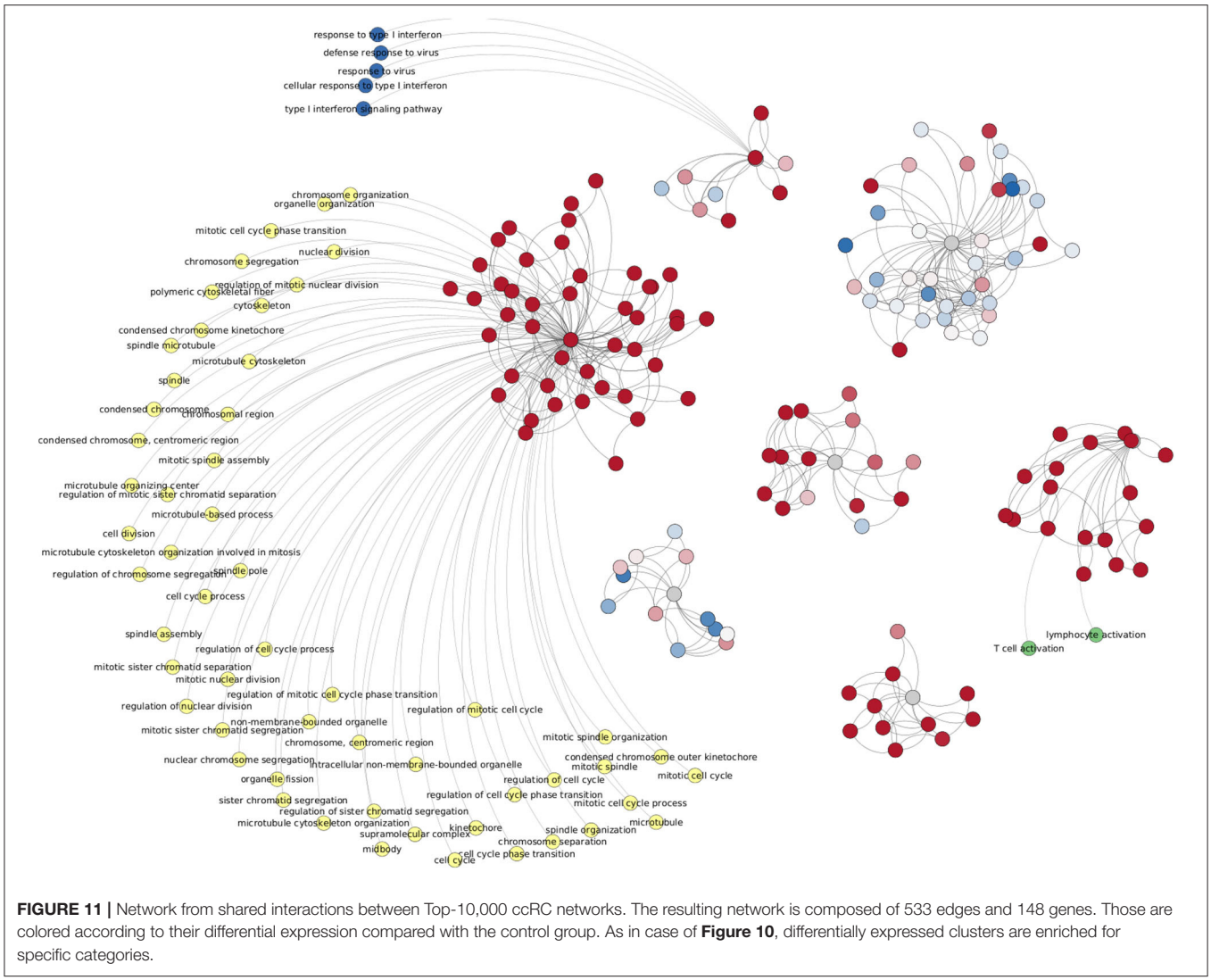

FIGURE 11 | Network from shared interactions between Top-10,000 ccRC networks. The resulting network is composed of 533 edges and 148 genes. Those are colored according to their differential expression compared with the control group. As in case of **Figure 10**, differentially expressed clusters are enriched for specific categories.

malignancy of those phenotypes. Since the loss of long-range co-expression was observed in breast cancer and more remarkably in the Basal subtype (the one with worst prognosis), our working hypothesis was the more advanced the cancer stage, the higher the *cis-* ratio.

After breast cancer network analysis reported previously, clear cell renal carcinoma is the second cancer in which we observe a remarkable difference between *cis-* and *trans-* interactions, showing an important decrease in inter-chromosome gene-gene co-expression interactions in cancer networks.

Unexpectedly, the progression stage does not correlate with *cis-* ratio. This was observed not only in the top-10,000 edges networks, but also in a rank of five orders of magnitude. This could imply that *cis-* ratio is not a parameter to distinguish progression stages, at least for ccRC.

By observing the discrepancy between the *cis-* rate of ccRC progression stages with those observed in breast cancer molecular subtypes (García-Cortés et al., 2020), regarding that high proportion of intra-chromosome interactions are observed in those phenotypes with a worst prognosis we may argue the following:

- The fact that *cis-* rate does not coincide with progression stages, may reflect that high proportion of intra-chromosomal interactions are not a parameter to take into account to differentiate cancer progression, at least in clear cell renal carcinoma.
- A high *cis-* rate does not imply malignancy or worst prognosis in a cancerous network, but a different co-expression program in which gene-interactions are favored to physically close genes.
- The mechanisms behind the preferential co-expression to neighbor genes must imply epigenetic factors, such as micro-RNAs, lncRNAs, methylation profiles, tridimensional structure of DNA, chromatin modifications, CTCF binding sites, etc. (For a profound revision of spatial regulation of DNA in the oncogenic process, see Hernández-Lemus et al., 2019).

We want to stress that kidney cancers are fundamentally different from breast cancers in many forms (Hoadley et al., 2018). For the latter, topological similarities between breast cancer and ccRC co-expression networks must be taken carefully. However, it is remarkable that in both tissues (clear cell and breast carcinomas), as well as in separated instances (progression stages and molecular subtypes), the effect of loss of long-range co-expression is a common feature of cancer.

Here, we have focused on two main molecular signatures, namely the expression and the co-expression landscapes. In the first layer, we have observed that the differential expression profile is very similar between progression stages, even between stage I and stage IV, which may indicate that the expression profile is somehow acquired once cancer has started. However, certain genes appear to replicate the progression of oncogenic process, such as the case of SLC6A19 and PLG (underexpression), and SAAC2-SAAC4 and CXCL13 (overexpression). It is worth mentioning that none of these genes have been previously reported as progressively differentiated in renal carcinoma.

On the other hand, the similitude observed at the expression level, was not observed at the co-expression network level. Actually, the number of shared links is really low. We argue that the differential expression profiles are indeed insufficient to properly describe gene expression regulation, but the way that those genes interact in time and space is what ultimately determine the establishment of tumor phenotype.

In the case of **Figure 6**, the fact that chromosome Y is the only one with a higher *cis-* rate in control network than in ccRC stages may imply that, for this chromosome and its genes, local co-expression is crucial to maintain a proper functionality. It is widely known that two thirds of all ccRC cases correspond to men (Aron et al., 2008; Woldrich et al., 2008; Qu et al., 2015; Zaitsu et al., 2020). Since Chr Y is directly linked to gender, one may argue that an imbalance in the *cis-/trans-*proportion may be implicated in gender-bias on clear cell renal carcinoma.

Despite the high differences between control and stage networks, and even between stages, there are some conserved gene co-expression relationships independent of the phenotype. An instance of this is shown **Figure 10**. Those interactions shared among the five phenotypes show very few common links, but clustered in biologically relevant genesets. Those genesets are important for cell maintenance (that is perhaps, the reason for which they appear in the control network). At the same time, these genesets are overexpressed, thus indicating an exacerbated process in the cancer stages, as in the case of apical and tight junction assembly, or extracellular matrix remodeling.

Conversely, the immune response cluster is depleted, thus indicating that the immune system response may be decreased at any moment in the course of the carcinogenic process.

Analogously, in **Figure 11** we may observe the shared interactions between cancer-only networks. This subset of interactions may result of the utmost relevance, since it represents those gene-gene co-expression relationships that are exclusive of clear cell renal carcinoma. These interactions are highly enriched for very specific biological processes, which means that these interactions may have repercussion in cell functionality. Another point to remark regarding ccRC-only intersection is that the enriched functions are preserved at five orders of magnitude network sizes.

The fact that topological and functional analyses show similar results at five orders of magnitude in network sizes, have implications in at least two main issues: (a) *cis-* rate is invariant to the cut-off, and (b) enriched categories do not depend of the cut-off value. Here, we have provided a methodology to discover functional characteristics of gene-co-expression networks that are intrinsic to the phenotype and not depend on the network cut-off.

We are aware that gene co-expression may be strongly influenced by several factors: micro-RNAs, long non-coding RNAs, methylation patterns, copy number alterations, 3D-structure of DNA, CTCFs binding sites, to mention but a few. More research is thus needed for a better understanding of the delicate interplay between gene expression and co-expression. This is a first approach to draw close both worlds in an integrative manner.

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

JZ-F performed computational analyses, developed and implemented programming code, performed pre-processing and low-level data analysis, made the figures, and drafted the manuscript. EH-L contributed to design the theoretical and modeling analysis and contributed and supervised the writing of the manuscript. JE-E conceived and designed the project, supervised the project, performed biological analyses, and drafted the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by CONACYT (267236 Ph.D. student scholarship to JZ-F, 285544/2016, and 2115/2018 to JE-E), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). Additional support has been granted by the National Laboratory of Complexity Sciences (232647/2014 CONACYT). JE-E was recipient of the 2018 Miguel Alemán Fellowship in Health Sciences. EH-L was a recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences. JZ-F was a doctoral student from the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). This work was part of his Ph.D. Thesis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.578679/full#supplementary-material

Supplementary Material 1 | Quality control for gene expression in the five phenotypes. This zip file contains five folders with the quality control pre and post normalization of gene expression data. They include length bias correction, GC-content correction, and PCA for the five groups.

Supplementary Material 2 | Differential expression values for each ccRC stage vs. control samples. This zip file contains the four differential expression analyses, including Log_2FC value, p-value, adjusted p-value, and B-statistic for each gene in all stages. HTML files for volcano plots are also provided.

Supplementary Material 3 | Heatmaps for intersections and differences in all phenotypes with MI cut-offs of 100, 1,000, 10,000, 100,000, and 1,000,000 interactions. Venn diagrams for intersections of all phenotypes with the aforementioned cut-off values.

Supplementary Material 4 | Complete list of enriched processes for network intersections at different cut-off values. These files contain the enriched categories for both sets, all-phenotypes (control and the four progression stages), as well as ccRC-only intersections. In all cases, the list of enriched categories was performed over network communities, not over the whole geneset. Files are separated based on the cut-off of networks. Visualization of enriched process in Networks of 100,000 and 1,000,000 MI cuts.

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. Physiol*. 8:915. doi: 10.3389/fphys.2017.00915

Alcalá-Corona, S. A., Espinal-Enríquez, J., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2018). The hierarchical modular structure of HER2+ breast cancer network. *Front. Physiol*. 9:1423. doi: 10.3389/fphys.2018.01423

Alcalá-Corona, S. A., Velázquez-Caldelas, T. E., Espinal-Enríquez, J., and Hernández-Lemus, E. (2016). Community structure reveals biologically functional modules in MEF2C transcriptional regulatory network. *Front. Physiol*. 7:184. doi: 10.3389/fphys.2016.00184

Arjumand, W., and Sultana, S. (2012). Role of VHL gene mutation in human renal cell carcinoma. *Tumor Biol*. 33, 9–16. doi: 10.1007/s13277-011-0257-3

Aron, M., Nguyen, M. M., Stein, R. J., and Gill, I. S. (2008). Impact of gender in renal cell carcinoma: an analysis of the seer database. *Eur. Urol*. 54, 133–142. doi: 10.1016/j.eururo.2007.12.001

Braga, E., Khodyrev, D., Loginov, V., Pronina, I., Senchenko, V., Dmitriev, A., et al. (2015). Methylation in the regulation of the expression of chromosome 3 and microRNA genes in clear-cell renal cell carcinomas. *Russ. J. Genet*. 51, 566–581. doi: 10.1134/S1022795415050026

Braga, E. A., Fridman, M. V., Loginov, V. I., Dmitriev, A. A., and Morozov, S. G. (2019). Molecular mechanisms in clear cell renal cell carcinoma: Role of miRNAs and hypermethylated miRNA genes in crucial oncogenic pathways and processes. *Front. Genet*. 10:320. doi: 10.3389/fgene.2019.00320

Cowey, C. L., and Rathmell, W. K. (2009). VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy. *Curr. Oncol. Rep*. 11, 94–101. doi: 10.1007/s11912-009-0015-5

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* 1695, 1–9.

de Anda-Jáuregui, G., Alcalá-Corona, S. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019a). Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Appl. Netw. Sci*. 4:22. doi: 10.1007/s41109-019-0129-0

de Anda-Jáuregui, G., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019b). Spatial organization of the gene regulatory program: an information theoretical approach to breast cancer transcriptomics. *Entropy* 21:195. doi: 10.3390/e21020195

de Anda-Jáuregui, G., Fresno, C., García-Cortés, D., Enríquez, J. E., and Hernández-Lemus, E. (2019c). Intrachromosomal regulation decay in breast cancer. *Appl. Math. Nonlinear Sci*. 4, 223–230. doi: 10.2478/AMNS.2019.1.00020

Dmitriev, A. A., Rudenko, E. E., Kudryavtseva, A. V., Krasnov, G. S., Gordiyuk, V. V., Melnikova, N. V., et al. (2014). Epigenetic alterations of chromosome 3 revealed by noti-microarrays in clear cell renal cell carcinoma. *BioMed Res. Int*. 2014. doi: 10.1155/2014/735292

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, 1–23. doi: 10.1007/s41109-020-00291-1

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of EMT and met micro-RNA regulation in breast cancer. Scientific reports 7, 1–17. doi: 10.1038/s41598-017-13903-1

Edge, S. B., Byrd, D. R., Carducci, M. A., Compton, C. C., Fritz, A., Greene, F., et al. (2010). *AJCC Cancer Staging Manual*, Vol. 7. New York, NY: Springer.

Espinal-Enriquez, 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, 1–19. doi: 10.1038/s41598-017-01314-1

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* 13, 397–406. doi: 10.1074/mcp.M113.035600

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernandez-Lemus, E., and Espinal-Enriquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol*. 10:1232. doi: 10.3389/fonc.2020.01232

Hernández-Lemus, E., Reyes-Gopar, H., Espinal-Enríquez, J., and Ochoa, S. (2019). The many faces of gene regulation in cancer: a computational oncogenomics outlook. *Genes* 10:865. doi: 10.3390/genes10110865

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.

Jaffe, E. S., Harris, N. L., Stein, H., and Vardiman, J. W. (eds.). (2001). *World Health Organization Classification of Tumours, Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues*. Lyon: IARC Press.

Jiao, F., Sun, H., Yang, Q., Sun, H., Wang, Z., Liu, M., et al. (2020). Association of cxcl13 and immune cell infiltration signature in clear cell renal cell carcinoma. *Int. J. Med. Sci*. 17:1610. doi: 10.7150/ijms.46874

Jung, M., Mollenkopf, H.-J., Grimm, C., Wagner, I., Albrecht, M., Waller, T., et al. (2009). MicroRNA profiling of clear cell renal cell cancer identifies a robust signature to define renal malignancy. *J. Cell. Mol. Med*. 13, 3918–3928. doi: 10.1111/j.1582-4934.2009.00705.x

Kaelin, W. G. (2004). The von Hippel-Lindau tumor suppressor gene and kidney cancer. *Clin. Cancer Res*. 10:6290S–6295S. doi: 10.1158/1078-0432.CCR-sup-040025

Li, M., Wang, Y., Song, Y., Bu, R., Yin, B., Fei, X., et al. (2015). MicroRNAs in renal cell carcinoma: a systematic review of clinical implications. *Oncol. Rep*. 33, 1571–1578. doi: 10.3892/or.2015.3799

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. doi: 10.1186/s13059-014-0550-8

Luo, T., Chen, X., Zeng, S., Guan, B., Hu, B., Meng, Y., et al. (2018). Bioinformatic identification of key genes and analysis of prognostic values in clear cell renal cell carcinoma. *Oncol. Lett*. 16, 1747–1757. doi: 10.3892/ol.2018.8842

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7:S7. doi: 10.1186/1471-2105-7-S1-S7

Moch, H. (2013). An overview of renal cell cancer: pathology and genetics. *Semin. Cancer Biol*. 23, 3–9. doi: 10.1016/j.semcancer.2012.06.006

Moch, H., Cubilla, A. L., Humphrey, P. A., Reuter, V. E., and Ulbright, T. M. (2016). The 2016 who classification of tumours of the urinary system and male genital organs–part A: renal, penile, and testicular tumours. *Eur. Urol.* 70, 93–105. doi: 10.1016/j.eururo.2016.02.029

Neely, B. A., Wilkins, C. E., Marlow, L. A., Malyarenko, D., Kim, Y., Ignatchenko, A., et al. (2016). Proteotranscriptomic analysis reveals stage specific changes in the molecular landscape of clear-cell renal cell carcinoma. *PLoS ONE* 11:e0154074. doi: 10.1371/journal.pone.0154074

Nueda, M. J., Ferrer, A., and Conesa, A. (2012). ARSYN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics* 13, 553–566. doi: 10.1093/biostatistics/kxr042

Qu, Y., Chen, H., Gu, W., Gu, C., Zhang, H., Xu, J., et al. (2015). Age-dependent association between sex and renal cell carcinoma mortality: a population-based analysis. *Sci. Rep.* 5:9160. doi: 10.1038/srep09160

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. doi: 10.1093/nar/gkz369

Redova, M., Svoboda, M., and Slaby, O. (2011). MicroRNAs and their target gene networks in renal cell carcinoma. *Biochem. Biophys. Res. Commun.* 405, 153–156. doi: 10.1016/j.bbrc.2011.01.019

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., et al. (2016). g:profiler–a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 44, W83–W89. doi: 10.1093/nar/gkw199

Reimand, J., Arak, T., and Vilo, J. (2011). g:profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* 39(Suppl_2):W307–W315. doi: 10.1093/nar/gkr378

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 35(Suppl_2):W193–W200. doi: 10.1093/nar/gkm226

Ricketts, C. J., De Cubas, A. A., Fan, H., Smith, C. C., Lang, M., Reznik, E., et al. (2018). The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* 23, 313–326. doi: 10.1016/j.celrep.2018.03.075

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinformatics* 12:480. doi: 10.1186/1471-2105-12-480

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123. doi: 10.1073/pnas.0706851105

Sayagués, J. M., Corchete, L. A., Gutiérrez, M. L., Sarasquete, M. E., del Mar Abad, M., Bengoechea, O., et al. (2016). Genomic characterization of liver metastases from colorectal cancer patients. *Oncotarget* 7:72908. doi: 10.18632/oncotarget.12140

Schulten, H.-J., Hussein, D., Al-Adwani, F., Karim, S., Al-Maghrabi, J., Al-Sharif, M., et al. (2016). Microarray expression profiling identifies genes, including cytokines, and biofunctions, as diapedesis, associated with a brain metastasis from a papillary thyroid carcinoma. *Am. J. Cancer Res.* 6:2140.

Serrano-Carbajal, E. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2020). Targeting metabolic deregulation landscapes in breast cancer subtypes. *Front. Oncol.* 10:97. doi: 10.3389/fonc.2020.00097

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Tarazona, S., García, F., Ferrer, A., Dopazo, J., and Conesa, A. (2011). NOISeq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet J.* 17, 18–19. doi: 10.14806/ej.17.B.265

The Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499:43. doi: 10.1038/nature12222

The Cancer Genome Atlas Research Network (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* 374, 135–145. doi: 10.1056/NEJMoa1505917

Wang, C., Wu, C., Yang, Q., Ding, M., Zhong, J., Zhang, C.-Y., et al. (2016). MIR-28-5p acts as a tumor suppressor in renal cell carcinoma for multiple antitumor effects by targeting RAP1B. *Oncotarget* 7:73888. doi: 10.18632/oncotarget.12516

Woldrich, J. M., Mallin, K., Ritchey, J., Carroll, P. R., and Kane, C. J. (2008). Sex differences in renal cell cancer presentation and survival: an analysis of the national cancer database, 1993-2004. *J. Urol.* 179, 1709–1713. doi: 10.1016/j.juro.2008.01.024

Zaitsu, M., Toyokawa, S., Takeuchi, T., Kobayashi, Y., and Kawachi, I. (2020). Sex-specific analysis of renal cell carcinoma histology and survival in Japan: a population-based study 2004 to 2016. *Health Sci. Rep.* 3:e142. doi: 10.1002/hsr2.142

Zhang, Z., Lin, E., Zhuang, H., Xie, L., Feng, X., Liu, J., et al. (2020). Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma. *Cancer Cell Int.* 20, 1–18. doi: 10.1186/s12935-020-1113-6

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 December 2020
doi: 10.3389/fgene.2020.595912

Inference of Genetic Networks From Time-Series and Static Gene Expression Data: Combining a Random-Forest-Based Inference Method With Feature Selection Methods

Shuhei Kimura[1]*, *Ryo Fukutomi*[2], *Masato Tokuhisa*[1] and *Mariko Okada*[3]

[1] Faculty of Engineering, Tottori University, Tottori, Japan, [2] Graduate School of Sustainability Science, Tottori University, Tottori, Japan, [3] Laboratory of Cell Systems, Institute of Protein Research, Osaka University, Osaka, Japan

OPEN ACCESS

Edited by:
Kimberly Glass, Brigham and Women's Hospital and Harvard Medical School, United States

Reviewed by:
Frank Emmert-Streib, Tampere University, Finland
Jesús Espinal-Enríquez, Instituto Nacional de Medicina Genómica (INMEGEN), Mexico

***Correspondence:**
Shuhei Kimura
kimura@tottori-u.ac.jp

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

Received: *17 August 2020*
Accepted: *23 November 2020*
Published: *15 December 2020*

Citation:
Kimura S, Fukutomi R, Tokuhisa M and Okada M (2020) Inference of Genetic Networks From Time-Series and Static Gene Expression Data: Combining a Random-Forest-Based Inference Method With Feature Selection Methods. Front. Genet. 11:595912. doi: 10.3389/fgene.2020.595912

Several researchers have focused on random-forest-based inference methods because of their excellent performance. Some of these inference methods also have a useful ability to analyze both time-series and static gene expression data. However, they are only of use in ranking all of the candidate regulations by assigning them confidence values. None have been capable of detecting the regulations that actually affect a gene of interest. In this study, we propose a method to remove unpromising candidate regulations by combining the random-forest-based inference method with a series of feature selection methods. In addition to detecting unpromising regulations, our proposed method uses outputs from the feature selection methods to adjust the confidence values of all of the candidate regulations that have been computed by the random-forest-based inference method. Numerical experiments showed that the combined application with the feature selection methods improved the performance of the random-forest-based inference method on 99 of the 100 trials performed on the artificial problems. However, the improvement tends to be small, since our combined method succeeded in removing only 19% of the candidate regulations at most. The combined application with the feature selection methods moreover makes the computational cost higher. While a bigger improvement at a lower computational cost would be ideal, we see no impediments to our investigation, given that our aim is to extract as much useful information as possible from a limited amount of gene expression data.

Keywords: FANTOM5, gene expression, feature selection, random forest, genetic network inference

1. INTRODUCTION

The dynamic behavior of gene expression determines a variety of cell functions. Our understanding of biological systems requires the study of complex patterns of gene regulation, as the regulation among genes determines how genes are expressed. One promising approach developed for the analysis of gene regulation is the inference of genetic networks. In a genetic network inference problem, mutual regulations among genes are inferred from gene expression data measured by

biological technologies, such as DNA microarrays, RNA-seq using next generation sequencers, and so on. The inferred network models can serve as ideal tools to help biologists generate hypotheses and facilitate the design of their experiments. Many researchers have thus taken an interest in the inference of genetic networks.

A number of genetic network inference methods have been proposed (Larrañaga et al., 2006; Meyer et al., 2008; Chou and Voit, 2009; Hecker et al., 2009; de Matos Simoes and Emmert-Streib, 2012; Emmert-Streib et al., 2012; Glass et al., 2013). Among them, random-forest-based methods show promise for their excellent performance (Huynh-Thu et al., 2010; Maduranga et al., 2013; Petralia et al., 2015; Huynh-Thu and Geurts, 2018; Kimura et al., 2019). Some of these inference methods also have a useful ability to analyze both time-series and static gene expression data (Petralia et al., 2015; Huynh-Thu and Geurts, 2018; Kimura et al., 2019). The time-series data are a series of sets of gene expression levels measured at successive time points after a stimulation. The static data are sets of gene expression levels measured under steady-state conditions. The random-forest-based inference methods analyze gene expression data by assigning confidence values to all of the candidate regulations. While many genetic network inference methods try to find regulations that are actually contained in the target network, the random-forest-based methods only rank the candidates by assigning every candidate a confidence value. When biologists try to perform experiments for confirming the inferred regulations of genes, the confidence values computed by the random-forest-based methods could be used to determine the order of the experiments. The random-forest-based inference methods would become much more useful, however, if they had the ability to detect genes that actually regulated a gene of interest.

By combining the random-forest-based inference method with some feature selection method, we have been able to detect regulations that are actually contained in the target genetic network. Feature selection, a procedure studied in the computational intelligence field, removes input variables irrelevant to the output in an approximation task or a classification task (Guyon and Elisseeff, 2003; Cai et al., 2018). We found, however, in preliminary experiments, that a combined method integrating the random-forest-based method with one of the existing feature selection methods often fails to detect genes that weakly affect a gene of interest. The main purpose of the existing feature selection methods might explain this failure, as the methods were developed not to detect all of the input variables that actually affect the output, but to find input variables that maximize the predicting performance of the obtained model. More recently, our group developed a new feature selection method whose purpose is to find all of the input variables that actually affect the output and to remove as many of the irrelevant input variables as possible (Kimura and Tokuhisa, 2020).

In this manuscript, we propose a method to remove unpromising candidate regulations by combining the random-forest-based inference method with the new feature selection method we developed in Kimura and Tokuhisa (2020), along with two modified versions of the same. The feature selection methods used in this study are effective in not only removing several irrelevant input variables, but also in assigning confidence values to the input variables to show the likelihood that they actually affects the output. In our combined method, we can therefore use the confidence values computed by the feature selection methods to adjust the confidence values assigned to all of the candidate regulations by the random-forest-based method.

The remainder of this manuscript is organized as follows. In the section 2, we introduce the random-forest-based inference method used in this study. In the section 3, we describe the feature selection methods, and then explain a way to combine them with the inference method. We confirm the effectiveness of the proposed combined method through numerical experiments using artificial and biological gene expression data in the sections 4 and 5, respectively. Finally, in the section 6, we conclude with our future work.

2. RANDOM-FOREST-BASED INFERENCE METHOD

As mentioned previously, this study combines the random-forest-based inference method with a series of feature selection methods. While any random-forest-based method can serve this purpose, in this study we apply an inference method (Kimura et al., 2019) that is capable of analyzing both time-series and static gene expression data. This section briefly describes the inference method.

2.1. Model for Describing Genetic Networks

The inference method applied in this study describes a genetic network using a set of differential equations of the form

$$\frac{dX_n}{dt} = F_n(\mathbf{X}_{-n}) - \beta_n X_n, \qquad (n = 1, 2, \cdots, N), \qquad (1)$$

where $\mathbf{X}_{-n} = (X_1, \cdots, X_{n-1}, X_{n+1}, \cdots, X_N)$, X_m $(m = 1, 2, \cdots, N)$ is the expression level of the m-th gene, N is the number of genes contained in the target network, β_n (> 0) is a constant parameter, and F_n is a function of arbitrary form.

When using this model, we infer a genetic network by obtaining a function F_n and a parameter β_n $(n = 1, 2, \cdots, N)$ that produce time-courses consistent with the observed gene expression levels. The following section presents a way to obtain them.

2.2. Obtaining F_n and β_n

The inference method (Kimura et al., 2019) divides an inference problem of a genetic network consisting of N genes into N subproblems, each of which corresponds to each gene. By solving the n-th subproblem, the method obtains a reasonable approximation of the function F_n and a reasonable value for the parameter β_n. The remainder of this section will describe the n-th subproblem.

2.2.1. Problem Definition

The inference method used in this study obtains an approximation of the function F_n and a value for the

parameter β_n through the optimization of the following one-dimensional function.

$$S_n(\beta_n) = \sum_{k=1}^{K_T} \frac{w_k^T}{\beta_n} \left[\left. \frac{dX_n}{dt} \right|_{t_k} - \hat{F}_n \left(\mathbf{X}_{-n}|_{t_k} ; \beta_n \right) + \beta_n \, X_n|_{t_k} \right]^2 + \sum_{k=1}^{K_S} \frac{w_k^S}{\beta_n} \left[\left. \frac{dX_n}{dt} \right|_{s_k} - \hat{F}_n \left(\mathbf{X}_{-n}|_{s_k} ; \beta_n \right) + \beta_n \, X_n|_{s_k} \right]^2, \quad (2)$$

where $\mathbf{X}_{-n}|_{t_k} = (X_1|_{t_k}, \cdots, X_{n-1}|_{t_k}, X_{n+1}|_{t_k}, \cdots, X_N|_{t_k})$, $\mathbf{X}_{-n}|_{s_k} = (X_1|_{s_k}, \cdots, X_{n-1}|_{s_k}, X_{n+1}|_{s_k}, \cdots, X_N|_{s_k})$, and $X_m|_{t_k}$ and $X_m|_{s_k}$ $(m = 1, 2, \cdots, N)$ are the expression levels of the m-th gene at the k-th measurement in time-series and steady-state experiments, respectively. K_T (≥ 2) and K_S (≥ 0) are the numbers of measurements performed in the time-series and steady-state experiments, respectively. Note that the expression levels $X_m|_{t_k}$ and $X_m|_{s_k}$ are measured using biochemical techniques in the genetic network inference. $\left. \frac{dX_n}{dt} \right|_{t_k}$ and $\left. \frac{dX_n}{dt} \right|_{s_k}$ are the time derivatives of the expression level of the n-th gene at the k-th measurement in the time-series and steady-state experiments, respectively. The time derivatives of the expression level of the n-th gene in the time-series experiments, i.e., $\left. \frac{dX_n}{dt} \right|_{t_k}$'s, are directly estimated from the measured time-series of the gene expression levels using a smoothing technique, such as a spline interpolation (Press et al., 1995), a local linear regression (Cleveland, 1979), a modified Whittaker's smoother (Vilela et al., 2007), or the like. On the other hand, the time derivatives of the expression level of the n-th gene in the steady-state experiments, i.e., $\left. \frac{dX_n}{dt} \right|_{s_k}$'s, are all set to zero. w_k^T and w_k^S are weight parameters for the k-th measurements in the time-series and steady-state experiments, respectively. Kimura et al. (2019) showed that the performance of the random-forest-based inference method improves by discounting the weight values of the measurements that were obtained under states similar to each other. $\hat{F}_n(\,\cdot\,; \beta_n)$ is an approximation of the function F_n trained under the given β_n. The inference method (Kimura et al., 2019) obtains an approximation of the function F_n using a random forest (Breiman, 2001). The section 2.2.2 below will describe a way to obtain $\hat{F}_n$ using a random forest. The inference method described here uses the golden section search (Press et al., 1995) to minimize the objective function (2).

2.2.2. Approximation of F_n

The computation of the objective function (2) requires an approximation of the function F_n, i.e., $\hat{F}_n$. As described previously, an approximation of the function F_n is obtained using a random forest. In the inference method (Kimura et al., 2019), the random forest that approximates the function F_n is trained based on training data consisting of the following set of input-output pairs,

$$\left\{ \left(\mathbf{X}_{-n}|_{t_k}, \left. \frac{dX_n}{dt} \right|_{t_k} + \beta_n \, X_n|_{t_k} \right) \middle| \, k = 1, 2, \cdots, K_T \right\} \cup \left\{ \left(\mathbf{X}_{-n}|_{s_k}, \left. \frac{dX_n}{dt} \right|_{s_k} + \beta_n \, X_n|_{s_k} \right) \middle| \, k = 1, 2, \cdots, K_S \right\}.$$

Note that a value for the parameter β_n is always given when computing a value for the objective function (2). Therefore, we can train the random forest using the training data described above. Note also that, in order to keep consistency with the objective function (2), the random forest used in the method (Kimura et al., 2019) tries to obtain an approximation of the function F_n that minimizes a weighted sum of the squared errors between the given output values and the values computed from the model.

2.3. Assigning Confidence Values to the Regulations

As is done in other random-forest-based inference methods, the inference method described in this section uses a variable importance measure defined in tree-based machine learning techniques, such as a random forest, to evaluate the confidence values of all of the candidate regulations. Note again, however, that the random forest used in the inference method tries to minimize the weighted sum of the squared errors. The confidence value of the regulation of the n-th gene from the m-th gene, $C_{n,m}$, is thus computed by

$$C_{n,m} = \frac{1}{Sq_{w0}} \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} \sum_{v \in V_i(m)} I(v), \quad (3)$$

where

$$Sq_{w0} = \sum_{k=1}^{K_T} w_k^T (y_{t_k} - \bar{y}_{w0})^2 + \sum_{k=1}^{K_S} w_k^S (y_{s_k} - \bar{y}_{w0})^2, \quad (4)$$

$$\bar{y}_{w0} = \frac{1}{N_{w0}} \left[\sum_{k=1}^{K_T} w_k^T y_{t_k} + \sum_{k=1}^{K_S} w_k^S y_{s_k} \right], \quad (5)$$

$$N_{w0} = \sum_{k=1}^{K_T} w_k^T + \sum_{k=1}^{K_S} w_k^S, \quad (6)$$

$$y_{t_k} = \left. \frac{dX_n}{dt} \right|_{t_k} + \beta_n^* \, X_n|_{t_k}, \quad (7)$$

$$y_{s_k} = \left. \frac{dX_n}{dt} \right|_{s_k} + \beta_n^* \, X_n|_{s_k}, \quad (8)$$

$$I(v) = N_w(v) Sq_w(v) - N_w(v_L) Sq_w(v_L) - N_w(v_R) Sq_w(v_R), \quad (9)$$

$$Sq_w(v) = \sum_{k \in T(v)} w_k^T \left[y_{t_k} - \bar{y}_w(v) \right]^2 + \sum_{k \in S(v)} w_k^S \left[y_{s_k} - \bar{y}_w(v) \right]^2, \quad (10)$$

$$\bar{y}_w(\nu) = \frac{1}{N_w(\nu)}\left[\sum_{k\in T(\nu)} w_k^T y_{t_k} + \sum_{k\in S(\nu)} w_k^S y_{s_k}\right], \quad (11)$$

$$N_w(\nu) = \sum_{k\in T(\nu)} w_k^T + \sum_{k\in S(\nu)} w_k^S, \quad (12)$$

N_{tree} is the number of trees in the random forest $\hat{F}_n^*$, and $V_i(m)$ is a set of nodes that use the expression levels of the m-th gene to split the training examples in the i-th decision tree of $\hat{F}_n^*$. ν_L and ν_R are the left and right children nodes of the node ν, respectively, and $T(\nu)$ and $S(\nu)$ are sets of indices of the training examples generated from time-series and static gene expression data, respectively, and allocated to the node ν. $\hat{F}_n^*$ and β_n^* are the approximation of the function F_n and the value for the parameter β_n, respectively, obtained through the optimization of the function (2).

3. COMBINING A RANDOM-FOREST-BASED INFERENCE METHOD WITH FEATURE SELECTION METHODS

As mentioned previously, any existing feature selection method can be combined with a random-forest-based inference method. We found however that the combination of the random-forest-based method and an existing feature selection method often degrades the quality of the inferred genetic network. This degradation might be explained by the purpose for which the existing feature selection methods were designed, namely, to select input variables that maximize the predicting performance of the approximated function. More recently, however, Kimura and Tokuhisa (2020) proposed a new feature selection method that seeks to find all of the input variables that actually affect the output. In this study, we combine the random-forest-based inference method described in the previous section with this new feature selection method (Kimura and Tokuhisa, 2020), along with two modified versions of the same.

In this section, we first describe the new feature selection method (Kimura and Tokuhisa, 2020) as originally proposed and several modified forms, and then propose a way to combine them with the random-forest-based inference method.

3.1. Feature Selection Methods Based on Variable Importance Measure

The feature selection method (Kimura and Tokuhisa, 2020), we apply uses a variable importance measure to check whether or not each input variable actually affects the output. If a certain input variable is relevant to the output, its importance score is likely to be larger than that of a random variable. The feature selection methods described here are designed based on this idea.

Assume that a set of K input-output pairs $\{(\mathbf{x}_k, y_k)|k = 1, 2, \cdots, K\}$ is given, where $\mathbf{x}_k = (x_{1,k}, x_{2,k}, \cdots, x_{N,k})$, $x_{i,k}$ is the value for the i-th input variable at the k-th observation, and y_k is the output value at the k-th observation. Then, the feature selection method (Kimura and Tokuhisa, 2020) tries to find all of the input variables relevant to the output according to the following procedure.

1. Construct a new training dataset $\{(\mathbf{z}_k, y_k)\big| k = 1, 2, \cdots, K\}$ based on the given dataset $\{(\mathbf{x}_k, y_k)\big| k = 1, 2, \cdots, K\}$, where $\mathbf{z}_k = (x_{1,k}, x_{2,k}, \cdots, x_{N,k}, x_{1,k}^{pmt}, x_{2,k}^{pmt}, \cdots, x_{N,k}^{pmt})$, and $x_{i,k}^{pmt}$ is the value for the i-th permuted input variable at the k-th observation. The values for the i-th permuted input variable in this algorithm, $x_{i,k}^{pmt}$'s, are obtained by randomly permuting those for the i-th original input variable, $x_{i,k}$'s.
2. Train a random forest using the training dataset constructed in the step 1.
3. In order to statistically check whether or not input variables are relevant to the output, construct N_{RF} different random forests by repeating the steps 1 and 2.
4. When a value for $\frac{C_i - N_{RF}/2}{\sqrt{N_{RF}/4}}$ exceeds the α_s-quantile of the standard normal distribution, conclude that the i-th input variable actually affects the output, where C_i is the number of random forests in which the importance score of the i-th original input variable is greater than that of the i-th permuted input variable. Note here that this study also uses a probability defined by $1 - \frac{C_i}{N_{RF}}$ as a confidence value that the i-th input value actually affects the output.

To give the original and permuted input variables even chances of being selected for the splitting of the training examples, the feature selection method uses a slightly modified training algorithm for the random forest. See Kimura and Tokuhisa (2020) for more detailed information about the modification.

While the feature selection method described above is capable of detecting input variables that weakly affect the output, each irrelevant input variable is erroneously concluded to be relevant with a probability of about 0.5. In this study, we overcome the poor specificity of the feature selection method by constructing two other feature selection methods based on the same design concept (Kimura and Tokuhisa, 2020) and then combining all three of the methods together. To be specific, the two newly constructed feature selection methods respectively use Extra-Trees (Geurts et al., 2006) and VR-Trees (Liu et al., 2008), instead of the random forest, in the algorithm described above. Extra-Trees and VR-Trees are variants of the random forest. The method used to combine the three feature selection methods is described in the next section. While the original feature selection method uses a modified training algorithm for the random forest, note that the two newly constructed methods use the training algorithms for Extra-Trees and VR-Trees without any modification. In this paper, we refer to these three methods as the feature selection methods using the random forest, Extra-Trees, and VR-Trees, respectively.

3.2. Algorithm of the Combined Method

As mentioned earlier, we combined the random-forest-based inference method (Kimura et al., 2019) with the feature selection methods described in the previous section. In addition to removing unpromising regulations, the combined method improves the confidence values of all of the candidate regulations.

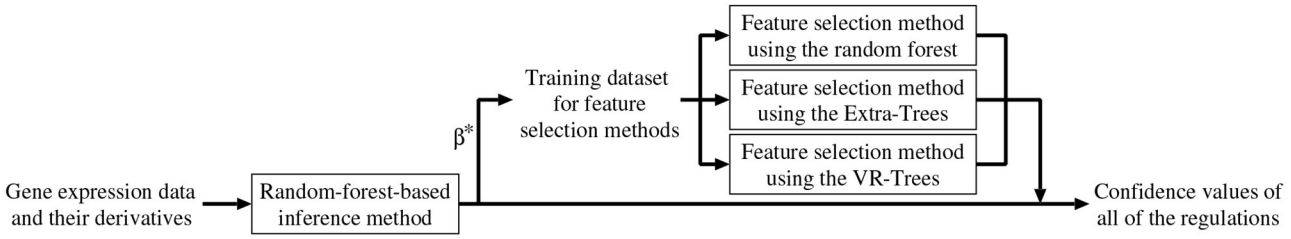

FIGURE 1 | A framework of the proposed method.

Below, we explain the algorithm of the combined method (see also **Figure 1**).

1. Set a counter n to 1.
2. Perform the random-forest-based inference method (Kimura et al., 2019) for the n-th subproblem, then obtain an approximation of the function F_n and a value for the parameter β_n. Here, we represent them as $\hat{F}_n^*$ and β_n^*, respectively.
3. By applying $\hat{F}_n^*$ and β_n^* to the Equation (3), compute the confidence value of the regulation of the n-th gene from the m-th gene, $C_{n,m}$ $(m = 1, 2, \cdots, N, m \neq n)$.
4. Construct a training dataset of input-output pairs,

$$\left\{ \left(\mathbf{X}_{-n}|_{t_k}, \left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n^* X_n|_{t_k} \right) \middle| k = 1, 2, \cdots, K_T \right\} \cup \left\{ \left(\mathbf{X}_{-n}|_{s_k}, \left.\frac{dX_n}{dt}\right|_{s_k} + \beta_n^* X_n|_{s_k} \right) \middle| k = 1, 2, \cdots, K_S \right\},$$

   and then apply the feature selection methods using the random forest, Extra-Trees, and VR-Trees to the constructed dataset. Note that the random-forest-based inference method used in this study trains models that consider the weight values, w_k^T's and w_k^S's, assigned to the given gene expression data. Therefore, our feature selection methods also consider these weight values when training the random forests, Extra-Trees, and VR-Trees used in these methods.
5. If one or more of the three feature selection methods conclude that the m-th gene does not regulate the n-th gene, set $C_{n,m}$ to zero. In this study, a confidence value $C_{n,m}$ equal to zero indicates that the proposed method infers no regulation of the n-th gene from the m-th gene. Otherwise, adjust the confidence value $C_{n,m}$ according to

$$C_{n,m} \leftarrow pC_{n,m} + (1-p) \min \left\{ D_{n,m}^{RF}, D_{n,m}^{ET}, D_{n,m}^{VT} \right\},$$

   where p $(0 \leq p \leq 1)$ is a mixing parameter. The mixing parameter represents the degree to which our combined method relies on the confidence values computed by the random-forest-based inference method. $D_{n,m}^{RF}$, $D_{n,m}^{ET}$, and $D_{n,m}^{VT}$ are the confidence values of the regulation of the n-th gene from the m-th gene, obtained from the feature selection methods using the random forest, Extra-Trees and VR-Trees, respectively. As mentioned in the section 3.1, the feature selection methods used in this study often falsely conclude an irrelevant input variable to be relevant. In this step, therefore, we adopt the worst estimate among the estimates obtained from the three feature selection methods in order to reduce the number of irrelevant regulations falsely concluded to be relevant.
6. $n \leftarrow n + 1$. If $n \leq N$, return to the step 2.
7. Output all of the confidence values, i.e., $C_{n,m}$'s $(m, n = 1, 2, \cdots, N, m \neq n)$.

4. EXPERIMENTS WITH ARTIFICIAL GENE EXPRESSION DATA

This section describes experiments conducted with artificial genetic network inference problems to evaluate the performance of the proposed method.

4.1. Analysis Using DREAM3 Data

To investigate the effect of the mixing parameter p on the performance of the proposed method, we first performed the experiment with a series of DREAM3 problems.

4.1.1. Experimental Setup

The proposed method was applied to five artificial genetic network problems obtained from the DREAM3 *in silico* network challenges (http://dreamchallenges.org/): Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3. The target networks of these problems consisted of 100 genes each ($N = 100$) and were designed based on actual biochemical networks.

Each problem used here contained both time-series and static expression data of all 100 genes. The time-series data were 46 datasets consisting of time-series of gene expression levels obtained by solving a set of differential equations on the target network, and were polluted by internal and external noise (Schaffter et al., 2011). The time-series datasets began from randomly generated initial values, and each gene in each set was assigned 21 observations, with time intervals of 10 between two adjacent observations. The static data consisted of wild-type, knockout and knockdown data. The wild-type data contained the steady-state gene expression levels of the unperturbed network. The knockout and knockdown data contained the steady-state

expression levels of every single-gene knockout and every single-gene knockdown, respectively. When trying to solve the n-th subproblem corresponding to the n-th gene, however, we removed the static data of the knockout and the knockdown of the n-th gene. The number of measurements in the time-series experiment, K_T, was therefore $46 \times 21 = 966$, while that of the steady-state experiment, K_S, was $1 + 100 + 100 - 2 = 199$. Noisy time-series data were provided as the observed data, so we smoothed them using a local linear regression (Cleveland, 1979), a data smoothing technique. The same local linear regression was used to estimate the time derivatives of the gene expression levels. The genetic network of 100 genes was inferred solely from the smoothed time-series of the gene expression levels, their estimated time derivatives, and the static gene expression data.

The number of trees in the random forest (N_{tree}), the number of input variables to be considered in each internal node of each tree (N_{test}), and the maximum height of each tree (N_{hmax}) were set to 1,000, $\lceil \frac{N-1}{3} \rceil$, and 32, respectively, according to the recommended parameter values for the random-forest-based inference method (Kimura et al., 2019). Because the parameter to be estimated, β_n, was positive, we searched for an optimum value in a logarithmic space. The search area of $\log \beta_n$ was $[-10, 5]$. The inference method used in the proposed method must give values for the weight parameters for the gene expression data, i.e., w_k^T's and w_k^S's. The weight parameters for the measurements in each of the 46 time-series datasets were set at the values used by Kimura et al. (2019), namely, 0.6674 for the 10th measurement, 0.3348 for the 11th measurement, and 0.002174 for the last 10 measurements. The weight parameters for the other measurements in the time-series datasets and for the measurements in the static dataset were set to 1.0 and 1.1, respectively.

The number of random forests constructed (N_{RF}), the number of trees in each random forest, and the significance level of the statistical test (α_s) were set to 100, 100, and 0.01, respectively, for the feature selection method using the random forest, as well as for the feature selection methods using Extra-Trees and VR-Trees. Again, the recommended values were used for the other parameters for the feature selection methods: the numbers of input variables to be considered in each internal node of each tree in the random forest and in Extra-Trees were set to $\lceil \frac{N-1}{3} \rceil \times 2$ and $(N-1) \times 2$, respectively, and α, the parameter that controls the probability that the deterministic test-selection will be selected over the random test-selection in VR-Trees, was set to 0.5.

Another parameter, namely, the mixing parameter p, must also be assigned a value in the proposed method. In this study, we investigated how the parameter p affected the performance of our method by running a series of experiments with different mixing parameter values. As the proposed method is a stochastic algorithm, we applied the method with each of the parameter settings to each of the five problems ten times.

4.1.2. Results

We tested the performance of the proposed method using the area under the recall-precision curve (AURPC), a performance measure that increases from 0 to 1 as the numbers of false-positive and false-negative regulations decrease. The recall-precision curve of an algorithm was obtained by checking the recalls and precisions. The recall and the precision are defined as

$$\text{recall} = \frac{TP}{TP + FN}, \quad \text{precision} = \frac{TP}{TP + FP},$$

where TP, FP, and FN are the numbers of true-positive, false-positive, and false-negative regulations, respectively. The recall and precision were computed by constructing a network of regulations whose confidence values exceeded a threshold, and then comparing it with the target network. Note that the proposed method assigns confidence values to all of the candidate regulations. Next, the recall-precision curve of the algorithm was obtained by changing the threshold for the confidence value. Auto-regulations/auto-degradations were disregarded in the evaluation of the performance.

Table 1 lists the AURPCs of the proposed method with different mixing parameter values in the five problems. The table also shows the performance of the random-forest-based inference method (Kimura et al., 2019), a method equivalent to that proposed here without the feature selection. As described in the section 3.2, the proposed method removes unpromising candidate regulations and then adjusts the confidence values of the remaining the candidates. When the mixing parameter p is set to 1.0, however, our method omits this adjustment of the confidence values. The experimental results thus show that the removal of the unpromising candidate regulations improves the performance of the inference method only to a slight degree. Note that our method removed 268.4, 235.8, 208.9, 73.0, and 107.6 candidate regulations, on average, in Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3, respectively. Given that Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3 have $N \times (N-1) = 9,900$ candidate regulations each, and 125, 119, 166, 389, and 551 actual regulations, respectively, we see that the numbers of regulations removed by the proposed method were very small. Hence, the limited improvement in the performance might be explained by the small number of unpromising candidate regulations removed in the five problems solved.

When the mixing parameter p is set to 0.0, on the other hand, the proposed method outputs the confidence values of the regulations computed only on the basis of the values provided by the feature selection methods. The experimental results of our method with $p = 0.0$ indicate that the confidence values computed by the feature selection methods are unreliable. As the table shows, however, we can improve the performance of the proposed method by combining the confidence values computed by the random-forest-based inference method with those computed by the feature selection methods. Our method seems to perform at its best when the parameter p is set to around 0.5. The standard deviations of the AURPCs, on the other hand, widened as the value for parameter p fell from 0.9 to 0.1. As a result, the network inferred by the proposed method with a smaller parameter p was likely to be of a lower quality than that inferred by the method without the feature selection methods. In the remaining experiments in this study, we thus

TABLE 1 | The performance of the proposed method with different values for the mixing parameter p on the DREAM3 problems.

		Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3
		AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
Proposed method	($p = 1.0$)	0.41910 ±0.00390	0.54478 ±0.00586	0.50084 ±0.00287	0.39486 ±0.00344	0.31297 ±0.00224
Proposed method	($p = 0.9$)	0.42143 ±0.00378	0.54539 ±0.00563	0.50594 ±0.00289	0.40047 ±0.00370	0.32093 ±0.00221
Proposed method	($p = 0.8$)	0.42307 ±0.00380	0.54607 ±0.00530	0.50825 ±0.00301	0.40261 ±0.00371	0.32234 ±0.00245
Proposed method	($p = 0.7$)	0.42422 ±0.00395	0.54674 ±0.00523	0.50984 ±0.00314	0.40384 ±0.00377	0.32256 ±0.00263
Proposed method	($p = 0.6$)	0.42493 ±0.00420	0.54736 ±0.00537	0.51055 ±0.00347	0.40446 ±0.00384	0.32223 ±0.00272
Proposed method	($p = 0.5$)	0.42532 ±0.00437	0.54772 ±0.00542	0.51060 ±0.00395	0.40419 ±0.00393	0.32151 ±0.00278
Proposed method	($p = 0.4$)	0.42520 ±0.00465	0.54767 ±0.00579	0.50996 ±0.00446	0.40321 ±0.00393	0.32054 ±0.00291
Proposed method	($p = 0.3$)	0.42410 ±0.00533	0.54689 ±0.00658	0.50856 ±0.00472	0.40179 ±0.00378	0.31975 ±0.00297
Proposed method	($p = 0.2$)	0.42216 ±0.00663	0.54514 ±0.00766	0.50655 ±0.00489	0.40059 ±0.00386	0.31922 ±0.00293
Proposed method	($p = 0.1$)	0.42034 ±0.00757	0.54344 ±0.00813	0.50332 ±0.00507	0.40046 ±0.00390	0.31881 ±0.00265
Proposed method	($p = 0.0$)	0.07094 ±0.00195	0.07486 ±0.00206	0.09892 ±0.00252	0.13139 ±0.00232	0.13949 ±0.00284
Random-forest-based inference method (Kimura et al., 2019)		0.41918 ±0.00388	0.54477 ±0.00586	0.50083 ±0.00285	0.39482 ±0.00344	0.31291 ±0.00223

The performance of the random-forest-based inference method (Kimura et al., 2019) is also shown. AVG and STD represent the averaged AURPC and its standard deviation, respectively.

set the mixing parameter p to 0.9. The networks inferred by the proposed method with $p = 0.9$ were better than those inferred by the method without the feature selection in 49 of the 50 ($= 5 \times 10$) trials performed on the DREAM3 problems.

The proposed method has a much higher computational cost than the random-forest-based inference method (Kimura et al., 2019), as the random forest, Extra-Trees, and VR-Trees must be trained many times. As described earlier, we divided the inference problem of a genetic network consisting of 100 genes into 100 subproblems. While the random-forest-based inference method (Kimura et al., 2019) required an average of 30.3 min to solve a single subproblem, the proposed method required an average of 127.9 min to solve a subproblem on the same workstation (Xeon Gold 6150 2.7GHz). Though inconvenient, we do not see high computational cost of the proposed method as a hindrance to our study, given that our primary aim is to extract as much useful information as possible from a limited amount of gene expression data. Moreover, the computation time required by our method can be easily shortened by performing the calculations in parallel.

4.2. Analysis Using DREAM4 Data

Our next step was to compare the proposed method with the other genetic network inference methods on the DREAM4 problems.

4.2.1. Experimental Setup

For our next experiment, we applied the proposed method to five problems from the DREAM4 *in silico* network challenges. Similar to the DREAM3 problems, the target networks in these problems consisted of 100 genes, and were designed based on actual biochemical networks. These networks were described using a model identical to that of the DREAM3 networks (Schaffter et al., 2011).

Each problem contained both the time-series and static expression data of all 100 genes. The time-series data were 10 datasets of time-series of gene expression levels. Each dataset consisted of the expression levels at 21 time points, and was polluted by internal and external noise. A dataset was constructed by applying a perturbation to the network at the first time point and removing the perturbation at the 11-th time point. The perturbation affected the transcription rates of a different set of several genes in each dataset. The static data consisted of wild-type, knockout, and knockdown data.

To take the perturbations into account explicitly, we added 10 elements to the gene expression data, each corresponding to one of the perturbations. The i-th added element had a value of 1 for the measurements between the 1st and 10th time points in the i-th time-series dataset generated by adding the i-th perturbation, and a value of 0 for the other measurements. The number of

TABLE 2 | The AURPCs of the proposed method with $p = 0.9$ on the DREAM4 problems.

	Network1	Network2	Network3	Network4	Network5
	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
Proposed method ($p = 0.9$)	0.44629	0.31188	0.35118	0.35700	0.28935
	±0.00351	±0.00364	±0.00369	±0.00366	±0.00399
Random-forest-based inference method	0.42797	0.28656	0.33930	0.34079	0.27199
(Kimura et al., 2019)	±0.00312	±0.00300	±0.00397	±0.00347	±0.00415
dynGENIE3	0.34	0.22	0.32	0.34	0.22
(Huynh-Thu and Geurts, 2018)	–	–	–	–	–
MCZ	0.48	0.38	0.38	0.36	0.17
(Greenfield et al., 2010)	–	–	–	–	–
dynGENIE3 + MCZ	0.60	0.43	0.47	0.52	0.37
	–	–	–	–	–
iRafNet	0.552	0.337	0.414	0.421	0.298
(Petralia et al., 2015)	–	–	–	–	–

The table also shows the performances of the random-forest-based inference method (Kimura et al., 2019), dynGENIE3 (Huynh-Thu and Geurts, 2018), MCZ (Greenfield et al., 2010), a combination of dynGENIE3 and MCZ, and iRafNet (Petralia et al., 2015).

elements, N, was therefore $100 + 10 = 110$. When trying to solve the n-th subproblem corresponding to the n-th gene, we also removed the static data of the knockout and the knockdown of the n-th gene. The numbers of measurements of the time-series and steady-state experiments, i.e., K_T and K_S, were thus $10 \times 21 = 210$ and $1 + 100 + 100 - 2 = 199$, respectively. The local linear regression (Cleveland, 1979) was used to smooth the given time-series data and to estimate the time derivatives of the gene expression levels. We inferred a genetic network using only the smoothed time-series of the gene expression levels, their estimated time derivatives, and the static gene expression data.

The 6th, 7th, 8th, 9th, and 10th measurements in each of the time-series datasets were all assigned weight values of 0.2 (Kimura et al., 2019). The 17th, 18th, 19th, 20th, and 21th measurements were all assigned weight values of 0.02. The 4th, 5th, 15th, and 16th measurements were assigned weight values of 0.7333, 0.4667, 0.6733, and 0.3466, respectively. The values for the remaining w_k^T's and for w_k^S's were set to 1.0 and 1.1, respectively. As described in the section 4.1.2, the mixing parameter p was set to 0.9. The other experimental conditions were unchanged from those used in the section 4.1.

4.2.2. Results

We also used the area under the recall-precision curve (AURPC) to quantify the performance of the inference method in this experiment. Although we inferred the regulations of the 100 genes from these genes and the 10 additional elements representing 10 perturbations, we disregarded the regulations of the genes from the additional elements for the evaluation of the performance. Auto-regulations/auto-degradations were also disregarded in the evaluation of the performance. **Table 2** shows the AURPCs of the proposed method on the five problems, along with the AURPCs of the original random-forest-based inference method (Kimura et al., 2019), dynGENIE3 (Huynh-Thu and Geurts, 2018), MCZ (Greenfield et al., 2010), a combination of dynGENIE3 and MCZ, and iRafNet (Petralia et al., 2015). The AURPCs of dynGENIE3, MCZ, and the combination of dynGENIE3 and MCZ are taken from Huynh-Thu and Geurts (2018), while the AURPCs of iRafNet are taken from Petralia et al. (2015).

As the table illustrates, the use of the feature selection methods improved the quality of the inferred network. The improvements brought about by the feature selection methods were larger than the improvements obtained in the experiment performed in the section 4.1. The better performance obtained might have partly stemmed from the larger number of unpromising regulations removed by the proposed method on the DREAM4 problems. Our method removed an average of 2075.6, 1676.1, 1797.8, 1652.8, and 1559.9 regulations from $100 \times 109 = 10900$ candidate regulations in Network1, Network2, Network3, Network4, and Network5, respectively.

The proposed method, however, failed to outperform the other inference methods in some cases, as the table shows. Note however that dynGENIE3 and iRafNet are both designed based on the random forest. As such, we could modify these inference methods to improve the performance by applying the proposed idea. Remember also that, when using MCZ, we must provide static data for every single-gene knockout if we are to obtain a reasonable genetic network. The use of static data for every single-gene knockout might partly explain the excellent performance of the combination of dynGENIE3 and MCZ. The excellent performance of iRafNet seems to stem from a similar cause. Data of this type, however, are difficult to measure, which puts a limit to their practical use.

5. ANALYSIS OF BIOLOGICAL GENE EXPRESSION DATA

In the final experiment of this study, we used the proposed method to analyze actual gene expression data.

TABLE 3 | The measurement conditions of the time-series datasets used in this study.

Cell name	Stimulus	Measured time (min.)
Saos-2 cells	Ascorbic acid and BGP	0, 15, 30, 45, 60, 80, 100, 120, 150, 180, 240
MCF-7 cells	EGF1	0, 15, 30, 45, 60, 80, 100, 120, 150, 180, 210, 240, 300, 360, 420, 480
MCF-7 cells	HRG	0, 15, 30, 45, 60, 80, 100, 120, 150, 180, 210, 240, 300, 360, 420, 480
ARPE-19 cells	TGF-β and TNF-α	0, 15, 30, 45, 60, 80, 100, 120, 150, 180, 210, 240, 300
Lymphatic endothelial cells	VEGF	0, 15, 30, 45, 60, 80, 100, 120, 150, 180, 210, 240, 300, 360, 420, 480
Mesenchymal stem cells	IBMX, DEX and insulin	0, 15, 30, 45, 60, 80, 100, 120, 150, 180
Aortic smooth muscle cells	FGF-2	0, 15, 30, 45, 60, 120, 180, 240, 300, 360
Aortic smooth muscle cells	IL-1B	0, 15, 30, 45, 60, 120, 180, 240, 300, 360

TABLE 4 | The top 20 regulations ranked by the confidence values computed by the proposed method.

Rank	Result from original data	Result from modified data
1	**EGR1 ← FOS**	**EGR1 ← FOS**
2	**EGR2 ← FOS**	*FOS ← HRG*
3	*ATF3 ← TGF-β and TNF-α*	*ATF3 ← TGF-β and TNF-α*
4	**JUNB ← FOSB**	*EGR2 ← HRG*
5	EGR3 ← EGR2	**JUNB ← FOSB**
6	**FOSL1 ← ATF3**	EGR3 ← EGR2
7	**MYC ← FOS**	**EGR3 ← FOS**
8	EGR1 ← EGR2	**FOSL1 ← ATF3**
9	**EGR3 ← FOS**	**EGR2 ← FOS**
10	**FOSB ← JUNB**	EGR1 ← EGR2
11	**JUNB ← EGR2**	**MYC ← FOS**
12	EGR3 ← EGR1	**JUNB ← EGR2**
13	**FOS ← EGR2**	EGR3 ← EGR1
14	ETS2 ← EGR2	**FOSB ← JUNB**
15	**JUN ← FOSB**	**JUN ← VEGF**
16	**EGR2 ← MYC**	ETS2 ← EGR2
17	**JUN ← VEGF**	**JUN ← FOSB**
18	EGR2 ← EGR1	FOSL1 ← FOSB
19	FOSL1 ← FOSB	**FOSB ← EGR2**
20	**FOSB ← EGR2**	**ATF3 ← JUN**

The rankings are obtained from an analysis of original data identical to those of Kimura et al. (2019), and the modified data constructed by considering the decomposition of the chemical compounds used for the stimulation of the cells. The regulations written in boldface and italic fonts have reportedly been confirmed in human and/or other species and are accordingly assumed to be reasonable.

5.1. Experimental Setup

In this experiment, we analyzed the expression data of 11 immediate early genes related to transcription, i.e., ATF3, EGR1, EGR2, EGR3, ETS2, FOS, FOSB, FOSL1, JUN, JUNB, and MYC. The time-series and static gene expression levels were obtained from FANTOM5 data (http://fantom.gsc.riken.jp/5/) (FANTOM Consortium et al., 2014). The time-series datasets consisted of sets of expression levels of the genes measured in Saos-2, MCF-7, ARPE-19, lymphatic endothelial, mesenchymal stem, and aortic smooth muscle cells at successive time points after several kinds of external stimuli were applied. **Table 3** presents detailed information on the time-series datasets used in this study. Two types of static data were used for the experiment. The first were sets of gene expression levels for the Saos-2 and mesenchymal stem cells introduced as untreated controls. The second were the measurements taken at time 0 in the respective time-series datasets. The numbers of measurements contained in the time-series and static data in this experiment, K_T and K_S, were $11 + 16 + 16 + 13 + 16 + 10 + 10 + 10 = 102$ and $2 + 8 = 10$, respectively. Eight elements corresponding to the stimuli applied to the cells were added to the gene expression data, in order to take the external stimuli explicitly into account: "ascorbic acid and BGP," "EGF1," "HRG," "TGF-β and TNF-α," "VEGF," "IBMX, DEX and insulin," "FGF-2," and "IL-1B." An added element had a value of 1 for the measurements in the time-series dataset obtained by applying the stimulus corresponding to the element, and a value of 0 for the other measurements. The total number of elements, N, was therefore $11+8 = 19$. By applying the proposed method to the gene expression data described here, we inferred regulations of the 11 selected genes from both the 11 genes and the 8 additional elements. These gene expression data were also analyzed in Kimura et al. (2019).

The following weight values for the expression data were determined according to Kimura et al. (2019). The weight values corresponding to the 11th, 12th, 13th, 14th 15th, and 16th measurements in the time-series dataset of the lymphatic endothelial cells were set to 0.75, 0.5, 0.25, 0.25, 0.25, and 0.25, respectively. The weight values for the 8th, 9th, 10th, and 11th measurements in the time-series dataset of the Saos-2 cells, and for the 7th, 8th, 9th, and 10th measurements in the two time-series datasets of the aortic smooth muscle cells, were set to 0.8333, 0.6667, 0.5, and 0.5, respectively. The weight values for the two measurements in the steady-state experiments with the Saos-2, MCF-7, mesenchymal stem, and aortic smooth muscle cells were set to 0.55. The weight values for the other measurements in the time-series and static datasets were set to 1.0 and 1.1, respectively. The other experimental settings were identical to those used in the previous experiment.

5.2. Results

Table 4 lists the top 20 regulations with respect to the confidence values computed by the proposed method. The correct structure of the target network, however, is still unknown. We thus compared the inferred regulations with those obtained from the STRING database (https://string-db.org/) (Szklarczyk et al., 2014) of protein-protein interactions. The comparison results suggest that 13 of the 20 regulations (boldface font in the table) are reasonable, as the interactions between the proteins

corresponding to the genes have been confirmed in human and/or other species. Moreover, the regulation of ATF3 from the external stimulus "TGF-β and TNF-α" (italic font in the table) seems to be reasonable because TGF-β has been confirmed to induce ATF3 (Yin et al., 2010).

The proposed method, on the other hand, concluded that 28 candidate regulations were unpromising, and set their confidence values to zero. While the regulations of EGR1 from the external stimuli "FGF-2" and "IL-1B" were among the 28 removed regulations, the protein-protein network obtained from the STRING database suggested that these two regulations should not be removed. As described in the section 5.1, this study represented the existence and absence of an external stimulus as 1 and 0, respectively. This simple representation might help to explain the erroneous conclusion that the two regulations just mentioned were unpromising.

Our next step, therefore, was to obtain a more reasonable genetic network by making the representation of the external stimuli more realistic. To do so, we first had to consider the decomposition of the chemical compounds used for stimulating the cells. When preparing the gene expression data, we set the value for each of the 8 added elements corresponding to the external stimuli to $0.9^{\frac{t}{48}}$, instead of 1, where t was the time (min.) elapsed after the stimulation of the cells. We then applied the proposed method to the modified gene expression data. **Table 4** also shows the top 20 regulations ranked by the confidence values obtained in the additional experiment with the modified data. To check the effect of the modification of the data, we compared the inferred regulations with those contained in the protein-protein network obtained from the STRING database. The comparison indicated that 12 of the 20 regulations were reasonable (boldface font in the table), as the interactions between the corresponding proteins were reportedly confirmed. We could also conclude, for the reason mentioned previously, that the regulation of ATF3 from the external stimulus "TGF-β and TNF-α" was reasonable. The regulations of FOS and EGR2 from the external stimulus "HRG" (italic font in the table) appeared to be reasonable as well, given the suggestion from Yuan et al. (2008) and Martine-Moreno et al. (2017) that these regulations existed. In the top 20 regulations inferred in the additional experiment, the number of reasonable regulations was larger, and the ranks of the unreasonable regulations seemed to be slightly lower. The regulations of EGR1 from the external stimuli "FGF-2" and "IL-1B," meanwhile, were erroneously removed in the experiment with the original gene expression data, as mentioned earlier. These two regulations remained in the inferred regulations in this additional experiment, although the number of removed regulations decreased to 18.

As mentioned earlier, the improvement in performance brought about by combining the random-forest-based inference method with the feature selection methods is often small. In the experiments in this section, therefore, the top 20 regulations obtained by the proposed method were completely identical to those of the original random-forest-based method (Kimura et al., 2019). Moreover, the numbers of regulations removed by the proposed method were also modest. By comparing the removed regulations with those now known, however, we can check the validity of the inferred network. This feature of the proposed method could be useful, when we try to analyze actual gene expression data.

6. CONCLUSION

Several random-forest-based inference methods have been proposed. While these methods show promise, they are only of use in ranking all of the candidate regulations by assigning them confidence values. They are of no use in removing unnecessary regulations. In this study, we propose a new method to remove unpromising candidate regulations by combining the random-forest-based inference method (Kimura et al., 2019) with the original feature selection method (Kimura and Tokuhisa, 2020) and two modifications of that method. By using the outputs from the feature selection methods, the proposed method also adjusts the confidence values of the candidate regulations. Numerical experiments performed with artificial gene expression data showed that the combination of the inference method with the feature selection methods slightly improved the quality of the inferred genetic network. Though its computational cost is high, we believe that the proposed method is useful for our chief purpose of extracting as much useful information as possible from a limited amount of gene expression data. Through experiments with actual data, we showed that the removal of unpromising regulations is a useful feature for confirming the validity of an inferred genetic network. The number of regulations removed by the proposed method, however, was often very small. In future work, we plan to search for strategies to detect larger numbers of unpromising regulations.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SK developed the method and performed the experiments. RF and MT designed some parts of the proposed algorithm. MO supervised the biological aspect of this work. All authors read and approved the manuscript.

FUNDING

This work was partially supported by JSPS KAKENHI Grant Number 18H04031.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.595912/full#supplementary-material

REFERENCES

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cai, J., Kuo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: a new perspective. *Neurocomputing* 300, 70–79. doi: 10.1016/j.neucom.2017.11.077

Chou, I. C., and Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math. Biosci.* 219, 57–83. doi: 10.1016/j.mbs.2009.03.002

Cleveland, W. S. (1979). Robust locally weight regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 79, 829–836. doi: 10.1080/01621459.1979.10481038

de Matos Simoes, R., and Emmert-Streib, F. (2012). Bagging statistical network inference from large-scale gene expression data. *PLoS ONE* 7:e33624. doi: 10.1371/journal.pone.0033624

Emmert-Streib, F., Glazko, G. V., Altay, G., and de Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.* 3:8. doi: 10.3389/fgene.2012.00008

FANTOM Consortium, RIKEN PMI, and CLST (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013). Passing messages between biological networks to refine predicted interactions. *PLoS ONE* 8:e64832. doi: 10.1371/journal.pone.0064832

Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. (2010). DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS ONE* 5:e13397. doi: 10.1371/journal.pone.0013397

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models – a review. *BioSystems* 96, 86–103. doi: 10.1016/j.biosystems.2008.12.004

Huynh-Thu, V.A., and Geurts, P. (2018). dynGENIE3: Dynamical GENIE3 for the inference of gene networks from time series expression data. *Sci. Rep.* 8:3384. doi: 10.1038/s41598-018-21715-0

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 5:e12776. doi: 10.1371/journal.pone.0012776

Kimura, S., and Tokuhisa, M. (2020). "Detection of weak relevant variables using random forests," in *Proceedings of SICE Annual Conference 2020*, 838–845.

Kimura, S., Tokuhisa, M., and Okada, M. (2019). Inference of genetic networks using random forests: assigning different weights for gene expression data. *J. Bioinform. Comput. Biol.* 17:1950015. doi: 10.1142/S021972001950015X

Larrañaga, R., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Brief. Bioinform.* 7, 86–112. doi: 10.1093/bib/bbk007

Liu, F. T., Ting, K. M., Yu, Y., and Zhou, Z.-H. (2008). Spectrum of variable-random trees. *J. Artif. Intell. Res.* 32, 355–384. doi: 10.1613/jair.2470

Maduranga, D. A. K., Zheng, J., Mundra, P. A., and Rajapakse, J. C. (2013). Inferring gene regulatory networks from time-series expression using random forests ensemble. *Pattern Recogn. Bioinform.* 13–22. doi: 10.1007/978-3-642-39159-0_2

Martine-Moreno, M., O'Shea, T. M., Zepecki, J. P., Olaru, A., Ness, J. K., Langer, R., et al. (2017). Regulation of peripheral myelination through transcriptional buffering of Egr2 by an antisense long non-coding RNA. *Cell Rep.* 20, 1950–1963. doi: 10.1016/j.celrep.2017.07.068

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: a R/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinform.* 9:461. doi: 10.1186/1471-2105-9-461

Petralia, F., Wang, P., Yang, J., and Tu, Z. (2015). Integrative random forest for gene regulatory network inference. *Bioinformatics* 31, i197–i205. doi: 10.1093/bioinformatics/btv268

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1995). *Numerical Recipes in C, 2nd Edn.* Cambridge: Cambridge University Press.

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: *in silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics* 27, 2263–2270. doi: 10.1093/bioinformatics/btr373

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Vilela, M., Borges, C. C. H., Vinga, S., Vanconcelos, A. T. R., Santos, H., Voit, E. O., et al. (2007). Automated smoother for the numerical decoupling of dynamics models. *BMC Bioinform.* 8:305. doi: 10.1186/1471-2105-8-305

Yin, X., Wolford, C. C., Chang, Y.-S., McConoughey, S. J., Ramsey, S.A., Aderem, A., et al. (2010). ATF3, an adaptive-response gene, enhances TGFβ signaling and cancer-initiating cell features in breast cancer cells. *J. Cell Sci.* 123, 3558–3565. doi: 10.1242/jcs.064915

Yuan, G., Qian, L., Song, L., Shi, M., Li, D., Yu, M., et al. (2008). Heregulin-β promotes matrix metalloproteinase-7 expression via HER2-mediated AP-1 activation in MCF-7 cells. *Mol. Cell Biochem.* 318, 73–79. doi: 10.1007/s11010-008-9858-6

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 January 2021
doi: 10.3389/fgene.2020.603264

Generating Ensembles of Gene Regulatory Networks to Assess Robustness of Disease Modules

James T. Lim[1], Chen Chen[2], Adam D. Grant[3] and Megha Padi[1,3]**

[1] Department of Molecular and Cellular Biology, The University of Arizona, Tucson, AZ, United States, [2] Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, The University of Arizona, Tucson, AZ, United States, [3] University of Arizona Cancer Center, The University of Arizona, Tucson, AZ, United States

The use of biological networks such as protein–protein interaction and transcriptional regulatory networks is becoming an integral part of genomics research. However, these networks are not static, and during phenotypic transitions like disease onset, they can acquire new "communities" (or highly interacting groups) of genes that carry out cellular processes. Disease communities can be detected by maximizing a modularity-based score, but since biological systems and network inference algorithms are inherently noisy, it remains a challenge to determine whether these changes represent real cellular responses or whether they appeared by random chance. Here, we introduce Constrained Random Alteration of Network Edges (CRANE), a method for randomizing networks with fixed node strengths. CRANE can be used to generate a null distribution of gene regulatory networks that can in turn be used to rank the most significant changes in candidate disease communities. Compared to other approaches, such as consensus clustering or commonly used generative models, CRANE emulates biologically realistic networks and recovers simulated disease modules with higher accuracy. When applied to breast and ovarian cancer networks, CRANE improves the identification of cancer-relevant GO terms while reducing the signal from non-specific housekeeping processes.

Keywords: network, community significance, community robustness, network community, community detection, regulatory network, community structure, cancer

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer, University of Oslo, Norway

Reviewed by:
Xiaoning Qian, Texas A&M University, United States
Ozan Ozisik, Aix Marseille University, Inserm, MMG, France

****Correspondence:***
James T. Lim
jtlim@email.arizona.edu
Megha Padi
mpadi@arizona.edu

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

Received: *06 September 2020*
Accepted: *23 December 2020*
Published: *14 January 2021*

Citation:
Lim JT, Chen C, Grant AD and Padi M (2021) Generating Ensembles of Gene Regulatory Networks to Assess Robustness of Disease Modules. Front. Genet. 11:603264. doi: 10.3389/fgene.2020.603264

INTRODUCTION

Finding the underlying molecular mechanisms that drive complex disease remains a difficult problem. Complex diseases appear to be caused by many perturbations scattered around the gene regulatory network, which creates a considerable amount of variability in disease susceptibility (Schadt et al., 2009; Califano et al., 2012; Pickrell, 2014). Network analysis has therefore become a popular approach to model molecular interactions in the cell and prioritize candidate disease genes (Greene et al., 2015; Marbach et al., 2016; Santolini and Barabasi, 2018). Many of these methods capitalize on the idea that biological networks are composed of "communities," or modules, of genes that work in concert to carry out cellular functions and cause a disease (Hartwell et al., 1999; Menche et al., 2015; Platig et al., 2016). A module in a biological network typically refers to a set of genes that is densely interconnected in the network, function together, or are co-regulated (Girvan and Newman, 2002; Segal et al., 2003; Ghiassian et al., 2015). Identifying the changes in

network structure associated with disease onset can reveal more mechanistic insights than standard approaches like differential expression analysis; this approach is often called "differential network biology" (Ideker and Krogan, 2012). A wide variety of tools have been developed to identify the changes in network edges and network structure that accompany disease (Gill et al., 2010; Tesson et al., 2010; Gambardella et al., 2013; Van Landeghem et al., 2016).

However, evaluating the robustness and significance of changes in network structure remains a challenge. Gene regulatory networks are often inferred from transcriptomic data using imperfect inference tools, with no easy way of assessing their underlying variance (Lancichinetti et al., 2011; Menche et al., 2015; Choobdar et al., 2019; Palowitch, 2019). Moreover, community detection algorithms can lead to multiple solutions corresponding to local optima of the fitness function (Newman, 2006; Blondel et al., 2008; Campigotto et al., 2014). Two types of approaches are often used to judge the quality of network communities: consensus clustering and statistical significance (Lancichinetti et al., 2011; Lancichinetti and Fortunato, 2012; Menche et al., 2015; Zitnik and Leskovec, 2018; Palowitch, 2019). The consensus approach combines multiple solutions from the optimization algorithm to find the most likely assignment of genes to communities (Lancichinetti and Fortunato, 2012; Choobdar et al., 2019). Alternatively, the statistical significance of individual communities can be estimated by comparing them with a null distribution derived from randomized networks with the same degree characteristics as the original network (Ideker et al., 2002; Emmert-Streib, 2007; Lancichinetti et al., 2011; Mall et al., 2017; Kojaku and Masuda, 2018; Newman, 2018). Network randomization is typically carried out using generative models.

In the present study, we set out to rank the most robust disease-driven changes in the community structure of gene regulatory networks. We first inferred weighted bipartite networks by integrating transcription factor (TF) binding motifs and gene expression data, and then optimized a modularity-based score to identify candidate modules more active in disease conditions than in matched controls (Padi and Quackenbush, 2018). Other approaches for differential network analysis could be used, including DiffCoEx, DINA, DNA, and Diffany (Gill et al., 2010; Tesson et al., 2010; Gambardella et al., 2013; Van Landeghem et al., 2016), but these methods are limited to either identifying individual correlation-based edges or examining pre-defined gene sets and network features, making them less generalizable to multiple types of questions and networks. Modularity optimization methods can help reveal new biological insights across multiple contexts, but they typically result in multiple solutions and cannot provide information about which disease modules are the most robust or significant.

We tried applying existing methods to rank the most significant genes within our candidate disease modules. Consistent with previous observations, consensus clustering led to a loss of resolution and an inability to detect smaller gene sets annotated to more informative biological pathways (Lancichinetti and Fortunato, 2012; Jeub et al., 2018). Next, we estimated the significance of the disease modules relative to a null distribution for the control network created using two popular generative models – the configuration model (Gabrielli et al., 2019) and the stochastic block model (SBM) (Aicher et al., 2015). However, these models could not realistically simulate the characteristics of a gene regulatory network. Transcriptional regulation is strongly constrained by the fact that any given TF regulates a limited number of genes, depending on TF binding sites, activators/repressors, and epigenetic state (Roeder, 1996; Ptashne and Gann, 1997; Lee and Young, 2000; Teif and Rippe, 2009; Gerstein et al., 2012). Both the configuration model and SBM ignore this restriction and assume that each TF node can influence all genes in the network (configuration) or all genes in a community (SBM), which leads to improper sampling of edge weight variance.

Therefore, we identified a need for a new, computationally efficient generative model that accounts for the known constraints of gene regulation (Proulx et al., 2005; Bansal et al., 2009; Sah et al., 2014; Fosdick et al., 2018). It is challenging to randomize weighted networks while imposing multiple constraints, because each modification propagates to the rest of the network, leading to extreme edge weights if they are not properly controlled. There is no accepted method for generating ensembles of weighted bipartite networks with fixed node strengths (the total weight of edges adjoining each node) (Fosdick et al., 2018). Here we present a new algorithm for network randomization called Constrained Random Alteration of Network Edges (CRANE). CRANE can produce ensembles of unipartite or bipartite weighted networks with fixed node strengths that resemble gene regulatory networks. These ensembles can be used as null distributions to evaluate the importance of genes and regulators in candidate disease modules. To demonstrate the utility of CRANE, we apply it to simulated disease modules, as well as transcriptional networks derived from angiogenic ovarian tumors and hormone receptor-positive breast cancers. In simulations, CRANE performs better than all comparable approaches in finding the "true" disease module. When applied to breast and ovarian cancer networks, several methods are able to improve identification of cancer-related processes in specific cases, but CRANE is the only one that consistently reveals biological insights across multiple networks and conditions while also reducing background noise from non-specific housekeeping processes. Our study demonstrates that CRANE can evaluate candidate disease modules to identify a subset of genes that is robustly associated to the disease.

MATERIALS AND METHODS

General Workflow

To rank significant nodes in disease modules, we use the following general procedure (**Figure 1A**): we first construct disease and matched control networks from gene expression data (e.g., RNA-seq) using a *network inference algorithm*. Next, we identify disease-specific network features (e.g., disease modules) using *network analysis methods*. Our main goal is to evaluate the significance of these disease-specific features. To do this, we compare their associated scores in the disease network to a null distribution created from the control network using a *network*

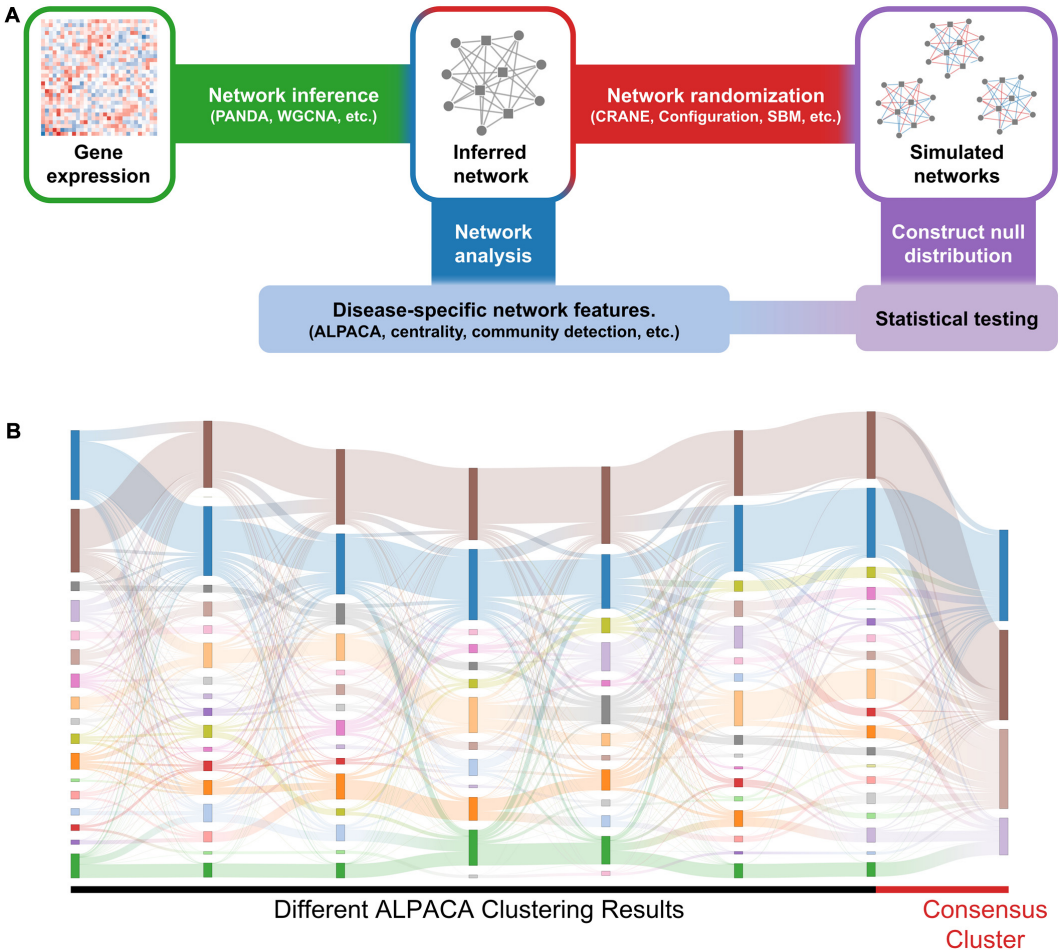

FIGURE 1 | General workflow and results of consensus clustering. **(A)** We identify significant changes in network structure by comparing disease-specific network features against a null distribution. We first construct disease-specific and matched control networks from gene expression data (e.g., RNA-seq) using a network inference algorithm. Next, we compare the networks to extract disease-specific network features. Independently, we apply network randomization to the control network to create a null distribution. The disease-specific network features are then compared against the null distribution to evaluate statistical significance. **(B)** The Sankey plot shows community assignments from seven separate runs of ALPACA on angiogenic vs. non-angiogenic ovarian cancer networks. Each column represents an ALPACA solution with the far-right column showing the result from consensus clustering of 1,000 different ALPACA solutions. The height of each box and ribbons indicates the size of each module and the number of shared nodes between corresponding modules, respectively.

randomization algorithm. For large data sets ($n > 300$), a "true" null distribution can be generated by subsetting the expression profiles for the matched controls ($n = 50$ for each subsample) and constructing independent "replicate" control networks.

Using CRANE or Other Methods to Evaluate Disease Modules: Basic Procedure

Here, we describe the basic procedure for ranking significant genes in disease modules using network randomization algorithms. We provide callouts to other subsections of Section "Materials and Methods" where more details can be found.

We first create disease and matched control networks, either from gene expression data using a *network inference algorithm* (see "Data preprocessing and network inference"), or by simulating networks with artificial disease modules (see "Simulated networks"). Second, we compare the two networks using ALPACA with default parameters, as implemented in the R package (freely available[1]), to identify candidate disease modules (Padi and Quackenbush, 2018). We choose to use ALPACA, but one can use any network analysis tool that groups nodes together, including standard community detection techniques like modularity maximization and differential analysis tools like DiffCoEx, DINA, DNA, and Diffany (Gill et al., 2010; Tesson et al., 2010; Gambardella et al., 2013; Van Landeghem et al., 2016). ALPACA outputs a vector consisting of a module assignment K^0_{node} and differential modularity score S^0_{node} for each node. S^0_{node} quantifies how much the node contributes to the global change in modularity between the disease and control networks. Next, we use a *network randomization algorithm* – either CRANE with

[1]github.com/PadiLab/ALPACA

α in the range of 0.1 to 0.4 (see section "CRANE Algorithm"), SBM (see section "Stochastic Block Model"), configuration model (see section "Configuration Model"), random edge weight permutation (see section "Random Edge Weight Permutation") or data subsampling – on the control network $\left(N^0\right)$ to obtain a null distribution for S^0_{node}. To do this, we perform the following steps:

(i) Use *network randomization algorithm* to generate a perturbed network (N^{P}) starting from the control network $\left(N^0\right)$.
(ii) Compute a "null" differential modularity matrix D^{P}_{ij} for each N^{P} by comparing N^{P} to the original control network N^0 using an in-built function in the ALPACA R package.
(iii) Score each node according to its contribution to the differential modularity D^{P}_{ij} of the module defined by K^0_{node}, to get its score $S^{\text{P}}_{\text{node}}$ under the null hypothesis that the observed change in network structure is only due to measurement noise.

We repeat (i)–(iii) for 1,000 perturbed networks and use the resulting vector of $S^{\text{P}}_{\text{node}}$ values to fit a null distribution T_{node} for each node. Finally, we compute the *p*-value for each node (i.e., its significance as a member of a true disease module) by assuming T_{node} follows a normal distribution:

$$P - value = 1 - \Phi\left(\frac{S^0_{\text{node}} - mean(T_{\text{node}})}{sd(T_{\text{node}})}\right)$$

where Φ is the normal cumulative distribution function, and *mean* and *sd* represent the mean and standard deviation of the node score distribution. To evaluate the results of each network randomization algorithm, we ranked all the genes in the network by their *p*-value and either statistically compared this ranking against the true disease genes (in the case of simulated networks; see section "Simulated Networks"), or evaluated the top-ranked genes within each module for functional enrichment (in the case of real cancer data). In the latter case, in order to make our conclusions threshold-independent, we evaluated the top 25, 50, 75, etc., up to 500 core genes (for example, in one typical module, this would correspond to genes with adjusted *p*-values less than 10^{-23}, 10^{-21}, 10^{-19}, etc. up to 0.01) from each module to identify enriched GO terms at different cutoffs. All significant GO terms ($P_{adj} < 0.05$) across all cutoffs were included in the final result for each module (see "Module-Specific Functional Enrichment Analysis" section below for more details). GO term enrichment was calculated using the R package GOstats (v2.54.0), with the following parameters: the gene universe is the set of all possible target genes in the initial networks and the *p*-value calculation is conditioned on the GO hierarchy structure. In each module, the GOstats *p*-values were adjusted for multiple testing using the Benjamini–Hochberg method. We note that all genes can be ranked together by their *p*-values (with adjusted *p*-value < 0.05 as significant) to combine signals from all modules across the whole network, or top-ranked genes from each module can be kept separate and interpreted as smaller sets of tightly interacting genes.

In addition to this basic procedure, we motivated the development of CRANE by performing consensus clustering on multiple stochastic runs of ALPACA, which uses the Louvain method for community detection (see section "Consensus Clustering" for more details). We also quantified the similarity between networks created by CRANE by computing the normalized mutual information (see section "Computing NMI" for more details).

Module-Specific Functional Enrichment Analysis

The following steps were taken to evaluate and compare the performance of each network randomization method at uncovering module-specific disease-relevant GO terms in cancer networks:

(i) Take all genes assigned to one module and rank them by their network randomization score (e.g., CRANE *p*-value).
(ii-a) Extract the top 25 genes (e.g., genes with CRANE-derived adjusted *p*-value $< 10^{-23}$) and compute the adjusted *p*-value for overlap with a disease relevant GO term (e.g., "blood vessel development" for angiogenic ovarian cancer) using a hypergeometric test.
(ii-b) Repeat (ii-a) with top 50 genes (e.g., genes with CRANE-derived adjusted *p*-value $< 10^{-21}$).
(iii) Repeat (ii) iteratively until top 500 genes (e.g., genes with CRANE-derived adjusted *p*-value < 0.01).
(iv) All GO term *p*-values across all thresholds (20 *p*-values per GO term) are collected and the average of the corresponding $-1 * log_{10}$ *p*-value is reported.

Data Preprocessing and Network Inference

Batch-corrected and normalized ovarian PanCancer TCGA RNA-seq values were downloaded from cBioPortal (Cancer Genome Atlas Research Network, 2011; Cerami et al., 2012; Gao et al., 2013). Low-expressing genes were removed by keeping only genes with at least 1 count per million in at least half of the total samples using the R package edgeR (v3.26.5) and processed with the voom function within the R package limma (v3.38.3) using TMM normalization. Angiogenic ($n = 124$) and non-angiogenic ($n = 166$) tumors were grouped as described in Glass et al. (2015). Preprocessed METABRIC breast cancer expression data was downloaded from cBioPortal (Cerami et al., 2012; Curtis et al., 2012; Gao et al., 2013; Pereira et al., 2016), along with estrogen receptor negative (ER−; $n = 445$) and estrogen receptor positive (ER+; $n = 1449$) status as measured by immunohistochemistry.

Many methods are available to infer gene regulatory networks from transcriptomic data, including ARACNE, CLR, MERLIN, PANDA, and WGCNA, but there is no clear winner across all contexts (Zhang and Horvath, 2005; Margolin et al., 2006; Marbach et al., 2012; Glass et al., 2013; Zhang et al., 2016; Siahpirani and Roy, 2017). For our analyses we chose to

use PANDA (Passing Attributes between Networks for Data Assimilation) and WGCNA.

PANDA

We chose to use PANDA to construct our gene regulatory network because it can integrate known transcription factor (TF) binding sites, and because it does not use TF mRNA level as a proxy for TF activity, instead inferring this latent variable from target gene co-expression, making it particularly appropriate for mammalian contexts where TFs are often regulated by post-translational modification, competitive binding, or localization. Expression data from each subtype was integrated with transcription factor binding sites using the network inference algorithm PANDA with default parameters to create subtype-specific regulatory networks (Glass et al., 2013). Subsampled networks were inferred by selecting random subsets of 50 subjects without replacement from the gene expression of each respective subtype. A prior network of binding sites for 730 TFs was defined as the occurrence of the corresponding motif in a [−750,+250] bp window around the transcription start site (Sonawane et al., 2017). The following formula was applied for analyses requiring exponentially transformed PANDA edge weights (Sonawane et al., 2017), where w_{ij} are the initial z-score edge weights output by PANDA, and W_{ij} are the final transformed edge weights:

$$W_{ij} = \ln\left(e^{w_{ij}} + 1\right)$$

WGCNA

We constructed signed weighted gene co-expression networks using the R package WGCNA (v1.69) (Zhang and Horvath, 2005). Input for the co-expression network consisted of normalized expression values from 1,000 randomly selected genes and random subsets of 50 subjects chosen without replacement from the ER+ METABRIC breast cancer expression data. For all subsampled WGCNA networks, a soft thresholding power of eight was used.

CRANE Algorithm

CRANE takes a weighted network as input and provides a perturbed version of that network as output. In the following, we will describe the procedure for bipartite networks. We first compute the strength of node i as the sum of the edge weights adjoining that node, or $S_i = \sum_j w_{ij}$ (As an optional step to increase network variance further, noise can be added to the original sequence of node strengths by adding normally distributed random numbers with mean 0 and standard deviation estimated from subsampled networks). Given m is the total number of TFs and n is the total number of genes, A_{ij} is the $m \times n$ adjacency matrix of the input network where rows (TFs) and columns (genes) are ordered randomly. We create an empty $m \times n$ adjacency matrix B_{ij} that will become the perturbed network. The first row (first TF) of B_{ij} is initialized with edge weights from the first row of A_{ij}. Then for each TF_l, where $l = [1, \ldots, m-1]$, we apply the following steps: we perturb the current (lth) row B_{lj} by adding normally distributed random numbers with mean 0 and standard deviation computed from the original edge weights for TF_l, i.e., $sd(A_{lj})$. This perturbation is multiplied by a parameter α, giving the user the ability to adjust the magnitude of the perturbation. The B_{lj} edge weights are multiplied by a factor of $\sum_{j=1}^{n} A_{lj} / \sum_{j=1}^{n} B_{lj}$ to ensure the TF strength in B_{lj} is equal to A_{lj}. We compute initial values for $B_{l+1,j}$ (edge weights for the next TF) by computing $\sum_{i=1}^{l+1} A_{ij} - \sum_{i=1}^{l} B_{ij}$, thus keeping the node strengths in B_{ij} equal to the node strengths in A_{ij}. After the initial $B_{l+1,j}$ have been determined, we check if any edge weights within $B_{l+1,j}$ fall outside of the global maximum or minimum of the original edge weights in A_{ij}. For any values in $B_{l+1,j}$ greater than $max(A_{ij})$ edge weight, we add the difference in value between $B_{l+1,j}$and $max(A_{ij})$ to the corresponding B_{lj}. For any values $B_{l+1,j}$ less than $min(A_{ij})$ we subtract the difference in value between $B_{l+1,j}$ and $min(A_{ij})$ to the corresponding B_{lj}. Then the modified edge weights in B_{lj} are normalized to maintain the correct TF strength and a new set of $B_{l+1,j}$ are computed. We repeat the correction process until all values are within the range of A_{ij}.

Note that α is the only user-adjustable parameter in CRANE; in the Results section, we provide a robustness analysis and guidance for choosing an appropriate value of α. A more detailed description of CRANE, including pseudocode, can be found in the **Supplementary Methods**. A unipartite version of CRANE is also available for use. CRANE is freely available as an R package at https://github.com/PadiLab/CRANE.

Configuration Model

The configuration model is a method for generating random networks from a given node degree or strength sequence (Garlaschelli, 2009; Mastrandrea et al., 2014; Gabrielli et al., 2019). For weighted networks, the configuration model is typically constructed as an exponential random graph. To fit the configuration model to the PANDA network, we transformed z-score edge weights to positive weights using the formula given in the "Data Preprocessing and Network Inference" section. Based on the fact that PANDA is a fully connected graph, the configuration model can be written as described in Gabrielli et al. (2019), i.e.,

$$P(A) = \prod_{ij} e^{-(\theta_i + \theta_j) a_{ij}} (\theta_i + \theta_j)$$

where P is the probability of a network with adjacency matrix A and edge weights given by its entries a_{ij}, and the θ parameters are Lagrange multipliers that need to be estimated. The Maximal Likelihood (ML) function then constrains the θ parameters by the given node strength sequence:

$$\sum_{j} \left(\theta_i + \theta_j\right)^{-1} = \sum_{j} a_{ij} = S_i$$

where S_i represents the strength of node i, which is defined as the summation of edge weights adjoining node i. We used Barzilai–Borwein spectral methods for directly solving this ML system of equations using the R package *BB* (v 2019.10.1) (Varadhan and Gilbert, 2009).

Stochastic Block Model

The stochastic block model (SBM) is a random graph generative model. The SBM defines a probability distribution over networks by assuming pre-existing communities or "blocks" where the intracommunity edges are stronger (larger edge weights) than intercommunity edges. This probability distribution can be used to produce random graphs with pre-defined inter- and intra-community edge densities. Fitting a stochastic block model (SBM) to the PANDA network is very time consuming (Aicher et al., 2015). To efficiently test the performance of SBM, we introduced a strong assumption of equivalence between modularity optimization and SBM maximum likelihood (Newman, 2016). Thus, the network community structure found by CONDOR (Complex Network Description of Regulators) (Platig et al., 2016) – a modularity maximization method for weighted bipartite networks – was directly used to generate the block structure in the SBM. We assumed a normal distribution for the edge weights as the PANDA network edge weights represent z-scores. The parameters for every block can then be estimated directly using the sample mean and sample variance of the corresponding edge bundles.

Random Edge Weight Permutation

We wanted to compare CRANE against a naïve method of randomizing a gene regulatory network by permuting its edge weights. Fully permuting the network leads to unrealistic results due to destruction of prior motif information and community structure. To retain as much of the prior biological information as possible, the edges in the network were first divided into motif-positive and motif-negative groups based on whether they were included in the prior network of binding sites for 730 TFs. Next, communities were detected using CONDOR (Platig et al., 2016). Finally, the inter- and intra-community edge weights were grouped together by motif status and randomly shuffled.

Simulated Networks

To simulate disease modules, we first took a random subset of 50 subjects out of 445 subjects from the estrogen receptor negative (ER-) METABRIC breast cancer expression data and constructed a baseline PANDA network. We then inserted high edge weights (edge weight = 5) between randomly selected TFs and genes to create a simulated disease network. The new module consisted of between 3 and 20 TFs, and five times as many genes as TFs. The simulated disease network was compared to a second "replicate" baseline network inferred from an independent random subset of 50 subjects from the ER- breast tumors. We applied a panel of methods – including ALPACA, consensus clustering, CRANE (α = 0.1–0.4), configuration model, SBM, and random edge weight permutation – and evaluated the results of each method by comparing the ranks of true positives (the known genes in the disease module) against a background consisting of genes not in the disease module. Kolmogorov–Smirnov and Wilcoxon rank-sum tests were used to compute the p-value for the difference in the distribution of the ranks. Both tests gave similar results, and so in the figures, we present the Wilcoxon p-values. F-scores were also computed to evaluate the accuracy of each method using the following formula:

$$F = \frac{True\ Positives}{True\ Positives + 0.5(False\ Positives + False\ Negatives)}$$

Positives were defined as the top 1% of ranked nodes.

Consensus Clustering

To generate consensus clusters, we first repeated ALPACA 1,000 times on the same pair of transcriptional networks, as described in Padi and Quackenbush (2018) but with the n nodes ordered randomly in each iteration of the Louvain algorithm. We combined the 1,000 resulting partitions to create an $n \times n$ consensus matrix C with each entry C_{ij} indicating the number of partitions in which nodes i and j of the network were assigned to the same cluster, divided by the total number of partitions (1,000). For the final step, we applied the Louvain algorithm (R package igraph v1.2.4.1) on C to find the consensus cluster membership for each node (Csardi and Nepusz, 2006; Blondel et al., 2008).

Computing NMI

The algorithm CONDOR with default parameters was used to detect the community structure of weighted bipartite networks (Greene et al., 2015; Platig et al., 2016). Using CONDOR community assignments as input, the normalized mutual info (NMI) score between two networks was computed using the "compare" function in the R package igraph (v1.2.4.1) (Danon et al., 2005; Csardi and Nepusz, 2006).

ALPACA and CRANE Pipeline Implementation

To implement the ALPACA and CRANE analysis pipeline presented, first gene regulatory networks for the disease and the control conditions should be inferred from gene expression using PANDA (Glass et al., 2013). The "alpaca.crane" function within the R package CRANE[2] will automatically run ALPACA (also available separately[3]) to compare the two networks and output the module membership and the significance of the nodes.

RESULTS

Existing Methods for Evaluating Significance of Disease Modules

We tried applying the most popular available methods for identifying significant changes in community structure – namely, consensus clustering and comparing against randomized networks – on cancer networks. To apply these methods, we first need to define the networks and candidate disease modules (**Figure 1A**). PANDA was applied as described in the "Materials and Methods" section ("Data preprocessing and network inference") to TCGA data from angiogenic and non-angiogenic ovarian tumors (Cancer Genome Atlas Research Network, 2011)

[2]https://github.com/PadiLab/CRANE

[3]https://github.com/PadiLab/ALPACA

and to METABRIC breast cancer data (Curtis et al., 2012; Pereira et al., 2016) to produce weighted bipartite networks.

We next needed to find a set of candidate disease modules, or groups of genes that interact more with each other in one cancer subtype than expected. To do this, we used ALPACA, a method we previously developed that optimizes a differential modularity score (DMS) to identify groups of nodes exhibiting higher inter-node connectivity in a disease (e.g., angiogenic) network than in a matched control (e.g., non-angiogenic) network (Padi and Quackenbush, 2018). Although we chose to use PANDA and ALPACA (other choices are described in section "Materials and Methods"), we note that the following analyses – including consensus clustering, comparison against randomized networks, and CRANE – can be carried out for any networks (inferred using any method) and any subset of nodes (or disease module) that have stronger interactions in the disease network than in the matched control.

To implement consensus clustering, we merged one thousand partitions from individual ALPACA solutions derived by comparing angiogenic vs. non-angiogenic ovarian tumor data to generate a consensus co-membership matrix (see section "Materials and Methods" for details). We then applied the Louvain method to the consensus matrix to determine consensus community assignment (Blondel et al., 2008). Consistent with previous observations in the literature, we found that consensus clustering led to a significant loss of resolution (**Figure 1B**) (Lancichinetti and Fortunato, 2012; Choobdar et al., 2019) and the inability to detect more specific disease pathways with richer biological interpretations (Jeub et al., 2018).

Next, we used leading network generative models to create a null distribution by randomizing the control network, against which we can compare the disease module scores and estimate their significance. We chose the configuration model (Gabrielli et al., 2019) and the stochastic block model (SBM) (Aicher et al., 2015) as they both have rigorous mathematical descriptions and are two of the most commonly used generative models (Saul and Filkov, 2007; Sah et al., 2014; Baum et al., 2019). The configuration model constrains the expectation value of the node strengths to match the original network, and assumes an exponential distribution for the edge weights (Garlaschelli, 2009; Mastrandrea et al., 2014; Gabrielli et al., 2019). The stochastic block model defines a probability distribution over networks by matching the pre-existing communities or "blocks" of closely connected nodes found in the original network (Aicher et al., 2015). To evaluate the accuracy of these generative models, we chose to analyze METABRIC breast cancer expression data, one of the few diseases in which there are enough expression profiles to subsample eight independent sets of 50 baseline (ER−) expression profiles and generate "biological replicate" PANDA networks. We applied ALPACA to compare ER+ vs. ER− tumors and identify candidate ER+ modules. We next constructed a "true" null distribution of differential modularity scores for each gene in the candidate modules using the eight "biological replicate" networks. We then compared the characteristics of this true null against ensembles of randomized ER- networks produced by SBM and the configuration model (**Figure 2**; see section "Materials and Methods" for details).

We found that both SBM and the configuration model failed to accurately recapitulate the true null distribution computed from the subsampled networks (**Figure 3A**). Both methods appear to overestimate the edge weight variance, probably because they ignore the physical constraints (e.g., TF binding motifs or chromatin accessibility patterns) by which cells specify patterns of gene regulation (Aicher et al., 2015; Gabrielli et al., 2019). To check whether this observation could hold more generally, we repeated the analysis using a different, commonly used network inference method called WGCNA (weighted gene co-expression network analysis) which generates a matrix of gene co-expression values and applies soft-thresholding to impose a scale-free topology criterion (Zhang and Horvath, 2005). This thresholding procedure converts the co-expression to a new value that can be interpreted as a connection weight. We found that the configuration model also fails to fit subsampled WGCNA breast cancer networks (**Supplementary Figure 1**), likely because it puts equal emphasis on all the edges and does not properly conserve the highest-confidence regulatory interactions.

CRANE: New Method for Sampling Weighted Networks

We developed a new algorithm, Constrained Random Alteration of Network Edges (CRANE), that samples weighted networks with fixed node strengths while retaining the underlying gene regulatory structure. Fixing the node strengths preserves the module resolution while creating more realistic variance in edge weights and reducing bias from promiscuous hub TFs and genes that seed modules associated with disease-independent housekeeping processes (see section "Materials and Methods" for details). We found that CRANE is better able to mimic the "true" null distribution of differential modularity scores arising from subsampled PANDA networks than the configuration model and SBM (**Figures 3A,B**). Similarly, CRANE better estimates the edge weight variance in WGCNA networks than the configuration model (**Supplementary Figure 1**). In particular, the mean of the CRANE-generated distribution remains in close proximity to the "true" subsampled null distribution, while other generative models have large deviations across multiple moments of the distribution.

The magnitude of the perturbations in the network created by CRANE is governed by a user-defined parameter α. To choose this parameter appropriately, we compared the properties of networks generated with different α values with the subsampled networks from the previous section. We focused on the distribution of differential modularity scores, the variance in edge weights, and the similarity in community structure (as measured by the normalized mutual information, or NMI) between the original network and the randomized networks. As expected, increasing the parameter α leads to decreasing NMI score (or similarity) (**Figure 3C**) and increasing edge weight variance (**Figure 3D**), but there is no single value of α that exactly mimics the subsampled networks (**Figure 3B**). We decided to use a range of values (α from 0.1 to 0.4) that provide a reasonably good fit to the breast cancer patient data and test the robustness and sensitivity to the exact value below. However, other values of α

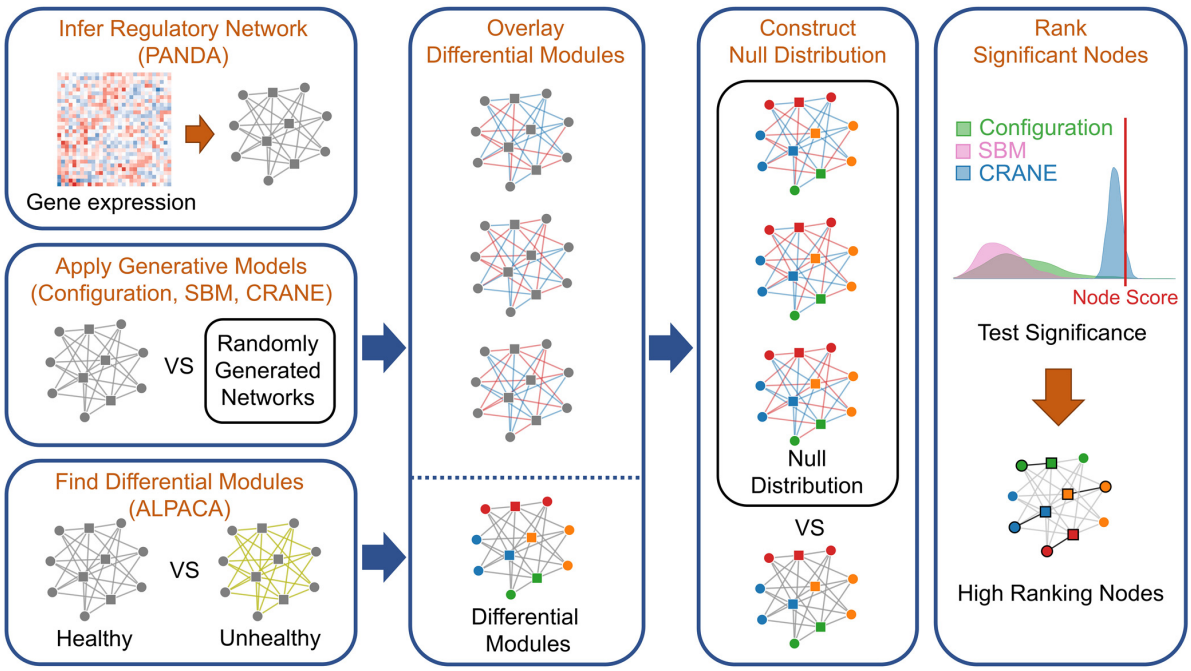

FIGURE 2 | Workflow for applying network generative models to rank genes in disease modules. To integrate ALPACA with generative models (network randomization), we first construct transcriptional networks representing the "control" and "disease" networks by integrating known TF binding sites with gene expression data. We then use ALPACA to identify putative disease modules and compute the differential modularity scores (DMS) for each node. We construct the null distribution of the DMS by comparing the control network to randomized networks generated using the configuration model, SBM, permutation, or CRANE. The *p*-value is calculated by comparing the true node DMS to the null distribution.

may be more appropriate in other contexts, depending on the uncertainty in gene expression data and expression correlations.

Using CRANE to Identify Simulated Disease Modules

We tested whether CRANE could find artificially created disease modules in settings resembling real weighted biological networks. To simulate the effect of measurement noise, we created two independent sets of randomly subsampled (n = 50) gene expression data from the same baseline condition, estrogen receptor negative (ER−) breast cancer (BC), and used them to infer two gene regulatory networks, BCN1 and BCN2. Keeping BCN1 as the baseline network, an artificial disease module was created in BCN2 by increasing the edge weights between randomly selected subsets of transcription factors and genes, ranging from 3 to 20 TFs in size, and five times as many genes. We then applied a large panel of methods – namely, ALPACA, consensus clustering, random edge weight permutation, SBM, CRANE, and subsampling – to find differential modules and rank the nodes according to either their differential modularity score (DMS), or by the *p*-value representing how much their scores deviate from the generated null distribution. A Wilcoxon rank-sum test was used to evaluate how highly each method ranked the genes in the true disease module. In order to include the configuration model in this panel, we also performed a second test after applying an exponential transformation on the network edge weights, since the configuration model requires positive edge weights (Gabrielli et al., 2019).

We found that, although ALPACA by itself can successfully recover artificial modules of size greater than 48 nodes (**Figure 4A**), CRANE was able to dramatically improve performance, as indicated by more significant Wilcoxon *p*-values, showing that the simulated "disease" genes were ranked higher by CRANE than by ALPACA; CRANE also increased F-scores computed using the top 1% ranked nodes in each method (**Supplementary Figure 2**). Consensus clustering improved performance in recovering a single added module but embeds the artificial module within a much larger community, reducing the resolution (**Figure 4B**), whereas CRANE maintained high resolution. By the same metrics, CRANE was more successful than random edge weight permutation, the configuration model, and the SBM. This performance gain in CRANE was preserved across α-parameter values ranging from 0.1 through 0.4, suggesting that, within this range, the exact value of α is not critical (**Supplementary Figure 3**). As expected, the "true" subsampled distribution performed best out of all the methods. We observed a similar trend in performance whether or not the exponential transformation was applied to the network edge weights (**Supplementary Figure 4**).

Applying CRANE to Cancer Data

To determine if CRANE can be used to increase the detection of network alterations in complex diseases, we applied CRANE

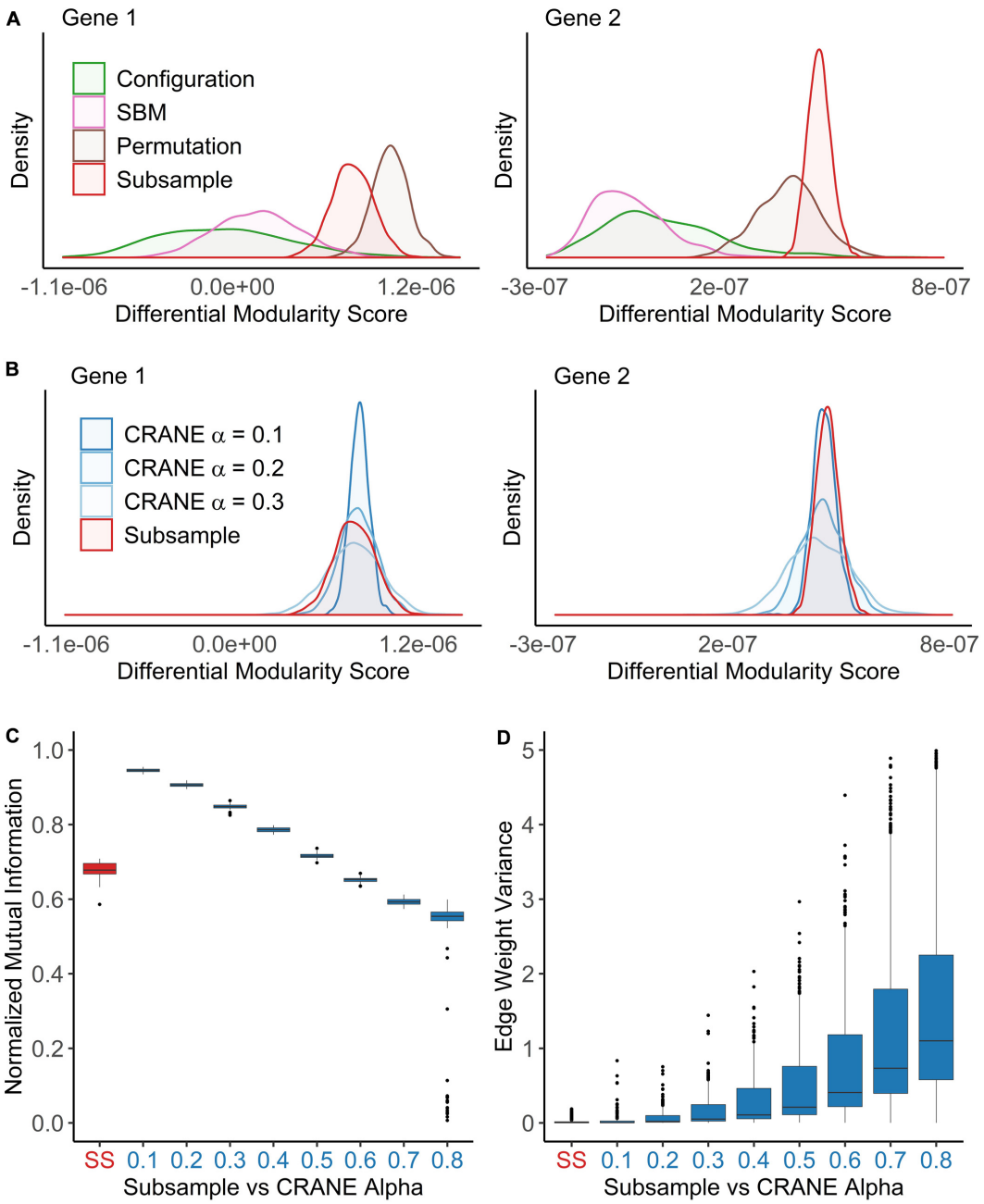

FIGURE 3 | CRANE can generate networks that resemble subsampled data while maintaining control of key network properties. Using the breast cancer transcriptional network as a reference, network ensembles were generated using configuration model, SBM, permutation, and CRANE. As a "true" null distribution, eight PANDA networks were inferred by subsampling ($n = 50$) the gene expression data without replacement from the estrogen receptor negative subtypes. **(A,B)** Density plots showing the null distribution of the differential modularity score (x-axis) computed using different methods for two example genes. **(C,D)** The boxplots show the impact of the CRANE alpha parameter on **(C)** community structure and **(D)** edge weight variance, as compared to subsampled (SS) networks. **(C)** Plot showing the normalized mutual information (y-axis) between the reference network and CRANE-generated networks for different values of alpha (x-axis). **(D)** The edge weight variance (y-axis) among subsampled or CRANE-generated networks at different values of alpha (x-axis).

to real biological data. Since there is no "ground truth" dataset for disease modules in transcriptional networks, there is no straightforward way to count false positives and false negatives and compute the precision and accuracy of our results. Instead, we quantified the extent to which highly ranked genes from CRANE are statistically enriched in biological functions driving two well-understood disease processes – angiogenesis in ovarian cancer and estrogen response in breast cancer – using Fisher's exact test p-values. Using our simulation study as a guide (**Supplementary Figure 3**), we chose $\alpha = 0.1$ as a value that

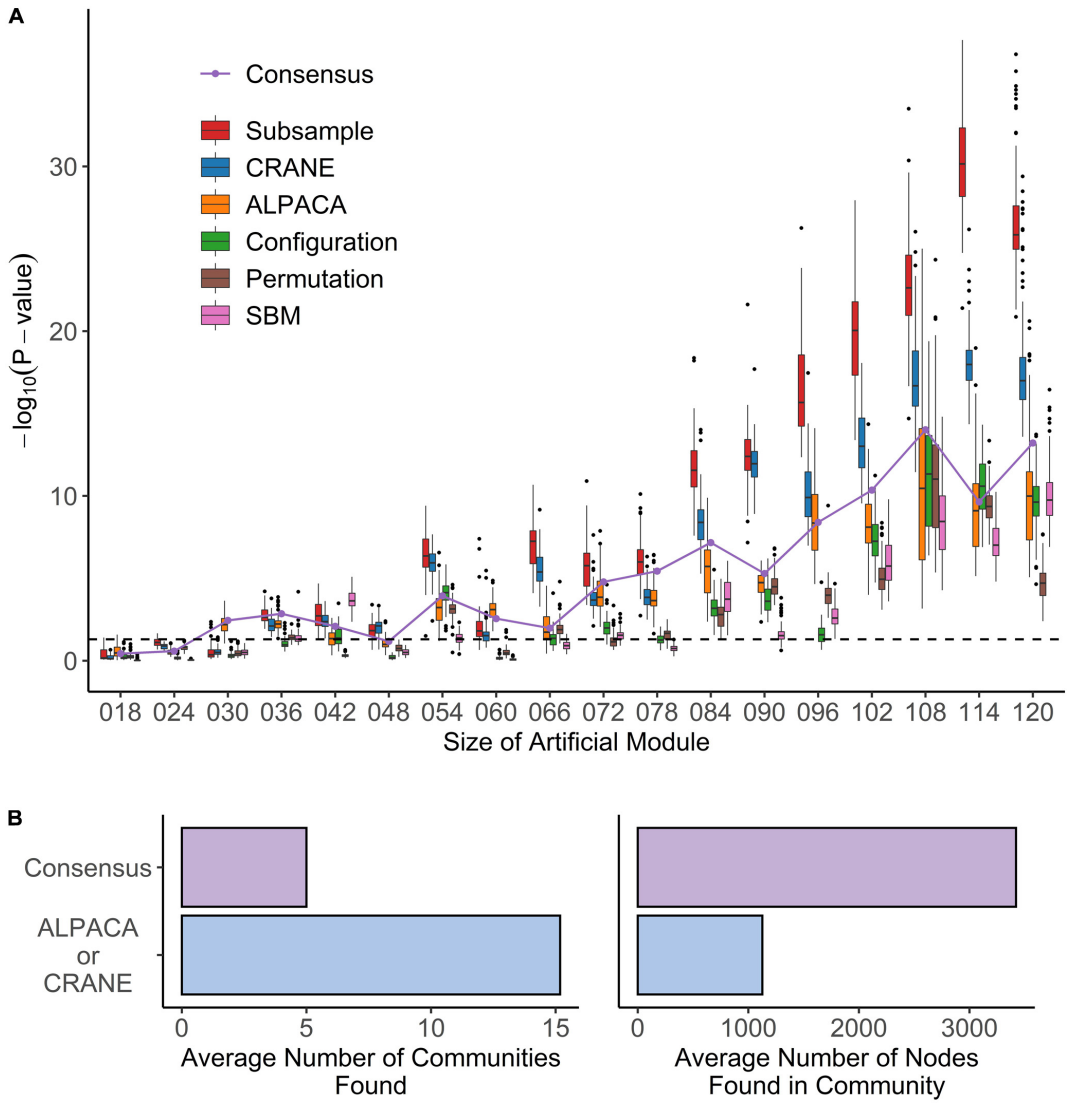

FIGURE 4 | Performance of seven methods on identifying artificial modules in simulated disease networks. **(A)** Box plot shows performance of each method – subsample, CRANE, ALPACA, configuration model, permutation, or SBM – on network simulations with exponentially transformed edge weights. *P*-values (*y*-axis) for various sizes of artificial modules (*x*-axis) are computed using Wilcoxon rank-sum test. **(B)** The boxplots show the total number of modules and module size across all simulation trials for ALPACA versus consensus clustering.

would provide the biggest performance increase with the least amount of computational cost (the run time of CRANE increases with α due to the deviation correction step).

Ovarian Cancer

Ovarian cancer is one of the leading causes of death among women in the developed world (Arend et al., 2013; Bowtell et al., 2015; Reid et al., 2017). Ovarian cancer is divided into many histologic subtypes based on cellular origin, pathogenesis, molecular alterations, and gene expression (Reid et al., 2017). In particular, an angiogenesis gene signature can categorize ovarian cancer patients into a poor-prognosis subtype (Bentink et al., 2012). To test CRANE on angiogenic ovarian cancer, we first applied PANDA to infer ovarian cancer gene regulatory networks from Pan-Cancer TCGA RNA-seq data (Cancer Genome Atlas Research Network, 2011). Normalized RNA-seq profiles were classified into 124 angiogenic and 166 non-angiogenic ovarian cancer tumors as described in Glass et al. (2015). We then applied the same panel of methods as above, ranked the top-scoring genes, and evaluated their functional enrichment for biological processes.

We first checked the performance of consensus clustering compared to ALPACA to see how the reduction in community resolution would impact the biological interpretation. Consistent with our previous work, ALPACA discovers finer community structure enriched for GO terms that are specific to the angiogenic ovarian cancer phenotype such as "blood vessel development" and "cardiovascular system development"

(**Figure 5A**) (Padi and Quackenbush, 2018). In comparison, consensus clustering results in loss of community resolution (**Figure 5B**), which in turn leads to the lack of enrichment in more specific GO terms. Instead, communities are enriched for general processes such as "RNA splicing," "monoubiquitinated protein deubiquitination," "translation initiation," and "ribosome biogenesis" (**Supplementary Table 1**).

We then applied CRANE ($\alpha = 0.1$) and found that it showed good performance and resolution in recovering disease-specific processes (**Figure 6A**). CRANE-ranked genes were statistically enriched for expected GO terms such as "angiogenesis" and "positive regulation of angiogenesis," with *p*-values similar to the "true" subsampled distribution, and exhibited mild improvement (i.e., more significant Fisher's exact test *p*-values) over ALPACA (**Figure 6A**). Although the improvement in GO term detection in the individual modules was modest, we noticed that both CRANE and subsampling increased the ranking of "blood vessel development" related genes, when genes were ranked across the whole network instead of in a module-specific manner (**Supplementary Figure 5**). This is because the blood vessel development genes are split across two modules which allows them to be masked by other enriched processes present in the same modules, such as inflammation pathways (**Supplementary Figure 6**). Interestingly, compared to ALPACA and consensus clustering, CRANE reduces signals from non-specific housekeeping processes, like "RNA transport" and "RNA processing." The permutation and SBM methods performed poorly in uncovering the disease-specific GO terms, as these methods had a tendency to overestimate the DMS distribution while underestimating the variance (**Figure 6C**).

CRANE and subsampling also consistently identified communities that represent inflammation and immune response. Genes in Module 1 deemed most significant by CRANE were enriched for interferon response, interleukins, cytokine signaling, and inflammation, consistent with the theory that chronic inflammation is associated with risk of cancer (Hanahan and Weinberg, 2011) (**Supplementary Table 2**). Specifically, immunomodulators and interferon gamma have been proposed as a therapeutic target in ovarian cancer (Wall et al., 2003; Cohen et al., 2016). The enrichment in inflammation and immune response was not readily detectable using ALPACA, permutation, and SBM (**Figure 6A** and **Supplementary Tables 4–6**). CRANE is therefore able to uncover additional communities enriched with processes relevant to the disease phenotype.

We also tested our methods after exponentially transforming the edge weights and found that neither CRANE nor the "gold standard" subsampling method improve the recovery of angiogenesis related processes compared to ALPACA (**Figure 6B**). The exponentiation process leads to a change in community structure in the PANDA networks (NMI = 0.69) that results in most of the blood vessel development genes being concentrated in a single giant differential module (**Supplementary Figure 7**). The embedment of the angiogenesis genes in a large module along with overall increase in edge weight variance leads to reduction in CRANE performance, whereas other methods have inflated node *p*-values due to a tendency to underestimate the null distribution (**Figure 6D**).

Breast Cancer

Breast cancer is the second most common cancer and a leading cause of death for women worldwide (Bray et al., 2018). Although breast cancer is highly heterogeneous, one of its most important risk factors is overexpression of the estrogen receptor (ER+) leading to increased cell growth (Garcia-Closas et al., 2008; Ahmed et al., 2009; Dunning et al., 2009). Cellular networks in ER+ breast tumors should therefore exhibit increased estrogen signaling.

We used PANDA to infer ER+ (1449 subjects) and ER− (445 subjects) gene regulatory networks from microarray data collected by the METABRIC consortium (Curtis et al., 2012; Pereira et al., 2016). We compared the ER+ network to the ER− network using the same panel of methods as before, and we analyzed the top-ranked genes from each method for enrichment in GO terms. Consensus clustering and ALPACA both failed to detect estrogen-specific pathways (**Figure 7A** and **Supplementary Tables 6, 7**). Similar to the results from ovarian cancer, general biological processes such as RNA localization, mRNA splicing, protein catabolic process, and chromosome organization were highly enriched after consensus clustering (**Supplementary Table 6**).

On the contrary, reranking the nodes using CRANE ($\alpha = 0.1$) effectively uncovered estrogen specific GO terms such as "cellular response to estrogen" and "positive regulation of intracellular estrogen receptor signaling pathway" with more significant *p*-values than ALPACA, consensus, and the permutation method (**Figure 7A** and **Supplementary Table 8**). Similar to the ovarian cancer analysis, CRANE decreased the significance of non-specific housekeeping processes. The "true" subsampled distribution performed even better than CRANE, reinforcing our hypothesis that real disease pathways are robust relative to the underlying noise in regulatory networks.

We also applied the full panel of methods on exponentially transformed breast cancer PANDA networks. The exponential transformation decreased the discovery of estrogen related processes compared to the non-exponentiated network (**Figures 7A,B**). Nevertheless, all methods showed improvements in the significance level of "cellular response to estrogen stimulus" compared to ALPACA and consensus clustering. The configuration model again had a tendency to underestimate the null distribution of differential modularity scores (DMS), leading to a general inflation of GO term significance (**Figure 7D**). The permutation method performed well in discovering estrogen-related GO terms. This is likely because for this specific dataset, edge permutation produces a DMS null distribution close to the subsampled distribution. However, over all the analyses we performed, the configuration model, SBM, and permutation methods generally exhibited larger deviations from the subsampled distribution than CRANE, leading to unreliability in their performance (**Figures 6C,D**, **7C,D**). In summary, we found that different generative models may be useful in specific networks, contexts, and conditions, but only CRANE provides reliable and consistent performance across

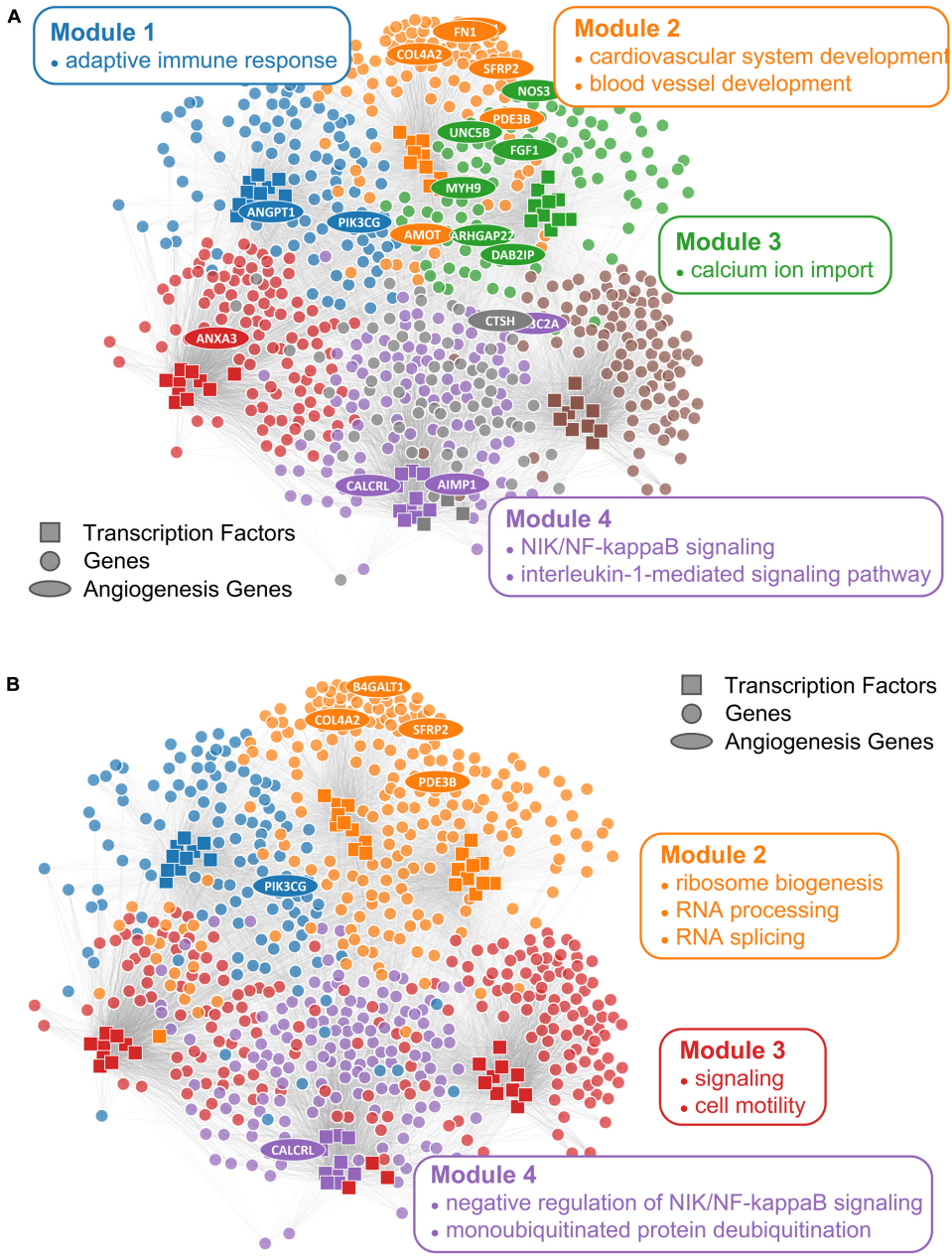

FIGURE 5 | Comparison of ALPACA modules and consensus clustering in angiogenic ovarian tumors. Network with **(A)** ALPACA solution has seven modules while **(B)** the consensus clustering results in four modules. Top 10 core TFs and 100 core genes were extracted from each module based on DMS from ALPACA. The consensus community or ALPACA membership was then overlaid on top by coloring the nodes. The angiogenesis genes (ellipse) were labeled based on whether they were ranked within the top 100 genes in the respective methods. Network is annotated with representative enriched GO terms in each module with $P_{adj} < 0.05$.

multiple settings in identifying genes statistically enriched for disease-related processes rather than housekeeping functions.

DISCUSSION

Phenotypic transitions like disease are often driven by the appearance of new groups of genes, or communities, that carry out relevant cellular processes. However, most methods for detecting these new communities rely on maximizing a modularity-based score, and there is no easy way of determining whether the solutions represent true disease modules or whether they could have appeared in healthy tissue due to measurement noise. Consensus clustering offers an effective way of finding stable communities; however, the loss of community resolution leads to a reduction in interpretability. Comparing disease modules with randomized versions of a matched control network could help identify genes that are significantly associated with

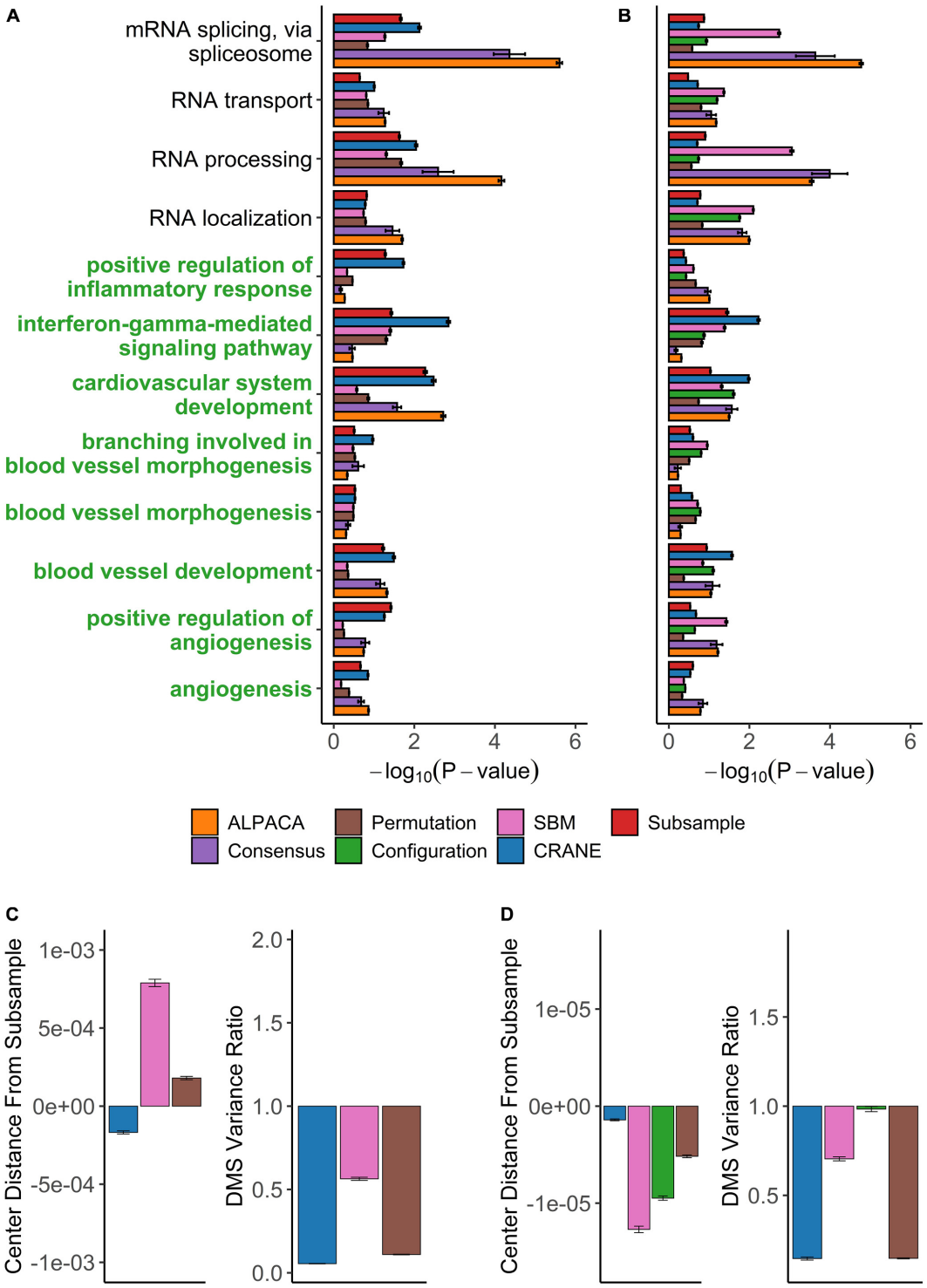

FIGURE 6 | Performance of seven methods on discovering disease-relevant modules associated with angiogenic ovarian tumors. The top five-hundred genes in each module discovered by each method were extracted and subjected to GO term enrichment analysis. **(A,B)** Horizontal bar plots show a curated set of GO terms and their average $-\log_{10}P$-values over 100 different ALPACA runs. The GO terms (*y*-axis) colored in green are disease-relevant terms while black terms represent general biological processes. **(C,D)** The left vertical bar plot shows the average center distance and the right bar plot shows the average ratio between the mean of the null distribution created from subsampled networks and the mean of the null distribution generated from the indicated methods. Negative distance indicates that the specific method underestimates the center of the "true" subsample distribution. For the variance ratio, values less than 1 represent greater variance in the subsample distribution compared to the indicated methods. The GO term enrichment analysis and the distribution analysis were performed on networks with either **(A,C)** PANDA edge weights or **(B,D)** exponentially transformed edge weights. The error bars represent mean ± S.E.M.

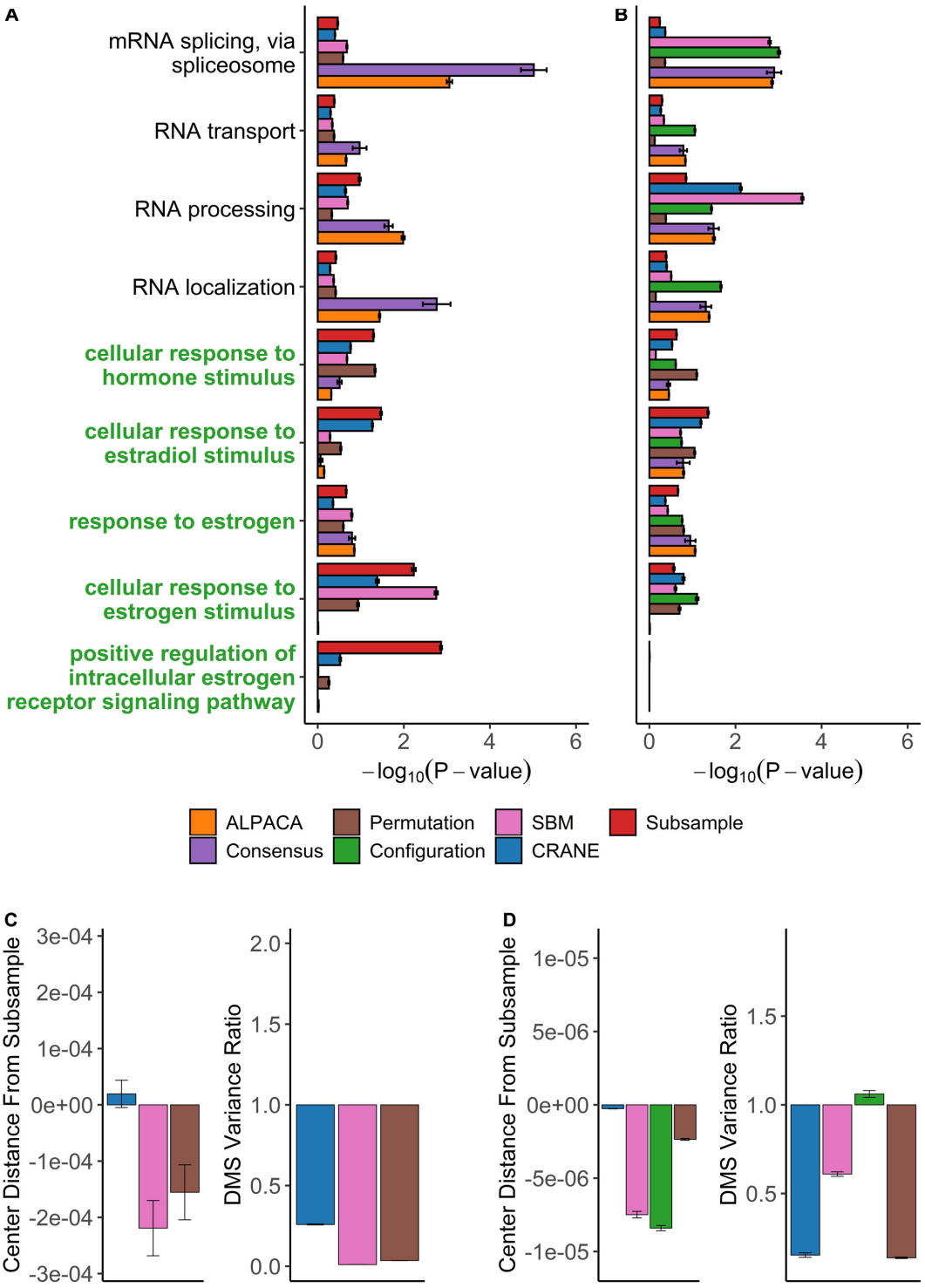

FIGURE 7 | Performance of seven methods on discovering disease-relevant modules associated with ER-positive breast tumors. The top five-hundred genes in each module discovered by each method were extracted and subjected to GO term enrichment analysis. **(A,B)** Horizontal bar plots show a curated set of GO terms and their average $-\log_{10}P$-values over 100 different ALPACA runs. The GO terms (*y*-axis) colored in green are disease-relevant terms while black terms represent general biological processes. **(C,D)** The left vertical bar plot shows the average center distance and the right bar plot shows the average ratio between the mean of the null distribution created from subsampled networks and the mean of the null distribution generated from the indicated methods. Negative distance indicates that the specific method underestimates the center of the "true" subsample distribution. For the variance ratio, values less than 1 represent greater variance in the subsample distribution compared to the indicated methods. The GO term enrichment analysis and the distribution analysis were performed on networks with either **(A,C)** PANDA edge weights or **(B,D)** exponentially transformed edge weights. The error bars represent mean ± S.E.M.

disease. However, available network generative models are unable to randomize gene regulatory networks while properly controlling the sparsity and edge weight variance. Additionally, biological experiments are resource-limited and do not typically generate enough data to empirically estimate network variance for statistical testing.

We therefore devised CRANE, an algorithm for generating more realistic null distributions of gene regulatory networks by maintaining node strengths and the underlying "hard-wired" structure. We compared CRANE against a "true" null distribution created by down-sampling a large breast cancer dataset to make multiple independent replicate networks. The strength parameter α in CRANE can be used to alter the variance in the edge weights, community structure, and modularity score of the randomized networks. However, our analysis showed that there is no single value of α that fully recapitulates the "true" null distribution. This may be because CRANE independently perturbs all edges while subsampled networks retain correlations between network edges. When applied to cancer networks, CRANE was more accurate at reproducing the center of the subsampled distribution but less accurate at reproducing the variance (**Figures 6C,D**, **7C,D**). We hypothesize that better modeling of the variance of the null distribution would further improve the performance of CRANE.

We used simulated networks with artificial disease modules to evaluate the accuracy and statistical significance of the ranking of disease genes by CRANE. CRANE was consistently more successful in identifying the real disease genes than network generative models and edge weight permutation (**Figure 4A**). This is likely due to the stricter constraints in CRANE that ensure the randomized networks mimic the original network structure, while other methods deviate due to their looser constraints (**Figures 3A,B**). We note that the "true" subsampled distribution performed best out of all the methods, suggesting that there is room to further improve CRANE's ability to capture all the properties of gene regulatory networks.

CRANE also achieved more robust discovery of disease specific processes in cancer regulatory networks. Comparing angiogenic to non-angiogenic ovarian tumors, we found that CRANE leads to a mild improvement in detecting differences in expected pathways like blood vessel development and inflammatory processes. Additionally, CRANE was able to minimize noise from housekeeping processes that are present in all living cells. Comparing ER+ to ER− subtypes of breast cancer, we found that running a modularity maximization method like ALPACA or consensus clustering failed to identify expected changes in estrogen signaling. In contrast, ranking genes by their significance using CRANE revealed that estrogen-related modules were robustly activated in ER+ breast cancer.

The superior performance of CRANE in breast cancer relative to ovarian cancer is likely rooted in differences in the performance of ALPACA in the two datasets. In ovarian cancer, the angiogenesis genes had high ALPACA scores and re-ranking them by significance did not make a big difference (**Supplementary Figures 6, 8**); in breast cancer, the estrogen genes had lower ALPACA scores to begin with, providing CRANE with more room for improvement. We also found that CRANE performs poorly after exponentiating edge weights, because the exponential transformation leads to a reduction in ALPACA resolution (**Supplementary Figures 7, 9**). Therefore, exploring other edge weight transformation methods that retain finer community structure while controlling the influence of negative (low-confidence) PANDA edge weights may improve the performance of ALPACA and CRANE.

CRANE assisted differential network analysis minimally requires the user to provide (i) a pair of disease and control networks, and (ii) a list of nodes that defines a candidate module. Although we have applied it in conjunction with PANDA and ALPACA, other network inference and module identification algorithms could also be used in principle. CRANE is designed for weighted networks, with approximately normally distributed edge weights, that incorporate sparsity; in general, network inference methods that use a combination of data-driven correlations and prior information, partial thresholding, or other constraints could be compatible with CRANE. Binary networks with edges either present or absent – e.g., protein–protein interactions measured by IP-MS or Y2H – may require a different statistical treatment. The user-defined candidate module should be more strongly interconnected (higher total edge weight) in the disease network than in the control network, but otherwise could be identified using any method.

In summary, CRANE is a flexible algorithm that can be applied to both weighted unipartite (e.g., WGCNA) and bipartite (e.g., PANDA) gene regulatory networks to generate biologically realistic null distributions. We have demonstrated that this null distribution can be used to better rank the genes that significantly drive disease pathways. In the future, we anticipate that CRANE could be used to evaluate the significance of other features (e.g., information flow or betweenness centrality) of disease networks that are built around a "skeleton" of prior information, like TF binding sites or interaction databases. As gene regulatory networks become an increasingly common framing device for multi-omics data, CRANE provides a robust approach to identify what aspects of these networks are truly altered in disease.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cbioportal.org.

AUTHOR CONTRIBUTIONS

MP and JL conceived of the project. JL developed the algorithm, performed analyses, and wrote the manuscript. CC performed comparison with network generative models. AG and MP helped to refined the analyses. All authors helped to writing the manuscript.

FUNDING

This publication was supported by Institutional Research Grant number IRG-16-124-37 from the American Cancer Society.

ACKNOWLEDGMENTS

We thank Jiawen Yang for useful discussions. This manuscript has been released as a pre-print at bioRxiv 2020.07.12.198747 (Lim et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.603264/full#supplementary-material

REFERENCES

Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., et al. (2009). Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* 41, 585–590. doi: 10.1038/ng.354

Aicher, C., Jacobs, A. Z., and Clauset, A. (2015). Learning latent block structure in weighted networks. *J. Complex Netw.* 3, 221–248. doi: 10.1093/comnet/cnu026

Arend, R. C., Londono-Joshi, A. I., Straughn, J. M. Jr., and Buchsbaum, D. J. (2013). The Wnt/beta-catenin pathway in ovarian cancer: a review. *Gynecol. Oncol.* 131, 772–779. doi: 10.1016/j.ygyno.2013.09.034

Bansal, S., Khandelwal, S., and Meyers, L. A. (2009). Exploring biological network structure with clustered random networks. *BMC Bioinform.* 10:405. doi: 10.1186/1471-2105-10-405

Baum, K., Rajapakse, J. C., and Azuaje, F. (2019). Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models. *F1000Research* 8:465. doi: 10.12688/f1000research.18705.2

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J. B., Hirsch, M. S., Holton, K., et al. (2012). Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PLoS One* 7:e30269. doi: 10.1371/journal.pone.0030269

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Statist. Mechan. Theory Exper.* 2008:10008.

Bowtell, D. D., Bohm, S., Ahmed, A. A., Aspuria, P. J., and Bast, R. C. Jr. (2015). Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat. Rev. Cancer* 15, 668–679. doi: 10.1038/nrc4019

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

Califano, A., Butte, A. J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* 44, 841–847. doi: 10.1038/ng.2355

Campigotto, R., Céspedes, P. C., and Guillaume, J.-L. (2014). A generalized and adaptive method for community detection. *arXiv* [Preprint], Available online at: https://arxiv.org/abs/1406.2518 (accessed May 2020).

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

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

Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., et al. (2019). Assessment of network module identification across complex diseases. *Nat. Methods* 16, 843–852. doi: 10.1038/s41592-019-0509-5

Cohen, C. A., Shea, A. A., Heffron, C. L., Schmelz, E. M., and Roberts, P. C. (2016). Interleukin-12 immunomodulation delays the onset of lethal peritoneal disease of ovarian cancer. *J. Interf. Cytokine Res.* 36, 62–73. doi: 10.1089/jir.2015.0049

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Inter. J. Complex Syst.* 1695, 1–9.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *J. Statist. Mech. Theory Exper.* 2005:9008.

Dunning, A. M., Healey, C. S., Baynes, C., Maia, A. T., Scollen, S., Vega, A., et al. (2009). Association of ESR1 gene tagging SNPs with breast cancer risk. *Hum. Mol. Genet.* 18, 1131–1139. doi: 10.1093/hmg/ddn429

Emmert-Streib, F. (2007). The chronic fatigue syndrome: a comparative pathway analysis. *J. Comput. Biol.* 14, 961–972. doi: 10.1089/cmb.2007.0041

Fosdick, B. K., Larremore, D. B., Nishimura, J., and Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *SIAM Rev.* 60, 315–355. doi: 10.1137/16m1087175

Gabrielli, A., Mastrandrea, R., Caldarelli, G., and Cimini, G. (2019). Grand canonical ensemble of weighted networks. *Phys. Rev. E* 99:030301.

Gambardella, G., Moretti, M. N., de Cegli, R., Cardone, L., Peron, A., and di Bernardo, D. (2013). Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics* 29, 1776–1785. doi: 10.1093/bioinformatics/btt290

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:pl1. doi: 10.1126/scisignal.2004088

Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A., et al. (2008). Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.* 4:e1000054. doi: 10.1371/journal.pgen.1000054

Garlaschelli, D. (2009). The weighted random graph model. *New J. Phys.* 11:073005. doi: 10.1088/1367-2630/11/7/073005

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100. doi: 10.1038/nature11245

Ghiassian, S. D., Menche, J., and Barabasi, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinform.* 11:95. doi: 10.1186/1471-2105-11-95

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G. C. (2013). Passing messages between biological networks to refine predicted interactions. *PLoS One* 8:e64832. doi: 10.1371/journal.pone.0064832

Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B., and Yuan, G. C. (2015). A network model for angiogenesis in ovarian cancer. *BMC Bioinform.* 16:115. doi: 10.1186/s12859-015-0551-y

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. doi: 10.1038/ng.3259

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402, C47–C52. doi: 10.1038/35011540

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8:565. doi: 10.1038/msb.2011.99

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1), S233–S240. doi: 10.1093/bioinformatics/18.suppl_1.s233

Jeub, L. G. S., Sporns, O., and Fortunato, S. (2018). Multiresolution Consensus Clustering in Networks. *Sci. Rep.* 8:3259. doi: 10.1038/s41598-018-21352-7

Kojaku, S., and Masuda, N. (2018). A generalised significance test for individual communities in networks. *Sci. Rep.* 8:7351. doi: 10.1038/s41598-018-25560-z

Lancichinetti, A., and Fortunato, S. (2012). Consensus clustering in complex networks. *Sci. Rep.* 2:336. doi: 10.1038/srep00336

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLoS One* 6:e18961. doi: 10.1371/journal.pone.0018961

Lee, T. I., and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Ann. Rev. Genet.* 34, 77–137. doi: 10.1146/annurev.genet.34.1.77

Lim, J. T., Chen, C., Grant, A. D., and Padi, M. (2020). Generating ensembles of gene regulatory networks to assess robustness of disease modules. *bioRxiv* [Preprint], doi: 10.1101/2020.07.12.198747

Mall, R., Cerulo, L., Bensmail, H., Iavarone, A., and Ceccarelli, M. (2017). Detection of statistically significant network changes in complex biological networks. *BMC Syst. Biol.* 11:32. doi: 10.1186/s12918-017-0412-6

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366–370. doi: 10.1038/nmeth.3799

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

Mastrandrea, R., Squartini, T., Fagiolo, G., and Garlaschelli, D. (2014). Enhanced reconstruction of weighted networks from strengths and degrees. *New J. Phys.* 16:043022. doi: 10.1088/1367-2630/16/4/043022

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Newman, M. (2018). *Networks*. Oxford: Oxford university press.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Newman, M. E. (2016). Community detection in networks: modularity optimization and maximum likelihood are equivalent. *arXiv* [Preprint], Available online at: https://arxiv.org/abs/1606.02319 (accessed May 2020).

Padi, M., and Quackenbush, J. (2018). Detecting phenotype-driven transitions in regulatory network structure. *NPJ Syst. Biol. Appl.* 4:16. doi: 10.1038/s41540-018-0052-5

Palowitch, J. (2019). Computing the statistical significance of optimized communities in networks. *Sci. Rep.* 9:18444. doi: 10.1038/s41598-019-54708-8

Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K., Provenzano, E., Bardwell, H. A., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* 7:11479. doi: 10.1038/ncomms11479

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* 94, 559–573. doi: 10.1016/j.ajhg.2014.03.004

Platig, J., Castaldi, P. J., DeMeo, D., and Quackenbush, J. (2016). Bipartite Community Structure of eQTLs. *PLoS Comput. Biol.* 12:e1005033. doi: 10.1371/journal.pcbi.1005033

Proulx, S. R., Promislow, D. E., and Phillips, P. C. (2005). Network thinking in ecology and evolution. *Trends Ecol. Evol.* 20, 345–353. doi: 10.1016/j.tree.2005.04.004

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. *Nature* 386, 569–577. doi: 10.1038/386569a0

Reid, B. M., Permuth, J. B., and Sellers, T. A. (2017). Epidemiology of ovarian cancer: a review. *Cancer Biol. Med.* 14, 9–32. doi: 10.20892/j.issn.2095-3941.2016.0084

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* 21, 327–335. doi: 10.1016/s0968-0004(96)10050-5

Sah, P., Singh, L. O., Clauset, A., and Bansal, S. (2014). Exploring community structure in biological networks with random graphs. *BMC Bioinform.* 15:220. doi: 10.1186/1471-2105-15-220

Santolini, M., and Barabasi, A. L. (2018). Predicting perturbation patterns from the topology of biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 115, E6375–E6383. doi: 10.1073/pnas.1720589115

Saul, Z. M., and Filkov, V. (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics* 23, 2604–2611. doi: 10.1093/bioinformatics/btm370

Schadt, E. E., Friend, S. H., and Shaywitz, D. A. (2009). A network view of disease and compound screening. *Nat. Rev. Drug Discov.* 8, 286–295. doi: 10.1038/nrd2826

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176. doi: 10.1038/ng1165

Siahpirani, A. F., and Roy, S. (2017). A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res.* 45:2221. doi: 10.1093/nar/gkw1160

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088.

Teif, V. B., and Rippe, K. (2009). Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Res.* 37, 5641–5655. doi: 10.1093/nar/gkp610

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinform.* 11:497. doi: 10.1186/1471-2105-11-497

Van Landeghem, S., Van Parys, T., Dubois, M., Inze, D., and Van de Peer, Y. (2016). Diffany: an ontology-driven framework to infer, visualise and analyse differential molecular networks. *BMC Bioinform.* 17:18. doi: 10.1186/s12859-015-0863-y

Varadhan, R., and Gilbert, P. (2009). BB: an R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *J. Statist. Softw.* 32, 1–26.

Wall, L., Burke, F., Barton, C., Smyth, J., and Balkwill, F. (2003). IFN-γ induces apoptosis in ovarian cancer cells in vivo and in vitro. *Clin. Cancer Res.* 9, 2487–2496.

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

Zhang, L., Feng, X. K., Ng, Y. K., and Li, S. C. (2016). Reconstructing directed gene regulatory network by only gene expression data. *BMC Genomics* 17(Suppl. 4):430. doi: 10.1186/s12864-016-2791-2

Zitnik, M., and Leskovec, J. (2018). Prioritizing network communities. *Nat. Commun.* 9, 1–9.

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

GLOSSARY OF STAND-ALONE METHODS

PANDA (network inference algorithm): infers a weighted bipartite gene regulatory network by iteratively integrating target gene co-expression with transcription factor binding site occurrence (Glass et al., 2013).

WGCNA (network inference algorithm): constructs weighted networks that obey the scale-free topology criterion using soft thresholding of gene correlations (Zhang and Horvath, 2005).

ALPACA (network analysis method): finds candidate disease modules by optimizing a differential modularity score defined as the difference in edge density between disease and matched control networks (Padi and Quackenbush, 2018).

CONDOR (network analysis method): algorithm for detecting community structure in weighted bipartite networks (Platig et al., 2016).

CRANE (network randomization algorithm): perturbs network edges while fixing node strengths and maintaining realistic features of gene regulatory networks.

Configuration model (network randomization algorithm): an exponential random graph model for generating networks from a given node degree or strength sequence (Gabrielli et al., 2019).

SBM (network randomization algorithm): a generative model that defines a probability distribution between node pairs by assuming pre-existing densely connected communities or "blocks" (Aicher et al., 2015).

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 February 2021
doi: 10.3389/fgene.2021.618089

Abiotic Stress-Responsive miRNA and Transcription Factor-Mediated Gene Regulatory Network in *Oryza sativa*: Construction and Structural Measure Study

Rinku Sharma[1], Shashankaditya Upadhyay[2], Sudeepto Bhattacharya[3] and Ashutosh Singh[1]**

[1] Department of Life Sciences, Shiv Nadar University, Gautam Buddha Nagar, India, [2] Department of Electrical Engineering, Indian Institute of Technology, New Delhi, India, [3] Department of Mathematics, Shiv Nadar University, Gautam Buddha Nagar, India

Climate changes and environmental stresses have a consequential association with crop plant growth and yield, meaning it is necessary to cultivate crops that have tolerance toward the changing climate and environmental disturbances such as water stress, temperature fluctuation, and salt toxicity. Recent studies have shown that trans-acting regulatory elements, including microRNAs (miRNAs) and transcription factors (TFs), are emerging as promising tools for engineering naive improved crop varieties with tolerance for multiple environmental stresses and enhanced quality as well as yield. However, the interwoven complex regulatory function of TFs and miRNAs at transcriptional and post-transcriptional levels is unexplored in *Oryza sativa*. To this end, we have constructed a multiple abiotic stress responsive TF-miRNA-gene regulatory network for *O. sativa* using a transcriptome and degradome sequencing data meta-analysis approach. The theoretical network approach has shown the networks to be dense, scale-free, and small-world, which makes the network stable. They are also invariant to scale change where an efficient, quick transmission of biological signals occurs within the network on extrinsic hindrance. The analysis also deciphered the existence of communities (cluster of TF, miRNA, and genes) working together to help plants in acclimatizing to multiple stresses. It highlighted that genes, TFs, and miRNAs shared by multiple stress conditions that work as hubs or bottlenecks for signal propagation, for example, during the interaction between stress-responsive genes (TFs/miRNAs/other genes) and genes involved in floral development pathways under multiple environmental stresses. This study further highlights how the fine-tuning feedback mechanism works for balancing stress tolerance and how timely flowering enable crops to survive in adverse conditions. This study developed the abiotic stress-responsive regulatory network, APRegNet database (http://lms.snu.edu.in/APRegNet), which may help researchers studying the roles of miRNAs and TFs. Furthermore, it advances current understanding of multiple abiotic stress tolerance mechanisms.

Keywords: ***Oryza sativa*, microRNA, transcription factor, regulatory network, post-transcriptional regulation, target mimics**

OPEN ACCESS

Edited by:
Maud Fagny,
UMR 7206 Eco-Anthropologie et Ethnobiologie (EAE), France

Reviewed by:
Tatiana Belova,
University of Oslo, Norway
Daisuke Todaka,
The University of Tokyo, Japan

****Correspondence:***
Ashutosh Singh
ashutosh.bio@gmail.com
Sudeepto Bhattacharya
sudeepto.bhattacharya@snu.edu.in

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

Received: *16 October 2020*
Accepted: *19 January 2021*
Published: *12 February 2021*

Citation:
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

INTRODUCTION

Abiotic stresses such as drought, cold, and salt can reduce the productivity and yield of plants with a direct adverse impact on global food security (Myers et al., 2017). In the spontaneously changing climate scenarios of recent years, there has been an increase in episodes of occurrence and the severity of these stresses (Lesk et al., 2016). Rice (*Oryza sativa*) is the most imperative crop across the globe, grown in over a hundred countries (including India), with a production rate greater than 700 million tons per annum (Londo et al., 2006). Researchers have estimated that around 1% of enhancement per annum in the *O. sativa* yield is required to fulfill increasing population demands (Normile, 2008). The literature on this subject includes several studies on rice, examining ways to enhance its nutritional quality and tolerance toward many diseases (Grover and Minhas, 2000). Increasing the yield of rice to meet this increasing demand is one of the most challenging aspects of this research, as various abiotic stresses adversely affect production (Mantri et al., 2012).

Plants have developed dynamic responses at the morphological, physiological, and biochemical levels that allow them to escape and/or adapt to calamitous environmental conditions. Plants regulate these responses at the molecular level through a series of complex interwoven network of events. Transcription factors (TFs) and microRNAs play an important role in regulating the activity of the genes at transcriptional and post-transcriptional levels, respectively, involving a complex series of events (Hobert, 2008; Li and Zhang, 2016; O'Brien et al., 2018).

Transcription factors belong to multi-gene families and have DNA-binding and protein-protein interaction domains through which they interact with cis-elements of their target genes and oligomerize with other TFs or with other regulators, respectively (Boeva, 2016). They aid in the regulatory system in several ways, by managing stress-responsive gene expression at the correct time and place, and controlling developmental and defense responses (Kissoudis et al., 2014; Shao et al., 2015; Samad et al., 2017). A single TF can control the expression of several genes in a particular pathway. Furthermore, recent studies have firmly linked the expression of a gene to the expression of TFs. For example, miR169 at the mRNA level controls the expression of the NFYA5 TF. It was shown in a transgenic plant experiment that suppression of *NFYA5* gene expression leads to susceptibility toward drought stress in plants (Li et al., 2008). This proves them as potential targets for the manipulation of desired traits in plants.

MicroRNAs (miRNA) are a major class of small endogenous RNAs of length 20–24 nucleotides. They assemble with ARGONAUTE (AGO) proteins forming an RNA-induced silencing complex (RISC) in the cytoplasm. AGO protein has PAZ and PIWI domains. PIWI domain creates an RNaseH-like fold that helps in cleaving RNA targets complementary to the miRNA strand assembled with the AGO in RIS-complex. It is often found that miRNAs regulate the target genes at the protein level without causing major change at the mRNA level. These findings suggest the capability of plant miRNAs to control gene expression at mRNA and protein levels (Voinnet, 2009; O'Brien et al., 2018). Recent studies have revealed the role of miRNAs in attenuating plant growth and development under the influence of several environmental stresses. Induction of miRNA expression under stress leads to repression of target genes, whereas, their repression leads to the expression of target genes under stressful conditions. The miRNAs play a central role in complex gene regulatory networks and are studied as a novel target for plant improvement, including improved tolerance to various environmental stresses. For example, over-expression of miR156 enhances tolerance to heat stress in *Arabidopsis thaliana* (Stief et al., 2014) and increases biomass in switchgrass (Fu et al., 2012). Over-expression of miR402 brings more tolerance to salinity, drought, and cold stress in *A. thaliana* (Kim et al., 2010). TFs and miRNAs play an essential role in multiple stress conditions. Recent studies have revealed that several TFs and miRNAs show similar expression patterns in response to multiple stresses (Zhang, 2015; Wang et al., 2016). This indicates that they could be targets for multiple abiotic stress-tolerant variety development.

In the literature on this subject, many miRNAs, TFs, and mRNAs were reported in response to multiple stress conditions in different plant species using computational programs and deep-sequencing techniques. Researchers have also studied the potential role of miRNAs, and TFs in gene expression control through several experimental methods. Over the past few years, however, there has been a shift in interest toward deciphering the complex interwoven regulatory networks operating in plants. The studies conducted either describe transcriptional or post-transcriptional regulation of genes, but a comprehensive study is lacking (Chen et al., 2018; Sun and Dinneny, 2018; Haque et al., 2019). Several online plant resources also exist for searching transcriptional or post-transcriptional regulation of genes. These include- (a) PlantRegMap (Jin et al., 2017), which contains information about TFs and its direct target genes for 135 plant species integrated from literature mining, TF ChIP-seq, and prediction combined TF binding motifs and regulatory elements; (b) AtRegNet (Palaniswamy et al., 2006), which provides information about TFs and the direct target genes of *A. thaliana* integrated from experimental methods like- EMSA, ChIP-seq, yeast one-hybrid analysis, etc.; and, (c) AtmiRNet contains information about the transcriptional regulation of miRNAs based on experimental methods like—EMSA, yeast one-hybrid analysis, transgenic plant expressing an inducible TF-GR (glucocorticoid receptor) fusion protein experiments, etc. PASmiR and miRNEST are comprehensive literature curated databases for stress-responsive miRNA and its targets (Zhang et al., 2013; Szcześniak and Makabowska, 2014). However, no database or repository provides a global view of transcriptional and post-transcriptional integrated regulation of gene expression under abiotic stress. To acknowledge this current issue, the present study constructed abiotic stress-responsive gene regulatory networks for *O. sativa* and studied them in-depth using the network theoretic approach. We developed a comprehensive database, APRegNet database,[1] whose construction is discussed for abiotic responsive gene regulatory networks at both transcriptional and post-transcriptional level for *A. thaliana*,

[1] http://lms.snu.edu.in/APRegNet

O. sativa, and *Zea mays* in response to drought, cold, salt, and waterlogging stress. This database provides a valuable resource for contemporary researchers.

MATERIALS AND METHODS

Data Sources

We retrieved the small RNA and mRNA (GPL2025) expression high-throughput datasets of cold, drought, and salt stress conditions (**Supplementary Table 1**) from the public domain, GEO (Gene Expression Omnibus[2]) and ArrayExpress Archive.[3]

The *O. sativa* AGO1-associated sRNA HTS (high-throughput sequencing) datasets GSM455962, GSM455963, and GSM455964 were downloaded from the GEO database (**Supplementary Table 1**).

The *O. sativa* degradome sequencing datasets – GSE17398 and GSE19050 (**Supplementary Table 1**) were downloaded from the GEO database.

The miRNA sequences were downloaded from miRBase (release 21[4]) (Kozomara and Griffiths-Jones, 2014). We retrieved the transcripts of *O. sativa* genes and the gene annotations from the *O. sativa* Genome Annotation Project (RGAP), available at ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/ (Kawahara et al., 2013).

Finding Dysregulated Genes and miRNAs Related to Abiotic Stress

Differentially Expressed Genes Identified by Transcriptome Meta-Analysis

The GCRMA R package (Wu et al., 2004) was used to normalize the raw expression data and outlier samples were detected by the ArrayQualityMetrics R package (Kauffmann et al., 2009). Thereafter, transcriptome meta-analysis was performed for differentially expressed genes (DEGs) identified using function RPadvance in the Bioconductor package (Hong et al., 2006) and pathway analysis using the KEGG database (Kanehisa et al., 2017).

Differentially Expressed miRNAs Identification

The HTS raw reads were pre-processed, which comprise adaptor trimming, low-quality tags removal, and determining sequence quality check. We also summarized clean tag length distribution and common and specific sequences between samples. The sequences of rRNA, scRNA, snRNA, tRNA, exon, intron, and repeat sequence tags were removed using GenBank[5] and Rfam (12.2) database.[6] The cleaned reads were then aligned to reference genome (MSU7) RGAP, available at ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/.

The sequence coordinates were then compared to *O. sativa* miRNA GFF file and different measures of the expression level were generated, such as, the read count (total number of reads assigned to the reference RNA), adjusted read count (read count normalized by the number of times that the read maps to the library or the genome) and normalized RPM (reads per million) which was done by miRanalyzer (Hackenberg et al., 2009). The EdgeR and DeSeq R package were utilized for differential expression analysis of miRNA of drought, cold and salt stress conditions.

AGO1-Enrichment Analysis of Differentially Expressed miRNA

The differentially expressed miRNAs from various abiotic stresses were subjected to AGO1-enrichment analysis by applying the following rules: (1) the miRNA should be detectable in at least one of the AGO1-associated sRNA HTS datasets and (2) its normalized accumulation levels should be three RPM or higher.

Abiotic Stress Responsive TF-miRNA Induced Gene Regulatory Network Construction and Network Measure Calculation

The regulatory network covered five types of regulatory relationships: TF-gene, TF-miRNA, TF-TF, miRNA-gene, and target_mimics-miRNAs. We extracted the TF regulatory information from EGRINs (Environmental gene regulatory influence networks) for *O. sativa* (Wilkins et al., 2016). The miRNA regulatory activity is also controlled by a kind of ribo-regulator known as target_mimics. A 23-nucleotide sequence conserved in plant species in a family of non-coding RNAs resembles a cleavable miRNA target site, however, the site is not cleaved and instead negatively regulates miRNA activity through mimicry (Franco-Zorrilla et al., 2007). The target_mimics regulatory information were obtained from PeTMbase (plant endogenous target mimics database) database (Karakülah et al., 2016) and miRNA regulation of genes was predicted by the psRNATarget tool (Dai and Zhao, 2011) using default parameters and we validated the predicted targets through degradome sequencing data analysis, which was performed by CleaveLand v4.4.4 (Addo-Quaye et al., 2009).

We then examined certain local and global properties, namely; the scale-free behavior (Clauset et al., 2009), small-world-ness, assortative mixing, strongly connected components, and network centralities (degree, closeness, and betweenness) according to methods outlined in previous studies (Upadhyay et al., 2017; Sharma et al., 2020).

APRegNet: Database Construction

Data Sources

We captured both transcriptional and post-transcriptional regulation of genes under abiotic stress in *A. thaliana*, *O. sativa*, and *Z. mays*. We considered TFs, miRNAs, and target_mimics

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

[3]https://www.ebi.ac.uk/arrayexpress/

[4]http://www.mirbase.org/

[5]http://www.ncbi.nlm.nih.gov/GenBank/

[6]http://rfam.sanger.ac.uk/

as regulators. **Figure 1** shows the fundamental regulatory interactions among – TF, miRNA, target gene, and target_mimics and also the essential steps of transcriptional and post-transcriptional regulation of gene expression.

We performed transcriptome meta-analysis for DEGs and miRNAs identification [for *O. sativa* discussed in this manuscript, for *A. thaliana* (Sharma et al., 2020) and *Z. mays* (unpublished)]. Thereafter we fetched the transcriptional regulation of genes, miRNAs, and other transcription factors and post-transcriptional regulation of miRNA by target_mimics from reliable databases restricted to experimentally validated interactions only (**Table 1**). We validated the miRNA regulation of target gene expression through degradome sequence meta-analysis [for *O. sativa* discussed in this manuscript, for *A. thaliana* (Sharma et al., 2020) and *Z. mays* (unpublished)].

KEGG pathway,[7] PlantTFDB,[8] INTERPRO,[9] and DAVID[10] sources were also integrated for **Supplementary Information**.

Database Implementation and Web User Interface Design

We developed an apprehensible and user-friendly web interface, APRegNet (see text footnote 1) for users to query and download the regulatory relationships and networks. APRegNet focuses on providing better navigation through individual sections to increase data discoverability. There are five tabs provided at the top of the interface ("Home," "About," "Source," "Browser," "Download," and "Contact Us") through which users can navigate and explore the required information. It runs on a XAMPP web server with an MYSQL database in the backend for data storage and management. Text query box is provided at the top of each page to search by various types of components (i.e., by TF, miRNA, or gene in the regulatory networks), by stress (cold, drought, salt, and waterlogging), and by species (*A. thaliana*, *O. sativa*, and *Z. mays*). The query result page shows results in three sections:

(1) **Network properties:** The network properties (in-degree, out-degree, closeness, and betweenness) of the queried component (gene/TF/miRNA), confined to selected species and stress-specific regulatory network;
(2) **Functional annotation:** This section includes Gene Model, Primary gene Symbol, GO term: Biological process, GO term: Molecular Function, GO term: Cellular Component, Transcription factor family, and KEGG pathway.
(3) **Interaction:** Displays first interacting patterns of the query component and type of interaction between source and target.

The user can download the complete regulatory network in CSV file format for each species stress-wise from the download page. **Figure 2** illustrates the snapshots of the database.

Query processing scripts are written in PHP and SQL. The database is tested and works well with commonly available web browsers, such as Mozilla Firefox, Google Chrome, Safari, and Microsoft Internet Explorer. The schema of the database is given in **Figure 3**.

[7]https://www.genome.jp/kegg/pathway.html
[8]http://planttfdb.cbi.pku.edu.cn/
[9]https://www.ebi.ac.uk/interpro/
[10]https://david.ncifcrf.gov/

RESULTS

In this study, miRNAs and transcription factors were woven into a complex inter-regulatory network within the cell liable for reprogramming gene expression in response to abiotic stresses in *O. sativa* deciphered through transcriptome (both coding and non-coding) meta-analysis. We further looked into the structural perspective of the proposed abiotic responsive networks which elucidated the significant role of some genes/TF/miRNA in the abiotic stress response mechanism.

Abiotic Stress-Responsive Genes Identification by Meta-Analysis

We performed a meta-analysis of 15 studies of gene expression in response to abiotic stresses (cold, drought, and salt stress) (**Supplementary Table 1**) for each of the 51,279 genes using the RankProd method. The meta-analysis of 15 individual transcriptome profiling studies identified 5,255 genes showing significant differential expression in response to at least one of the abiotic stresses under investigation compared to control conditions in *O. sativa* [PFP (percentage of false positives) <0.01; **Figures 4A,B**, **Supplementary Figure 2A**, and **Supplementary Table 2**]. The comparative study of DEGs across the three stresses showed that 201 genes were commonly expressed across the three stresses among which 82 genes have conserved expression patterns. By contrast, a stress pair-wise comparison of the DEG lists found that drought and cold stress share the maximum number of DEGs, and over 75% of them showed similar expression patterns under both stresses. Functional annotation showed that among the conserved genes group, carbohydrate, nitrogen, and chlorophyll metabolism-related genes are under-expressed but that universal stress response proteins, hormonal signal transduction pathway, transport, energy production, and conservation-related genes were over-expressed across all the three stresses. The results showed that 73% (3,859) of the total DEGs were stress-specific (**Figures 4A,B**, **Supplementary Figure 2A**, and **Supplementary Table 2**).

Abiotic Stress-Responsive miRNAs Identification

In the present study, with the aid of publically available abiotic stress-specific sRNA HTS data for *O. sativa*; we identified drought, cold, and salt stress-responsive miRNAs. **Supplementary Table 3** shows the number of miRNAs identified in control and each stress sample. Differential expression analysis also showed that 129 miRNAs were differentially expressed in response to at least one abiotic stress under study (**Supplementary Table 4**, **Figures 5A,B**, and **Supplementary Figure 2B**). A comparative study revealed that in response to salt stress, the highest numbers of miRNAs differentially expressed (68: 26 up-regulated and 42 down-regulated), followed by drought stress (60: 6 up-regulated and 54 down-regulated),

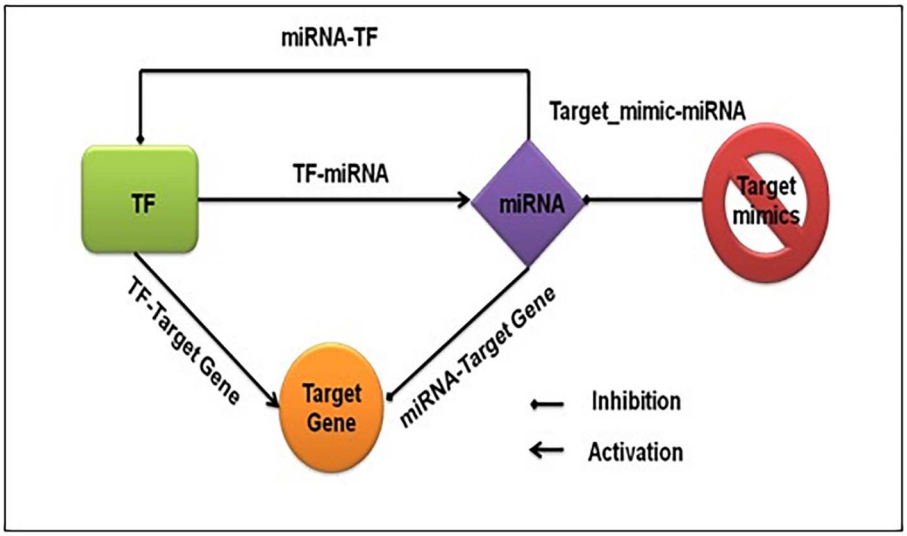

FIGURE 1 | Fundamental regulatory interactions at the transcriptional and post-transcriptional level. The transcription factor regulating the expression of functional genes, other transcription factors, and miRNA, and at the post-transcriptional level miRNA regulating the expression of functional genes and transcription factors. The target mimics regulates miRNA expression.

TABLE 1 | Data source: for abiotic stress (cold, drought, salt, and waterlogging) responsive genes and miRNAs regulatory transcription factors relationships extracted from various databases.

Source	Description	Species	Link
ATRM: *Arabidopsis thaliana* transcriptional regulatory map	ATRM database is a curated a high-confidence *Arabidopsis thaliana* transcriptional regulatory map derived by a systematic literature mining.	*Arabidopsis thaliana*	http://atrm.cbi.pku.edu.cn/
AtmiRNET: resconstructing regulatory networks of *Arabidopsis thaliana*	AtmiRNET database contain manually curated information about transcriptional regulation of *Arabidopsis thaliana* miRNAs derived from literature	*Arabidopsis thaliana*	http://atmirnet.itps.ncku.edu.tw/home.php
PlantRegMap: plant transcriptional regulatory map	PlantRegMap database contains genome-wide transcriptional regulatory interactions curated from literature and inferred by combining TF binding motifs and regulatory elements.	132 plant species	http://plantregmap.gao-lab.org/
Environmental gene regulatory influence networks (EGRINs)	This research article provided information about gene regulation network for *Oryza sativa* in response to high temperatures, water deficit, and agricultural field conditions by systematically integrating time-series transcriptome data, patterns of nucleosome-free chromatin, and the occurrence of known cis-regulatory elements.	*Oryza sativa*	PMCID: PMC5134975, doi: 10.1105/tpc.16.00158
ArrayExpress	Contain gene expression data from Array and sequencing techniques	–	https://www.ebi.ac.uk/arrayexpress/
PlantTFDB (plant transcription factor database)	Contain highly curated information about 320,370 transcription factors from 165 plant species	Plants	http://planttfdb.cbi.pku.edu.cn/index.php
mirBase	An archive of miRNA sequences and annotation	–	http://www.mirbase.org/
NCBI GEO (Gene Expression Omnibus)	Contain gene expression data from Array and sequencing techniques	–	https://www.ncbi.nlm.nih.gov/geo/
TAIR (The *Arabidopsis thaliana* Information Resource)	TAIR database provides genetic and molecular biology data for *Arabidopsis thaliana*	*Arabidopsis thaliana*	https://www.arabidopsis.org/
Oryza sativa Genome Annotation Project	Contain genome sequence from the Nipponbare subspecies of *Oryza sativa* and annotation of the 12 *Oryza sativa* chromosomes.	*Oryza sativa*	http://rice.plantbiology.msu.edu/
Gramene-*Zea mays*	A curated, open-source, integrated data resource for *Zea mays*	*Zea mays*	http://ensembl.gramene.org/Zea_mays/Info/Index
PeTMbase	A database of plant endogenous target mimics (eTMs)	Plants	http://petmbase.org

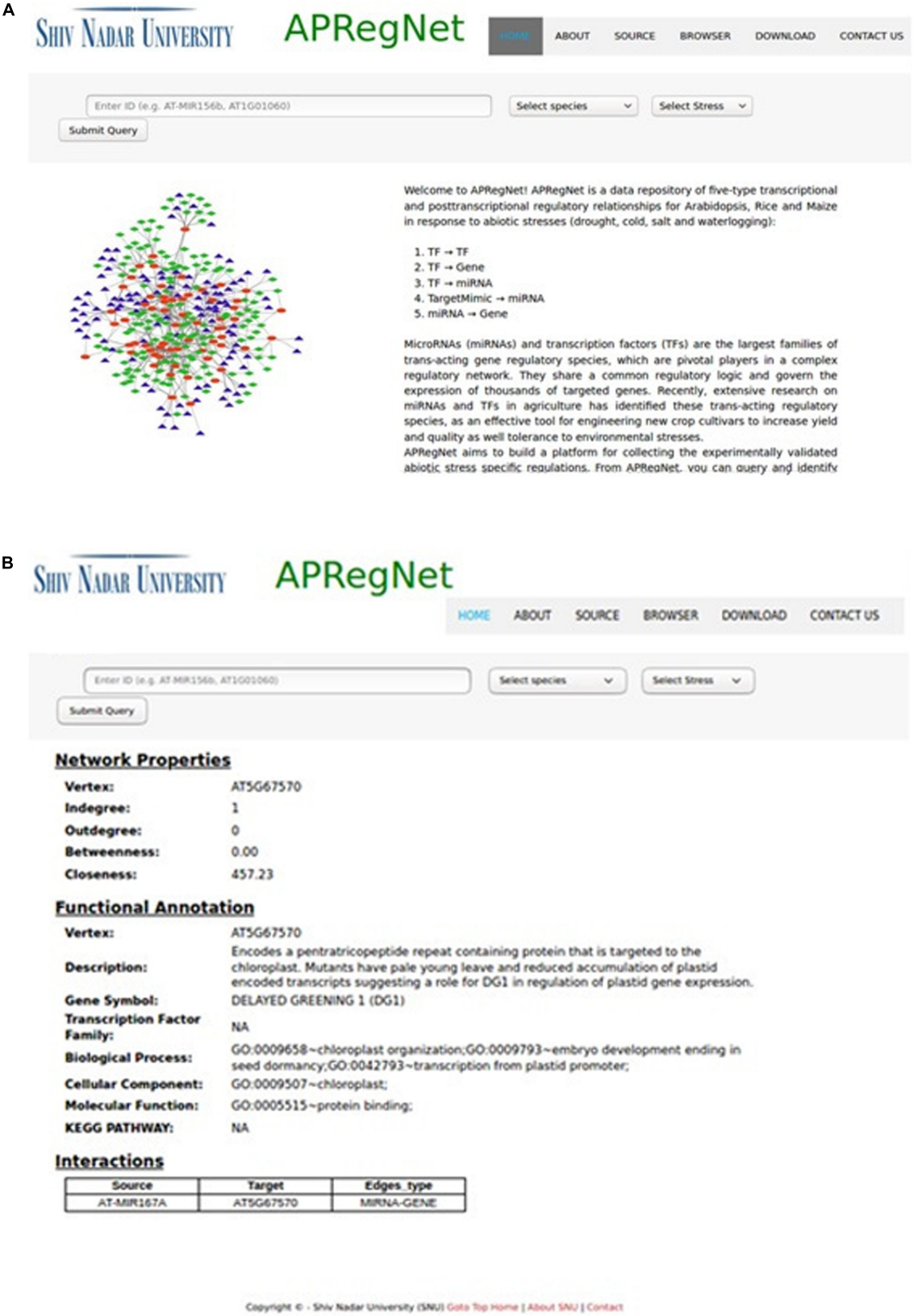

FIGURE 2 | HomePage and result page of APRegNet Database. **(A)** Depicts the interface of the home page of APRegNet where the user can search information by typing the relevant keywords in the search tab and **(B)** shows the search result page.

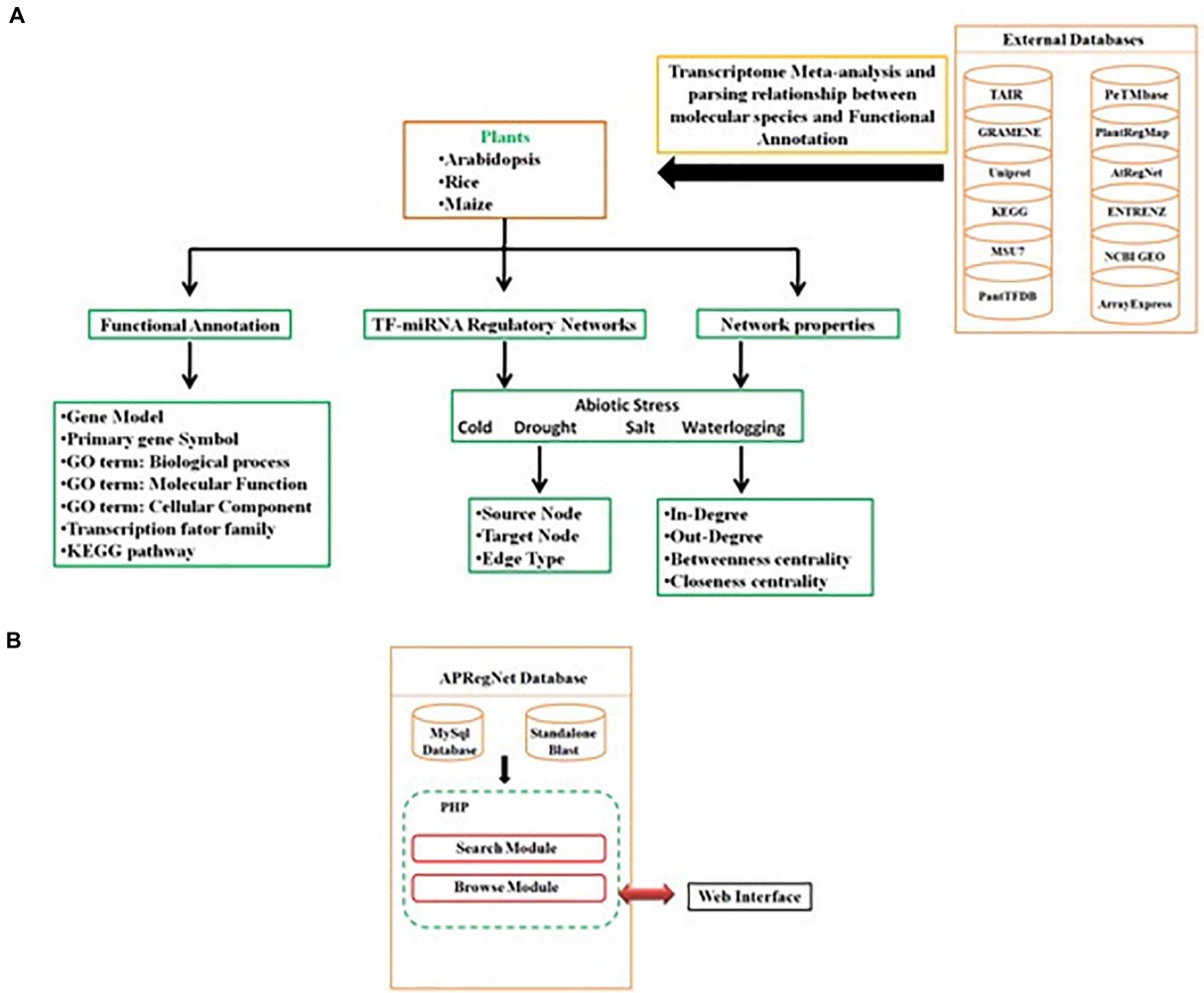

FIGURE 3 | Schema of APRegNet Database. **(A,B)** Explains the method and data source for regulatory network generation and arrangement of various files in the database.

and then cold stress (59: 6 up-regulated and 53 down-regulated). We found that under all three stresses, 12 DE miRNAs were common, having the same expression direction and stress-pair-wise comparison, which revealed that drought and cold stress shared maximum DE miRNAs (36) with conserved expression pattern followed by drought and salt stress (19) and then cold and salt stress (16). We observed that common miRNAs also have conserved expression patterns across the stresses. Several miRNAs showed differential expression unique to the abiotic stress category.

Argonaute 1 Enrichment Analysis

The miRNAs, at the post-transcriptional level, regulate the target gene expression by target cleavage. They escort RISCs to the target mRNA for degradation or translational repression. A recent study has reported that ARGONAUTE protein (AGO1) selectively binds with miRNAs and short interfering RNAs in plants and catalyzes the cleavage of target mRNAs (Baumberger and Baulcombe, 2005; Arribas-Hernández et al., 2016). The present study identified AGO1-enriched miRNAs and found 91 AGO1-enriched miRNAs in response to abiotic stresses for *O. sativa* (**Supplementary Table 5**).

AGO1-Enriched miRNAs Target Genes Identification

The targeted genes for abiotic stress-responsive AGO1-enriched miRNAs were initially predicted using psRNAtarget, a web-based online search tool under default parameters. As a result, 15,698 target binding sites predicted for 91 AGO1-enriched miRNAs expressed under abiotic stresses (**Supplementary Table 6**). These predicted target genes for miRNAs were validated through the meta-analysis of *O. sativa* degradome sequencing datasets, which is a high-throughput technique for large-scale validation of the miRNA-target duplex interactions (German et al., 2009). In the present analysis, six degradome sequencing datasets of

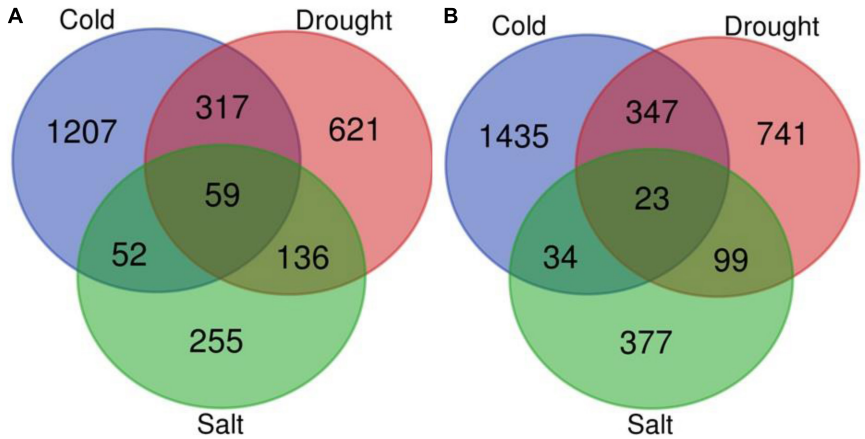

FIGURE 4 | Venn diagram for differentially expressed genes. Venn diagram depicting the number of up **(A)** and down-regulated **(B)** genes under drought, cold, and salt stress in *Oryza sativa*.

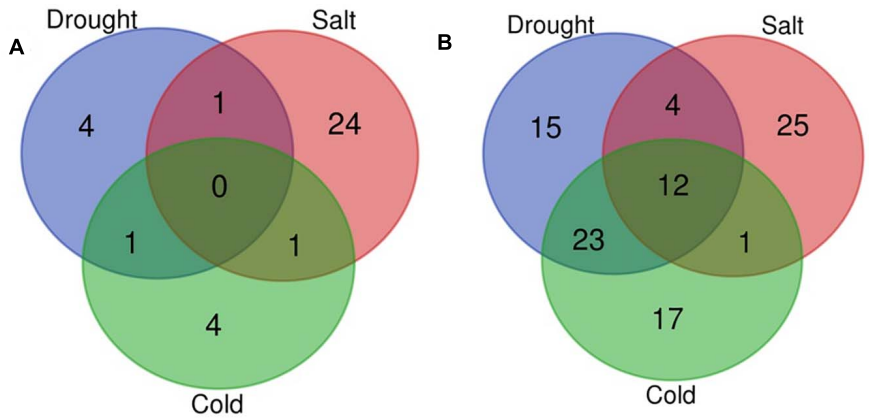

FIGURE 5 | Venn diagram for differentially expressed miRNAs. Venn diagram depicting the number of up **(A)** and down-regulated **(B)** miRNAs under drought, cold, and salt stress in *Oryza sativa*.

O. sativa were downloaded from the public domain and used for performing a comprehensive validation of the predicted gene targets of abiotic stress-responsive miRNAs using CleaveLand4 version 4.4 pipeline (Addo-Quaye et al., 2009). It plots the sequenced tag abundance on each transcript and further grouped the cleaved target transcripts into five categories based on the relative abundance of the degradome tags mapping at the miRNA target site through the height of the degradome peak at each occupied transcript position (Categories 0, 1, 2, 3, and 4) (**Supplementary Table 7**).

A total of 538 psRNAtarget predicted (AGO1-enriched) miRNA-target interaction was validated by degradome sequencing data analysis from which 63 were classified as Category 0, 2 as Category 11, 82 as Category 2, 9 as Category 3, and 373 as Category 4. Category 0: >1 raw tags at the position, abundance at the position was equal to the maximum on the transcript. There was only one maximum on the transcript; Category 1: >1 raw tag at the position, abundance at the position was equal to the maximum on the transcript. There was over one maximum position on the transcript; Category 2: >1 raw tag at the position, abundance at the position was less than maximum but higher than the median for the transcript; Category 3: >1 raw tag at the position, abundance at the position was equal to or less than the median for transcript; Category 4: only 1 raw tag at the position. The validated interactions involved 445 genes and 80 miRNAs highly abundant under the influence of abiotic stresses (**Supplementary Table 8**).

Abiotic Stress-Specific TF-miRNA-Gene Network Construction

We constructed the abiotic stress-responsive miRNA-TF gene regulatory networks for *O. sativa* in response to drought, cold, and salt stress. The networks contain five types of regulatory information: miRNA regulating genes, TF regulating genes, TF regulating TF, TF regulating miRNA genes, and target-mimics regulate miRNAs. The transcription factor regulation of abiotic responsive genes and miRNAs were derived from EGRINs,

target-mimics regulation of abiotic stress-responsive miRNAs was derived from the publically available database: PeTMbase. The abiotic stress-responsive miRNAs validated targeted genes through degradome sequence data analysis. After assembly of all these relations, we built large directed graphs comprising (1352, 13546), (1063, 9901), and (949, 5920) nodes-edges pairs showing the relationship between genes, miRNA, and transcription factors in response to cold, drought, and salt stress, respectively. To study the nature of the constructed networks (a random or complex or regular type of network), the degree distribution was studied in the context of the power law using the method described by Clauset et al. (2009). The parameters that qualify the in-degree and out-degree sequences for the networks shown as a **Supplementary Table 11**. We found that each network in-degree and out-degree distribution follows power law as shown by the straight line in log-log plots, which is a distinctive nature of a complex network with non-random degree distribution, possibly a scale-free network (**Supplementary Figure 1**). Further, assortativity of all the networks computed and results displayed negative values of assortative mixing for in-in, out-out, in-out, and out-in degree pairs across the networks that revealed the interaction between nodes having the higher degree with nodes having a lower degree (**Table 2**). This showed the disassortative nature of all the present study networks which revealed the importance of hub nodes, i.e., nodes with the highest connectivity or degree, in case of failure of these hub nodes the network is likely to become disconnected which leads to no flow of information in the system and thus the system is disrupted.

We examined whether the networks had small-world properties or not. We computed transitivity (T_G) and average shortest path length (ASL_G) for three TF-miRNA induced stress-specific networks and compared their values with transitivity T_{ER} and average shortest path length ASL_{ER} of the Erdos–Renyi random network of the same order and sizes. We then calculated small-world-ness (S) for cold, drought, and salt stress-specific regulatory networks. The results showed that all the networks had a value >1, which suggested the small-world nature of all three networks (**Table 3**).

TABLE 2 | Topological attributes of the *Oryza sativa* abiotic stress-responsive TF-miRNA-gene networks.

Topological attributes	Rice		
	Cold	Drought	Salt
Number of edges	13,546	9,901	5,920
Number of nodes	1,352	1,063	949
Largest SCC size	42	43	43
Graph diameter	12	12	12
Characteristic path length	3.708	3.704	3.718
The average number of neighbors	18.993	18.087	11.926
Assortativity (in-in)	−0.523	−0.529	−0.457
Assortativity (in-out)	0.013	0.017	0.014
Assortativity (out-in)	−0.305	−0.249	0.017
Assortativity (out-out)	−0.470	−0.473	−0.437

TABLE 3 | Calculations of small-world-ness of the *Oryza sativa* miRNA-TF-gene regulatory networks.

Network	*Oryza sativa*				
	Transitivity_ Net	Transitivity_ Rand	ASD_Net	ASD_Rand	Sw-ness
Drought	0.027	0.017	2.842	2.716	1.533
Salt	0.042	0.012	3.329	3.011	3.118
Cold	0.022	0.014	3.020	2.755	1.403

Various network centrality measures calculations prioritized the molecular species (miRNA/TF/gene) of the networks. The details of the top 10 nodes of each stress-specific network of O. *sativa* are given in **Supplementary Table 9**.

In the directed graph like- regulatory networks of the present study, the two nodes "x" and "y" belong to the same strongly connected component, if there are directed paths both from "x" and "y" and from "y" and "x" (Jeong and Berman, 2008). We found that one SCC is present in drought, cold, and salt stress-specific regulatory networks. **Supplementary Table 10** contains details of the TFs present in SCC. Functional analysis revealed the role of the SCC components in multiple abiotic stress responses, hormonal signal transduction, flower development, and cell fate specification.

APRegNet Database

The APRegNet database developed by integrating the experimentally validated regulatory interactions between TFs, miRNAs, genes, and target_mimics from various sources; as a comprehensive repository for genome-wide regulatory networks operating in *A. thaliana*, *O. sativa*, and *Z. mays* in response to abiotic stresses (cold, drought, salt, and waterlogging). It contains regulatory relationships at both transcriptional and post-transcriptional levels as well as interaction among TF/miRNAs and their targets with easily downloadable options. It also provides the data source information for the regulatory interactions. **Table 4** lists the basic statistics of the regulatory networks in APRegNet. This database contains regulatory information for 4,063, 2,026, and 7,152 genes/TFs/miRNAs for *A. thaliana*, *O. sativa*, and *Z. mays*, respectively.

The query result page displays few network centrality measures (namely, degree centrality, closeness centrality, and betweenness centrality) of the query component (gene/miRNA/TF) related to specific species stress network, which quantifies the importance of the component in the network for the quick and efficient flow of information.

DISCUSSION

Oryza sativa Abiotic Stress Responsive miRNA-TF-Gene Regulatory Networks

Cells use signal transduction pathways and regulatory mechanisms to coordinate multiple processes, allowing them

TABLE 4 | Basic statistics of the database: number of nodes, edges, TF (transcription factors), miRNAs (microRNA), stress-responsive genes, and various relationships among them (C, cold stress; D, drought stress; S, salt stress; and W, waterlogging stress).

	Arabidopsis thaliana				*Oryza sativa*			*Zea mays*			
Elements	**D**	**C**	**S**	**W**	**D**	**C**	**S**	**D**	**C**	**S**	**W**
Nodes	1,971	2,058	1,865	251	1,063	1,352	949	1,129	3,182	2,523	2,760
Edges	3,807	3,720	3,787	401	9,901	13,546	5,920	9,247	24,829	16,955	21,287
TF	369	354	393	163	340	390	337	307	378	386	464
miRNA	24	31	45	12	34	30	52	56	29	100	41
Gene	1,578	1,673	1,427	76	689	932	560	766	2,775	2,037	2,255
TF→TF	410	386	432	195	1,454	1,450	1,428	2,105	2,045	2,113	2,049
TF→Gene	3,251	3,134	3,091	133	8,299	11,917	4,098	6,148	22,279	13,172	18,521
TF→miRNA	76	91	128	40	3	38	30	431	231	678	288
TargetMimic→miRNA	22	45	47	9	16	5	28	151	97	283	137
miRNA→Gene	48	64	89	24	129	136	336	412	177	709	292

to respond to and adapt to an ever-changing environment. The biological system of components that interact with or regulate each other can be represented by a mathematical object called a graph (Bollobás and Cockayne, 1979), which comprises nodes and edges. Since, in a biological system, the flow of information among the components is directional; the edges are therefore directed. The understanding of the features that emerge from the entire cellular function requires an integrated, theoretical description of the relationships between different cellular components. The *O. sativa* abiotic stress-responsive miRNA and transcriptional factor gene regulatory networks in the present study were constructed and quantitatively described using the network theoretical approach. The observed topologies of the regulatory networks provide clues about the influence of the organization on the function and dynamic responses of the plant toward multiple abiotic stresses.

A study of the structural measures of the constructed regulatory networks revealed the scale-free and small-world nature of the abiotic stress-responsive miRNA-TF-gene regulatory networks. This implies that the networks are highly tolerant to random failures within the system, a quick and efficient flow of environmental stress signal from the source to sink takes place, which helps the plant adjust to a stressful environment (Barabási and Albert, 1999; Newman, 2003; Humphries and Gurney, 2008; Ghoshal and Barabási, 2011; Sharma et al., 2020). The networks are stable and invulnerable to scale change, and even the removal of a significant number of non-hub nodes (TF/miRNA/gene) cannot affect global behavior, but continuity in signal flow may be affected if hub nodes are inactivated. As in small-world networks, the fraction of non-hub nodes is larger than hub nodes, meaning that instances of hub node failure barely happen (Albert et al., 2000).

The analysis of centrality measures showed various MIKC_MADS transcription factor family members as the best-ranked nodes under all three stress (namely, cold, drought, and salt) responsive regulatory networks in *O. sativa* (**Supplementary Table 9**). This TF family member plays a crucial role in flowering time, floral organ identity determination, and fruit ripening (Theissen and Melzer, 2007; Li et al., 2016). According to the ABCDE flower development model, A, B, C, D, and E are different classes of genes, and interaction in a specific combination of these gene classes specifies different floral organs- Class A + E genes specify sepals, A + B + E specify petals, B + C + E specify stamens, C + E specify carpels, and C + D + E specify ovules (Zahn et al., 2006; Silva et al., 2016). According to centrality measures, in each of the three stress-responsive networks, these four classes of genes were among the top-ranked nodes. For example, Class A genes: OsMADS14, OsMADS15, and OsMADS18; Class B gene: OsMADS2; Class C genes: OsMADS3 and OsMADS58; and Class E gene: OsMADS6 (**Supplementary Table 9**). Recent studies have shown that these classes of genes are important not only in plant growth and development but also in connection with abiotic stress responses in *O. sativa*, wheat, and brachypodium (Arora et al., 2007; Wei et al., 2014; Ma et al., 2017). When the interacting partners of these genes were explored in the networks under study, we found that they interact with other TFs and stress-responsive genes (such as NAC, HSF, SNF1, WRKY, bHLH, PP2C, chaperones, etc.) having a significant role in multiple abiotic stress. Further SCC analysis also revealed that the MIKC_MADS TF family genes are present, along with other abiotic stress-responsive TFs in the strongly connected component of the networks under all three stresses (**Supplementary Table 10**). This shows they may play a significant role in managing timely floral development under multiple stress conditions in *O. sativa*.

APRegNet Database

Contemporary investigation of literature about transcription factors and miRNAs mediated regulation associated with plant abiotic stresses indicate that transcription factors and miRNAs act as key regulators in plant physiological adaption mechanisms during the response to various stress conditions. Information regarding the transcriptional and post-transcriptional regulation by TFs and miRNAs during plant responses to abiotic stress is distributed over numerous

recent studies. The APRegNet database (see text footnote 1) construction provides a comprehensive repository of abiotic stress-responsive transcriptional and post-transcriptional regulatory networks.

This database contains knowledge-based abiotic stress-specific regulatory networks in *A. thaliana*, *O. sativa*, and *Z. mays*, and was developed by incorporating various data sources. It is a comprehensive collection of the interactions among TFs, miRNAs, and genes, occurring in response to abiotic stress, reconstructed for public access. The established regulatory networks from APRegNet provide genome-wide regulatory interactions that lay an initial foundation and establish a prior background network to identify or verify molecular and functional regulations in pathways.

Within a plant species under multiple abiotic stresses, certain miRNAs or transcription factors show common expression. The study of such miRNAs and TFs using a systematic, comprehensive database approach like APRegNet (see text footnote 1), will expedite the enhancement of understanding of the stress-specific transcriptional and post-transcriptional regulatory role of transcription factors and miRNAs and their functional evolutionarily relationship in various plant species. The regulatory information stored in this database has promising uses for experimental biologists who intend to improve plant crop performance under multiple Abiotic stress environments.

We also include additional information such as links to other databases (Uniprot, TAIR, MSU7, NCBI Gene database, and MaizeGDB). We are planning to extend the APRegNet to include other plant species and capture information about more abiotic stress responses.

CONCLUSION

A complex interwoven network of multiple molecular species at various steps such as- transcription, post-transcription, translation, and post-translation control the expression of a gene. The alteration in expression of the gene in response to environmental cues enables the plant to acclimatize and survive in adverse environments, but this complex interwoven regulatory network is mostly unrevealed. In the present study, abiotic stress-responsive miRNA-TF-gene regulatory networks for *O. sativa* were reconstructed and analyzed to reveal information about how environmental stress induces these regulatory networks. Network structural measures were studied using the network theoretical approach, and deciphered several important features of the networks such as- scale-free, and small-world behavior that established the stable nature and quick, efficient transmission of signals (environmental cue). This process also highlighted the genes, TFs, and miRNAs shared by multiple stresses, working as a hub or bottleneck for signal propagation.

In crop plants, the right flowering time is pivotal for acclimatizing under the altered environmental condition and directly linked to grain yield (Gao et al., 2014). Under stress conditions, if flowering occurs prematurely, then seed set and grain-filling are adversely affected. If flowering is delayed, then there is the chance that plants may die without producing seeds. This study showed the interaction between stress-responsive genes (TFs/miRNAs/other genes) and genes involved in floral development pathways. However, this opens up further questions about how the fine-tuning feedback mechanism works in balancing stress tolerance and timely flowering to survive adverse conditions.

In summary, this study postulated some structural features of the miRNA and transcription factor regulated networks operating in a cell under the influence of multiple environmental stresses and prioritize the mainstream function of miRNAs and transcription factors in gene expression control.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

RS, SU, SB, and AS conceived and designed the study. RS and SU performed the analysis. RS wrote the manuscript. All authors read and approved the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

AS acknowledges support from Shiv Nadar University, India in providing infrastructure facilities and CSIR for providing financial support in the form of monthly stipend to RS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.618089/full#supplementary-material

Supplementary Figure 1 | Degree distribution plot

Supplementary Figure 2 | Venn diagram for up and down regulated differentially expressed **(A)** mRNA and **(B)** miRNAs

Supplementary Table 1 | Details of transcriptome raw data.

Supplementary Table 2 | List of differentially expressed genes under cold, drought, and salt stress in *Oryza sativa*.

Supplementary Table 3 | miRNA-seq data analysis result.

Supplementary Table 4 | List of differentially expressed miRNAs under cold, drought, and salt in *Oryza sativa*.

Supplementary Table 5 | AGO1-enriched miRNA differentially expressed under abiotic stresses.

Supplementary Table 6 | List of miRNAs targets predicted by psRNATarget.

Supplementary Table 7 | List of miRNAs targets identified by degradome sequencing data analysis.

Supplementary Table 8 | List of miRNAs targets validated by degradome sequencing data analysis results.

Supplementary Table 9 | Top 10 genes/TF/miRNA ranked by degree, betweenness, and closeness centrality in *Oryza sativa* abiotic stress-induced regulatory networks.

Supplementary Table 10 | Details of SCC found in three *Oryza sativa* abiotic stress-responsive miRNA-TF-gene regulatory networks.

Supplementary Table 11 | Scale free network analysis.

REFERENCES

Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25, 130–131. doi: 10.1093/bioinformatics/btn604

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382. doi: 10.1038/35019019

Arora, R., Agarwal, P., Ray, S., Singh, A. K., Singh, V. P., Tyagi, A. K., et al. (2007). MADS-box gene family in rice : genome-wide identification, development and stress. *BMC Bioinformatics* 8:242. doi: 10.1186/1471-2164-8-242

Arribas-Hernández, L., Kielpinski, L. J., and Brodersen, P. (2016). mRNA decay of most arabidopsis miRNA targets requires slicer activity of AGO1. *Plant Physiol.* 171, 2620–2632. doi: 10.1104/pp.16.00231

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 308, 639–641.

Baumberger, N., and Baulcombe, D. C. (2005). Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11928–11933. doi: 10.1073/pnas.0505461102

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. *Front. Genet.* 7:24. doi: 10.3389/fgene.2016.00024

Bollobás, B., and Cockayne, E. J. (1979). Graph-theoretic parameters concerning domination, independence, and irredundance. *J. Graph Theory* 3, 241–249. doi: 10.1002/jgt.3190030306

Chen, D., Yan, W., Fu, L.-Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. *Nat. Commun.* 9:4534. doi: 10.1038/s41467-018-06772-3

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law distributions in empirical data. *Soc. Ind. Appl. Math.* 51, 661–703. doi: 10.1137/070710111

Dai, X., and Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* 39, W155–W159. doi: 10.1093/nar/gkr319

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., et al. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* 39, 1033–1037. doi: 10.1038/ng2079

Fu, C., Sunkar, R., Zhou, C., Shen, H., Zhang, J. Y., Matts, J., et al. (2012). Overexpression of miR156 in switchgrass (*Panicum virgatum* L.) results in various morphological alterations and leads to improved biomass production. *Plant Biotechnol. J.* 10, 443–452. doi: 10.1111/j.1467-7652.2011.00677.x

Gao, H., Jin, M., Zheng, X.-M., Chen, J., Yuan, D., Xin, Y., et al. (2014). Days to heading 7, a major quantitative locus determining photoperiod sensitivity and regional adaptation in rice. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16337–16342. doi: 10.1073/pnas.1418204111

German, M. A., Luo, S., Schroth, G., Meyers, B. C., and Green, P. J. (2009). Construction of parallel analysis of RNA ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat. Protoc.* 4, 356–362. doi: 10.1038/nprot.2009.8

Ghoshal, G., and Barabási, A.-L. (2011). Ranking stability and super-stable nodes in complex networks. *Nat. Commun.* 2:394. doi: 10.1038/ncomms1396

Grover, A., and Minhas, D. (2000). Towards production of abiotic stress tolerant transgenic rice plants: issues, progress and future research needs. *Proc. Indian Natl. Sci. Acad.* 66, 13–32.

Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J. M., and Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* 37, W68–W76. doi: 10.1093/nar/gkp347

Haque, S., Ahmad, J. S., Clark, N. M., Williams, C. M., and Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. *Curr. Opin. Plant Biol.* 47, 96–105. doi: 10.1016/j.pbi.2018.10.005

Hobert, O. (2008). Gene regulation by transcription factors and microRNAs. *Science* 319, 1785–1786. doi: 10.1126/science.1151651

Hong, F., Breitling, R., Mcentee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). RankProd : a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 22, 2825–2827. doi: 10.1093/bioinformatics/btl476

Humphries, M. D., and Gurney, K. (2008). Network "small-world-ness": a quantitative method for determining canonical network equivalence. *PLoS One* 3:e0002051. doi: 10.1371/journal.pone.0002051

Jeong, J., and Berman, P. (2008). On cycles in the transcription network of Saccharomyces cerevisiae. *BMC Syst. Biol.* 2:12. doi: 10.1186/1752-0509-2-12

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Karakülah, G., Kurtoglu, K. Y., and Unver, T. (2016). PeTMbase: a database of plant endogenous target mimics (eTMs). *PLoS One* 11:e0167698. doi: 10.1371/journal.pone.0167698

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics — a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416. doi: 10.1093/bioinformatics/btn647

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-1

Kim, J. Y., Kwak, K. J., Jung, H. J., Lee, H. J., and Kang, H. (2010). MicroRNA402 affects seed germination of arabidopsis thaliana under stress conditions via targeting DEMETER-LIKE Protein3 mRNA. *Plant Cell Physiol.* 51, 1079–1083. doi: 10.1093/pcp/pcq072

Kissoudis, C., van de Wiel, C., Visser, R. G., and van der Linden, G. (2014). Enhancing crop resilience to combined abiotic and biotic stress through the dissection of physiological and molecular crosstalk. *Front. Plant Sci.* 5:207. doi: 10.3389/fpls.2014.00207

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, D68–D73. doi: 10.1093/nar/gkt1181

Lesk, C., Rowhani, P., and Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature* 529, 84–87. doi: 10.1038/nature16467

Li, C., Wang, Y., Xu, L., Nie, S., Chen, Y., Liang, D., et al. (2016). Genome-wide characterization of the MADS-box gene family in radish (*Raphanus sativus* L.) and assessment of its roles in flowering and floral organogenesis. *Front. Plant Sci.* 07:1390. doi: 10.3389/fpls.2016.01390

Li, C., and Zhang, B. (2016). MicroRNAs in control of plant development. *J. Cell. Physiol.* 231, 303–313. doi: 10.1002/jcp.25125

Li, W.-X., Oono, Y., Zhu, J., He, X.-J., Wu, J.-M., Iida, K., et al. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and post transcriptionally to promote drought resistance. *Plant Cell* 20, 2238–2251. doi: 10.1105/tpc.108.059444

Londo, J. P., Chiang, Y.-C., Hung, K.-H., Chiang, T.-Y., and Schaal, B. A. (2006). Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9578–9583. doi: 10.1073/pnas.0603152103

Ma, J., Yang, Y., Luo, W., Yang, C., Ding, P., Liu, Y., et al. (2017). Genome-wide identification and analysis of the MADS-box gene family in bread wheat (*Triticum aestivum* L.). *PLoS One* 12:e0181443. doi: 10.1371/journal.pone.0181443

Mantri, N., Patade, V., Penna, S., Ford, R., and Pang, E. (2012). "Abiotic stress responses in plants: present and future BT - abiotic stress responses in plants: metabolism, productivity and sustainability," in *Abiotic Stress Responses in Plants*, eds P. Ahmad and M. N. V. Prasad (New York, NY: Springer New York), 1–19. doi: 10.1007/978-1-4614-0634-1_1

Myers, S. S., Smith, M. R., Guth, S., Golden, C. D., Vaitla, B., Mueller, N. D., et al. (2017). Climate change and global food systems: potential impacts on food security and undernutrition. *Annu. Rev. Public Health* 38, 259–277. doi: 10.1146/annurev-publhealth-031816-044356

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* 45, 167–256. doi: 10.1137/s003614450342480

Normile, D. (2008). Reinventing rice to feed the world. *Science* 321, 330–333. doi: 10.1126/science.321.5887.330

O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol. (Lausanne).* 9:402. doi: 10.3389/fendo.2018.00402

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006). AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol.* 140, 818–829. doi: 10.1104/pp.105.072280

Samad, A. F. A., Sajad, M., Nazaruddin, N., Fauzi, I. A., Murad, A. M. A., Zainal, Z., et al. (2017). MicroRNA and transcription factor : key players in plant regulatory network. *Front. Plant Sci.* 8:565. doi: 10.3389/fpls.2017.00565

Shao, H., Wang, H., and Tang, X. (2015). NAC transcription factors in plant multiple abiotic stress responses : progress and prospects. *Front. Plant Sci.* 6:902. doi: 10.3389/fpls.2015.00902

Sharma, R., Upadhyay, S., Bhat, B., Singh, G., Bhattacharya, S., and Singh, A. (2020). Abiotic stress induced miRNA-TF-gene regulatory network: a structural perspective. *Genomics* 112, 412–422. doi: 10.1016/j.ygeno.2019.03.004

Silva, C. S., Puranik, S., Round, A., Brennich, M., Jourdain, A., Parcy, F., et al. (2016). Evolution of the plant reproduction master regulators LFY and the MADS transcription factors: the role of protein structure in the evolutionary development of the flower. *Front. Plant Sci.* 6:1193. doi: 10.3389/fpls.2015.01193

Stief, A., Altmann, S., Hoffmann, K., and Pant, B. D. (2014). Arabidopsis miR156 regulates tolerance to recurring environmental stress through SPL transcription factors. *Plant Cell* 26, 1792–1807. doi: 10.1105/tpc.114.123851

Sun, Y., and Dinneny, J. R. (2018). Q & A : how do gene regulatory networks control environmental responses in plants? *BMC Biol.* 16:38. doi: 10.1186/s12915-018-0506-7

Szcześniak, M. W., and Makabowska, I. (2014). miRNEST 2 . 0 : a database of plant and animal microRNAs. *Nucleic Acids Res.* 42, 74–77. doi: 10.1093/nar/gkt1156

Theissen, G., and Melzer, R. (2007). Molecular mechanisms underlying origin and diversification of the angiosperm flower. *Ann. Bot.* 100, 603–619. doi: 10.1093/aob/mcm143

Upadhyay, S., Roy, A., Ramprakash, M., Idiculla, J., Kumar, A. S., and Bhattacharya, S. (2017). A network theoretic study of ecological connectivity in Western Himalayas. *Ecol. Model.* 359, 246–257. doi: 10.1016/j.ecolmodel.2017.05.027

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669–687. doi: 10.1016/j.cell.2009.01.046

Wang, H., Wang, H., Shao, H., and Tang, X. (2016). Recent advances in utilizing transcription factors to improve plant abiotic stress tolerance by transgenic technology. *Front. Plant Sci.* 7:67. doi: 10.3389/fpls.2016.00067

Wei, B., Zhang, R., Guo, J., Liu, D., Li, A., Fan, R., et al. (2014). Genome-wide analysis of the MADS-box gene family in *Brachypodium distachyon*. *PLoS One* 9:e84781. doi: 10.1371/journal.pone.0084781

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., et al. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *Plant Cell* 28, 2365–2384. doi: 10.1105/tpc.16.00158

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* 99, 909–917. doi: 10.1198/016214504000000683

Zahn, L. M., Feng, B., and Ma, H. (2006). "Beyond the ABC—model:regulation of floral homeotic genes," in *Developmental Genetics of the Flower*, eds D. Soltis, J. Leebens-Mack, P. Soltis, and J. A. Callow (Cambridge, MA: Academic Press), 163–207. doi: 10.1016/S0065-2296(06)44004-0

Zhang, B. (2015). MicroRNA: a new target for improving plant tolerance to abiotic stress. *J. Exp. Bot.* 66, 1749–1761. doi: 10.1093/jxb/erv013

Zhang, S., Yue, Y., Sheng, L., Wu, Y., Fan, G., Li, A., et al. (2013). PASmiR : a literature-curated database for miRNA molecular regulation in plant response to abiotic stress. *BMC Plant Biol.* 13:33. doi: 10.1186/1471-2229-13-33

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 February 2021
doi: 10.3389/fgene.2021.617505

Gene-Microbiome Co-expression Networks in Colon Cancer

Irving Uriarte-Navarrete[1], *Enrique Hernández-Lemus*[1,2*] *and Guillermo de Anda-Jáuregui*[1,2,3*]

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Conacyt Research Chairs, National Council on Science and Technology, Mexico City, Mexico

It is known that cancer onset and development arise from complex, multi-factorial phenomena spanning from the molecular, functional, micro-environmental, and cellular up to the tissular and organismal levels. Important advances have been made in the systematic analysis of the molecular (mostly genomic and transcriptomic) within large studies of high throughput data such as The Cancer Genome Atlas collaboration. However, the role of the microbiome in the induction of biological changes needed to reach these pathological states remains to be explored, largely because of scarce experimental data. In recent work a non-standard bioinformatics strategy was used to indirectly quantify microbial abundance from TCGA RNA-seq data, allowing the evaluation of the microbiome in well-characterized cancer patients, thus opening the way to studies incorporating the molecular and microbiome dimensions altogether. In this work, we used such recently described approaches for the quantification of microbial species alongside with gene expression. With this, we will reconstruct bipartite networks linking microbial abundance and gene expression in the context of colon cancer, by resorting to network reconstruction based on measures from information theory. The rationale is that microbial communities may induce biological changes important for the cancerous state. We analyzed changes in microbiome-gene interactions in the context of early (stages I and II) and late (stages III and IV) colon cancer, studied changes in network descriptors, and identify key discriminating features for early and late stage colon cancer. We found that early stage bipartite network is associated with the establishment of structural features in the tumor cells, whereas late stage is related to more advance signaling and metabolic features. This functional divergence thus arise as a consequence of changes in the organization of the corresponding gene-microorganism co-expression networks.

Keywords: colorectal cancer, microbiome, tumor progression, probabilistic multilayer networks, information theory

OPEN ACCESS

Edited by:
Maud Fagny, UMR7206 Eco Anthropologie et Ethnobiologie (EAE), France

Reviewed by:
Joseph Nathaniel Paulson, Dana-Farber Cancer Institute, United States
Xiaowei Zhan, University of Texas Southwestern Medical Center, United States

***Correspondence:**
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx
Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx

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

Received: *14 October 2020*
Accepted: *22 January 2021*
Published: *15 February 2021*

Citation:
Uriarte-Navarrete I, Hernández-Lemus E and de Anda-Jáuregui G (2021) Gene-Microbiome Co-expression Networks in Colon Cancer. Front. Genet. 12:617505. doi: 10.3389/fgene.2021.617505

INTRODUCTION

Colon cancer is consistently ranked among the top five contributors to cancer deaths worldwide (Bray et al., 2018). Its incidence and mortality are rapidly rising in developing countries, possibly influenced by changes in lifestyle and socioeconomic conditions. It is expected that this trend will actually further increase according to recent studies (Arnold et al., 2017).

As with many other cancers, colon cancer is known to have a genetic component as well as environmental factors which further modulate or increase the risks. Its molecular determinants include genomic, regulatory, and epigenomic components (Raskov et al., 2020) whereas the environmental component is also multifactorial, ranging from toxicological exposure (Fernández-Martínez et al., 2020), physical activity (Friedenreich et al., 2020), dietary habits and more. A more recent factor that is an important research topic is the role that microbiome interactions may be playing at the molecular and patho-physiological levels.

Recent findings have pointed out to different, sometimes disparate phenomena, such as the influence of bacterial protein toxins (Fiorentini et al., 2020), altered microbiome composition (Xu et al., 2020), and the non-rational use of antibiotics (Simin et al., 2020). Among these, microbome-host interactions are hypothesized to modulate and integrate these diverse signals (Yang et al., 2020). For instance, experimental evidence has been found for functional alterations mediated by microorganisms involved in colon cancer progression (Yu et al., 2020). It is currently accepted that these complex biomolecular and organismal interactions can be better understood using a systems biology approach (Peñalver Bernabé et al., 2018).

In the context of oncology, network biology has proven to be a powerful tool for the integration of multiple high throughput technologies (de Anda-Jáuregui and Hernández-Lemus, 2020). Networks provide flexible frameworks to represent the relevant physio-pathological interactions present in the tumor environments. For instance, bipartite networks have been used to represent gene expression control by micro-RNAs; a strategy that allows not only to describe statistical associations, but also to identify putative functional associations (de Anda-Jáuregui et al., 2018, 2019).

In this work, we reconstructed bipartite networks that capture the statistical dependence between microorganism abundance and gene expression in *early* (stages I and II) and *late* (stages III and IV) colon cancer, using data from The Cancer Genome Atlas (TCGA). We analyze these networks to identify changes in the relative relevance of microorganisms between these conditions, in terms of their topological role in their respective networks. We analyzed genes associated to the highest ranked microorganisms in each network as a means to identify changes in associated biological functions. This work hence aims to provide novel insights into microorganism-mediated functional alterations potentially involved in colon cancer progression.

MATERIALS AND METHODS

For this work, we collected gene expression data from TCGA, along with microorganism quantification data that was generated by Poore et al. (2020), for the same 269 samples. We classified these samples into *early* ($n = 150$) and *late* ($n = 119$) colon cancer based on tumor stages as provided by TCGA metadata.

Interactions between each pair of measured microorganism and gene were detected using mutual information (MI) as a measure for statistical dependence. The highest ranked interactions were kept in order to reconstruct bipartite networks for each group. Downstream analyses included topological characterization and functional enrichment analysis. In **Figure 1**, we present a schematic representation of our analysis pipeline.

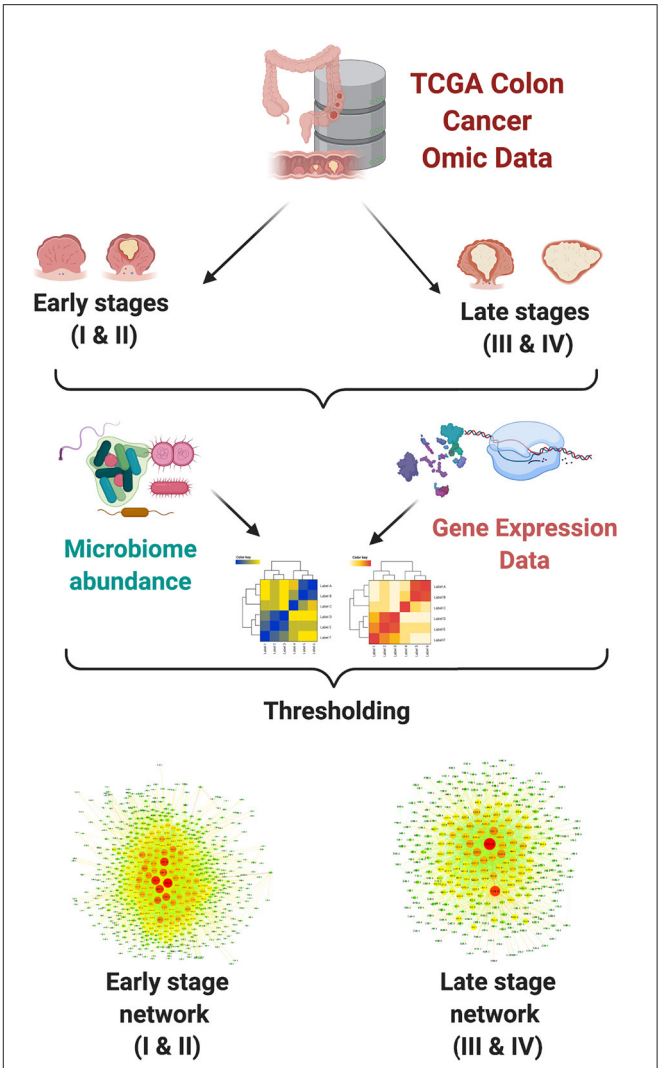

FIGURE 1 | Network analysis pipeline.

Gene Expression Data

We used data from TCGA, obtained through the Genome Data Commons portal. We used level three pre-processed gene expression data; the full analysis pipeline is documented at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/; briefly, RNA-seq data is aligned using *STAR* (Dobin et al., 2012), and reads mapped to each gene are counted using *HT-SEQ* (Anders et al., 2014); Read counts are normalized using the Fragments per Kilobase of transcript per Million mapped reads (FPKM) calculation, which divides counts by the gene length and the total number of reads mapped to protein-coding genes.

Based on the available metadata, samples with tumor stages I and II were grouped as early colon cancer, while samples with tumor stages III and IV were grouped as late colon cancer. Due to some samples being discarded from the microbiome quantification pipeline by the original authors (Poore et al., 2020, see next section), we ended up using 137 early stage and 64 late stage samples (see **Supplementary File 1** for the TCGA identificators of the used samples).

Microorganism Abundance Data

We used the public dataset generated in Poore et al. (2020) as our source for microorganism abundance data. Briefly, in said work the authors were able to quantify microorganism abundance in TCGA tumor samples via a novel bioinformatics approach. Briefly, They took raw whole genome sequencing (WGS) data and analyzed the nearly 0.9% of total sequencing reads were classified as non-human and assigned to bacteria, archaea, or viruses at the genus level using *Kraken* (Wood and Salzberg, 2014); which matches k-mers to taxa in a reference database. Normalization was performed considering sample number within a cancer type and sample type. To correct for batch effects, discrete taxonomical counts are converted to log-counts per million per sample using *Voom* (Law et al., 2014), and a secondary supervised normalization was performed to remove significant batch effects. Additionally, contamination concerns were addressed using the Bayesian source tracking model SourceTracker2 (Knights et al., 2011). Based on their quantification, we crossed microorganism abundance and gene expression data at the aliquot level, to ensure biological comparability between the datasets.

Microbiome-Gene Co-expression Quantification

Having matched gene expression and microorganism abundance data organized into *expression matrices*, we calculated mutual information for each pair of *microorganism* × *gene*. Mutual information is the maximum likelihood information theoretic measure of statistical dependence. Since it is capable to capture non-linear relationships between features, it has been successfully used for gene co-expression network reconstruction (de Anda-Jáuregui et al., 2016; He et al., 2017). It has also been previously used for bipartite network reconstruction of multiomic data (de Anda-Jáuregui et al., 2018, 2019). In this work, we calculated MI using the `infotheo` package in R.

Once MI values were calculated, we selected those interactions above the 99.5 quantile to be considered as links on a bipartite network: $\mathcal{B}(microorganism, gene)$; A bipartite graph (or bigraph) is a network whose nodes can be divided into two disjoint sets U and V such that each link connects a U-node to a V-node. Importantly, no links are found between two nodes belonging to the same set (Barabási et al., 2016). For mutual information calculation, data is discretized using the equal frequency method (Meyer 2008), which assigns each observation to one of N bins, with N being the number of observations. The discretized vectors are then used as the input for proper mutual information calculation, using an entropy estimation of the empirical probability distribution. Both of these calculations were performed using the infotheo package for R.

For completeness, the reconstructed networks contained all measured microorganisms ($N = 4,450$) and protein-coding genes ($N = 16,593$), even if they do not participate in any link (that is, they have connectivity degree $k = 0$). The threshold was selected based on previous analyses of multi-omic bipartite networks (de Anda-Jáuregui et al., 2018, 2019); we must acknowledge that by using this threshold we guarantee fair comparisons between the reconstructed networks; however, the structure and composition of these networks will not be comparable to networks generated through other methods (including the selection of a different threshold).

Network Analyses

We characterized the topology of each of the generated using a combination of the `igraph` (Csardi and Nepusz, 2006) in R and `networkx` (Hagberg et al., 2008) in Python. Additionally, we used `Cytoscape` (Shannon, 2003) to generate network visualizations. In this work, we focused mainly on centrality measures including degree, bipartite clustering coefficient, and redundancy coefficients (Latapy et al., 2008). Comparisons between appropriate distributions were evaluated using the Kolmogorov-Smirnov test.

Functional Enrichment of High-Degree Microorganism Gene Neighborhoods

We analyzed the neighborhoods of the highest ranked microorganisms (based on their degree) to identify host biological functions associated to these microorganisms. To do so, we performed over-representation analysis (ORA) via FDR-corrected hypergeometric tests for biological processes and molecular functions (as annotated in the Gene Ontology database) using the `WebgestaltR` (Liao et al., 2019) package. Parameters for ORA considered the full genome as the reference set, and a false discovery rate (FDR) threshold of 0.05. It should be noted that the enrichment is performed over the set of *genes* that conform the *neighborhood* of each *microorganism*; this is to identify biological functions from the host that can be associated to microorganisms through their co-expressed genes (see **Figure 2**). We further used natural language processing tools from the `tm` package in R (Meyer et al., 2008) to compare identified functions and processes, by tokenizing their names and descriptions and identifying the most mentioned keywords or tokens.

RESULTS AND DISCUSSION

Microorganism-Gene Co-expression Networks Are Topologically Similar in Early and Late Colon Cancer

By studying bipartite networks, we wanted to know what are some possible ways in which the presence of microorganisms may affect the host's response (as proxied by changes in gene expression highly correlated with microbial abundance) and vice versa. Clues to this may be provided by the microbe-gene links.

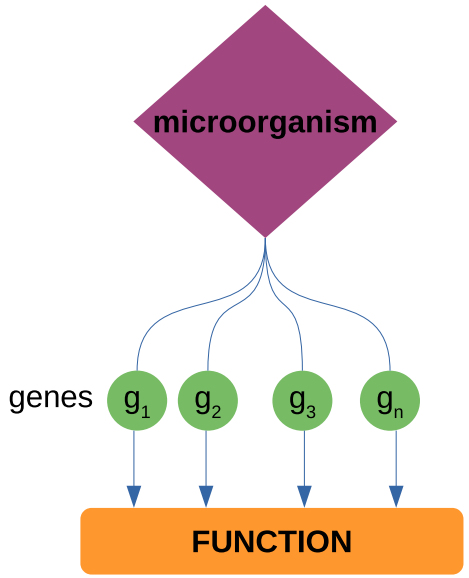

FIGURE 2 | Enrichment of host biological functions associated to microorganisms through their gene neighborhoods. Each microorganism has a set of neighbor genes in the bipartite network. This gene set is tested against a set of known biological functions (as annotated in the GO database) through the hypergeometric test. Through these procedure, we can associate known biological functions from the host to each of the measured microorganisms.

The reconstructed microorganism-gene co-expression networks for the early and late stages of colon cancer exhibit a similar global topology. They are both dominated by a giant connected component that contains all detected links. This giant connected component is composed of all measured microorganisms, and over 80% of measured genes. It should be noted that in the case of both genes and microorganisms, presence in the network is not directly correlated by the abundance in the original measurements, nor biased due to zero-inflation effects (see **Supplementary File 2**). **Figure 3** depicts these networks. **Table 1** presents the global topological features of these networks.

The bipartite degree distributions of these networks (seen in **Figure 4**) are quite similar between early and late stage. In this context, it is more informative to assess the degree distributions for each type of nodes (microorganisms and genes) separately. In this regard, we observe that in both networks, genes follow a heavy-tailed distribution (blue dots in **Figure 4**); that is, most genes are connected to few microorganisms, whereas a few genes are connected to many microorganisms. Meanwhile, microorganism nodes (red dots in **Figure 4**) exhibit a different pattern: a curve with no low-degree nodes; indicating that every detected microorganism has putative effects on the expression of a relatively large set of genes. In any case, the distributions for both genes and microorganisms are similar between early and late stages cancer networks.

We evaluated two other topological properties of the nodes in these networks: the clustering and redundancy coefficients (see **Figure 5**).

Network redundancy (sometimes called path degeneracy) is related to how many different paths or trajectories can be taken to go from one node to another. Unlike trees or loosely connected networks, complex networks (such as the ones discussed here) are characterized by being highly redundant. This means that there are multiple (sometimes many) different paths connecting two given nodes. For probabilistic networks this implies that the Markov blanket (the subset of the network with the useful connectivity information) spans much of the network. This in turn implies that to break up (percolate, in technical terms) the network to pieces, one must remove a large number of links. In the case of bipartite networks, the concept of redundancy has to be adapted, since neighborhood overlaps correspond to links obtained in several ways during projection which are not distinguishable. Then redundancy is caused by nodes that when removed from the bipartite graph, do not cause significant changes in the projection (Latapy et al., 2008).

The clustering coefficient is a quantitative measure of the tendency of nodes in a graph to cluster together. It is calculated for a node (local clustering coefficient), as the ratio of the number of "triangles" (technically "closed triplets") formed by links connected to this node, to all possible triangles that can be formed with this node and its immediate neighbors. The global clustering coefficient is a network quantity, which is indeed the average of the local clustering coefficient of all the nodes in connected components of the network. In the case of clustering coefficients for bipartite networks, these measure the probability that given four nodes with three links, they are actually all connected with four links (all the possible links in a bipartite configuration of four nodes) (Latapy et al., 2008).

In bipartite networks, these are measures of the contribution of a given node to the connectivity of nodes of the opposite type (Latapy et al., 2008). We observe that in the case of microorganisms (red curves in **Figure 5**), these exhibit low values: this indicates that there is no single microorganism through which most genes could interact. Meanwhile, genes (blue curves in **Figure 5**) exhibit higher values, meaning that gene-mediated connections between microorganisms are, on average, more likely to be redundant. **Table 2** shows the statistical differences between the evaluated distributions.

Despite these overall similarities, networks for early and late colon cancer exhibit notable differences in terms of their connections. Although the composition of the GCC is fundamentally similar in terms of the microorganisms and genes found in it, the way in which this are connected is completely dissimilar, with a Jaccard similarity for edges of only 0.28%.

This differences in connectivity in turn explain the different degree ranking of both microorganisms and genes. The ranked list of microorganisms and genes show poor correlation between the early and late stages (Spearman ρ of 0.015 for microorganisms and 0.269 for genes). Due to these differences, the highest ranked microorganisms are (a) different in the early and late stages of colon cancer and (b) have a different set of associated genes. With this in mind, we explored how these facts change the set of host biological functions associated to the most connected microorganisms.

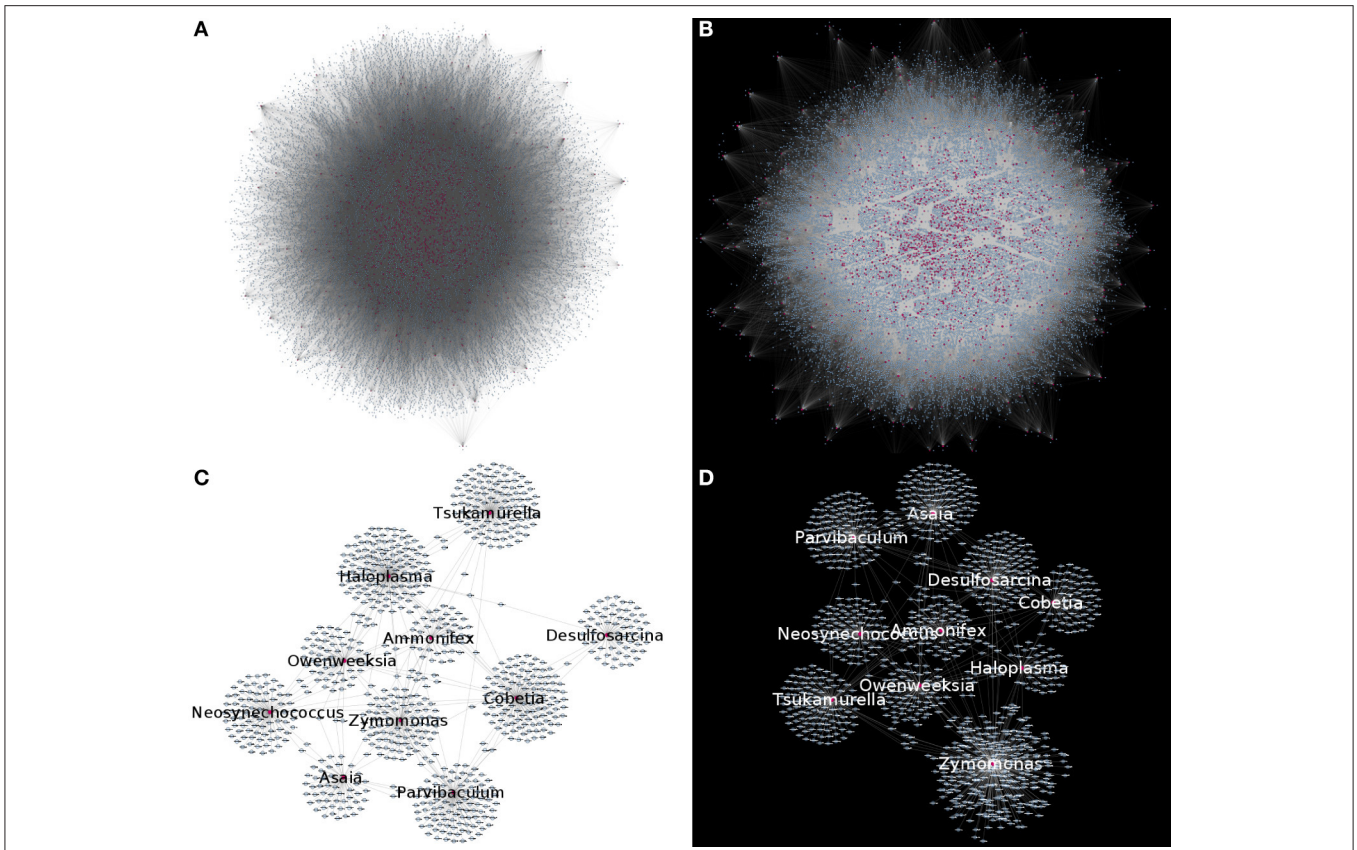

FIGURE 3 | Microorganism-gene co-expression networks for early **(A)** and late **(B)** colon cancer. In this visualizations, microorganisms are colored purple and genes are colored blue. Nodes with degree $k = 0$ are removed for visualization purposes, highlighting that in both networks, connected nodes form a single giant component. **(C,D)** Show a subset of the early **(C)** or late **(D)** networks, highlighting the most connected microorganisms.

TABLE 1 | General network descriptors.

	Early		Late
Genes ($k > 0$)	16,593		17,535
Microorganisms ($k > 0$)		1,464	
Edges	143,320		143,321
Giant connected component?		Yes	
GCC** size	18,057		18,999
GCC** node similarity*		91.79%	
Edge similarity*		0.28%	

*Similarity expressed as percentualized Jaccard index.
**GCC, giant connected component.

Regarding microorganisms, **Tables 3**, **4** present the top 10 highly connected microorganism (at the genus level) in the gene microorganism bipartite networks for early and late stage colon cancer, respectively.

By examining **Tables 3**, **4**, it may be surprising that most of the microbial species themselves have not been reported to be related with the onset and development of colon cancer. This of course may be explained by the fact that systematic high-throughput studies of the relationship between cancer and microbial dysbiosis are indeed still being developed. So the absence of evidence may not (yet) be taken as evidence of absence. However, in the next subsection we will see how, even though the organisms themselves may not sound that familiar, the statistically dependent gene neighborhoods of such microorganisms will recapitulate relevant functional features known in the biology of colon cancer.

Host Biological Functions Associated to Highly Connected Microorganisms Change With Colon Cancer Progression

We set to identify functions that could be linked to microorganisms detected in the early and late stage tumors. Since there is no annotation of human biological functions associated to microorganisms, we performed ORA on the gene neighborhoods of the top 10 highest ranked microorganisms by degree, searching for enrichment of biological processes and molecular functions annotated in Gene Ontology.

Enrichment Results for Biological Processes

The biological processes branch of the Gene Ontology is devoted to biologically relevant functional processes, some of these

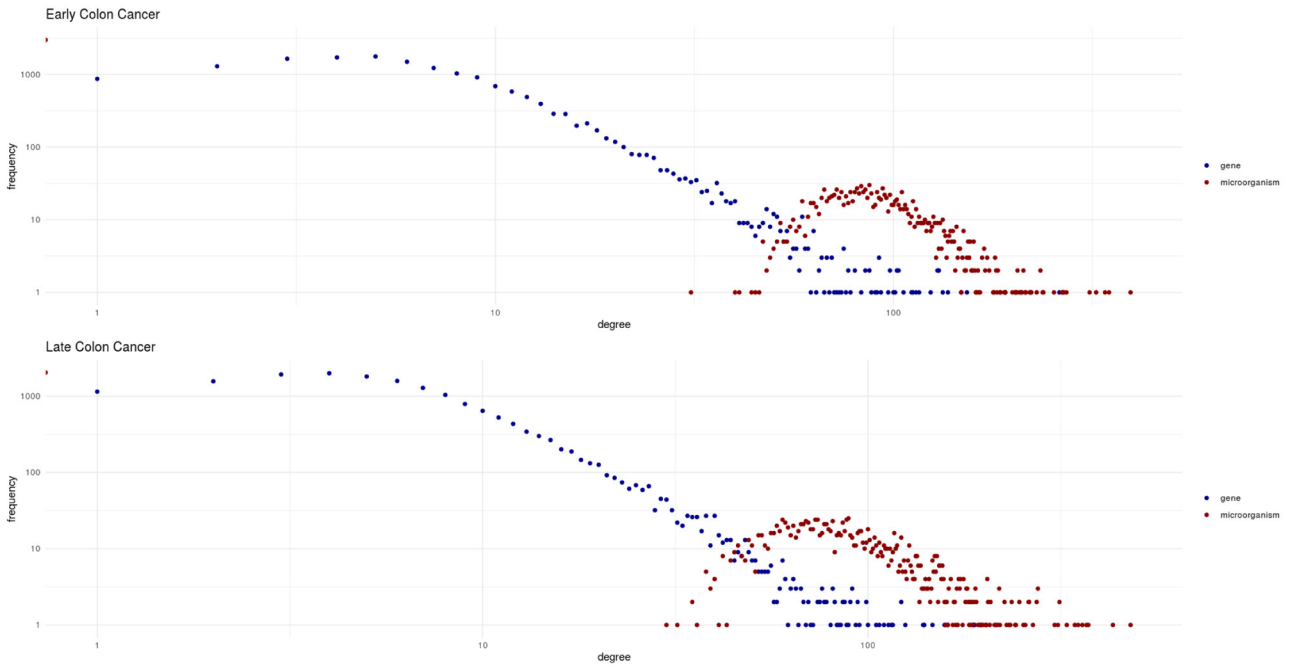

FIGURE 4 | Degree distributions for the early and late colon cancer networks. Values for microorganisms are shown in red, and values for genes are shown in blue. Notice how genes exhibit a heavy-tailed distribution, whereas a different behavior is observed for microorganisms in both networks.

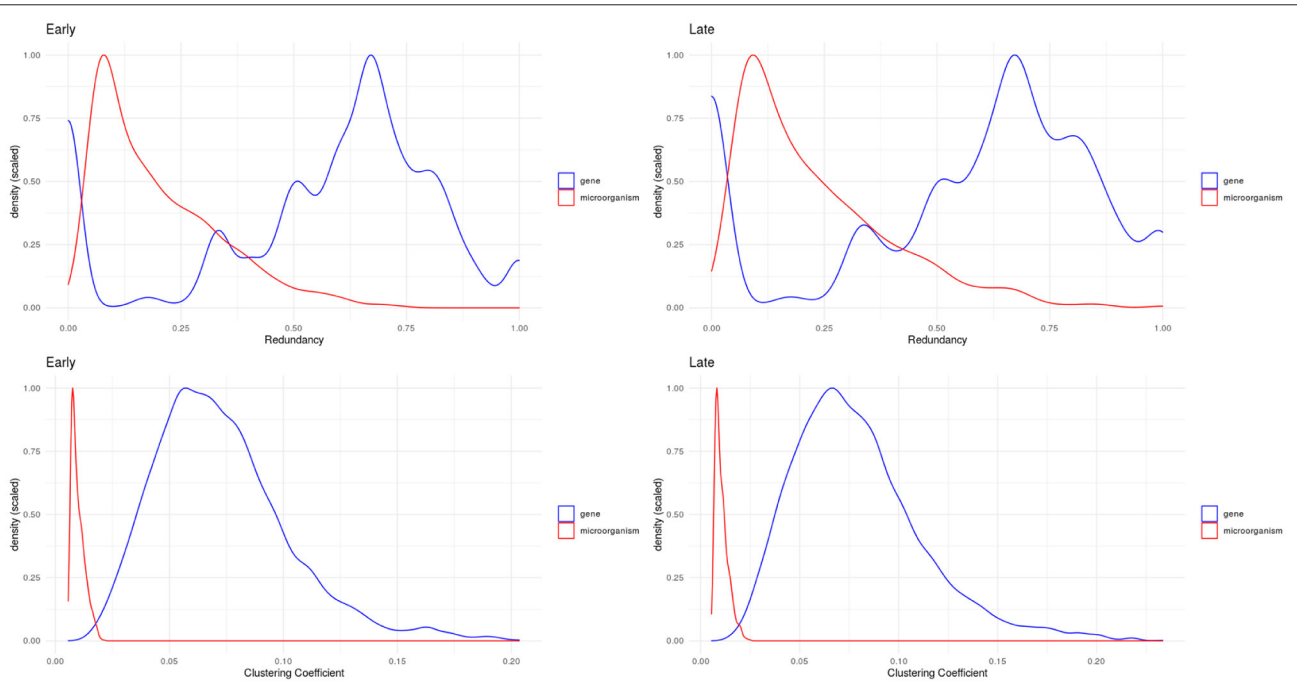

FIGURE 5 | Density plots for redundancy **(top)** and clustering **(bottom)** coefficients. We can see how microorganisms (red lines) are significantly less redundant and clustered than genes within these networks.

have clearly understood biomolecular mechanisms and some others are yet to be fully dissected. However, they allow for an advancement in our understanding of the molecular and cellular physiology behind gene and protein interactions.

Statistically enriched biological processes may represent functional processes in which the host-microbiome interactions are manifested. As we will see, some of these actually correspond to well-known hallmarks of cancer.

TABLE 2 | Distribution comparison (Kolmogorov-Smirnov test).

p-value, KS-test	Redundancy	Clustering coefficient
Microorganisms	6.037e-05	0.01244
Genes	0.01017	0.01838

TABLE 3 | Early colon cancer: top 10 highest ranked microorganism by degree.

Genus	Connectivity degree
Ilumatobacter	394
Rhodospirillum	348
Nitrosospira	340
Pontibacter	323
Shinella	311
Phaseolibacter	272
Vogesella	268
Azospirillum	267
Rubrivivax	265
Thermodesulfovibrio	253

TABLE 4 | Late colon cancer: top 10 highest ranked microorganism by degree.

Genus	Connectivity degree
Desulfurella	480
Nitriliruptor	432
Jeotgalicoccus	373
Actinocatenispora	369
Cryocola	360
Dactylosporangium	351
Pelomonas	344
Rhodovulum	328
Zymomonas	314
Methylomonas	314

In **Figures 6**, **7**, we present the results of these enrichment analyses as a heatmap. Notice that even if we performed the analyses for the 10 highest ranked microorganisms, only five genus were significantly associated to functions through their gene neighborhoods in each network.

Notably, higher enrichment values (in terms of FDR) are found in the early stage (**Figure 6**) than in the late stage (**Figure 7**). The interpretation is that biological functions are perhaps better mapped to the gene neighborhoods in the early colon cancer network—possibly indicating a more coordinated response to these microorganisms.

We identified only two biological processes appearing both in the early and late networks. These are *protein-containing complex localization* and *nuclear transport*. To better understand the functional differences identified, we tokenized the names of the detected biological processes and compared them between the early and late networks.

In **Figure 8**, we compare and contrast the terms associated to these biological processes. We observe in the early stages concepts associated to tumorigenesis such as *proliferation*, *biogenesis*, and (cell) *cycle*; as well as nucleic acids. Meanwhile, in the late stages, we observe terms that could be associated to late-stage cancer such as *migration* and *angiogenesis*. Concepts shared between both stages include *regulation*, *muscle*, and *protein*. For the full set of enrichment results, please refer to **Supplementary File 3**.

Enrichment Results for Molecular Functions

By recognizing that our understanding of the way microbiome-host interactions may be playing a role on the onset and development of cancer-associated biological processes is still quite incipient, we decided to also examine the molecular functions dimension of the Gene Ontology. This is so since molecular function refers to specific chemical and biochemical interactions of a more general nature that may be related to one, or more commonly to a large set of biological processes.

The rationale is that molecular species related to the entangled multi-microbial metabolism are possible interacting with the molecules involved in human (and in particular tumor and tumor micro-environment) cells.

Figures 9, **10** present the molecular function enrichment analysis for the early and late colon cancer networks. As in the case of biological processes, molecular functions are enriched on *different* microbial genus in the early and late stage networks. It is worth noticing that the more significant physico-chemical functions in the early stage correspond to structural features (particularly enriched for the gene-neighborhood of the Nitrosospira genus, see **Figure 9**) whereas the more enriched molecular functions in the late stage network corresponds to actin binding for genes in the network vicinity of the Pelomonas genus (**Figure 10**).

We can also notice in **Figure 10** that other microbial genuses' gene neighborhoods are highly enriched for molecular functions, such is the case of Jeotgallicoccus for actin binding, and to several types of oxido-reductase, as well as cytochrome-oxidase activity; and the case of Nitriliruptor for GTP-ase and nucleotide binding, and Desulphurella for ubiquitin and thyroid receptor activity.

As in the case of the Biological Processes enrichment analysis, **Figure 11** presents the results of natural language processing and tokenization of terms resulting in the statistically significant enrichment GO-categories. As it was mentioned, early stage molecular functions are somehow related to structural cellular features, whereas late stage are related to cellular metabolism and transport processes, being binding phenomena the common function at the intersection of both stage networks. For the full set of enrichment results, please refer to **Supplementary File 3**.

DISCUSSION

Topology of the Microbiome-Gene Co-expression Networks

Complex networks are characterized by their composition and global topological structure, that is by what are their elements and how are these connected in the networks. As presented

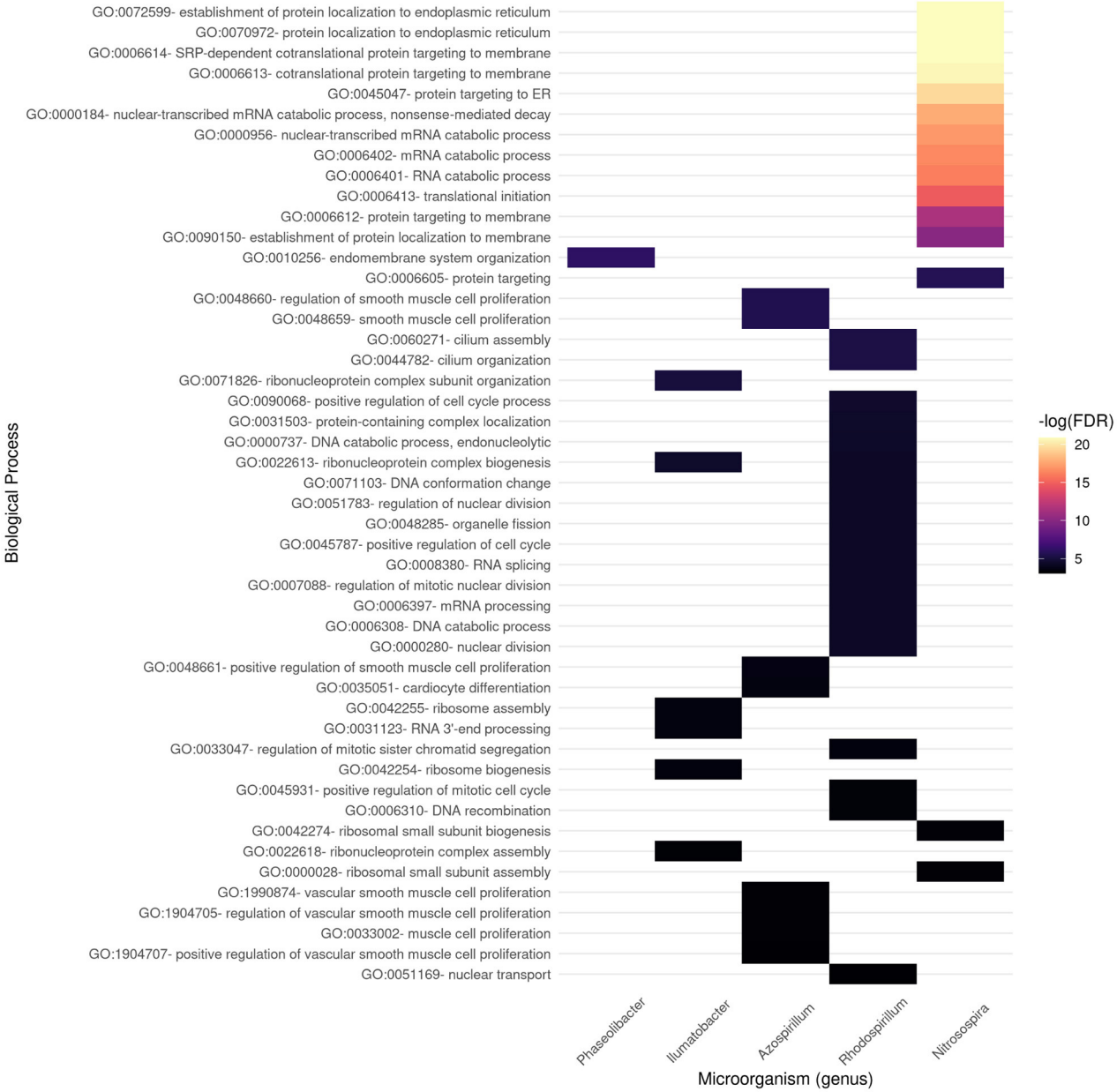

FIGURE 6 | Functional enrichment of highly connected microorganisms in the early colon cancer network—biological processes.

in **Figures 2–4** and **Table 1** in results, the global topological structures of early and late stage colorectal cancer bipartite networks are indeed quite similar. Approximately equal sizes in terms of number of nodes and edges. Similar size of their giant connected components and even a very high value of node similarity in their GCCs. However, as it can be seen in **Table 1** the edge similarity (a quantity proportional to the number of shared edges between the two networks) is actually extremely small (0.28%). This means that even if the elementary components of the networks (i.e., the genes and microorganisms) are almost the same and the global network features are so similar, the actual networks are indeed quite different, something unsurprising given that they represent two different biological scenarios.

Also noteworthy is the fact that by examining **Figure 4** we could notice that the two different types of nodes (genes and microorganisms) present striking differences in their degree connectivity probability distributions (blue dots representing genes and red dots microorganisms) and that the same patterns is observed for early and late stage colorectal cancer. The degree distributions for genes present long-tailed distributions that

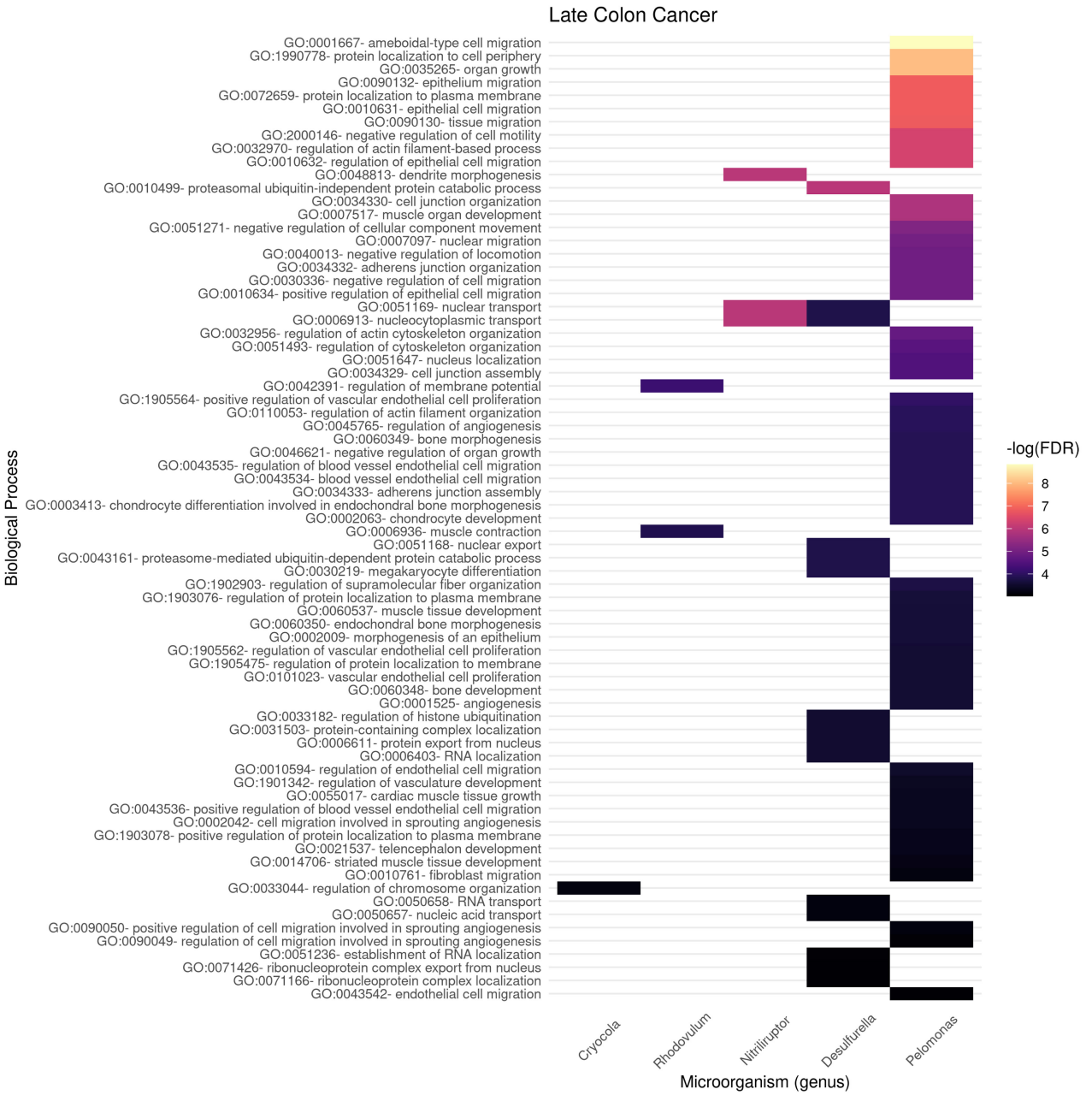

FIGURE 7 | Functional enrichment of highly connected microorganisms in the late colon cancer network—biological processes.

have been thoroughly characterized in complex biomolecular networks. In those long-tailed distributions one can notice how most genes have a relatively low number of connections whereas a few hub genes are densely connected in the networks.

Microorganisms, on the other hand present a rather different degree distribution scenario. In both networks, microorganisms show a more symmetric short-tailed distribution in which a most microorganisms are highly connected and present narrower variability in their connectivity degree. This difference perhaps represent that microbial communities somehow serve as integrating entities in the bipartite network. This, in turn, may be related with the low redundancy coefficients displayed by microorganisms in both networks as it can be seen in **Figure 5** (top row). Low redundancy of the specific microbial agents may prove later to have relevance for the design of microbiome-driven therapeutic strategies, though it is still very early to further speculate on this.

One relevant and complementary aspect to consider on the role that gene-microbial interactions may play can be glimpsed by looking at the probability density distributions for the clustering

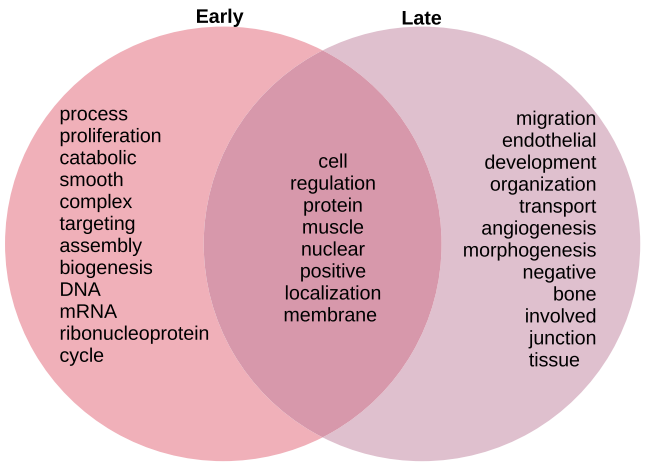

FIGURE 8 | Venn diagram of top 20 most mentioned concepts in biological processes associated to early and late colon cancer.

coefficient (**Figure 5** bottom row). We can see that in both networks (early and late stage) microorganisms present low values of clustering coefficient, whereas for genes there are wider probability distributions. Microorganisms are highly connected but not so-clustered. This in turn contributes to their being less redundant. This also may imply that the gene-microbiome co-expression program in the cancer networks is shaped by the full set of gene-microbial interactions and is not dominated by a few central players. This fact has been already discussed in the literature: physio-pathological phenomena related to microbial activity is, in general, influenced by microbiome dysbiosis rather than by the activity of a single or a few microorganisms.

Changes in Network Composition and Relative Importance

The latter points led us to discuss on how, even if the whole set of microorganisms is present in both, early and late stage colorectal cancer networks, their connectivity and importance in information processing within the networks vastly differ.

Consider **Tables 3**, **4**, for instance. There, we can see that the top 10 highly ranked microorganisms (that is, those with higher statistical dependencies and connectivity in the gene-microbial co-expression networks) are quite different. Indeed, no microorganism is present simultaneously at the top 10 of both networks, even at the, somewhat general, genus level presented here. This points out to a possible *reprogramming* of the gene-microbiome regulatory structure associated with the phenotypic differences between early and late stage colorectal cancer.

Regarding the highest ranked microorganisms associated to early stage colon cancer (**Table 3**), we have found that, in the case of Rhodospirillum, for instance, it is known to be able to produce molecules such as L-asparaginase which is a regulator of telomerase activity that has been found able to act on human cancer and immune cells (Zhdanov et al., 2017a,b; Plyasova et al., 2020). Nitrosospira is associated with processes related to ammonia oxidation (Kowalchuk and Stephen, 2001) in connection with colon cancer (Bingham et al., 1996; Bruce et al., 2000; Davis and Milner, 2009; O'keefe, 2016). Pontibacter has been found enriched in patients with gastric cancer and correlated with TNM severity (Dong et al., 2019).

In the case of Shinella, significant abundance has been found in mucosal associated microbiota in patients with severe irritable bowel syndrome (Li et al., 2018), and also is known to be involved in the production of N-nitrosonornicotine, a strong (group 1) carcinogen (Qiu et al., 2016). Vogesella dysbiosis has been recently found associated with gastric cancer (Coker et al., 2018; Rajilic-Stojanovic et al., 2020), as well as with changes in the endometrial microbiota associated with inflammatory cytokines in endometrial cancer (Lu et al., 2020), and with esophageal squamous cell carcinoma (Lv et al., 2020).

As regards to Rubrivivax, it is able to produce a molecule rubrivivaxin that is a cytotoxic agent and a COX-1 inhibitor (Kumavath et al., 2011). As is known COX-1 and COX-2 are relevant players in human colorectal cancer (Sano et al., 1995; Sinicrope and Gill, 2004; Pannunzio and Coluccia, 2018). Rubrivivax dysbiosis has also been found present in connection to lung cancer (Greathouse et al., 2018).

Thermodesulfovibrio has been recently discussed to play a role in the modulation of FOXP3 and IL-17 involved in immune tolerance in colon cancer (Bergsten et al., 2020). Sulfate reducing bacteria, also including Desulphurella are known to be associated with the pathogenesis of colorectal cancer (Kováč et al., 2017; Suri et al., 2019). Nitriliruptor has been reported to be involved colorectal cancer (Marzban et al., 2020), its dysbiosis has been mentioned also in connection to renal carcinomas (Wang et al., 2020) and severe cases of irritable bowel syndrome (Zhuang et al., 2018).

In connection with microorganisms associated with late stage colon cancer (**Table 4**), Jeotgalicoccus abundance has been found to be abnormal in the urinary microbiome in connection with bladder cancer (Hussein et al., 2021). It also has been included in a metagenomic panel screening for the diagnosis of ovarian cancer (Kim et al., 2020) and associated with antibiotic perturbation leading to accelerated tumor growth in breast cancer (Kirkup et al., 2019). Interestingly, Cryocola has been found to be increasingly abundant after *H. pylori* eradication in gastric cancer cells (Figueiredo and Castaño-Rodríguez, 2020) which may point out to second order competition effects. Dactylosporangium produces molecules such as macrolides that disrupt the mitochondrial membrane potentials in colorectal cancer cells HCT116 and HT29 (Tan et al., 2018) and belong to a class of microorganisms that are being considered as source of bioactive metabolites with pharmaceutical interest (Rangseekaew and Pathom-Aree, 2019).

In the case of Pelomonas, it has been recognized as involved in the onset of multifocal atrophic gastritis with intestinal metaplasia, a likely pre-malignant gastric lesion (Yang et al., 2016). It is also abundant in the tumor microenvironment of up to fifty percent of colorectal tumors in one study (Pierce et al., 2018). Pelomonas also has been found as one of the disrupted genera associated with bladder cancer (Liu et al., 2019; Mansour et al., 2020).

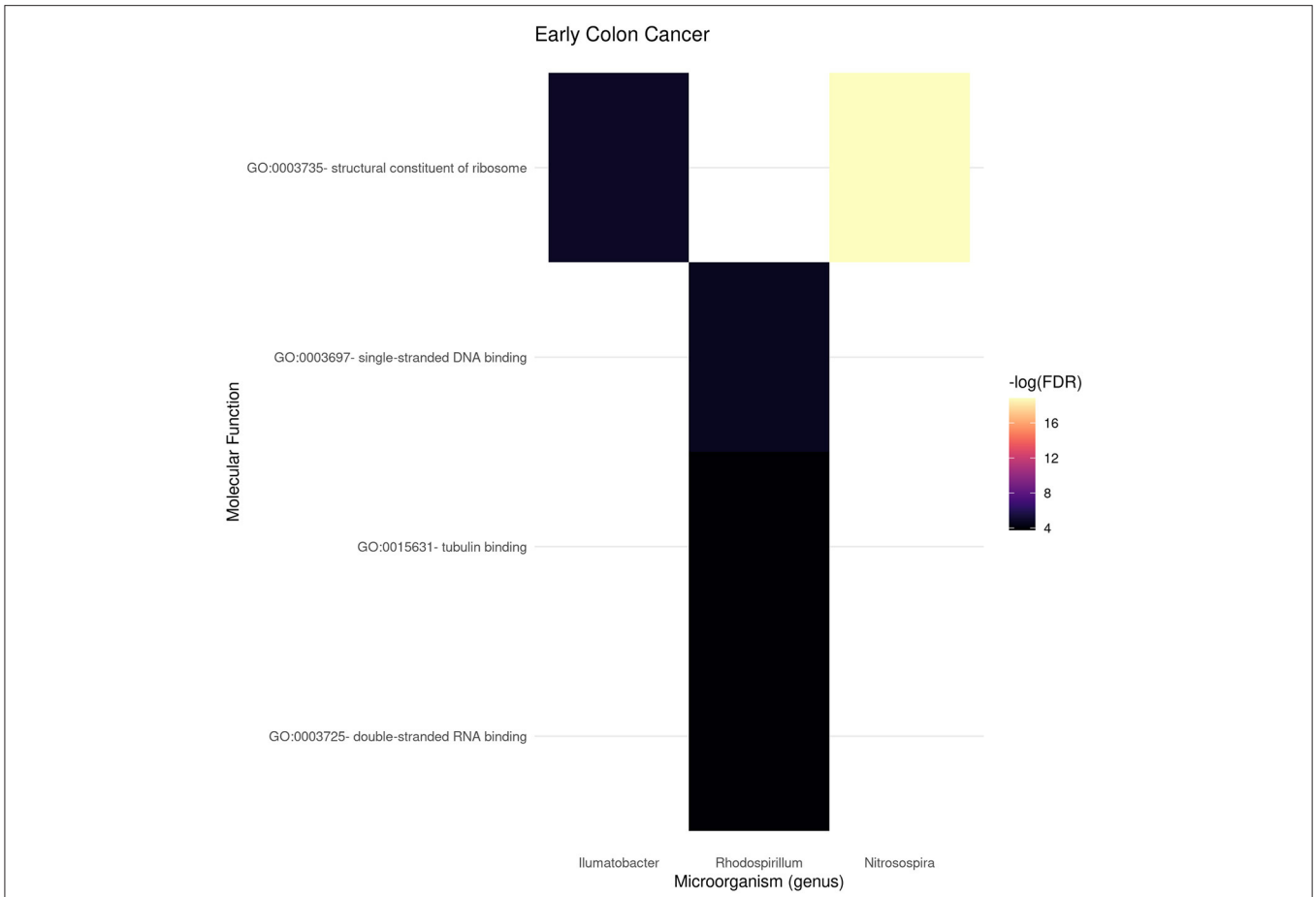

FIGURE 9 | Functional enrichment of highly connected microorganisms in the early colon cancer network—molecular functions.

Zymomonas have been recognized to play several roles in cancer. Zymomonas' levan is involved in MMP-9 activation and extracellular matrix remodeling and inflammation (Sturzoiu et al., 2011) and also to induce changes in oxidative states leading to antiproliferative and proapoptotic effects in MCF7 breast cancer cells (Queiroz et al., 2017). Similarly, Methylomonas have been found to be involved in the production of toxin genes that are functional drivers in human colorectal cancer (Dutilh et al., 2013) and in the production of azurin, a known cytotoxic factor regulating cell death (Chakrabarty et al., 2008).

It should be noticed, however, that confirmation studies, in particular functional intervention assays, are needed to establish more clearly the actual role of microbiome dysbiosis in connection with the onset and development of human malignancies in general and specially colon cancer.

Biological Functionality Associated to the Microbiome Changes With Progression

The concerted study of gene-microbial interactions is still at its infancy. It results challenging thus to ascertain or even hypothesize on the role that microbial communities play in the already complex and incomplete panorama of biomolecular interactions inside human cells and tissues. In order to advance, if just a little, in our understanding of how microorganisms and their joint metabolic fluxes and ecological interactions influence the molecular and cellular composition and functions, we have resorted to analyse the *gene-microorganism co-expression* networks. By looking at the known molecular players (genes) that present strong statistical dependencies with specific microbial species we may start by assigning those (via *guilt-by-association* schemes) a putative functional role in human (in this case, tumor) biology.

Gene enrichment analysis was used to indirectly *probe* associations with the microbiome by looking at the gene-neighborhood of highly connected microorganisms in early and late stage colorectal cancer bipartite networks. Gene Ontology Biological Processes (BP) and Molecular Function (MF) branches were considered as target databases for the statistical over-representation enrichment analysis as presented in **Figures 6**, **7** for BP, and **Figures 9**, **10** for MF in early/late colorectal tumor networks, respectively.

As presented in results, we were able to find functional differences between the early and late stage gene-microbiome

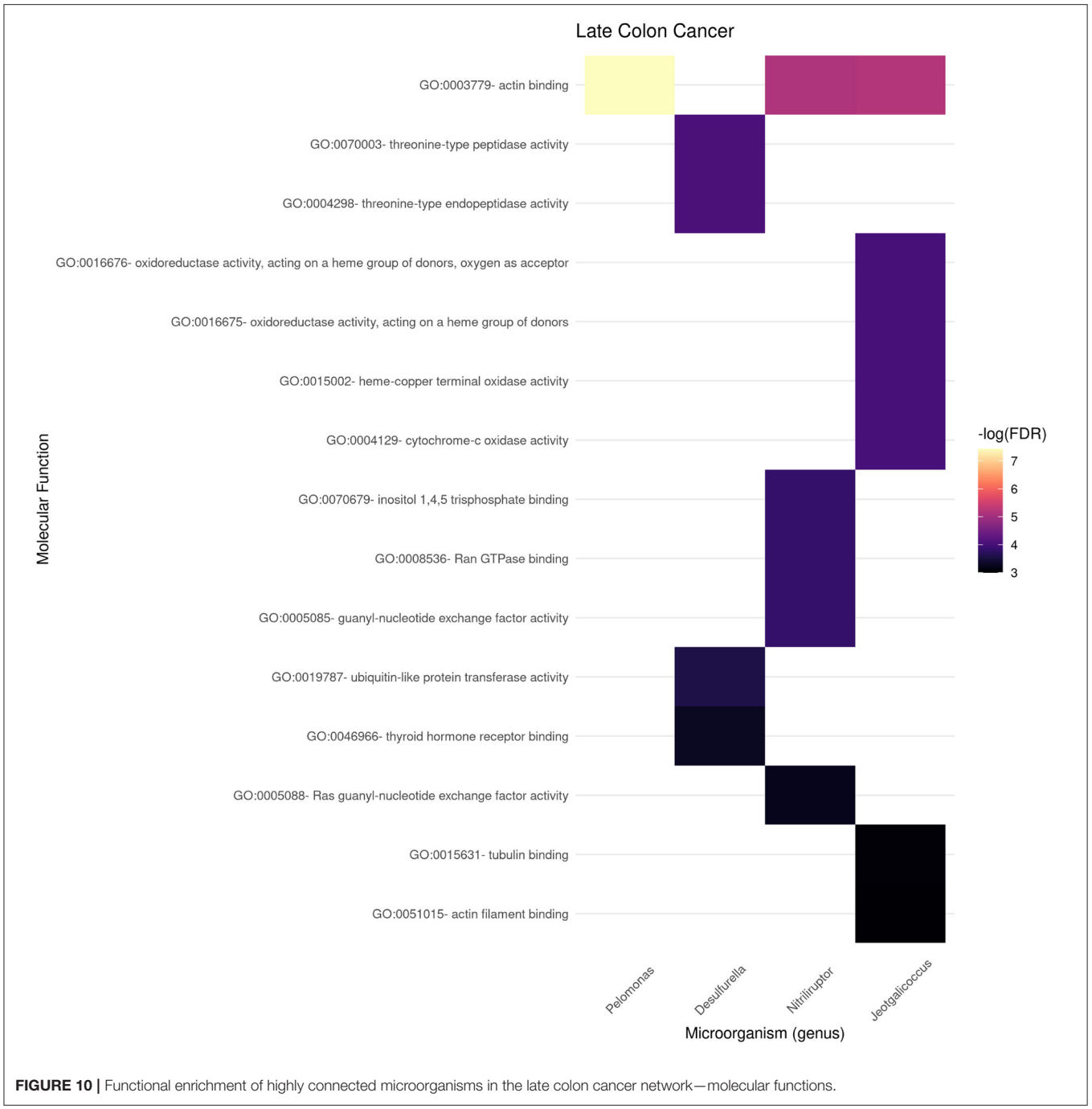

FIGURE 10 | Functional enrichment of highly connected microorganisms in the late colon cancer network—molecular functions.

co-expression programmes. A number of statistical significant processes and molecular functions are presented in the heatmaps in **Figures 5**, **6**, **8**, **9**. To present a summary of these findings, we used natural language processing tools on tokenized versions of the enrichment tables. **Figures 8**, **11** present Venn diagrams depicting highly mentioned tokens. We can see that in the case of BP (**Figure 8**), early stage networks are enriched for terms related to proliferation and cell growth, including structural elements and synthesis of biomaterials, whereas late stage is characterized by terms related to signaling and transport processes. Biochemical and physical regulation mechanisms are present in processes at the intersection of both networks.

Following a similar approach, tokens related to molecular functions associated with early and late stage colorectal cancer are presented in **Figure 11**. As in the case of biological processes, molecular functions associated with early tumors are related with structural features, late stage contains terms related to signaling and metabolic interactions, whereas the only molecular functions at the intersection of stages are related to binding.

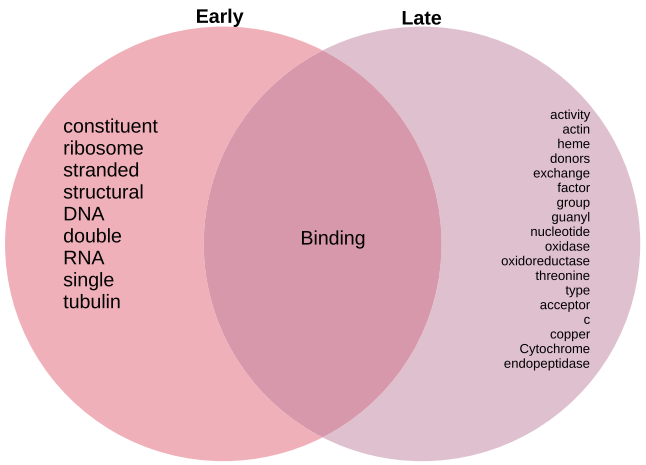

FIGURE 11 | Venn diagram of top 20 most mentioned concepts in molecular functions associated to early and late colon cancer.

By integrating these results some preliminary ideas may be drawn: first of all, it is becoming possible to analyse (albeit still in a somehow rudimentary way) the combined effect that the microbiome plays in conjunction with human tumor cells in the onset, establishment and development of colorectal cancer. These initial analyses, reveal differences in the functional features of the gene-microbiome bipartite co-expression networks, as inferred from probabilistic modeling of high-throughput genomic and transcriptomic experiments in large datasets. These differences, when supplemented with statistical enrichment analyses point out to a plausible scenario in which early stage colon cancer presents features related to the establishment of distinctive physical structures in the cells, that start to couple with biomolecular interactions at the cellular level, whereas advanced stages present an image of more complex signaling and metabolic processes occurring as the tumor keeps evolving to more advanced, malignant stages.

Scope and Limitations

In this work we identify changes in the co-expression/co-presence network connectivity found between colon cancer microbiome and its gene expression as the disease progresses. This type of studies are admittedly at their preliminary stages, but the integrative view they aim to provide seems promissory toward a better understanding of complex disease phenotypes. It is relevant, however, to acknowledge some limitations and assumptions of our current approach, in order to properly contextualize our findings and convey a balanced message.

One worth-mentioning constraint that may restrict the scope of our assertions is the following: Our work is based on experimental data coming from the TCGA colon cancer cohort. The volume of this cohort, as well as the availability of proper, well-curated, clinical metadata, makes it suitable for our (high-throughput, probabilistic-based) analyses. Furthermore, the open microbiome quantification strategy and the resulting data from Poore et al. (2020) allowed for a (relatively) high-confident network reconstruction. This is, however, the only cohort for which such suitable data is available, thus limiting our ability to replicate our findings in an independent cohort. While the sample size is adequate for probabilistic network reconstruction purposes, it can only capture as much of the microbiome heterogeneity as what was captured by the original authors. On a related topic, since access to the TCGA raw data required for the microbiome quantification data described in Poore et al. (2020) is controlled, we must rely on the quantification strategy as performed by the original authors—which is in turn influenced by sequencing depth and wet-lab procedure constraints from the original work.

Aside from these specific issues, some additional, general limitations should also be mentioned: although the methods used both in our work and in Poore et al. (2020) and even those in the TCGA original approach are all in the state of the art, there are still challenges. Even though the TCGA data has both, excellent depth and high quality sequencing, it was not intended as a metagenomic sequencing assay. Also, even the best metagenomic approaches rely on currently incomplete annotations. Pre-processing stages to consider multi-omic approaches, including metagenomic data are being developed so, these may not be as optimized and standardized as it will be desirable.

In spite of these clear limitations, we are convinced of the value of approaches such as the one presented here to start trying to answer these questions from an integrative data-centered view.

CONCLUSIONS

The progression of colon cancer involves changes in the interactions between cancer tissue and microbiome. In this work, we integrated microbiome quantification data with gene expression data using network models. These models describe the aforementioned changes in this interactions. We found that indeed, the set of microorganisms with a higher connectivity with host genes changes from the early to the late stages of colon cancer. Furthermore, reorganization is accompanied by changes in the associated set of biological functions, showing physiological adaptations associated to the tumor-microbiome relationships. To better understand and validate this findings, future experimental work is needed to properly characterize the mechanisms through which the microbiome may be mediating the observed tumor adaptations.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

IU-N organized data, performed calculations, and analyzed the data. EH-L co-designed the study, contributed to the methodological approach, analyzed data, discussed results,

reviewed the manuscript, and co-supervised the project. GdA-J envisioned the project, devised the methodological strategy, developed code, performed calculations, analyzed data, discuss results, drafted the manuscript, and 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 [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], 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. EH-L is recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

The authors want to thank Gabriela Graham for her support with language editing and proof-reading of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617505/full#supplementary-material

REFERENCES

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683–691. doi: 10.1136/gutjnl-2015-310912

Barabási, A.-L., et al. (2016). *Network Science*. Cambridge: Cambridge University Press.

Bergsten, E., Mestivier, D., Amiot, A., DeAngelis, N., Khazaie, K., and Sobhani, I. (2020). Immune tolerance to colon cancer is mediated by colon dysbiosis: human results and experimental *in vivo* validation. *J. Clin. Oncol.* 38:1. doi: 10.1200/JCO.2020.38.15_suppl.e16062

Bingham, S., Pignatelli, B., Pollock, J., Ellul, A., Malaveille, C., Gross, G., et al. (1996). Does increased endogenous formation of n-nitroso compounds in the human colon explain the association between red meat and colon cancer? *Carcinogenesis* 17, 515–523. doi: 10.1093/carcin/17.3.515

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

Bruce, W. R., Giacca, A., and Medline, A. (2000). Possible mechanisms relating diet and risk of colon cancer. *Cancer Epidemiol. Prevent. Biomark.* 9, 1271–1279.

Chakrabarty, A. M., Gupta, T. K. D., Punj, V., Zaborina, O., Hiraoka, Y., and Yamada, T. (2008). *Cytotoxic Factors for Modulating Cell Death*. US Patent App. 11/509,682. Boston, MA: Google Patents.

Coker, O. O., Dai, Z., Nie, Y., Zhao, G., Cao, L., Nakatsu, G., et al. (2018). Mucosal microbiome dysbiosis in gastric carcinogenesis. *Gut* 67, 1024–1032. doi: 10.1136/gutjnl-2017-314281

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJ. Complex Syst.* 1695, 1–9.

Davis, C. D., and Milner, J. A. (2009). Gastrointestinal microflora, food components and colon cancer prevention. *J. Nutr. Biochem.* 20, 743–752. doi: 10.1016/j.jnutbio.2009.06.001

de Anda-Jáuregui, G., Espinal-Enríquez, J., Drago-García, D., and Hernández-Lemus, E. (2018). Nonredundant, highly connected microRNAs control functionality in breast cancer networks. *Int. J. Genomics* 2018:9585383. doi: 10.1155/2018/9585383

de Anda-Jáuregui, G., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019). Highly-connected, non-redundant micrornas functional control in breast cancer molecular subtypes. *BiorXiv*. 1–10. doi: 10.1101/652354

de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Computational oncology in the multi-omics era: state of the art. *Front. Oncol.* 10:423. doi: 10.3389/fonc.2020.00423

de Anda-Jáuregui, G., Velázquez-Caldelas, T. E., Espinal-Enríquez, J., and Hernández-Lemus, E. (2016). Transcriptional network architecture of breast cancer molecular subtypes. *Front. Physiol.* 7:568. doi: 10.3389/fphys.2016.00568

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dong, Z., Chen, B., Pan, H., Wang, D., Liu, M., Yang, Y., et al. (2019). Detection of microbial 16s rRNA gene in the serum of patients with gastric cancer. *Front. Oncol.* 9:608. doi: 10.3389/fonc.2019.00608

Dutilh, B. E., Backus, L., van Hijum, S. A., and Tjalsma, H. (2013). Screening metatranscriptomes for toxin genes as functional drivers of human colorectal cancer. *Best Pract. Res. Clin. Gastroenterol.* 27, 85–99. doi: 10.1016/j.bpg.2013.03.008

Fernández-Martínez, N. F., Ching-López, A., Olry de Labry Lima, A., Salamanca-Fernández, E., Pérez-Gómez, B., Jiménez-Moleón, J. J., et al. (2020). Relationship between exposure to mixtures of persistent, bioaccumulative, and toxic chemicals and cancer risk: a systematic review. *Environ. Res.* 188:109787. doi: 10.1016/j.envres.2020.109787

Figueiredo, C., and Casta no-Rodríguez, N. (2020). The microbiome and gastric cancer: an update. *Microb. Health Dis.* 2:e627. doi: 10.26355/mhd_20206_267

Fiorentini, C., Carlini, F., Germinario, E. A. P., Maroccia, Z., Travaglione, S., and Fabbri, A. (2020). Gut microbiota and colon cancer: a role for bacterial protein toxins? *Int. J. Mol. Sci.* 21:6201. doi: 10.3390/ijms21176201

Friedenreich, C. M., Ryder-Burbidge, C., and McNeil, J. (2020). Physical activity, obesity and sedentary behavior in cancer etiology: epidemiologic evidence and biologic mechanisms. *Mol. Oncol.* 1–11. doi: 10.1002/1878-0261.12772

Greathouse, K. L., White, J. R., Vargas, A. J., Bliskovsky, V. V., Beck, J. A., von Muhlinen, N., et al. (2018). Microbiome-TP53 gene interaction in human lung cancer. *bioRxiv*. 273524. doi: 10.1101/273524

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring network structure, dynamics, and function using networkX," in *Proceedings of the 7th Python in Science Conference*, eds G. Varoquaux, T. Vaught, and J. Millman (Pasadena, CA), 11–15.

He, J., Zhou, Z., Reed, M., and Califano, A. (2017). Accelerated parallel algorithm for gene network reverse engineering. *BMC Syst. Biol.* 11:5. doi: 10.1186/s12918-017-0458-5

Hussein, A. A., Elsayed, A. S., Durrani, M., Jing, Z., Iqbal, U., Gomez, E. C., et al. (2021). Investigating the association between the urinary microbiome and bladder cancer: an exploratory study. *Urol. Oncol.* doi: 10.1016/j.urolonc.2020.12.011

Kim, Y.-K., Taesung, P., Song, Y. S., and Kim, S. I. (2020). *Method for Diagnosing Ovarian Cancer Through Microbial Metagenome Analysis*. US Patent App. 16/629,360. Boston, MA: Google Patents.

Kirkup, B. M., McKee, A., Makin, K. A., Paveley, J., Caim, S., Alcon-Giner, C., et al. (2019). Perturbation of the gut microbiota by antibiotics results in accelerated breast tumour growth and metabolic dysregulation. *BioRxiv*. 553602. doi: 10.1101/553602

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., et al. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* 8, 761–763. doi: 10.1038/nmeth.1650

Kováč, J., Kushkevych, I., et al. (2017). "New modification of cultivation medium for isolation and growth of intestinal sulfate-reducing bacteria," in *Proceeding of International PhD Students Conference MendelNet* (Brno), 702–707.

Kowalchuk, G. A., and Stephen, J. R. (2001). Ammonia-oxidizing bacteria: a model for molecular microbial ecology. *Annu. Rev. Microbiol.* 55, 485–529. doi: 10.1146/annurev.micro.55.1.485

Kumavath, R. N., Ramana, C. V., and Sasikala, C. (2011). Rubrivivaxin, a new cytotoxic and cyclooxygenase-i inhibitory metabolite from rubrivivax benzoatilyticus ja2. *World J. Microbiol. Biotechnol.* 27, 11–16. doi: 10.1007/s11274-010-0420-9

Latapy, M., Magnien, C., and Vecchio, N. D. (2008). Basic notions for the analysis of large two-mode networks. *Soc. Netw.* 30, 31–48. doi: 10.1016/j.socnet.2007.04.006

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29

Li, G., Yang, M., Jin, Y., Li, Y., Qian, W., Xiong, H., et al. (2018). Involvement of shared mucosal-associated microbiota in the duodenum and rectum in diarrhea-predominant irritable bowel syndrome. *J. Gastroenterol. Hepatol.* 33, 1220–1226. doi: 10.1111/jgh.14059

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). Webgestalt 2019: gene set analysis toolkit with revamped UIS and apis. *Nucleic Acids Res.* 47, W199–W205. doi: 10.1093/nar/gkz401

Liu, F., Liu, A., Lu, X., Zhang, Z., Xue, Y., Xu, J., et al. (2019). Dysbiosis signatures of the microbial profile in tissue from bladder cancer. *Cancer Med.* 8, 6904–6914. doi: 10.1002/cam4.2419

Lu, W., He, F., Lin, Z., Liu, S., Tang, L., Huang, Y., et al. (2020). Dysbiosis of the endometrial microbiota and its association with inflammatory cytokines in endometrial cancer. *Int. J. Cancer*. 1–12. doi: 10.1002/ijc.33428

Lv, W., Zuo, J., Wang, Y., Fan, Z., Feng, L., Wang, L., et al. (2020). The microbial characteristics of esophageal squamous cell carcinoma (ESCC) and healthy subjects. *J. Clin. Oncol.* 38:e16546. doi: 10.1200/JCO.2020.38.15_suppl.e16546

Mansour, B., Monyók, Á., Makra, N., Gajdács, M., Vadnay, I., Ligeti, B., et al. (2020). Bladder cancer-related microbiota: examining differences in urine and tissue samples. *Sci. Rep.* 10, 1–10. doi: 10.1038/s41598-020-67443-2

Marzban, M., Kashefian Naeeini, S., Ghazbani, A., and Karimi, Z. (2020). Systematic review of fecal and mucosa-associated microbiota compositional shifts in colorectal cancer. *Ann. Colorect. Res.* 8, 1–13. doi: 10.30476/ACRR.2020.46747

Meyer, D., Hornik, K., and Feinerer, I. (2008). Text mining infrastructure in R. *J. Stat. Softw.* 25, 1–54. doi: 10.18637/jss.v025.i05

O'keefe, S. J. (2016). Diet, microorganisms and their metabolites, and colon cancer. *Nat. Rev. Gastroenterol. Hepatol.* 13:691. doi: 10.1038/nrgastro.2016.165

Pannunzio, A., and Coluccia, M. (2018). Cyclooxygenase-1 (cox-1) and cox-1 inhibitors in cancer: a review of oncology and medicinal chemistry literature. *Pharmaceuticals* 11:101. doi: 10.3390/ph11040101

Peñalver Bernabé, B., Cralle, L., and Gilbert, J. A. (2018). Systems biology of the human microbiome. *Curr. Opin. Biotechnol.* 51, 146–153. doi: 10.1016/j.copbio.2018.01.018

Pierce, C. M., Hong, B. Y., Hoehn, H. J., Gomez, M. F., Melas, M., McDonnell, K., et al. (2018). Microbes in the tumor microenvironment: Bacterial influences on host immunity in colorectal cancer [abstract]. *Cancer Res.* 78(13 Suppl):Abstract nr 4746. doi: 10.1158/1538-7445.AM2018-4746

Plyasova, A. A., Pokrovskaya, M. V., Lisitsyna, O. M., Pokrovsky, V. S., Alexandrova, S. S., Hilal, A., et al. (2020). Penetration into cancer cells via clathrin-dependent mechanism allows l-asparaginase from rhodospirillum rubrum to inhibit telomerase. *Pharmaceuticals* 13:286. doi: 10.3390/ph13100286

Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 579, 567–574. doi: 10.1038/s41586-020-2095-1

Qiu, J., Li, N., Lu, Z., Yang, Y., Ma, Y., Niu, L., et al. (2016). Conversion of nornicotine to 6-hydroxy-nornicotine and 6-hydroxy-myosmine by *Shinella* sp. strain HZN7. *Appl. Microbiol. Biotechnol.* 100, 10019–10029. doi: 10.1007/s00253-016-7805-0

Queiroz, E. A., Fortes, Z. B., da Cunha, M. A., Sarilmiser, H. K., Dekker, A. M. B., Öner, E. T., et al. (2017). Levan promotes antiproliferative and pro-apoptotic effects in MCF-7 breast cancer cells mediated by oxidative stress. *Int. J. Biol. Macromol.* 102, 565–570. doi: 10.1016/j.ijbiomac.2017.04.035

Rajilic-Stojanovic, M., Figueiredo, C., Smet, A., Hansen, R., Kupcinskas, J., Rokkas, T., et al. (2020). Systematic review: gastric microbiota in health and disease. *Aliment. Pharmacol. Therap.* 51, 582–602. doi: 10.1111/apt.15650

Rangseekaew, P., and Pathom-Aree, W. (2019). Cave actinobacteria as producers of bioactive metabolites. *Front. Microbiol.* 10:387. doi: 10.3389/fmicb.2019.00387

Raskov, H., Søby, J. H., Troelsen, J., Bojesen, R. D., and Ggenur, I. (2020). Driver gene mutations and epigenetics in colorectal cancer. *Ann. Surg.* 271, 75–85. doi: 10.1097/SLA.0000000000003393

Sano, H., Kawahito, Y., Wilder, R. L., Hashiramoto, A., Mukai, S., Asai, K., et al. (1995). Expression of cyclooxygenase-1 and-2 in human colorectal cancer. *Cancer Res.* 55, 3785–3789. doi: 10.1016/0928-4680(94)90594-0

Shannon, P. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Simin, J., Fornes, R., Liu, Q., Olsen, R. S., Callens, S., Engstrand, L., et al. (2020). Antibiotic use and risk of colorectal cancer: a systematic review and dose-response meta-analysis. *Br. J. Cancer*. 1825–1832. doi: 10.1038/s41416-020-01082-2

Sinicrope, F. A., and Gill, S. (2004). Role of cyclooxygenase-2 in colorectal cancer. *Cancer Metast. Rev.* 23, 63–75. doi: 10.1023/A:1025863029529

Sturzoiu, C., Petrescu, M., Galateanu, B., Anton, M., Nica, C., Simionca, G. I., et al. (2011). Zymomonas mobilis levan is involved in metalloproteinases activation in healing of wounded and burned tissues. *Sci. Pap. Anim. Sci. Biotechnol.* 44, 453–458.

Suri, A., BAnSAl, S. K., Ammalli, P., and Karunanand, B. (2019). Role of microbiota in aetiopathogenesis of colorectal cancer. *J. Clin. Diagnost. Res.* 13, 1–5. doi: 10.7860/JCDR/2019/42445.13169

Tan, P. J., Lau, B. F., Krishnasamy, G., Ng, M. F., Husin, L. S., Ruslan, N., et al. (2018). Zebrafish embryonic development-interfering macrolides from streptomyces californicus impact growth and mitochondrial function in human colorectal cancer cells. *Process Biochem.* 74, 164–174. doi: 10.1016/j.procbio.2018.07.007

Wang, J., Li, X., Wu, X., Wang, Z., Zhang, C., Cao, G., et al. (2020). Uncovering the microbiota in renal cell carcinoma tissue using 16s rRNA gene sequencing. *J. Cancer Res. Clin. Oncol.* 481–491. doi: 10.1007/s00432-020-03462-w

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Xu, A. A., Hoffman, K., Gurwara, S., White, D. L., Kanwal, F., El-Serag, H. B., et al. (2020). Oral health and the altered colonic mucosa-associated gut microbiota. *Digest. Dis. Sci.* doi: 10.1007/s10620-020-06612-9

Yang, I., Woltemate, S., Piazuelo, M. B., Bravo, L. E., Yepez, M. C., Romero-Gallo, J., et al. (2016). Different gastric microbiota compositions in two human populations with high and low gastric cancer risk in colombia. *Sci. Rep.* 6:18594. doi: 10.1038/srep18594

Yang, Q., Wang, Y., Jia, A., Wang, Y., Bi, Y., and Liu, G. (2020). The crosstalk between gut bacteria and host immunity in intestinal inflammation. *J. Cell. Physiol.* 1–16. doi: 10.1002/jcp.30024

Yu, M. R., Kim, H. J., and Park, H. R. (2020). Fusobacterium nucleatum accelerates the progression of colitis-associated colorectal cancer by promoting EMT. *Cancers* 12:2728. doi: 10.3390/cancers12102728

Zhdanov, D. D., Pokrovsky, V. S., Pokrovskaya, M. V., Alexandrova, S. S., Eldarov, M. A., Grishin, D. V., et al. (2017a). Inhibition of telomerase activity and induction of apoptosis by rhodospirillum rubrum l-asparaginase in cancer jurkat cell line and normal human CD4+ t lymphocytes. *Cancer Med.* 6, 2697–2712. doi: 10.1002/cam4.1218

Zhdanov, D. D., Pokrovsky, V. S., Pokrovskaya, M. V., Alexandrova, S. S., Eldarov, M. A., Grishin, D. V., et al. (2017b). Rhodospirillum rubrum l-asparaginase targets tumor growth by a dual mechanism involving telomerase inhibition. *Biochem. Biophys. Res. Commun.* 492, 282–288. doi: 10.1016/j.bbrc.2017.08.078

Zhuang, X., Tian, Z., Li, L., Zeng, Z., Chen, M., and Xiong, L. (2018). Fecal microbiota alterations associated with diarrhea-predominant irritable bowel syndrome. *Front. Microbiol.* 9:1600. doi: 10.3389/fmicb.2018.01600

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

frontiers in Genetics

TECHNOLOGY AND CODE
published: 24 February 2021
doi: 10.3389/fgene.2021.630187

CoExp: A Web Tool for the Exploitation of Co-expression Networks

*Sonia García-Ruiz[1,2,3], Ana L. Gil-Martínez[1], Alejandro Cisterna[4], Federico Jurado-Ruiz[4], Regina H. Reynolds[1,2,3], NABEC (North America Brain Expression Consortium), Mark R. Cookson[5], John Hardy[6,7,8,9,10], Mina Ryten[1,2,3] and Juan A. Botía[3,4]**

[1] Institute of Neurology, University College London, London, United Kingdom, [2] NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London, United Kingdom, [3] Department of Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London, United Kingdom, [4] Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain, [5] Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, United States, [6] Department of Neurodegenerative Diseases, UCL Queen Square Institute of Neurology, London, United Kingdom, [7] Department of Neurodegenerative Disease, United Kingdom Dementia Research Institute at UCL, UCL Institute of Neurology, University College London, London, United Kingdom, [8] Reta Lila Weston Institute, UCL Queen Square Institute of Neurology, London, United Kingdom, [9] UCL Movement Disorders Centre, University College London, London, United Kingdom, [10] Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong, China

OPEN ACCESS

Edited by:
Kimberly Glass, Brigham and Women's Hospital and Harvard Medical School, United States

Reviewed by:
Margaret Woodhouse, Agricultural Research Service, United States
Barry Demchak, University of California, San Diego, United States

***Correspondence:**
Juan A. Botía
juanbotiablaya@gmail.com

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

Received: *16 November 2020*
Accepted: *03 February 2021*
Published: *24 February 2021*

Citation:
García-Ruiz S, Gil-Martínez AL, Cisterna A, Jurado-Ruiz. F, Reynolds RH, Cookson MR, Hardy J, Ryten M and Botía JA (2021) CoExp: A Web Tool for the Exploitation of Co-expression Networks. Front. Genet. 12:630187. doi: 10.3389/fgene.2021.630187

Gene co-expression networks are a powerful type of analysis to construct gene groupings based on transcriptomic profiling. Co-expression networks make it possible to discover modules of genes whose mRNA levels are highly correlated across samples. Subsequent annotation of modules often reveals biological functions and/or evidence of cellular specificity for cell types implicated in the tissue being studied. There are multiple ways to perform such analyses with weighted gene co-expression network analysis (WGCNA) amongst one of the most widely used R packages. While managing a few network models can be done manually, it is often more advantageous to study a wider set of models derived from multiple independently generated transcriptomic data sets (e.g., multiple networks built from many transcriptomic sources). However, there is no software tool available that allows this to be easily achieved. Furthermore, the visual nature of co-expression networks in combination with the coding skills required to explore networks, makes the construction of a web-based platform for their management highly desirable. Here, we present the CoExp Web application, a user-friendly online tool that allows the exploitation of the full collection of 109 co-expression networks provided by the CoExpNets suite of R packages. We describe the usage of CoExp, including its contents and the functionality available through the family of CoExpNets packages. All the tools presented, including the web front- and back-ends are available for the research community so any research group can build its own suite of networks and make them accessible through their own CoExp Web

application. Therefore, this paper is of interest to both researchers wishing to annotate their genes of interest across different brain network models and specialists interested in the creation of GCNs looking for a tool to appropriately manage, use, publish, and share their networks in a consistent and productive manner.

Keywords: co-expression network, guilt by association, web app for neuroscience, transcriptomics, brain

INTRODUCTION

Gene co-expression network analysis has been widely used to identify biologically important patterns in gene expression in a hypothesis-free and genome-wide manner (Miller et al., 2010; Forabosco et al., 2013; Uk Brain Expression Consortium (UKBEC), et al., 2016; de la Torre-Ubieta et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al, 2018; Bettencourt et al., 2019; Mencacci et al., 2020). The driving principle behind co-expression network analysis is that genes with highly correlated expression levels are also likely to share functional and biological relationships (Bakhtiarizadeh et al., 2018; Ma et al., 2018). With this in mind, gene co-expression networks (GCNs) can be viewed as models of how genes cluster together into modules of highly co-expressed genes through the use of graph-based approaches (Wolfe et al., 2005; Margolin et al., 2006; Langfelder and Horvath, 2008; Botía et al., 2017). Starting from an expression profile generated on a set of samples, the clustering process generates mutually exclusive gene sets (i.e., gene clusters). The resulting clusters are then annotated through a process which aims to describe them functionally and by their cellular specificity, amongst other properties. Using this approach, we can get a reasonably accurate summary of the input samples at the mRNA level. These models have value in themselves, as a means of efficiently describing the source gene expression profile. However, they can also be used to annotate external gene sets (Forabosco et al., 2013; Uk Brain Expression Consortium (UKBEC), et al., 2016; Salpietro et al., 2017; Bettencourt et al., 2019; Efthymiou et al., 2019) generated under different experimental settings and conditions. The vast majority of analyses in the literature using GCNs as an analytic tool, focus on isolated GCNs created from specific sample sets under specific conditions. However, GCNs are more powerful when we considered collectively. For example, if we want to study neuro-degenerative diseases at the gene level, and how specific genes behave in terms of their co-expression, it is much more useful to study the gene set of interest across different brain regions (i.e., those particularly vulnerable to disease, in comparison with those which are less or never affected), and in unrelated tissues (e.g., skin). With this in mind, it is tremendously useful to have a tool which enables gene sets to be studied across all conditions in a comparative manner, including predictions based on module membership about the genes' functions and cellular specificity across the conditions of interest.

Gene annotation is a basic task in bioinformatics. Whatever the process that led to identification of a gene set of interest (e.g., differential expression analysis, cellular screens, GWAS analysis or new gene discovery), a posterior annotation process is required. Thanks to the availability of manually curated databases of biological terms and their associated genes like the Gene Ontology (The Gene Ontology Consortium, 2017), it is possible to accurately annotate sets of genes with their predicted function (Conesa et al., 2005; Binns et al., 2009; Carbon et al., 2009; Eden et al., 2009; Supek et al., 2011). DAVID (Huang et al., 2009) and GSEA (Subramanian et al., 2005) represent examples of two different types of tools for gene annotation based on the available ontologies. DAVID and similar tools identify ontology terms which are enriched in the gene set of interest. Whereas, GSEA looks for significant overlaps between the gene set of interest and the gene sets found in MSigDB (Subramanian et al., 2005; Liberzon et al., 2015). This is a collection of manually curated and previously annotated gene sets organised under a variety of criteria. GSEA annotates genes based on how they are expressed across a phenotype but also on how they cluster together across all gene sets belonging to MSigDB. While the combination of GSEA and MSigDB form a general-purpose tool, CoExp is focused on providing gene sets that emerge from co-expression models. In CoExp, gene sets are grouped into a tissue-based hierarchy (e.g., gene sets discovered in neuropathologically normal putamen samples or gene sets discovered in frontal cortex samples originating from individual's with Alzheimer's disease). CoExp can show a user whether their genes of interest cluster in a significant manner within a given condition (see **Supplementary Table 1** for a list of available networks), the functional characterization of the gene cluster and whether it is enriched for any specific cell type.

All GCNs within CoExp have been created using a similar pipeline based on Weighted Gene Co-expression Network Analysis [WGCNA (Langfelder and Horvath, 2008)], optimised with k-means (Botía et al., 2017) and annotated for function and cellular specificity (see **Figure 1**). A network model is a data table, with an entry for each gene, that can be shared and used in the form of a text file. Managing networks as text files is not easy because it is a manual task and therefore prone to error. Furthermore, to use large text files for gene set annotation requires coding skills so that module annotation can be performed efficiently across many different network models (i.e., different networks built using different expression data sets). Furthermore, R's command-line environment reduces its usability in a world where the web-page format has become the most well-known and accepted way of browsing information. CoExp automates the annotation process and, most importantly visualisation of the underlying graph-based model on which co-expression networks are based, creating a more natural way to explore this type of data in a visual and interactive manner. Cytoscape

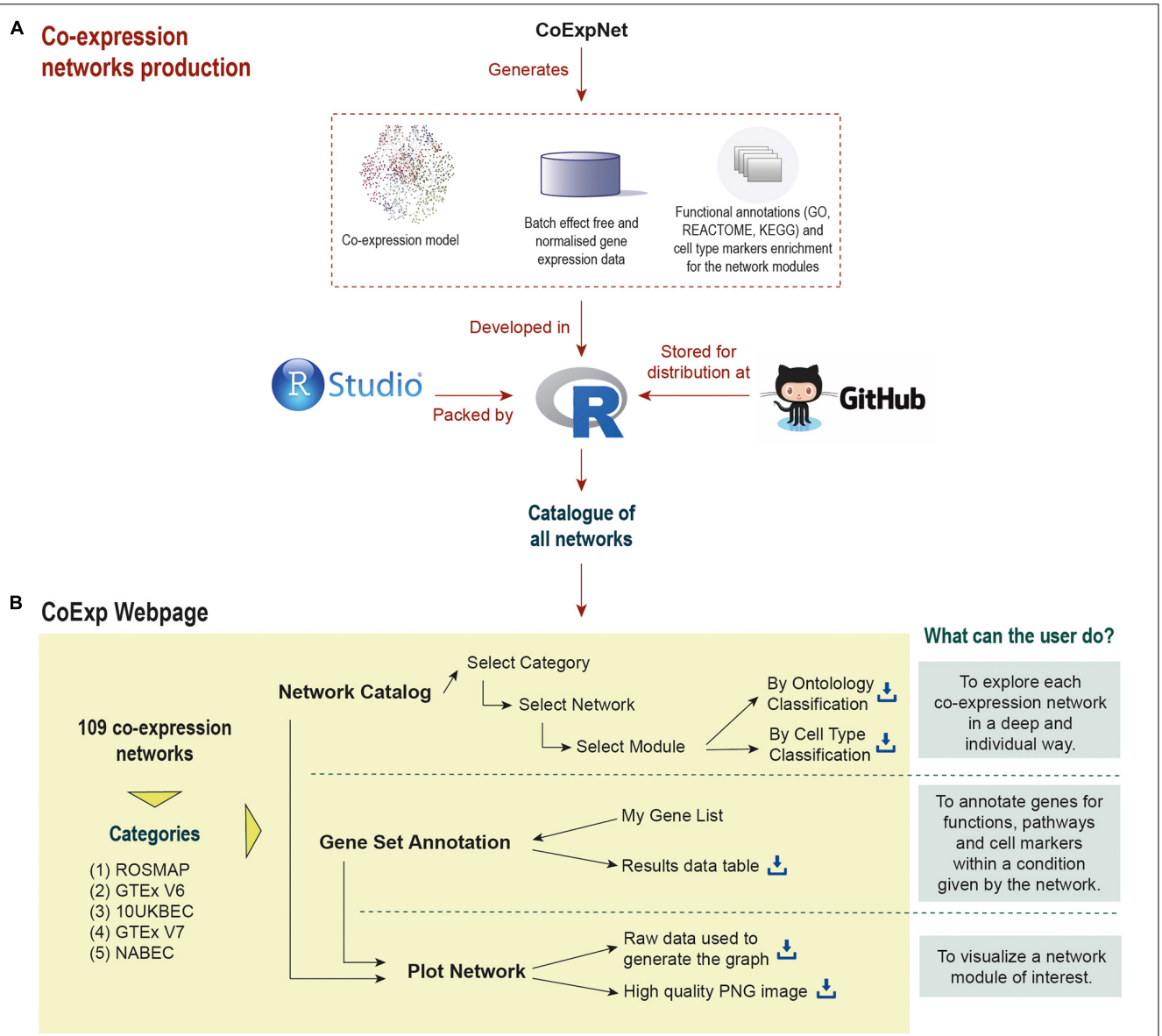

FIGURE 1 | Life cycle of a co-expression model in CoExp: **(A)** the network is built from an expression profiling data set using the CoExpNets package. This generates a network, a set of residuals as a result of data QC processes and annotations for gene module function and cellular specificity. All networks are integrated into R packages (with a package for each category, namely ROSMAP, GTEx V6 and V7, UKBEC, and NABEC). These are made available within GitHub (see the links at the "Availability and Implementation" section of this paper). The packages are then integrated into the Web application **(B)**. There they can be exploited and visualised.

(Shannon et al., 2003; Saito et al., 2012) is the pioneer stand-alone tool to provide generic network visualization, amongst many other functionalities including network generation. CoExp follows its approach, but tailored to GCNs.

In order to address these issues, we propose the CoExp web: a web-based application which aims to increase the usability and accessibility of co-expression network data. We illustrate the use of the CoExp web through the release of 109 different co-expression networks offered by the CoExpNets R package. We extend the functionality provided by the CoExpNets R package, through CoExp web's "Plot Network" option, which generates a directed graph to visualise the most important genes from a preferred module. Therefore, CoExp is a convenient way to deploy, share, and use any co-expression models.

METHODS

Co-expression Networks Generation

To construct co-expression models, we start from a gene expression profiling matrix E= M_{sxg} with samples listed as rows and genes as columns. Note that how the gene expression profiling was generated is not critical, i.e., we can either construct co-expression models from microarray or RNA-sequencing data.

In the current CoExp GCN catalogue, there are both microarray and RNA-seq based GCNs. The co-expression pipelines process this matrix to obtain an adjacency matrix, $A = M_{gxg}$ (see below) reflecting how adjacent to each other the genes are in terms of co-expression. *A* is then converted into a distance matrix $D = M_{gxg}$, required for the clustering process. As a result of this process, we obtain gene groups in the form of a gene partition *P*

$$P = \{P_i\}\,,\ 1 \leq i \leq k \ and \bigcup_{i=1}^{k} P_i = G,$$

such that the P_i are groups of genes, usually termed modules. The partition P can be disjunct, i.e., $\bigcap P_i = \emptyset$ or we may have a global membership function we can use on any gene and partition, $\mu(g,P_i)$ to generate values in [0,1] such that any gene may be a member of any partition to a certain degree. Disjunct partitions are easier to interpret and therefore they are more frequently used. In CoExp, genes belong only to a single group (i.e., module). In summary, given an expression profiling *E*, a GCN is a pair GCN(E) = (A, P), i.e., A is an adjacency matrix and P a partition.

Depending on the biological question we aim to study using GCNs, the *E* matrix can be treated differently. For example, we may try to correct *E* for any bias introduced by batch effects (Leek and Storey, 2007) or for any biological covariate whose probable effect will bias the models (e.g., sex or age). All the GCNs in CoExp have been corrected for batch [with the ComBat (Johnson et al., 2007) R package], for unknown latent effects with SVA and for gender and age (and post-mortem interval when this information is available) by regressing out the covariates through linear regression. In order to create all the GCNs, we first follow the standard WGCNA procedure: we identify the smoothing parameter that guarantees scale free topology for the network, we generate an adjacency matrix and the Topology Overlap Matrix (TOM). 1-TOM is used as the distance for hierarchical clustering. The gene clustering we obtain is subsequently refined using the k-means clustering algorithm (Botía et al., 2017). All networks include module membership for each gene (a measure of a gene's relevance within the cluster) and the eigengenes (the 1st PCA of a module's gene expression). Finally, the gene modules are annotated using gProfileR [see details here (Reimand et al., 2007)]. All networks are annotated for cellular specificity by performing a Fisher's Exact test on the overlap between selected gene markers (found in the CoExpNets package) and genes within each module.

The CoExp Web Software

The CoExp software architecture is depicted in **Figure 2**. The CoExp website has been implemented with the aim of connecting two different runtime environments, an R based and an ASP.NET based environment. The front end of CoExp has been developed following a Model View Controller (MVC) architecture implemented through the ASP.NET Core MVC framework, which is a cross-platform and open-source tool from the ASP.NET Core family of frameworks. Note that we have chosen the MVC architecture to enable division of the program logic into three main components: the model, the view and the controller. This facilitates the interpretation of the code by any external user interested in using or modifying it. This is essential for the maintainability of CoExp software over the coming years.

The back end of CoExp is based on a suite of five independent R packages corresponding to five network families (see section "Results") plus the CoExpNets package which provides the necessary code to generate new GCNs and with a unified API to all networks. To make the collection of CoExp R methods accessible to the front end, a web Application Programming Interface (API) was chosen to define the interactions between the two runtime environments. To build the API, the R back end list of methods were first published using the Plumber R package and then made accessible using the Swagger (OpenAPI) language-agnostic specification to describe them in a user-friendly interface. The corresponding documentation can be accessed here:[1]. The external API was built using REST Web services.

Finally, the web server we use is one from the Apache HTTP Server Project. The communication between the CoExp ASP.NET Core MVC libraries, which are natively served by a Kestrel server, and the APACHE server was made by using a reverse-proxy service, which acted as an intermediate layer connecting both servers.

To reduce the complexity of this architectural design and to make it possible for a user to install and use CoExp locally, both the front and back end have been encapsulated within two different Docker containers and made available on DockerHub[2].

The Suite of Packages to Store and Manage Networks

All GCNs made available within CoExp are organised into R Packages. Currently, there is one R package for each of the five categories of networks (see section "Results" and **Figure 2**). Together, they make the CoExpNets suite of packages. When a set of GCNs is created, they are encapsulated into a package in a predefined manner. Each package must include an R object with the network, and two CSV files with the functional and cell type enrichments, respectively.

In order to generate a new GCN, a gene expression profiling matrix is required with genes as columns and samples as rows. Columns must be named with the corresponding gene IDs and rows with the sample IDs. To create the GCN, we use the getDownstreamNetwork() R function at the CoExpNets package. This produces, as output, three files, including (1) a R object file (RDS type) with the network, (2) a csv file with the gProfileR output coming from annotation of all network modules with GO, REACTOME and KEGG terms, and (3) another csv file with the cell type enrichment signals. An additional and optional file, is the one including covariates of interest for the samples (e.g., age, sex, etc.). Therefore, a network component within a CoExp network suite R package is composed of these three files plus an additional file with the gene expression profile as it was used within the getDownstreamNetwork() call.

Let us suppose we have created a number of new GCNs using the procedure above. The next step toward their integration into CoExp includes packing all of them into an R package.

[1] https://rytenlab.com/swagger/index.html

[2] https://hub.docker.com/r/soniaruiz/coexp

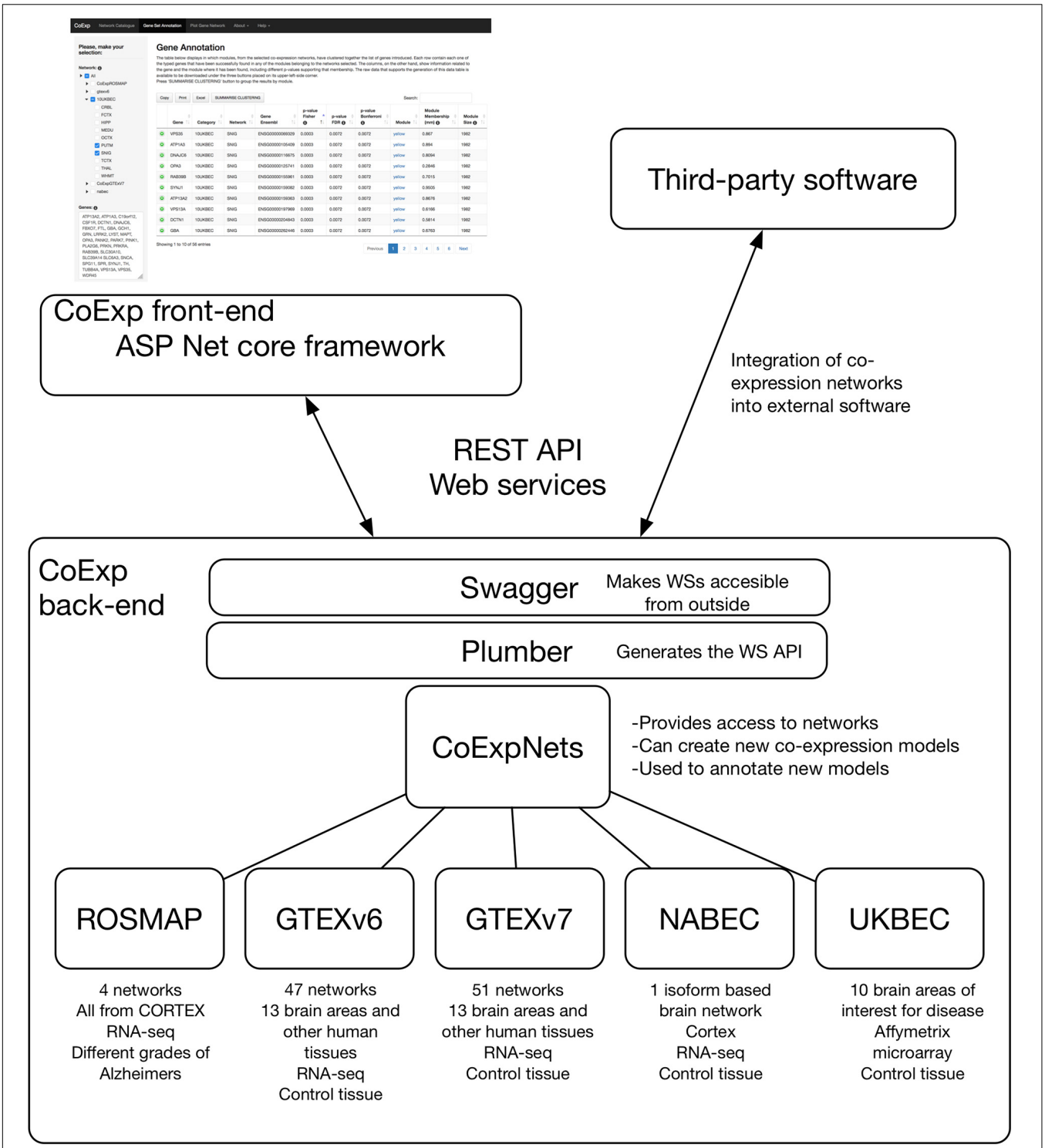

FIGURE 2 | The software architecture of CoExp. The back-end includes the CoExpNets suite of packages. These models are made accessible through Web services using Plumber and Swagger. The front-end is developed in ASP.NET but all models are also accessible for the research community through the same Web services API.

We recommend RStudio software to manage the package creation, and GitHub as the repository for making it available. We have followed the following reference (Wickham, 2015) to create all GCN suites currently available at CoExp. All include a README file explaining the network contents of the package, an R folder for the R code, a man folder for the

documentation of the functions to access the networks and an additional "i*nst*" folder with the files which together comprise each network. The R code accompanying each GCN R package must include an initDb() function which, when called installs the required file names in memory so CoExp knows which networks are available in each category and where to find each network file set when required. Two additional functions include getCovariates(), to obtain the sample covariates for each network, and generateModuleTOMs(), to create the matrix of distances between genes required to plot each network module.

All CGNs in all suites were created using gene expression profiles which were validated in their respective research projects. Moreover, only GCNs of high quality are included in all suites, i.e., they must include expressed genes above a threshold, they were checked for sample outliers, and all GCNs show abundant functional and cell type annotation across their modules.

RESULTS

The CoExp Web page consists of three separate tabs, corresponding to the three different ways of using the network models: (1) network catalogue browsing, (2) network-based annotation of gene sets, and (3) network module visualization through active graph plots.

All three tabs support the exploitation of the same collection of networks. This collection consists of 109 different co-expression networks (**Supplementary Table 1**) that belong to four different network groups: (1) the Religious Orders Study and Memory and Aging Project (ROSMAP) (Bennett et al., 2012a,b; De Jager et al., 2018) composed of four co-expression networks derived from post-mortem human frontal cortex originating from control individuals, as well as those with cognitive impairment and Alzheimer's disease; (2) The Genotype-Tissue Expression project (GTEx) V6 and V7 (The GTEx Consortium, 2015) composed of two suites of GCNs on 47 and 51 post-mortem control human tissue samples, respectively; (3) United Kingdom Brain Expression Consortium (UKBEC) (Forabosco et al., 2013; UK Brain Expression Consortium, et al, 2014) composed of 10 microarray-based gene expression profiling networks derived from post-mortem control human brain tissue; (4) North America Brain Expression Consortium (NABEC) (Dillman et al., 2017), composed of one gene co-expression network derived from post-mortem control human frontal cortex. Through CoExp, we and others have used these models to provide annotations for genes and gene sets in a variety of papers (Chelban et al., 2017, 2019; Salpietro et al., 2017, 2018; Efthymiou et al., 2019).

Many GCNs currently available within CoExp are brain-related. ROSMAP, UKBEC, and NABEC are all brain-specific GCN sets. The GTEx packages also include GCNs for 13 different brain areas. This GCN set also includes GCNs for a wide variety of human tissues. This makes it possible to compare gene sets across brain regions, but also to identify brain-specific phenomena (those not seen in alternative tissues). Furthermore, GTEx networks, enable investigation of gene sets outside of brain.

Network Catalogue Browser

The user can become familiar with the GCNs available by navigating through the catalogue. With this in mind, the first tab available on the upper menu corresponds to the "Network Catalogue" tab through which the user can inspect and download the whole network catalogue to obtain information about any network or any module within a network. To browse the catalogue, the first step consists of selecting a network category in the menu placed at the left-hand side of the webpage. The second step is the selection of a co-expression network of interest within that category. Finally, the user can select one of two different views: the 'Ontology Classification' or the "Cell Type Classification." The "Ontology Classification" view returns a data table in which each module from the selected network occupies one row. The columns provide summarized information about annotation terms enriched for the genes in the modules. The p-value column shows the enrichment obtained from gProfileR (Reimand et al., 2007), which incorporates data from well recognised ontologies, including Gene Ontology, REACTOME, and KEGG. The "Cell Type Classification" view, returns a data table in which the rows correspond to sets of gene markers relevant to specific brain cell types tested for enrichment (Fisher's Exact test) and each module occupies a column. Each cell within the table contains the Bonferroni corrected p-value for the enrichment of a set of cell type markers within a module. It is necessary to Bonferroni-correct (Benjamini and Hochberg, 1995) the Fisher's exact p-values for multiple testing, as each module is tested against all cell type marker sets. In all cases, the data table can be downloaded as an excel file, using the "Excel" button placed on the upper-left-side corner of the table. Note that it is possible to regenerate all these annotations for each network by using CoExpNets::annotate() function on the desired GCN, locally.

We can illustrate its use with an example. After clicking at the tab, let us select, for example, the 10UKBEC GCN category and then the SNIG (substantia nigra) GCN. After clicking "Accept," the "Gene Ontology" view returns a data table with a summary of the network clusters (**Figure 3**). It is notable that the "purple" module contains 498 genes expressed within UKBEC substantia nigra tissue, which are enriched for immune-related GO terms amongst others (**Figure 4**). This enrichment is unlikely to have occurred by chance based on the significant Bonferroni-corrected p-value of 8.55e-55 for the "immune system process" GO term. Similarly, after selecting the "Cell Type" view, we can see that the 498 genes within the "purple" module are enriched for microglial gene markers (p-value = 2.06e-91). Thus, navigating through only a couple of web interfaces we can explore any network module. Note that the Help section of the CoExp web has available a video illustrating how to use the Network catalogue browser under the section with the same name.

Gene Set Annotation

GCNs are often used to annotate a gene set of interest, in the context of a specific condition (e.g., a tissue of interest). "Gene Set Annotation" is the second tab within the main menu. Using this function, the user can investigate whether his/her own gene set

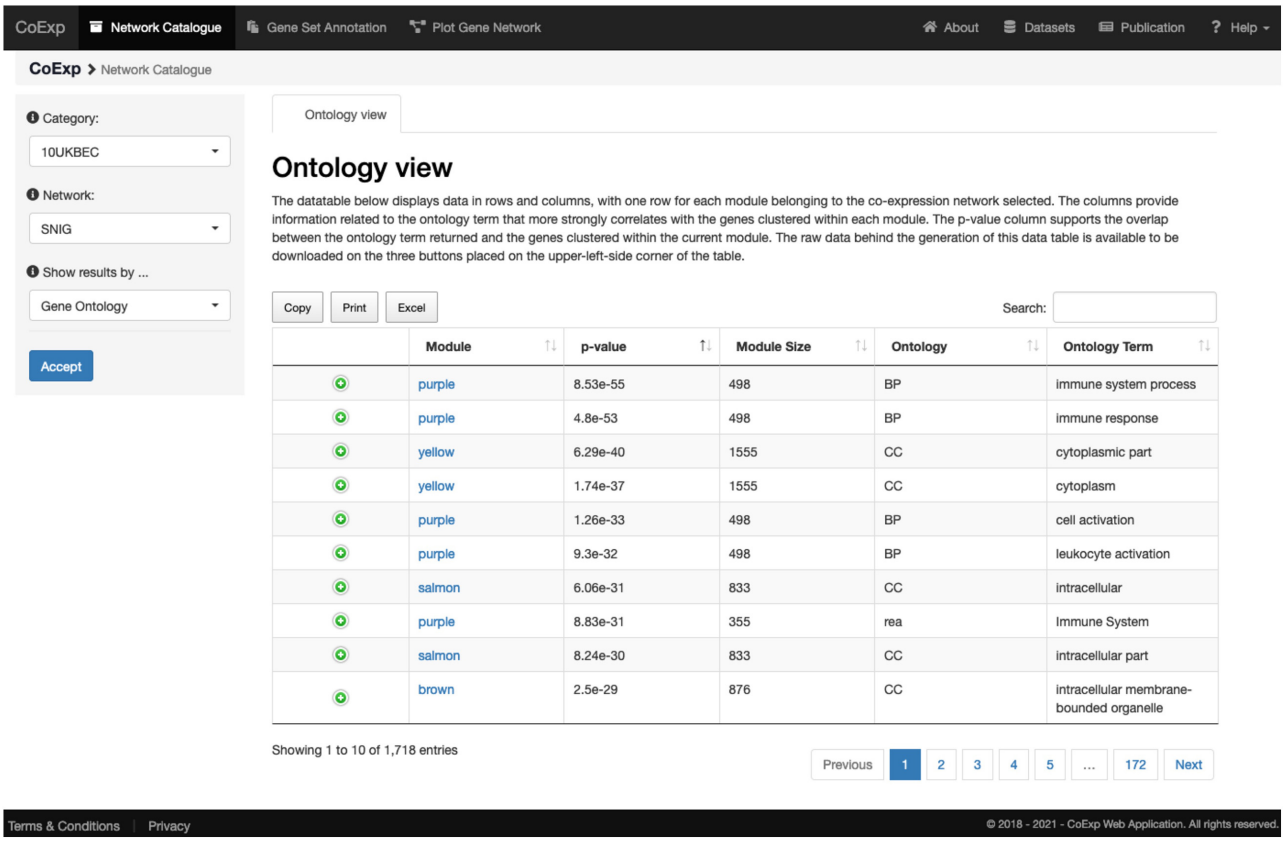

FIGURE 3 | View of the network catalogue when we select the substantia nigra (SNIG) network from the 10UKBEC network family and use the Ontology View. It shows the table corresponding to the functional enrichments obtained using gProfileR. Enriched terms are ordered by p-value by default. In this case, the first two terms provide evidence that the genes in the purple module are enriched for immune-related functions. All annotations can be downloaded in the form of an Excel file for later inspection. Furthermore, if the user is interested in a particular annotation term, he/she can use the search text frame at the top right, to look for terms matching specific keywords. A visual demonstration of all these functionalities is available, as a video, at the Help section of the CoExp web under the "Network Catalogue" section.

of interest is enriched within a single or multiple co-expression modules across all the co-expression networks from amongst those available in the catalogue. In this way, those genes can be annotated based on how they are distributed across the network modules and their biological context can be easily explored. If a gene of interest has not been found in any module a pop-up view will inform the user. The user is supplied a results table in which each row relates to a gene of interest which has been successfully found in any of the modules belonging to the network or networks selected. The columns provide information on the module in which the gene has been identified, including the statistical significance of the overlap between the input genes and the genes int the module, as well as a brief description of the module's function based on the top five most-significantly enriched GO terms. All the outputs associated with this type of analysis are available for download using the three buttons placed on the upper-left- corner of the table. Furthermore, the Gene Set Annotation tab has default values for all the choices the user has to make before proceeding with the annotation task.

Let us suppose we want to annotate 32 genes associated with Parkinson and complex parkinsonism as defined by Genomics England's PanelApp (Martin et al., 2019), and we want to study this gene set in a biologically relevant GCN such as the substantia nigra network (SNIG of the 10UKBEC network family). This would involve: (1) deciding which GCN or GCNs will be of interest, (2) using CoExp to see whether there are potentially interesting gene clusters (i.e., a subset of our genes cluster together in specific modules within the GCNs selected), (3) use the Network catalogue for a better characterization of the genes, and (4) generate network plots of the genes of interest. In order to select potential GCNs of interest, the user can inspect the network catalogue through the "Network Catalogue Tab" as we have shown in the last subsection. The step in which we obtain how genes cluster across the networks is exemplified in **Figure 5**. It demonstrates that a large proportion of the 32 genes of interest cluster together in the yellow module of the SNIG network. The next step in the analysis would be to obtain more details about the yellow module by selecting it and so navigating back to the Network Catalogue where we can access full details of the module including its functional annotation and enrichment for cellular specificity (**Figures 3**, **4**). Finally, how we obtain network plots is detailed in the following subsection. A video illustrating how

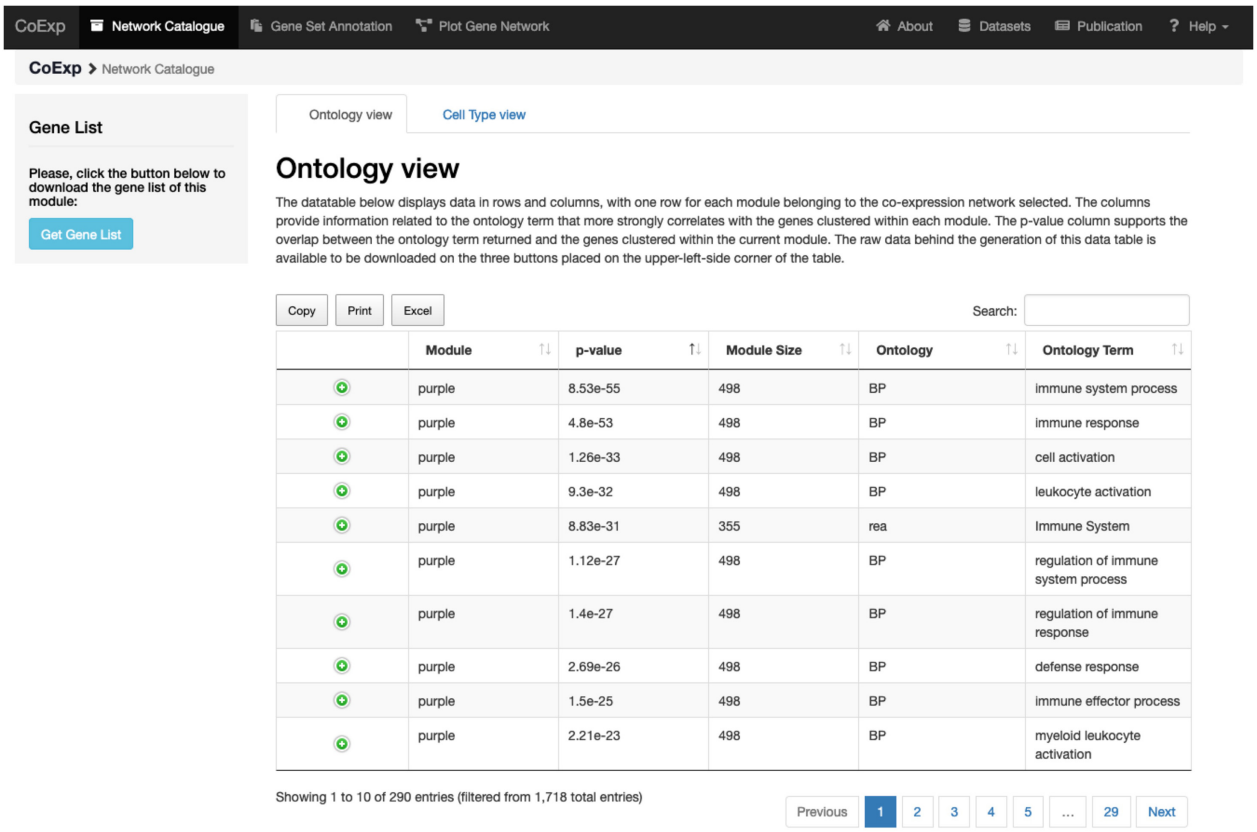

FIGURE 4 | Investigating a module of interest. By clicking on a module of interest, such as the purple module link, the user obtains a table with all the purple module-specific information on functional annotation. All annotations can be downloaded in the form of an Excel file for later inspection.

to use CoExp with this particular example is also available at the CoExp Help page, under the "Gene Set Annotation" section.

Plot Network

Once the user has decided which network is of interest, and which module within the network requires detailed visualization, the genes can be plotted. The third tab, "Plot Network," enables the graph-based visualisation of the genes within a module of interest, whether identified by browsing through the catalogue or because the user's gene set of interest significantly clusters within that module. The "Plot Network" tab generates an interactive directed graph formed by the hub genes within a module. The user can select how many of the most relevant genes will appear in the plot. The resulting plot is interactive in the sense that it can be zoomed, rotated, and the direct neighbours of any gene highlighted by just clicking on the gene of interest. Both the raw data and a high quality PNG image of the graph are available for download. See **Figure 6** for the result of selecting the ATP1A3 gene as the main gene of the network plot we want to obtain, for the running example of the 32 PD genes analysis on the SNIG 10UKBEC GCN. See also the available video for this very same example available at the CoExp web help page, under the "Plot Gene Network" section.

DISCUSSION

CoExp web is a web platform that enables the exploitation of co-expression networks. CoExp currently offers 109 co-expression models focused on brain transcriptomics with plans to expand its scope. It is a powerful, easy-to-use, and innovative tool for gene set annotation across a variety of brain-specific transcriptomic data sets, including also a variety of non-brain tissues that may be used as controls or on their own. CoExp makes co-expression models visually manageable, accessible, and easily exploitable by the scientific community. Everything is shareable in CoExp: all GCNs, the expression profiles from which they were created and their annotations are accessible within GitHub (see the links at the "Availability and Implementation" section of this paper). Furthermore, both the back and front-end software from which the CoExp Web application is generated are readily accessible such that any research laboratory can construct its own CoExp web site. This makes it a powerful tool for the wider research community interested in producing, using or sharing GCNs to support their research.

Future Works

We are extending the scope of CoExp in a number of areas. Firstly, we intend to expand the number of available GCNs.

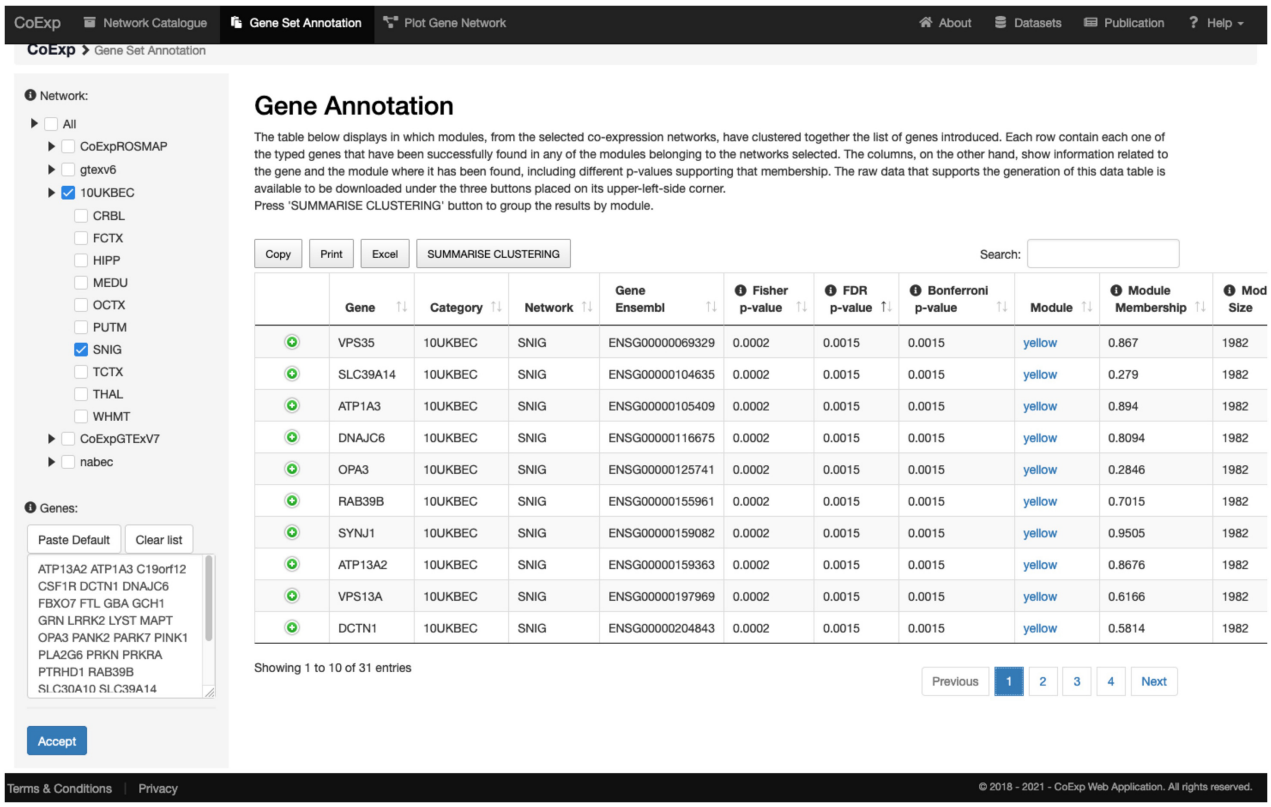

FIGURE 5 | Use of the gene set annotation tab. This screen capture shows the results of using the Gene Set Annotation tab with 32 genes associated with juvenile Parkinson's disease (visible within the text pane in the bottom left) and used as the default example when arriving at CoExp for the 1st time, together with default family and network values too (namely the SNIG network from UKBEC as used in **Figures 3**, **4**). The analysis yields the results on the right table. Each row corresponds to information about one gene out of the 32 genes used for this example. For example, in the 1st row we see that *VPS35* has a module membership of 0.867 within the SNIG yellow module. Furthermore, *VPS35* significantly clusters within the yellow module with many other juvenile Parkinson's disease genes. In fact, 10 out of the 32 genes cluster together within the yellow module. The module has 1,982 genes and the Fisher's exact test for the overlap yields a significant p-value (FDR 5% $P < 0.0072$). If we click on the "yellow" link then we can inspect this module as we did for the purple module described in **Figure 3**, **4**. In the case of the latter, we would see that that the module is enriched for multiple dopaminergic gene markers ($P < 2.78 \times 10^{-6}$) and that the most significantly enriched GO terms include transport ($P < 1.25 \times 10^{-19}$) and the establishment of localization ($P < 3.16 \times 10^{-18}$) amongst others. All annotations can be downloaded in the form of an Excel file for further inspection. For a visual demonstration of this same example, the interested user may access a video-tutorial available at the Help section of the CoExp web under the "Gene Set Annotation" section.

This will include incorporating GCNs generated using additional brain-related bulk RNA-sequencing data from projects, such as CommonMind (Hoffman et al., 2019) and PsyhcEncode (Wang et al., 2018). We aim to integrate them into CoExp as CGN suites. We are also working toward the generation of GCNs based on single-cell/single-nucleus transcriptomic datasets, including the single-nucleus RNA-sequencing data released by ROSMAP.

A second area of work is the integration of available sample covariates into the CGNs. These are very important to annotate models, as for example when a user wants to identify network modules which correlate with age of the samples, or with the case/control condition to identify disease-related modules. This will have an impact on the CoExp functions and Web interface.

A third expansion area is the inclusion of GCNs created by collaborators or any other members of the research community interested in publishing their networks. We are working on defining a CoExp network generation pipeline to satisfy a minimal level of quality for the CGN to be acceptable for publication at our Web. This particular area will also require a Web facility for automated network submission through the Web. In the meantime, we are happy to accept contributions in the form of new GCNs to be added to the CoExp catalogue. Any researcher willing to contribute to CoExp by submitting their own GCNs may contact the corresponding author and we will provide the necessary guidance.

Fourthly, the current implementation of CoExp runs on a single R environment at the back-end. This has direct impact on the number of concurrent users it can support. Concurrent requests to CoExp will be queued and attended to sequentially. We are currently working on enabling a multi-user CoExp environment thorough multiple docker images served through an HTTP proxy.

Finally, we plan to improve CoExp usability. To date, CoExp usability testing was conducted through the definition of use cases, represented as storyboards with each vignette designed to visualise input actions and the output of each interaction

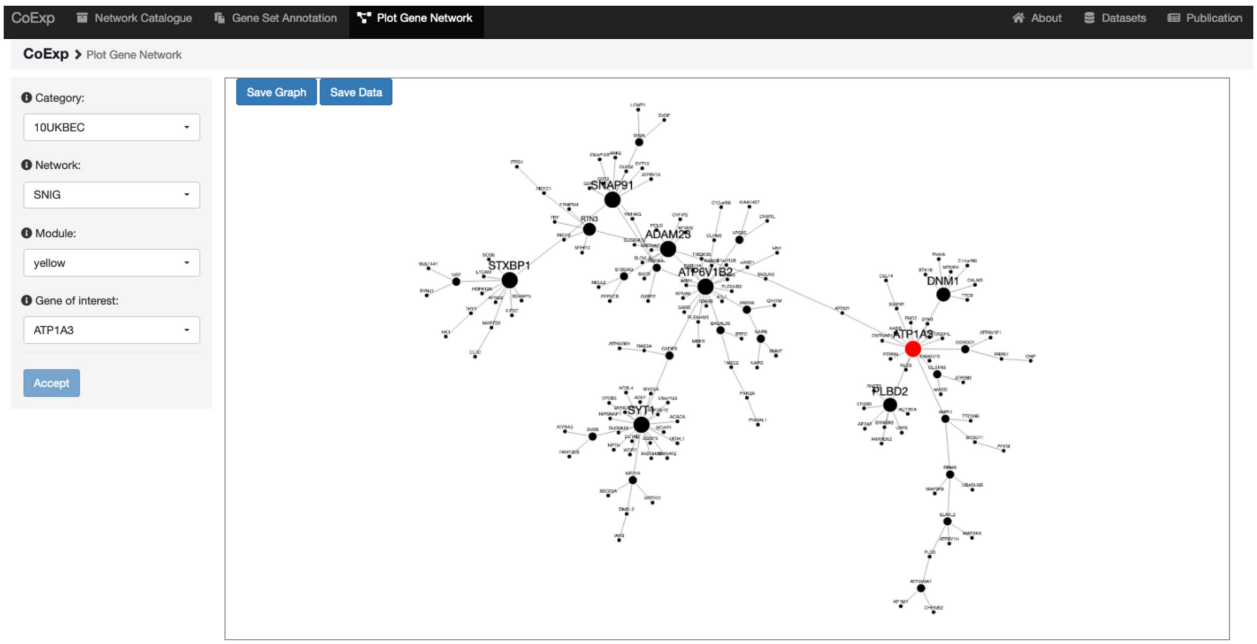

FIGURE 6 | Use of the Plot Gene Network Tab. This figure shows the view we get when we use the Plot Gene Network tab to visualize the gene *ATP1A3*. We chose this gene because it is associated with PD and is one of the genes with high module membership within the yellow module of the SNIG GCN within the 10UKBEC category (**Figure 5**). After clicking the "Accept" button, the plot shows, highlighted in red, the gene *ATP1A3* as a very important gene within its module (as indicated by the large node size). There are 16 other genes, which appear as its nearest neighbour and they can be used to further study the transcriptomic context of *ATP1A3*. Note that this example has been visually recreated in a demonstration video available at the Help section of the CoExp web under the "Plot Gene Network" section.

between the user and the Web application. These were later used at two hackathon meetings with specialised users from an international genomics consortium (the International Parkinson Disease Genetics Consortium), where there was an opportunity to work through sample questions to illustrate CoExp usage and to define possible analyses. We used the very valuable feedback gathered from the meetings to improve the CoExp user interface. Our future plans include the organization of more hackathons not only to improve CoExp usability, but also to expand its community of users.

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://rytenlab.com/coexp.

AUTHOR CONTRIBUTIONS

SG-R developed the CoExp Web tool. AG-M, AC, FJ-R, and RR participated in the creation of the gene co-expression networks and CoExpNets suite of packages. MC supervised the NABEC co-expression generation from scratch. JH participated in the project design. JB supervised the creation of the CoExpNets family of packages. MR and JB co-directed the whole project. All authors participated in the paper writing up and critically reviewed the manuscript.

FUNDING

This research was supported in part by the Intramural Research Program of the National Institute of Health, National Institute on Aging and by the Leonard Wolfson Doctoral Training Fellowship in Neurodegeneration. JH and MR were supported by the United Kingdom Medical Research Council (MRC), with JH supported by a grant (MR/N026004/) and MR through the award of a Tenure Track Clinician Scientist Fellowship (MR/N008324/1). JH was also supported by the United Kingdom Dementia Research Institute, The Wellcome Trust (202903/Z/16/Z), the Dolby Family Fund, the BRCNIHR Biomedical Research Centre, and the NIHR. AC was supported by the Science and Technology Agency, Séneca Foundation, Comunidad Autónoma Región de Murcia, Spain through the grant 20762/FPI/18. JB was supported by the same foundation through the research project 00007/COVI/20. AG-M was funded by Fundación Séneca (grant reference: 21230/PD/19).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.630187/full#supplementary-material

REFERENCES

Bakhtiarizadeh, M. R., Hosseinpour, B., Shahhoseini, M., Korte, A., and Gifani, P. (2018). Weighted gene co-expression network analysis of endometriosis and identification of functional modules associated with its main hallmarks. *Front. Genet.* 9:453. doi: 10.3389/fgene.2018.00453

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bennett, D. A., Schneider, J. A., Arvanitakis, Z., and Wilson, R. S. (2012a). Overview and findings from the religious orders study. *Curr. Alzheimer Res.* 9, 628–645. doi: 10.2174/156720512801322573

Bennett, D. A., Schneider, J. A., Buchman, A. S., Barnes, L. L., Boyle, P. A., and Wilson, R. S. (2012b). Overview and findings from the rush memory and aging project. *Curr. Alzheimer Res.* 9, 646–663. doi: 10.2174/156720512801322663

Bettencourt, C., Foti, S. C., Miki, Y., Botia, J., Chatterjee, A., Warner, T. T., et al. (2019). White matter DNA methylation profiling reveals deregulation of HIP1, LMAN2, MOBP, and other loci in multiple system atrophy. *Acta Neuropathol. (Berl.)* 139, 135–156. doi: 10.1007/s00401-019-02074-0

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., Apweiler, R., et al. (2009). QuickGO: a web-based tool for gene ontology searching. *Bioinformatics* 25, 3045–3046. doi: 10.1093/bioinformatics/btp536

Botía, J. A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., United Kingdom Brain Expression Consortium, Hardy, J., et al. (2017). An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Syst. Biol.* 11:47. doi: 10.1186/s12918-017-0420-6

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288–289. doi: 10.1093/bioinformatics/btn615

Chelban, V., Patel, N., Vandrovcova, J., Zanetti, M. N., Lynch, D. S., Ryten, M., et al. (2017). Mutations in NKX6-2 cause progressive spastic ataxia and hypomyelination. *Am. J. Hum. Genet.* 100, 969–977. doi: 10.1016/j.ajhg.2017.05.009

Chelban, V., Wilson, M. P., Warman Chardon, J., Vandrovcova, J., Zanetti, M. N., Zamba-Papanicolaou, E., et al. (2019). PDXK mutations cause polyneuropathy responsive to pyridoxal 5′-phosphate supplementation. *Ann. Neurol.* 86, 225–240.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., et al. (2018). A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* 5:180142.

de la Torre-Ubieta, L., Stein, J. L., Won, H., Opland, C. K., Liang, D., Lu, D., et al. (2018). The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* 172, 289–304.e18.

Dillman, A. A., Majounie, E., Ding, J., Gibbs, J. R., Hernandez, D., Arepalli, S., et al. (2017). Transcriptomic profiling of the human brain reveals that altered synaptic gene expression is associated with chronological aging. *Sci. Rep.* 7:16890.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48. doi: 10.1186/1471-2105-10-48

Efthymiou, S., Salpietro, V., Malintan, N., Poncelet, M., Kriouile, Y., Fortuna, S., et al. (2019). Biallelic mutations in neurofascin cause neurodevelopmental impairment and peripheral demyelination. *Brain* 142, 2948–2964. doi: 10.1093/brain/awz248

Forabosco, P., Ramasamy, A., Trabzuni, D., Walker, R., Smith, C., Bras, J., et al. (2013). Insights into TREM2 biology by network analysis of human brain gene expression data. *Neurobiol. Aging* 34, 2699–2714. doi: 10.1016/j.neurobiolaging.2013.05.001

Hoffman, G. E., Bendl, J., Voloudakis, G., Montgomery, K. S., Sloofman, L., Wang, Y. C., et al. (2019). CommonMind consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Sci. Data* 6:180.

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

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.

Leek, J. T., and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3:e161. doi: 10.1371/journal.pgen.0030161

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Ma, X., Zhao, H., Xu, W., You, Q., Yan, H., Gao, Z., et al. (2018). Co-expression gene network analysis and functional module identification in bamboo growth and development. *Front. Genet.* 9:574. doi: 10.3389/fgene.2018.00574

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7:S7. doi: 10.1186/1471-2105-7-S1-S7

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., et al. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.* 51, 1560–1565. doi: 10.1038/s41588-019-0528-2

Mencacci, N. E., Reynolds, R., Ruiz, S. G., Vandrovcova, J., Forabosco, P., Uk Brain Expression Consortium, et al. (2020). Transcriptomic analysis of dystonia-associated genes reveals functional convergence within specific cell types and shared neurobiology with psychiatric disorders. *Biorxiv* [preprint] doi: 10.1101/2020.01.31.928978

Miller, J. A., Horvath, S., and Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12698–12703. doi: 10.1073/pnas.0914257107

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 35, W193–W200.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., Lotia, S., et al. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9, 1069–1076. doi: 10.1038/nmeth.2212

Salpietro, V., Efthymiou, S., Manole, A., Maurya, B., Wiethoff, S., Ashokkumar, B., et al. (2018). A loss-of-function homozygous mutation in *DDX59* implicates a conserved DEAD-box RNA helicase in nervous system development and function. *Hum. Mutat.* 39, 187–192. doi: 10.1002/humu.23368

Salpietro, V., Zollo, M., Vandrovcova, J., Ryten, M., Botia, J. A., Ferrucci, V., et al. (2017). The phenotypic and molecular spectrum of PEHO syndrome and PEHO-like disorders. *Brain* 140:e49. doi: 10.1093/brain/awx155

Schizophrenia Working Group of the Psychiatric Genomics Consortium, Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* 50, 538–548. doi: 10.1038/s41588-018-0092-1

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

The GTEx Consortium (2015). The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660.

The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338.

UK Brain Expression Consortium, Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* 17, 1418–1428. doi: 10.1038/nn.3801

Uk Brain Expression Consortium (UKBEC), Ferrari, R., Forabosco, P., Vandrovcova, J., Botía, J. A., Guelfi, S., et al. (2016). Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. *Mol. Neurodegener.* 11:21.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362:eaat8464.

Wickham, H. (2015). *R Packages*. Sebastopol, CA: O'Reilly Media.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. *BMC Bioinformatics* 6:227. doi: 10.1186/1471-2105-6-227

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.

AVAILABILITY AND IMPLEMENTATION

The CoExp Web is accessible at https://snca.atica.um.es/coexp/.

The back and front ends code of CoExp Web application is fully available for download at GitHub at https://github.com/SoniaRuiz/CoExp_Web.

The docker images of CoExp web are available for download on Docker Hub at https://hub.docker.com/r/soniaruiz/coexp.

The CoExpNets suite of packages can be accessed in the following links:

http://github.com/juanbot/CoExpNets
http://github.com/juanbot/CoExpROSMAP
http://github.com/juanbot/CoExp10UKBEC
http://github.com/juanbot/CoExpGTEx
https://github.com/juanbot/CoExpGTExV7
http://github.com/juanbot/CoExpNABEC
Contact: juanbot@um.es

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fcell.2021.631982

Integrated Protein–Protein Interaction and Weighted Gene Co-expression Network Analysis Uncover Three Key Genes in Hepatoblastoma

Linlin Tian[1,2,3†], Tong Chen[2,3,4†], Jiaju Lu[2,3], Jianguo Yan[5], Yuting Zhang[1], Peifang Qin[1], Sentai Ding[2,3] and Yali Zhou[1,5*]*

[1] Department of Microbiology, Faculty of Basic Medical Sciences, Guilin Medical University, Guilin, China, [2] Department of Urology, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China, [3] Department of Urology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China, [4] Department of General Surgery, Shanghai Children's Hospital, Shanghai Jiao Tong University, Shanghai, China, [5] Key Laboratory of Tumor Immunology and Microenvironmental Regulation, Guilin Medical University, Guilin, China

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer, University of Oslo, Norway

Reviewed by:
Haitao Zhao, Peking Union Medical College Hospital (CAMS), China
Ravi Pandey, Jackson Laboratory for Genomic Medicine, United States

***Correspondence:**
Sentai Ding
dingsentai@126.com
Yali Zhou
zylmoli@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Cell and Developmental Biology

Received: *21 November 2020*
Accepted: *08 February 2021*
Published: *26 February 2021*

Citation:
Tian L, Chen T, Lu J, Yan J, Zhang Y, Qin P, Ding S and Zhou Y (2021) Integrated Protein–Protein Interaction and Weighted Gene Co-expression Network Analysis Uncover Three Key Genes in Hepatoblastoma. Front. Cell Dev. Biol. 9:631982. doi: 10.3389/fcell.2021.631982

Hepatoblastoma (HB) is the most common liver tumor in the pediatric population, with typically poor outcomes for advanced-stage or chemotherapy-refractory HB patients. The objective of this study was to identify genes involved in HB pathogenesis via microarray analysis and subsequent experimental validation. We identified 856 differentially expressed genes (DEGs) between HB and normal liver tissue based on two publicly available microarray datasets (GSE131329 and GSE75271) after data merging and batch effect correction. Protein–protein interaction (PPI) analysis and weighted gene co-expression network analysis (WGCNA) were conducted to explore HB-related critical modules and hub genes. Subsequently, Gene Ontology (GO) analysis was used to reveal critical biological functions in the initiation and progression of HB. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showed that genes involved in cell cycle phase transition and the PI3K/AKT signaling were associated with HB. The intersection of hub genes identified by both PPI and WGCNA analyses revealed five potential candidate genes. Based on receiver operating characteristic (ROC) curve analysis and reports in the literature, we selected CCNA2, CDK1, and CDC20 as key genes of interest to validate experimentally. CCNA2, CDK1, or CDC20 small interfering RNA (siRNA) knockdown inhibited aggressive biological properties of both HepG2 and HuH-6 cell lines *in vitro*. In conclusion, we identified CCNA2, CDK1, and CDC20 as new potential therapeutic biomarkers for HB, providing novel insights into important and viable targets in future HB treatment.

Keywords: CCNA2, CDC20, CDK1, hepatoblastoma, PPI, WGCNA

INTRODUCTION

Hepatoblastoma (HB) is caused by aberrant proliferation and/or differentiation of hepatic progenitor cell and represents a rare tumor that nevertheless accounts for most of liver tumors in infants and children (Allan et al., 2013). The majority of HB patients are diagnosed before 3 years of age, with a median age at diagnosis of 18 months (Spector and Birch, 2012). Over the past two decades, the incidence of HB has increased (Linabery and Ross, 2008; Bidwell et al., 2019), and

HB now accounts for several cases per million per year in the pediatric population (Tulla et al., 2015). Combined modality therapy, including complete surgical resection and adjuvant cisplatin-based chemotherapy, has significantly improved the prognosis for HB. However, the prognosis of patients with advanced-stage or chemotherapy-refractory HB remains poor, with a 3-year event-free survival of less than 50% (Perilongo et al., 2004; Hiyama, 2014). Therefore, it is vital to identify biomarkers that may aid the discovery of new therapeutic strategies and thus improve the clinical management of advanced-stage or chemotherapy-refractory HB.

As an increasingly popular method to detect genome-wide gene expression, the combination of expression profile data and bioinformatics analysis has become an effective modality for the identification of potential biomarkers and key pathways in various diseases. In particular, in the context of tumor research, public databases have been widely used for the analysis of gene expression data. However, previous studies have so far focused mainly on the identification of differentially expressed genes (DEGs) between tumor and normal tissue, which cannot directly unveil the associations between genes (Langfelder and Horvath, 2008), rather than identifying complex relationships between genes. Protein–protein interaction (PPI) and/or weighted gene co-expression network analysis (WGCNA) are key methodologies that enable the identification of interactions between genes (Yuan et al., 2017) and can thus further our understanding of complex biological mechanisms (Murakami et al., 2017). It has been demonstrated that critical genes and pathways of several human tumors can be identified through PPI and/or WGCNA analyses (Shi et al., 2020; Wang et al., 2020).

In the context of HB, there have been two studies investigating gene regulatory networks and interconnectivity of functionally related genes so far (He et al., 2016; Aghajanzadeh et al., 2020). He et al. (2016) preliminarily identified genes, microRNAs, and the associated pathways involved in HB. More recently, Aghajanzadeh et al. (2020) screened the DEGs using GEO2R and conducted functional enrichment analyses by the EnrichR. They constructed PPI network of the up-regulated genes and then detected the significant modules. However, neither preprocessing of the raw data nor WGCNA was conducted in their study. In addition, their study included one dataset and lacked experimental verification of the results. Based on two publicly available datasets, the present study aimed to identify highly related differential genes and hub genes as potential biomarkers for HB. A variety of R packages were utilized for a better visualization of the results. We preprocessed raw data and conducted batch effect correction. We identified DEGs between HB and normal liver tissue and subsequently conducted PPI analysis in order to detect densely connected modules and candidate key genes from PPI network. Additionally, we conducted WGCNA to detect the module displaying the highest association with HB as well as key genes. Based on the intersection of hub modules obtained from PPI or WGCNA, biological functions and molecular signaling pathways involved in HB were explored via Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses, respectively. These functional enrichment analyses were performed using the *clusterProfiler* R package. Moreover, we conducted an experimental verification of key genes by *in vitro* gene knockdown. Overall, our data may provide novel insights into important and viable targets for future HB treatment.

MATERIALS AND METHODS

Data Retrieval and Extraction

HB-related data were obtained and downloaded from the Gene Expression Omnibus (GEO[1]) database portal using the keyword "hepatoblastoma." The inclusion criteria for expression profile data were as follows: (a) the organism was *Homo sapiens*, (b) samples used for gene expression analysis included both HB tissue and normal liver tissue, (c) data for all samples were complete, and (d) HB and normal liver tissue samples could be clearly separated by principal component analysis (PCA). Only datasets that met all of the above criteria were included. Two datasets, GSE75271 (Sumazin et al., 2017) and GSE131329 (Kanawa et al., 2019), were therefore included for further analysis. GSE75271, consisting of 50 HB samples and five normal liver samples, was analyzed using GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array), while GSE131329, consisting of 53 HB samples and 14 normal liver samples, was analyzed via GPL6244 platform (Affymetrix Human Gene 1.0 ST Array).

Data Preprocessing and DEG Screening

Raw data files (*.CEL) from GSE75271 and GSE131329 were downloaded and processed. Data from GPL570 and GPL6244 platforms were imported using R packages *affy* (Gautier et al., 2004) and *oligo* (Carvalho and Irizarry, 2010), respectively. The gene expression profile probe names were transformed to gene symbols and Entrez IDs using the *hgu133plus2.db* R Bioconductor package and the *hugene10sttranscriptcluster.db* R Bioconductor package, respectively. If one gene symbol corresponded to different probes, combined average levels were considered for gene expression values (Barrett et al., 2013). All raw data were processed using data filtering, a base 2 log transformation, and quantile normalization. The imput.knn function in the *impute* R Bioconductor package was used for data filtering. After data merging, the ComBat function in the *sva* R package (Leek et al., 2012) was used for batch effect correction, and the results were verified by PCA. DEGs between HB and normal liver tissue samples were detected using the *limma* R package (Ritchie et al., 2015), with the following cut-off criteria for significance: adjusted $P < 0.05$ and $|\log2FC| > 1$.

Functional Gene and Pathway Enrichment Analysis

In order to explore the functional annotation of candidate genes, GO terms and KEGG pathway analyses were performed using the *clusterProfiler* R package (Yu et al., 2012). GO terms included biological process (BP), cellular component (CC), and molecular

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

function (MF). Adjusted *P* values below 0.05 were deemed significantly enriched.

PPI Network Establishment

We constructed a PPI network using the Search Tool for the Retrieval of Interacting Genes (STRING) online database (Szklarczyk et al., 2015), with an interaction score >0.4 set as the cut-off value. Subsequently, the Cytoscape software was utilized to visualize the PPI network (Shannon et al., 2003). The Molecular Complex Detection (MCODE) (Bandettini et al., 2012) plugin in Cytoscape was applied in order to extract densely connected modules from PPI network, with degree cut-off = 2, node score cut-off = 0.2, *K*-score = 2, and max depth = 100. The Cytoscape plugin CytoHubba (Chin et al., 2014) was utilized for the identification of key genes from the PPI analysis. We extracted the top 20 genes from both approaches, and the intersecting genes of all four approaches of CytoHubba ranking were deemed as hub genes. The four approaches of CytoHubba ranking used here were maximal clique centrality (MCC), edge percolated component (EPC), maximum neighborhood component (MNC), and node connect degree.

WGCNA

We constructed an unsigned weighted gene co-expression network using the *WGCNA* R package (Langfelder and Horvath, 2008). After data merging, batch effect correction, and exclusion of outlier samples, the complete gene expression matrix contained 8,204 genes across 116 samples. An expression matrix of 2,051 genes with the top 25% highest variance was used for WGCNA. We conducted hierarchical clustering of samples to remove outliers with a cut-off value of 80 to produce two stable clusters. Then, the soft threshold power β was determined in order to ensure a scale-free network. The resulting Pearson correlation matrix was converted to adjacency matrix via the power function, followed by transformation into a topological overlap matrix (TOM). The TOM was used to calculate corresponding dissimilarity. We carried out hierarchical clustering in order to cluster similar genes into the same module. The dynamic cutting algorithm was then used to detect the gene modules. Subsequently, we clustered the eigengenes according to the relationship and merged them into modules with the association >0.75. Module–trait association between each module and the phenotype was evaluated based on Pearson correlation. For each gene, module membership (MM) was characterized according to the association between module eigengene (ME) and its expression level. The association between gene expression and clinical phenotype represented gene significance (GS). After identifying a module of interest, GS and MM for each gene were computed in the given module. Finally, we performed GO and KEGG pathway analyses to illustrate potential biological functions of the identified module.

Identification and Verification of Critical Genes

Key genes were closely correlated genes in one module with a MM > 0.8 and a GS > 0.2. For subsequent analysis, intersecting genes identified from both PPI and the most significant modules were assessed. Based on GSE75271 and GSE131329 datasets, the expression values of key genes between HB and normal liver tissue samples were then compared.

Reagents and Antibodies

FBS (cat. no. 10099141), DMEM (cat. no. 11995065), PBS (cat. no. 10010023), and 0.25% Trypsin-EDTA (cat. no. 25200072) were purchased from Gibco (Grand Island, NY). Antibodies against CDK1 (cat. no. ab133327), CCNA2 (cat. no. ab181591), and β-actin (cat. no. ab8226) were obtained from Abcam (Cambridge, MA, United States). Antibodies against CDC20 (cat. no. 4823) and GAPDH (cat. no. 5174S) were procured from CST (Beverly, MA, United States).

Cell Culture

Human HB cell lines (HepG2 and HuH-6) were purchased from Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. Cells were cultured in DMEM supplemented with 10% FBS at 37°C/5% CO_2.

Western Blot Assay

Total proteins were extracted from HepG2 or HuH-6 cells using the RIPA buffer supplemented with a protease inhibitor cocktail. Lysates were separated using sodium dodecyl sulfate polyacrylamide gel electrophoresis and transferred onto a polyvinylidene fluoride membrane. The membrane was then blocked for 1 h using western blocking buffer and subsequently incubated using a primary antibody, followed by incubation for 2 h with IgG HRP-conjugated secondary antibody (Jackson ImmunoResearch, PA, United States). Proteins were detected using ChemiDoc-It system (Tanon, Shanghai, China). Band intensities were assessed using ImageJ. GAPDH or β-actin served as the loading control.

Small Interfering RNA

Small interfering RNAs (siRNAs) targeting CDK1, CCNA2, or CDC20, as well as non-targeting control siRNAs, were obtained from RiboBio (Guangzhou, China). siRNAs were transfected into HepG2 or HuH-6 cell lines according to the manufacturer's guidelines using Lipofectamine 2000. Transfection efficiency was confirmed via western blot (WB) 2 days after siRNA transfection.

Colony Formation Assay

HB cells were cultured in six-well plates containing media supplemented with 10% FBS to a density of 3×10^3 cells/well. The culture media were replaced by media containing 5% FBS the following day, and cells were cultured for 2 weeks. This step was followed by paraformaldehyde (PFA) fixing and staining with crystal violet. Subsequently, photos were taken. Cells were subsequently fixed using PFA, stained with crystal violet, and microscopic images were acquired.

Cell Viability Assay

Cell viability was assessed using Cell Counting Kit-8 (CCK-8, Dojindo, Japan). Briefly, cells were seeded into 96-well plates

(1×10^3 cells per well) and cultured for 4 h until adherence. The CCK-8 agent was added in each well at the indicated time-point, and the optical density at 450 nm was assessed after 1 h using a plate reader.

Transwell Invasion Assay

The transwell invasion assay was carried out using six-well plates containing transwell inserts (8-μm pore size; BD Biosciences) according to the manufacturer's guidelines. Matrigel purchased from BD Biosciences was added to serum-free media, transferred to the top chamber, and incubated for 5 h. Subsequently, cells were cultured in the top chamber supplemented with serum-free medium. The lower chamber was supplemented with 10% FBS, and cells were removed from the top chamber after 36-h incubation. For quantification of the cells in the lower chamber, membranes were PFA-fixed, stained with crystal violet, and invading cells were quantified using microscopy image analysis.

Statistical Analysis

Data analysis was carried out using R (version 3.6.3) and GraphPad Prism (version 8.0.1). Gene expression levels between HB and normal liver tissue samples were compared using the Student's *t*-test. To evaluate the predictive value of each hub gene for the distinction between HB and normal liver tissue, we applied the receiver operating characteristic (ROC) curve. An area under curve (AUC) > 0.90 and $P < 0.05$ indicated statistical significance.

RESULTS

DEG Screening and Functional Annotation

A detailed outline of our study is summarized in **Figure 1**. For our analysis, we combined two publicly available microarray gene expression datasets of HB and normal liver tissue samples. We carried out PCA to visualize data before and after batch effect correction, during which four outlier samples (GSM1948577, GSM1948562, GSM1948566, and GSM3770543) were removed (**Figures 2A–C**), resulting in a total of 99 HB samples and 19 normal liver samples after data preprocessing and quality control. We applied a filtering step (P value < 0.05 and $|\log2FC| > 1$) for the identification of DEGs, which resulted in a total of 856 DEGs. Among these DEGs, 350 were up-regulated, while 506 were down-regulated, with a volcano plot presented in **Figure 2D**. The heatmap of the top 100 genes is shown in **Figure 2E**. We conducted GO and KEGG pathway analyses to elucidate the biological functions and potential signaling pathways these genes may be involved in. GO analysis suggested that DEGs predominantly consisted of genes involved in small molecule catabolic processes, the collagen-containing extracellular matrix, and coenzyme binding (**Figures 3A–C**). KEGG analysis identified enrichment for PI3K-AKT signaling, cell cycle, and FoxO signaling (**Figure 3D**). Critical pathways, including the cell cycle, FoxO pathway, NF-kappa B signaling pathway, amoebiasis, and carbon metabolism, are presented along with their related genes in **Figure 3E**. For the term cell cycle, all enriched DEGs were up-regulated with the exception of GADD45B and GADD45D. It should be noted that we observed two genes concurrently enriched in three critical pathways: TGFB2 was enriched in the cell cycle, FoxO signaling, and amoebiasis pathways, while GADD45B was enriched in the cell cycle, FoxO signaling, and NF-kappa B signaling pathways.

PPI Network Establishment and Module Analyses

A PPI network containing 791 nodes and 9,054 edges was conducted using the Cytoscape software based on results of the STRING online database (**Figure 4A**). All four methods within the CytoHubba plugin were adopted, and the top 20 genes of each method were listed (**Table 1**). The intersecting genes that were concurrently listed in the four methods were regarded as hub genes (AURKA, AURKB, CDK1, CCNA2, CDC20, and PLK1) for PPI analysis. Nineteen clusters were obtained after module analysis using the MCODE plugin of Cytoscape, and we selected the top three modules as hub modules based on MCODE scores (**Figures 4B–D**). Notably, all six hub genes were found in module 1, which played an essential role in the constructed PPI network. Specifically, module 1 contained 59 nodes and 1,600 edges and had the highest MCODE score (55.172) of all modules. Another notable observation from module analysis was that all genes from module 1 exhibited up-regulation. Subsequently, we conducted GO and KEGG analyses of genes in module 1 using the R *clusterProfiler* package. For BP within the GO analysis, we found that genes in module 1 played a critical role in nuclear division, organelle fission, cell cycle transition, mitotic cell cycle transition, as well as chromosome segregation (**Figure 5A**). For CC within the GO analysis, we found that up-regulated genes were significantly enriched in the chromosomal region, condensed chromosome, and spindle (**Figure 5B**). The MF of GO analysis showed that genes were associated with ATPase activity, catalytic activity, action on DNA, protein serine/threonine kinase activity, and single-stranded DNA binding (**Figure 5C**). KEGG analysis showed that genes from module 1 were enriched for the cell cycle, DNA replication, as well as oocyte meiosis pathways (**Figure 5D**).

WGCNA and Hub Module Identification

During sample clustering, two samples were regarded as outliers and thus excluded (GSM1948574 and GSM3770517; **Supplementary Figure 1A**). Besides, we identified $\beta = 10$ and $R^2 = 0.88$ as the optimal soft threshold parameters to guarantee a scale-free network (**Supplementary Figures 1B,C**). We set clustering height cut-off to 0.25 in order to merge similar modules, which resulted in seven modules (**Figure 6A**). Specifically, blue, black, brown, pink, green, and magenta modules were identified as significant modules (**Figure 6B**). The blue module containing 259 genes appeared to be the most relevant module involved in HB. The top 100 genes of the blue module, ranked by gene significance for cancer, are listed in **Supplementary Table 1**. Subsequently, the module eigengenes and associations between eigengenes and sample types were

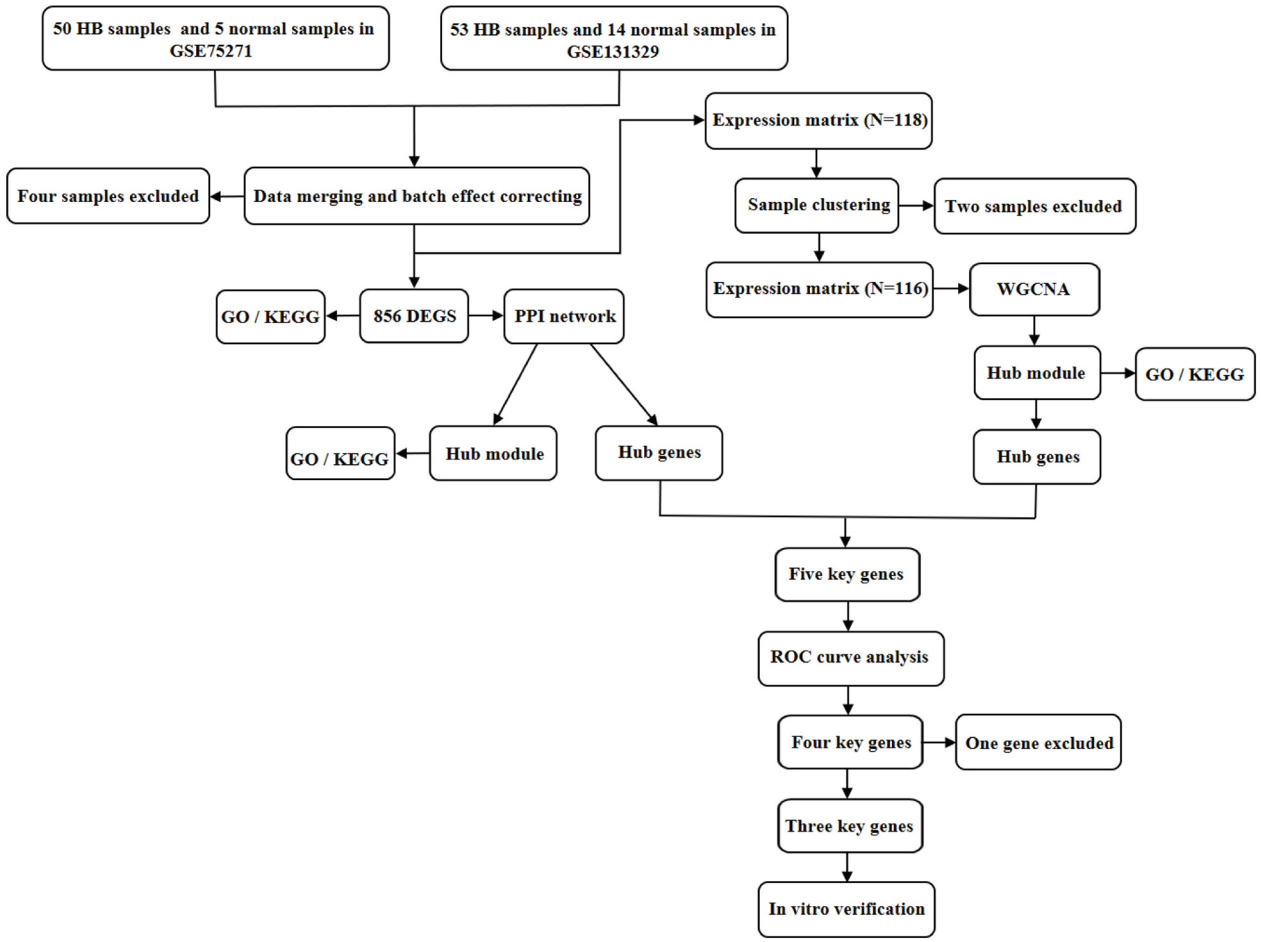

FIGURE 1 | Flowchart illustrating the study design. DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein–protein interaction; WGCNA, weighted gene co-expression network analysis; ROC, receiver operating characteristic.

computed. The module eigengene dendrogram was plotted, and the seven modules were divided into two clusters. Similar results were obtained from eigengene network heatmap (**Figure 6C**). Interestingly, the blue module was not only located close to cancer but also had a markedly positive association with cancer, meaning that genes in the blue module may be essential for tumor progression. Moreover, module–trait relationship analysis confirmed the highly positive correlation between the blue module and cancer (r = 0.64, P = 1e-14) (**Figure 6D**). When focusing on the blue module (**Figure 6E**), we substantiated a significantly positive association between MM and GS (r = 0.64, P = 6e-38). Consequently, the blue module was chosen for functional enrichment analysis, during which we aimed to elucidate potential biological processes involved in HB.

GO and KEGG Functional Enrichment Analyses of the Blue Module

In order to explore potential genes and pathways associated with HB growth, we conducted GO and KEGG analyses on the blue module identified by WGCNA. KEGG analysis indicated that genes in the blue module were markedly enriched for the cell cycle, oocyte meiosis, and DNA replication pathways (**Figure 6F**). Key pathways and their associated genes are shown in the heatmap (**Figure 6G**). Additionally, GO analysis demonstrated that genes in the blue module were primarily associated with organelle fission, nuclear division, chromosomal region, tubulin binding, and ATPase activity (**Figures 7A–C**). For the description of functionally enriched GO clusters, we utilized cnetplots to highlight the relationships between genes and critical pathways (**Figures 7D–F**).

Selection and Verification of Key Genes

Several key genes identified using PPI analysis were also included in the WGCNA blue module, with the intersecting genes being AURKA, AURKB, CDK1, CCNA2, and CDC20. The scoring of each hub gene in PPI and WGCNA is summarized in **Table 2**. For further validation of these potential key genes, we compared their expression values between HB and normal liver samples in the GSE75271 and GSE131329 datasets. Expression levels of these five key genes were markedly elevated in HB samples compared with normal liver samples (**Figure 8A**). ROC curve

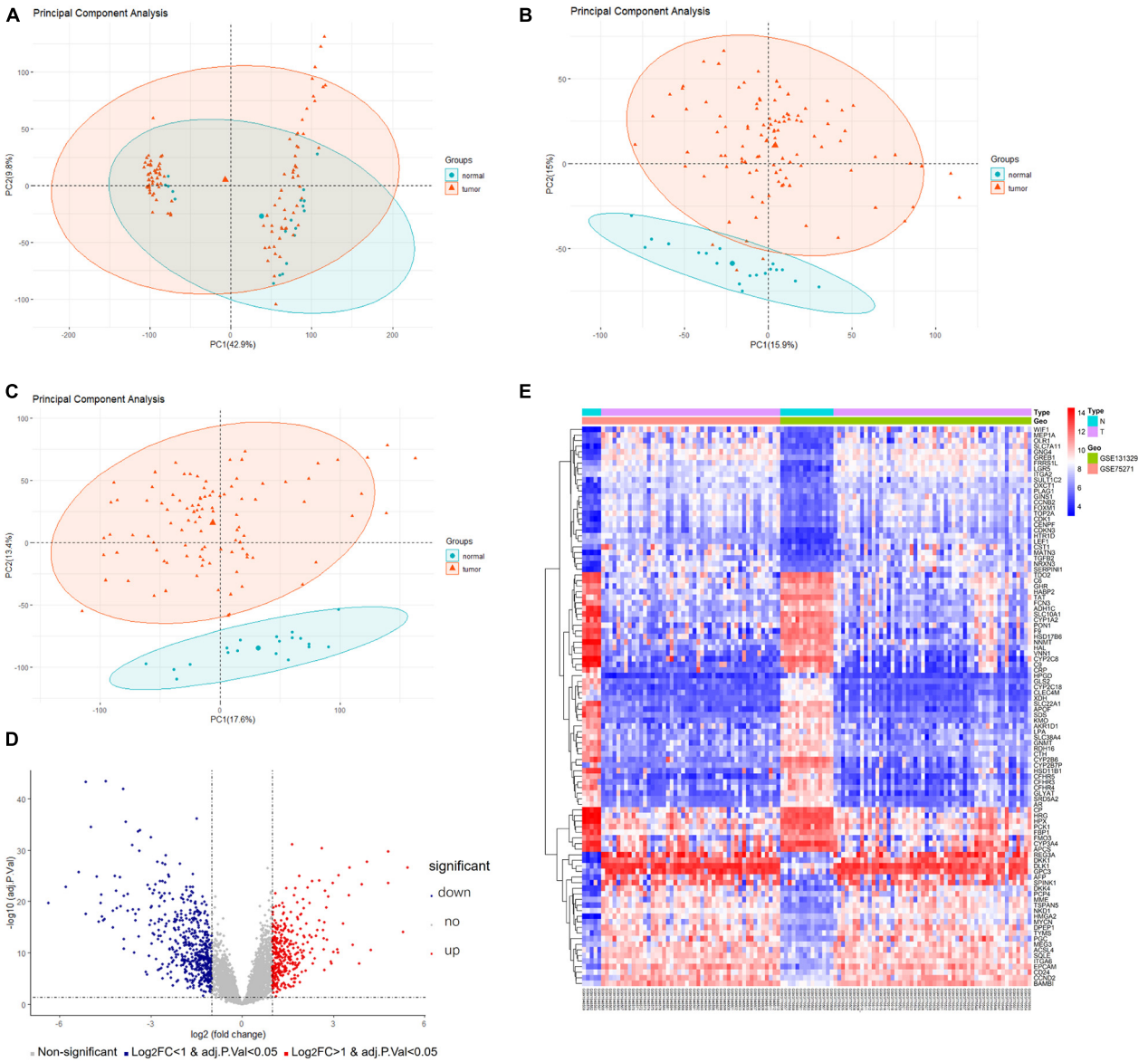

FIGURE 2 | Data preprocessing and DEG analysis of the GSE75271 and GSE131329 datasets. Principal component analysis indicating the overall profiles of two datasets **(A)** before and **(B)** after batch effect correction and data merging. **(C)** Principal component analysis after removal of outlier samples. **(D)** Volcano plots visualizing DEGs between HB and normal liver tissue samples from the two datasets. Red points represent up-regulation, while blue points indicate down-regulation; gray points represent normal expression. **(E)** Heatmap of the top 50 up-regulated and top 50 down-regulated DEGs with *P* value <0.05 and logFC > 1. Red points represent up-regulation; blue points indicate down-regulation. DEG, differentially expressed gene.

was utilized to evaluate the predictive value of each hub gene for the distinction between HB and normal liver tissue. The AUC of expression levels for four of the genes exceeded 0.90 in the ROC analysis (**Figure 8B**). Specifically, the AUC was 0.918 (95% CI, 0.865–0.970) for AURKA, 0.964 (95% CI, 0.933–0.994) for CDK1, 0.952 (95% CI, 0.915–0.990) for CCNA2, and 0.928 (95% CI, 0.882–0.973) for CDC20. A literature search revealed AURKA as a previously reported oncogenic gene in HB, and elevated expression levels of AURKA have been associated with an advanced COG stage as well as metastasis (Zhang et al., 2018; Tan et al., 2020). However, the role of CDK1, CCNA2, or CDC20 in HB growth has not been reported to date, and thus, we selected these three genes for subsequent experimental validation.

CDK1, CCNA2, or CDC20 Knockdown Inhibits Proliferative, Migrative, and Invasive Capacities of HB Cell Lines

In order to investigate the influence of CDK1, CCNA2, or CDC20 expression in HB cells, we knocked down CDK1, CCNA2, or CDC20 in HepG2 and HuH-6 cells via siRNAs. The knockdown efficiency of each hub gene was validated by WB analysis (**Figure 9A**). After transfection of CDK1-siRNA into

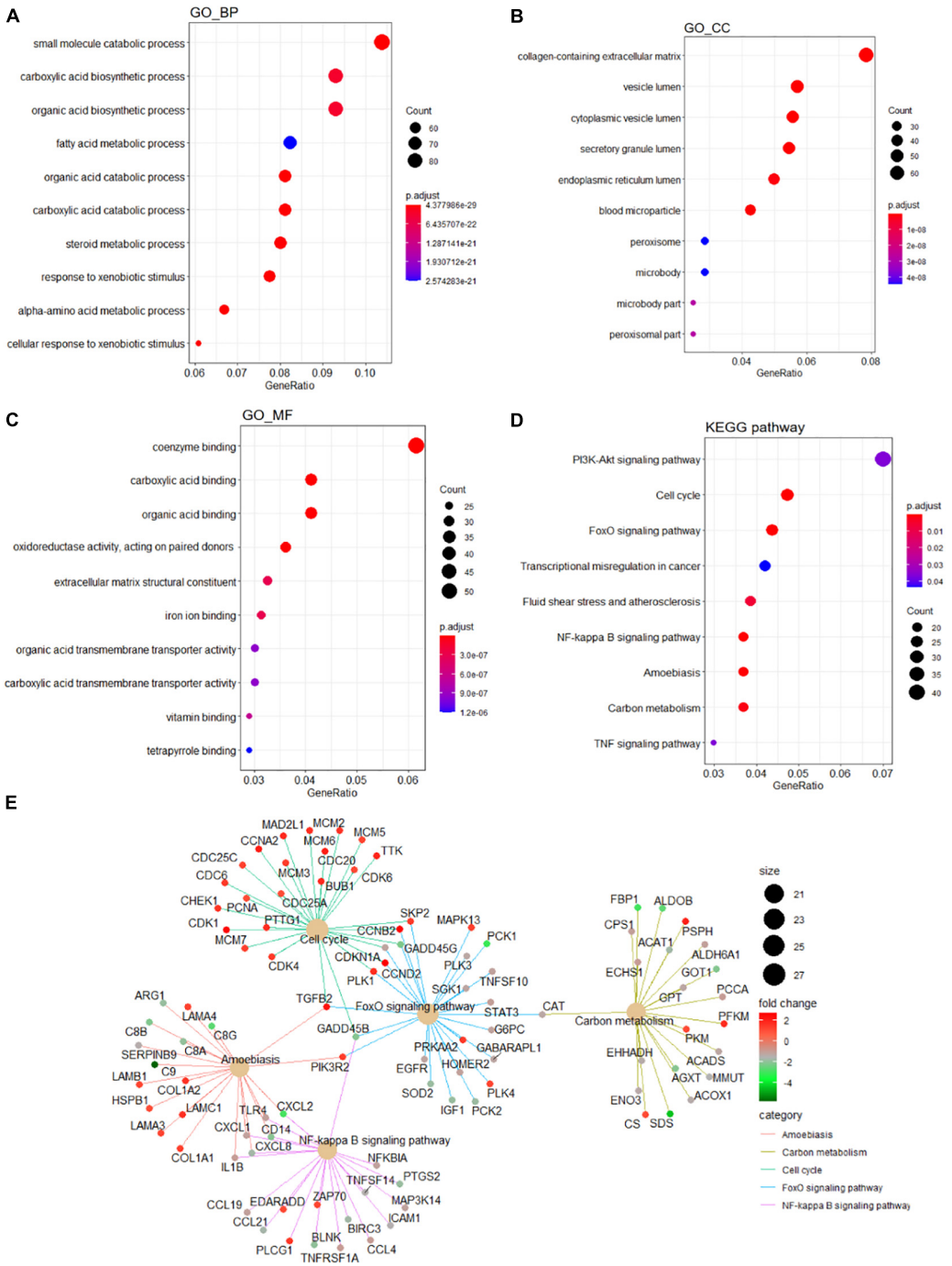

FIGURE 3 | Functional enrichment analyses of the DEGs. GO analysis containing **(A)** BP terms, **(B)** CC terms, and **(C)** MF terms. **(D)** KEGG pathway analysis of the DEGs. **(E)** The cnetplot of KEGG pathways showing genes enriched in different pathways. The symbol adjacent to nodes represents the specific gene. The color bar represents the fold change of genes in the respective pathways. DEGs, differentially expressed gene; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; CC, cellular component; MF, molecular function.

HepG2 and HuH-6 cell lines, the effect of CDK1 knockdown on cell proliferation was explored using both a CCK-8 assay (**Figure 9B**) and a colony formation assay (**Figures 9C,D**). These assays indicated a significantly lower proliferative ability of the CDK1-siRNA group compared to the control siRNA group. Similar effects were observed for CCNA2 or CDC20 knockdown in HB cells (**Figures 9B–D**). Next, we evaluated the effect of CDK1, CCNA2, or CDC20 knockdown on the invasive ability of HB cells using a transwell invasion assay (**Figures 9C,D**), which revealed a significantly decreased rate of the relative

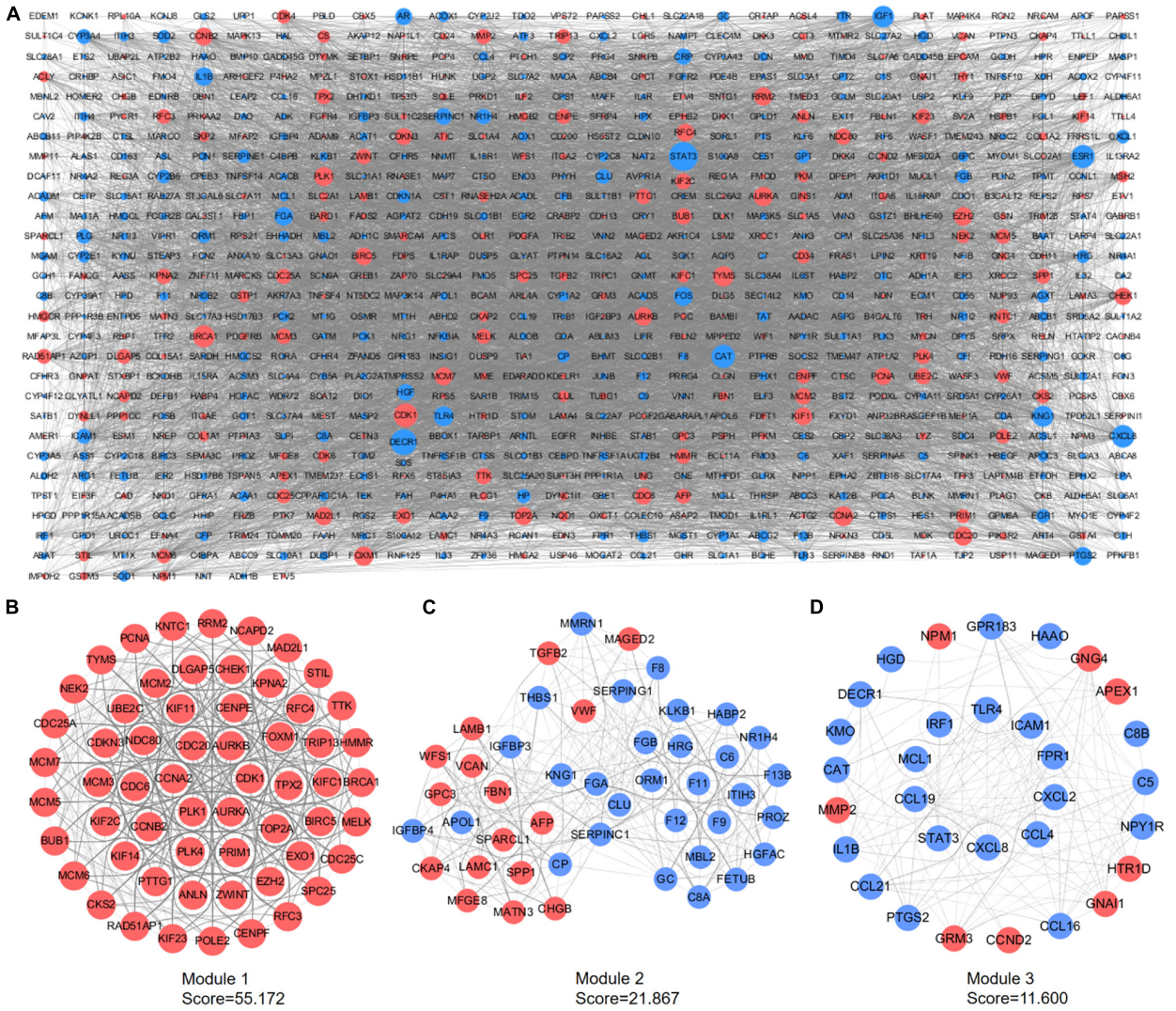

FIGURE 4 | PPI network construction and module analyses. **(A)** PPI network of DEGs was constructed in Cytoscape. Red points represent up-regulated genes, while blue points represent down-regulated genes. The node size depends on the degree of node connectivity; edges indicate straight associations. **(B)** Module 1 contains 59 nodes and 1,600 edges. **(C)** Module 2 contains 46 nodes and 492 edges. **(D)** Module 3 includes 31 nodes and 174 edges. Red nodes represent up-regulated genes; blue nodes represent down-regulated genes. DEG, differentially expressed gene; PPI, protein–protein interaction.

invasive cells relative to controls for all three knockdown models. Lastly, wound-healing assays revealed that CDK1, CCNA2, or CDC20 siRNA knockdown groups exhibited a markedly lower relative migration distance than the control did (**Figures 9E,F**). Taken together, these results demonstrated that knockdown of CDK1, CCNA2, or CDC20 inhibited proliferative, migratory, and invasive capabilities in both HepG2 and HuH-6 cell lines.

DISCUSSION

HB is the most common liver tumor in the pediatric population, and its incidence has been consistently increasing in the last years. Surgical resection and adjuvant cisplatin-based chemotherapy may severely affect the health-related quality of life of HB patients and their families, and the therapeutic efficacy in patients with advanced-stage or chemotherapy-refractory HB is unsatisfactory. Therefore, further exploring the molecular mechanisms of HB is essential for early diagnosis and better treatment strategy.

The results of our study showed that DEGs between HB and normal liver tissue samples were primarily associated with PI3K-AKT signaling, cell cycle, and FoxO signaling. Forty DEGs were associated with PI3K-AKT signaling, indicating that these genes may be critical for HB growth. Indeed, a previous study reported that inhibition of the PI3K/AKT signaling pathway resulted in suppressed proliferation and increased apoptosis of HB cells (Hartmann et al., 2009). Moreover, FoxO signaling has been reported as a key signaling pathway closely associated with PI3K/AKT signaling in many human tumors (Farhan et al., 2017), and inhibition of FoxO signaling has been shown to lead to cell

TABLE 1 | Hub genes identified using the Cytohubba plugin (Cytoscape).

Category	Ranking methods in the CytoHubba plugin			
	MCC	**EPC**	**MNC**	**Degree**
1	*KIF11*	*KIF11*	***AURKA***	***AURKA***
2	*RRM2*	*RRM2*	***AURKB***	***AURKB***
3	***AURKA***	***AURKA***	*FOXM1*	*TLR4*
4	*TTK*	***AURKB***	***CDK1***	*FOXM1*
5	***AURKB***	*MAD2L1*	***CCNA2***	*DECR1*
6	*MAD2L1*	*FOXM1*	*EZH2*	***CDK1***
7	*DLGAP5*	*TOP2A*	*TYMS*	***CCNA2***
8	*TOP2A*	***CDK1***	***CDC20***	*EZH2*
9	***CDK1***	***CCNA2***	***PLK1***	*TYMS*
10	***CCNA2***	*TYMS*	*BRCA1*	*KNG1*
11	*CCNB2*	*CCNB2*	*TLR4*	*IGF1*
12	*UBE2C*	*UBE2C*	*DECR1*	***CDC20***
13	***CDC20***	***CDC20***	*KNG1*	*FOS*
14	*BIRC5*	***PLK1***	*IGF1*	***PLK1***
15	***PLK1***	*MCM7*	*FOS*	*EGFR*
16	*MELK*	*CDC6*	*EGFR*	*STAT3*
17	*KIF23*	*TPX2*	*STAT3*	*ESR1*
18	*CDC6*	*RFC4*	*ESR1*	*CXCL8*
19	*TPX2*	*BRCA1*	*CXCL8*	*CAT*
20	*BUB1*	*CHEK1*	*CAT*	*BRCA1*

MCC, maximal clique centrality; EPC, edge percolated component; MNC, maximum neighborhood component; Degree, node connect degree. The bold values represent the genes that appear in all four ranking methods used here.

cycle arrest, apoptosis, and the suppression of PI3K/AKT/mTOR signaling in liver tumor (Carbajo-Pescador et al., 2014).

During WGCNA analysis, the blue module appeared to be the most relevant module involved in HB. The molecular function analysis revealed that genes in the blue module were enriched for the tubulin binding, microtubule binding, and microtubule motor activity pathways. Disrupted microtubule dynamics have previously been reported to modulate cell proliferation in several human tumors, including hepatocellular carcinoma (HCC) (Zhang et al., 2016; Aboubakr et al., 2017). The cellular component analysis revealed that genes in the blue module were predominantly associated with spindle, kinetochore, and mitotic spindle cellular components, which serve essential functions during mitosis (Sharp et al., 2000). Collectively, GO analysis showed that genes in the blue module were enriched for pathways such as organelle fission, cell cycle phase transition, chromosomal region, ATPase activity, catalytic activity, and acting on DNA. KEGG analysis indicated that genes in the blue module were enriched in the cell cycle, oocyte meiosis, and DNA replication pathways. Notably, the results from WGCNA GO/KEGG analysis were similar to the functional annotations of genes in the most significant module of PPI network.

The cell cycle is a set of organized and monitored stages through which a cell passes between cell divisions. Cells pass through the G0/G1, S, and G2 phases of interphase and subsequently directly enter the M phase, in which nuclear and cell division takes place (Norbury and Nurse, 1992). The progression from one stage of the cell cycle to another is controlled at checkpoints, which are regulated by interactions between cyclin-dependent kinases (CDKs) and their cyclin partners. Deregulation of the cell cycle may result in unscheduled proliferation, chromosome segregation defects, and ultimately the development of tumor (Bannon and Mc Gee, 2009). Indeed, cell cycle proteins are frequently overactive in tumor cells, and blocking cell cycle progression through inhibiting cell cycle proteins can lead to cell proliferation arrest in many tumor types. For instance, the retinoblastoma (Rb) gene encodes a tumor suppressor protein that is responsive to mitogenic signals to integrate the control of cell cycle (Hanahan and Weinberg, 2011). In tumor cells, defects in the Rb pathway give rise to the deregulation of the G1/S-phase cell cycle checkpoint, which in turn can lead to uncontrolled cell proliferation (Dyson, 1998). Using an approach combining bioinformatics analysis and subsequent experimental verification, we identified CDK1, CCNA2, and CDC20 as pivotal genes and potential biomarkers for future HB therapy. Interestingly, we found that all of these three key genes were involved in cell cycle (**Figures 3E**, **5D**, **6F,G**).

CCNA2 has previously been reported to be associated with chromosomal instability, epithelial–mesenchymal transition (EMT), and metastasis in tumors (Cheung et al., 2015). Specifically, CCNA2 binds to and activates CDK1 and CDK2, resulting in the formation of CDK/CCNA2 complex. It has been demonstrated that the CDK/CCNA2 complex drives S-phase progression (Girard et al., 1991; Yam et al., 2002), persists through the S and G2 phases, and is degraded upon entry into mitosis (den Elzen and Pines, 2001). Conversely, a decreased proliferative capacity of tumor cells has been observed after inhibition of the

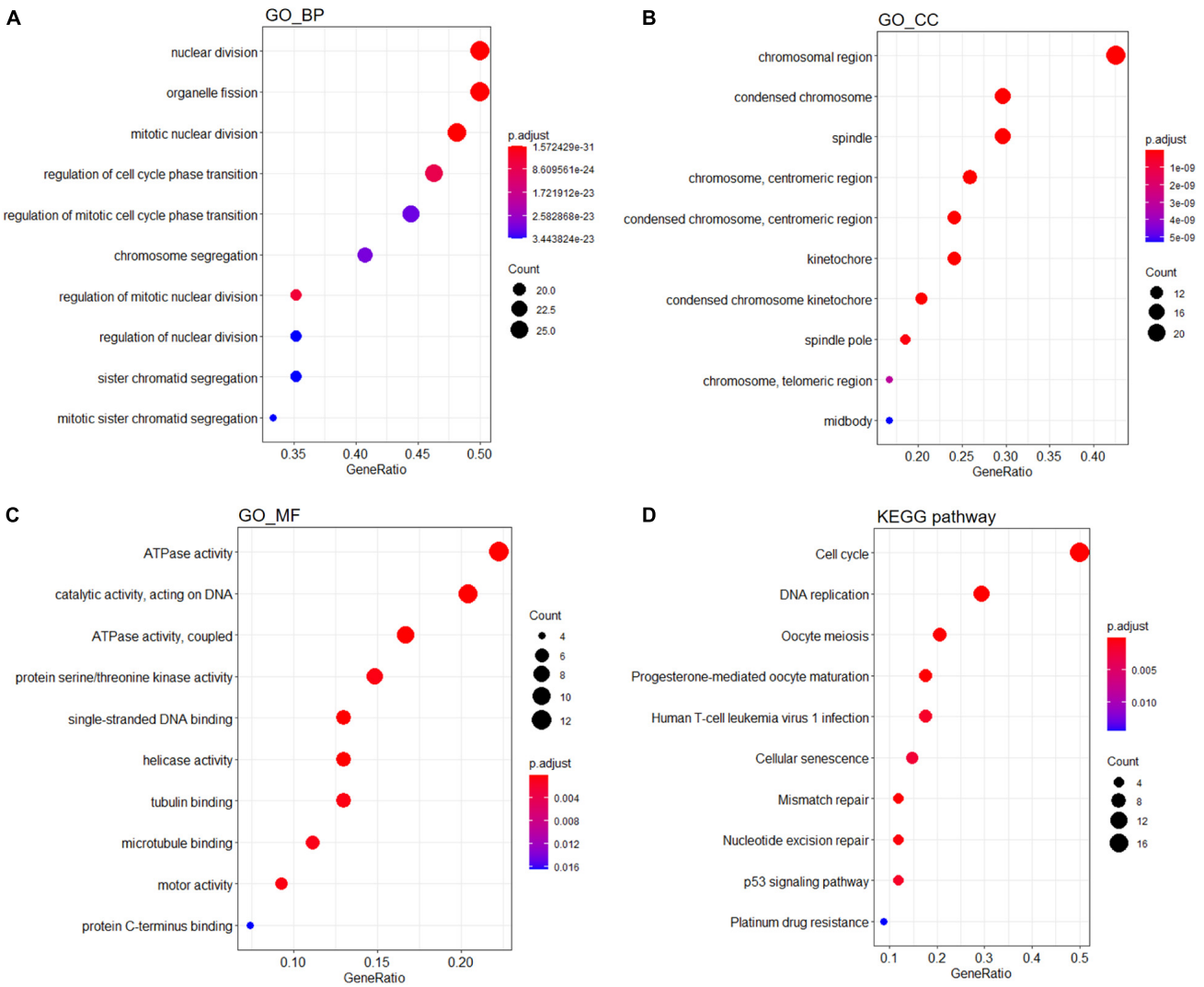

FIGURE 5 | Functional enrichment analyses of genes from module 1. GO analysis containing **(A)** BP, **(B)** CC, and **(C)** MF terms. **(D)** KEGG analysis of significantly enriched pathways of genes in module 1. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; CC, cellular component; MF, molecular function.

CDK/CCNA2 complex (Chen et al., 2004). Animal experiments indicated that a CCNA2 deficiency in hepatocytes may lead to the delayed formation of liver tumors (Gopinathan et al., 2014). At the cellular level, argininosuccinate lyase may promote HCC progression in association with CCNA2 (Hung et al., 2017). Our integrated microarray analysis revealed an upregulation of CCNA2 in HB tissues (**Figure 8A**), which is in line with results from a previous study (Shin et al., 2011). Moreover, *in vitro* experiments from our current study demonstrated, for the first time, that CCNA2 knockdown suppresses the proliferative, migrative, and invasive capacities of two HB cell lines.

In addition to regulation by CCNA2, the cell cycle is also modulated by CDKs via catalyzing phosphorylation of specific proteins (Ubersax et al., 2003). CDK1, one member of CDK family, is essential for mitosis, and inhibition of CDK1 has been shown to promote apoptosis in lymphomas and liver tumors in mice expressing MYC: in MYC-expressing HB transgenic mouse models, administration of a CDK1 inhibitor resulted in reduced tumor growth as well as extended survival (Goga et al., 2007). Taken together, these findings illustrate that CDK1 inhibition might specifically suppress the proliferative capacity of tumor cells. Similar to previous studies, in the present study, we demonstrate a significantly higher expression of CDK1 in HB tissue relative to normal liver tissue. Additionally, our functional assays indicated that CDK1 knockdown suppressed proliferative, migrative, and invasive properties of two HB cell lines.

In addition to CCNA2 and CDK1, our study also identified CDC20 as a key hub gene involved in HB growth, and subsequent experiments further demonstrated that aggressive biological behaviors of HB cell lines were inhibited after CDC20 knockdown. Previous studies have reported aberrantly high expression levels of CDC20 in oral squamous cell carcinoma

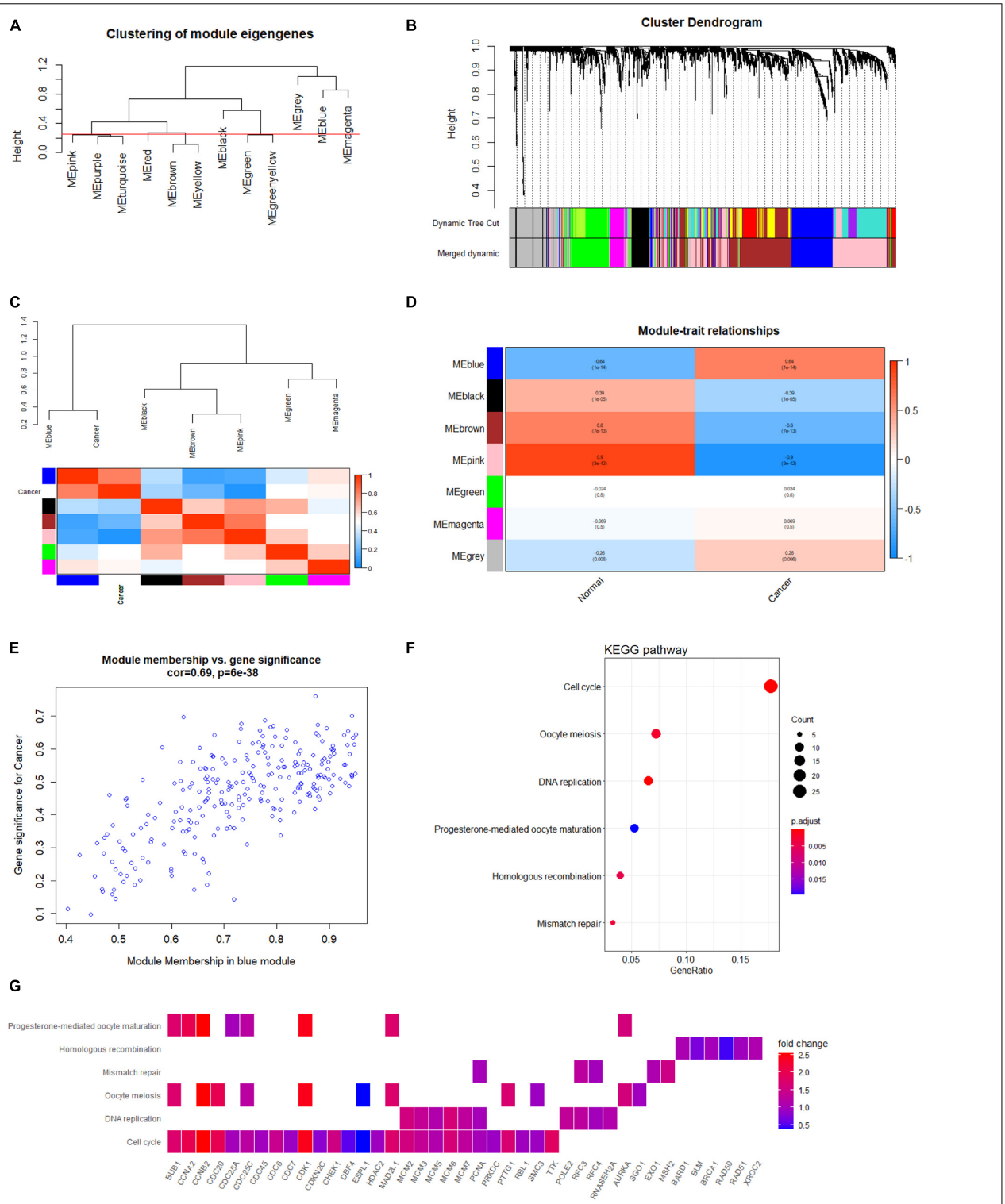

FIGURE 6 | Co-expression network analysis based on WGCNA. **(A)** Clustering of module eigengenes with a threshold of 0.25 height to identify similar modules. **(B)** Identification of HB-specific modules. Each branch represents an expression module of a highly interconnected groups of genes; each color indicates a corresponding co-expression module. **(C)** Heatmap of the eigengene network indicates correlations between different modules; tightly connected modules are clustered together. **(D)** Heatmap of associations among module eigengenes in normal liver and HB samples. **(E)** Scatter plots highlighting the association between GS and MM based on genes from the blue module. **(F)** KEGG analysis of significantly enriched pathways based on genes from blue module. **(G)** Heatmap of specific genes associated with each enriched key pathway. WGCNA, weighted gene co-expression network analysis; HB, hepatoblastoma; GS, gene significance; MM, module membership; KEGG, Kyoto Encyclopedia of Genes and Genomes.

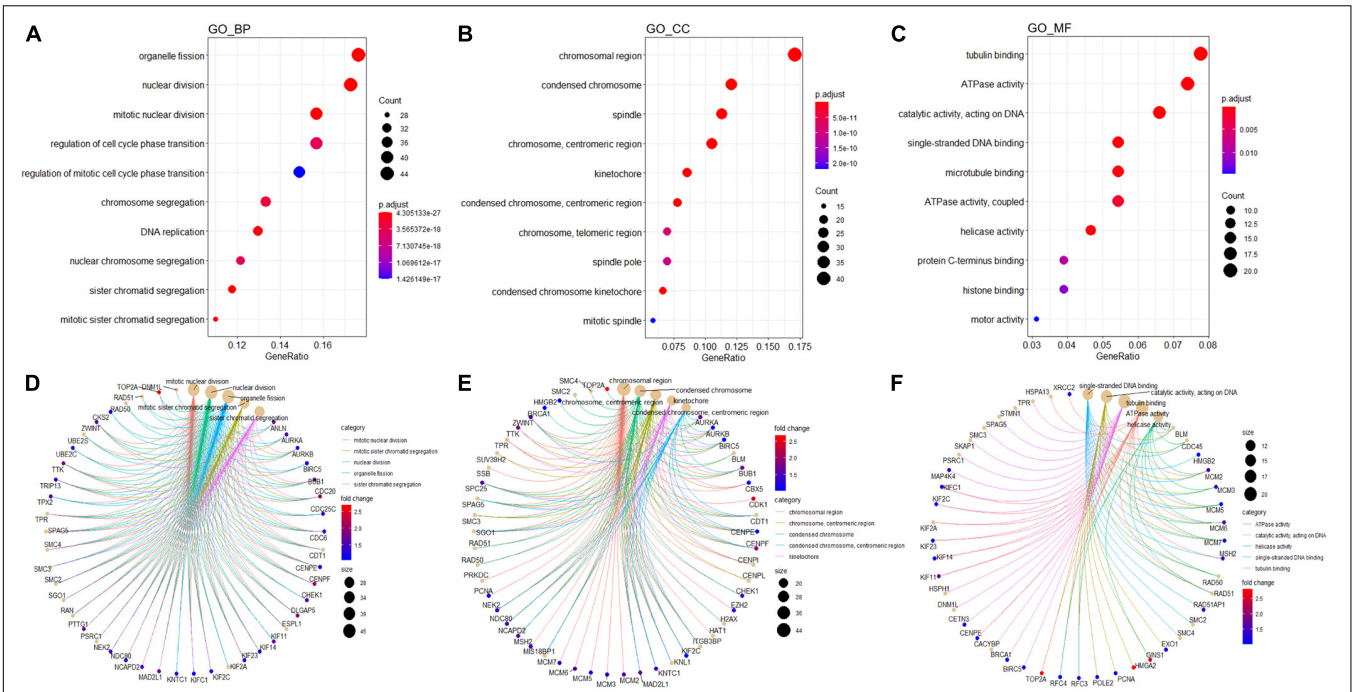

FIGURE 7 | GO analysis of genes from blue module. The significant GO BP **(A)**, CC **(B)**, and MF **(C)** terms after enrichment analysis of genes from the WGCNA blue module. Cnetplot indicating specific genes associated with enriched GO BP **(D)**, CC **(E)**, or MF **(F)** terms; the symbol adjacent to the node represents the specific gene. GO, Gene Ontology; BP, biological process; CC, cellular component; MF, molecular function.

TABLE 2 | Scores of five intersecting hub genes using different ranking methods in PPI. and WGCNA.

Entrez ID	Gene Symbol	PPI				WGCNA			
		MCC	EPC	MNC	Degree	GS	*p*.GS	MM	*p*.MM
6790	*AURKA*	9.22E + 13	96.26	83	84	0.549	1.64E-10	0.850	1.37E-33
9212	*AURKB*	9.22E + 13	93.997	80	80	0.415	3.54E-06	0.806	1.00E-27
983	*CDK1*	9.22E + 13	98.24	101	102	0.644	6.22E-15	0.950	1.52E-59
890	*CCNA2*	9.22E + 13	97.444	85	86	0.612	2.75E-13	0.934	4.36E-53
991	*CDC20*	9.22E + 13	97.548	82	82	0.559	6.80E-11	0.851	1.06E-33

PPI, protein–protein interaction; MCC, maximal clique centrality; EPC, edge percolated component; MNC, maximum neighborhood component; Degree, node connect degree; GS, gene significance with cancer; MM, module membership; p.GS, p value of gene significance with cancer; p.MM, p value of module membership.

(Mondal et al., 2007), gastric cancer (Kim et al., 2005), and lung adenocarcinoma (Liu et al., 2018). CDC20 knockdown has been shown to contribute to G2/M arrest, inhibiting tumor cell cycle progression (Kidokoro et al., 2008). Collectively, exploring therapeutic agents targeting the cell cycle via inhibition or modulation of CDK1, CCNA2, or CDC20 may be considered a promising therapeutic strategy for HB. In a recent study, Aghajanzadeh et al. (2020) identified 15 hub genes involved in HB based on bioinformatics analysis of GSE131329. CDK1 and CCNA2 were identified as hub genes in their study while CDC20 was not. This discrepancy could be due to the fact that different analytic methods were used and distinct datasets were assessed between their study and ours. Interestingly, using gene set enrichment and pathway analysis of the hub genes, the authors also identified cell cycle events as essential processes for HB development, which is in line with our findings.

The current study has some limitations. Experimental verification was only conducted *in vitro* at cellular level. In addition, the sample sizes for HB and normal liver tissue samples were asymmetrical, which may have potentially introduced bias in our analysis.

In conclusion, we conducted an integrative analysis of large-scale microarray gene expression profiling followed by experimental validation to investigate potential biomarkers and key genes involved in HB pathogenesis. By utilizing both PPI and WGCNA analyses, we identified CCNA2, CDK1, and CDC20 as hub genes in human HB. Subsequent *in vitro* experiments validated a potential oncogenic role for these three hub genes in two HB cell lines. Collectively, CCNA2, CDK1, and CDC20 may serve as promising biomarkers for HB and provide prospects for designing targeted therapies using synthetic inhibitors as anti-tumor agents.

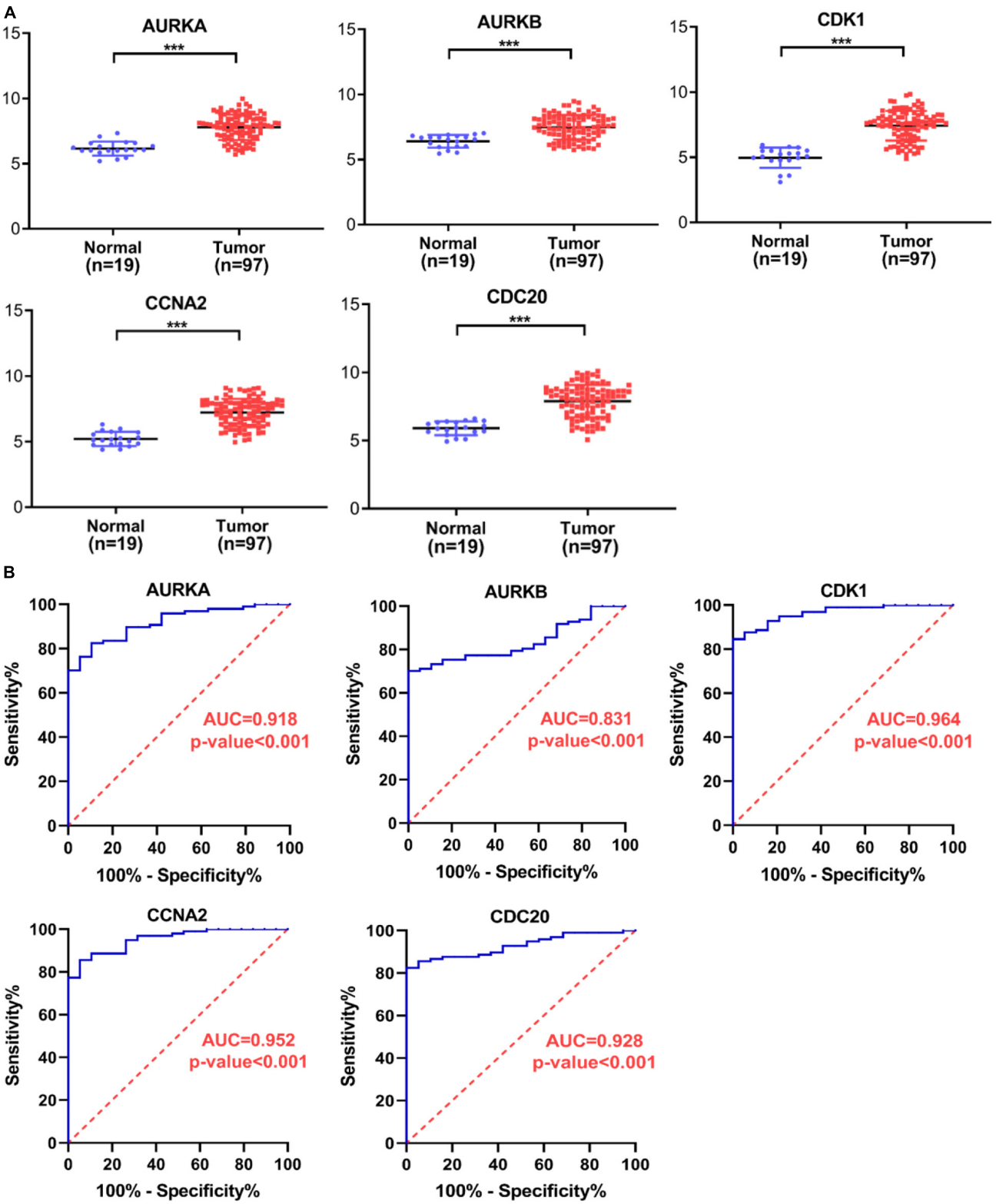

FIGURE 8 | Verification of hub gene expression levels and ROC curve analysis. **(A)** The expression levels of AURKA, AURKB, CDK1, CCNA2, and CDC20 mRNAs were markedly up-regulated in HB samples relative to normal liver samples. **(B)** ROC curve analysis of AURKA, AURKB, CDK1, CCNA2, and CDC20. ROC, receiver operating characteristic; AUC, area under the curve. ***$P < 0.001$.

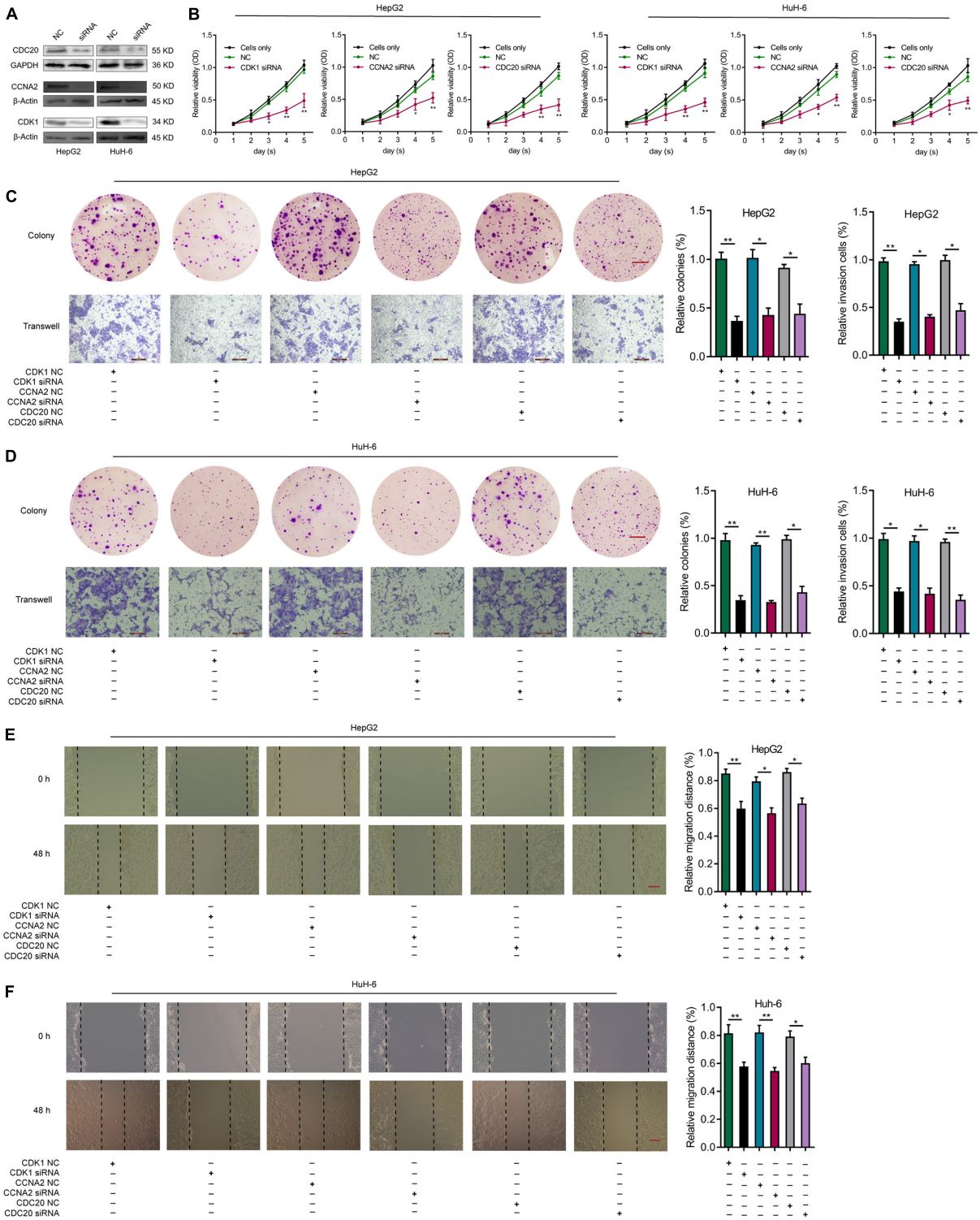

FIGURE 9 | Knockdown of CDK1, CCNA2, or CDC20 inhibits proliferative, migrative, and invasive capacities of HB cells *in vitro*. **(A)** WB analysis confirm the knockdown efficiency of CDK1, CCNA2, or CDC20 2 days after transfection with siRNAs for CDK1, CCNA2, or CDC20. **(B)** The CCK-8 assay illustrates the proliferative capacity of HB cells after siRNA transfection. After siRNA transfection of HepG2 **(C)** or HuH-6 **(D)** cells, the proliferative and invasive capacities of the respective cell lines were evaluated by colony formation assays (scale bars, 8 mm) and transwell invasion assays (scale bars, 200 μm), respectively. **(E,F)** Wound healing assay (scale bars, 500 μm) results that indicate the migrative capacities of HepG2 **(E)** or HuH-6 **(F)** cells after transfection with siRNA. $^{*}P < 0.05$, $^{**}P < 0.01$. ROC, receiver operating characteristic; HB, hepatoblastoma; WB, western blot; siRNAs, small interfering RNAs; CCK-8, Cell Counting Kit-8.

DATA AVAILABILITY STATEMENT

The original contributions presented in this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The medical research ethics committee of Basic Medicine School, Guilin Medical University identified and approved this study design.

AUTHOR CONTRIBUTIONS

YZ and SD created the study design. LT, TC, YZ, and JL performed the experiments and data analyses. LT, TC, PQ, and JY wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We appreciate the support from the Guangxi Universities Young and Middle-aged Teachers Basic Ability Improvement Project (2018KY0403), Guangxi Natural Science Foundation Youth Science Foundation Project (2018GXNSFBA138017), and Guangxi Science and Technology Base and Talent Project (Guike AD18281010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.631982/full#supplementary-material

Supplementary Figure 1 | Hierarchical clustering tree of samples and application of soft-threshold powers. **(A)** Hierarchical clustering of samples to detect outliers. **(B)** After network topology analysis for soft-threshold powers, the scale-free topology β = 10 was determined as soft threshold power. **(C)** Scale-free fitting exponential analysis of various soft threshold powers.

REFERENCES

Aboubakr, E. M., Taye, A., Aly, O. M., Gamal-Eldeen, A. M., and El-Moselhy, M. A. (2017). Enhanced anticancer effect of combretastatin A-4 phosphate when combined with vincristine in the treatment of hepatocellular carcinoma. *Biomed. Pharmacother.* 89, 36–46. doi: 10.1016/j.biopha.2017.02.019

Aghajanzadeh, T., Tebbi, K., and Talkhabi, M. (2020). Identification of potential key genes and miRNAs involved in hepatoblastoma pathogenesis and prognosis. *J. Cell Commun. Signal.* 14, 1–12. doi: 10.1007/s12079-020-00584-1

Allan, B., Parikh, P., Diaz, S., Perez, E., Neville, H., and Sola, J. (2013). Predictors of survival and incidence of hepatoblastoma in the paediatric population. *HPB (Oxford)* 15, 741–746. doi: 10.1111/hpb.12112

Bandettini, W. P., Kellman, P., Mancini, C., Booker, O. J., Vasu, S., Leung, S. W., et al. (2012). MultiContrast delayed enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. *J. Cardiovasc. Magn. Reson.* 14:83. doi: 10.1186/1532-429X-14-83

Bannon, J. H., and Mc Gee, M. M. (2009). Understanding the role of aneuploidy in tumorigenesis. *Biochem. Soc. Trans.* 37, 910–913. doi: 10.1042/BST0370910

Barrett, T., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bidwell, S., Peterson, C., Demanelis, K., Zarins, K., Meza, R., Sriplung, H., et al. (2019). Childhood cancer incidence and survival in Thailand: a comprehensive population-based registry analysis, 1990–2011. *Pediatr. Blood Cancer* 66:e27428. doi: 10.1002/pbc.27428

Carbajo-Pescador, S., Mauriz, J. L., García-Palomo, A., and González-Gallego, J. (2014). FoxO proteins: regulation and molecular targets in liver cancer. *Curr. Med. Chem.* 21, 1231–1246. doi: 10.2174/09298673216613228205703

Carvalho, B., and Irizarry, R. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26, 2363–2367. doi: 10.1093/bioinformatics/btq431

Chen, W., Lee, J., Cho, S. Y., and Fine, H. A. (2004). Proteasome-mediated destruction of the cyclin a/cyclin-dependent kinase 2 complex suppresses tumor cell growth in vitro and in vivo. *Cancer Res.* 64, 3949–3957. doi: 10.1158/0008-5472.Can-03-3906

Cheung, C. T., Bendris, N., Paul, C., Hamieh, A., Anouar, Y., Hahne, M., et al. (2015). Cyclin A2 modulates EMT via β-catenin and phospholipase C pathways. *Carcinogenesis* 36, 914–924. doi: 10.1093/carcin/bgv069

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8:S11. doi: 10.1186/1752-0509-8-s4-s11

den Elzen, N., and Pines, J. (2001). Cyclin A is destroyed in prometaphase and can delay chromosome alignment and anaphase. *J. Cell Biol.* 153, 121–136. doi: 10.1083/jcb.153.1.121

Dyson, N. (1998). The regulation of E2F by pRB-family proteins. *Genes Dev.* 12, 2245–2262. doi: 10.1101/gad.12.15.2245

Farhan, M., Wang, H., Gaur, U., Little, P. J., Xu, J., and Zheng, W. (2017). FOXO signaling pathways as therapeutic targets in cancer. *Int. J. Biol. Sci.* 13, 815–827. doi: 10.7150/ijbs.20052

Gautier, L., Cope, L., Bolstad, B., and Irizarry, R. (2004). affy–analysis of Affymetrix genechip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Girard, F., Strausfeld, U., Fernandez, A., and Lamb, N. J. (1991). Cyclin A is required for the onset of DNA replication in mammalian fibroblasts. *Cell* 67, 1169–1179. doi: 10.1016/0092-8674(91)90293-8

Goga, A., Yang, D., Tward, A. D., Morgan, D. O., and Bishop, J. M. (2007). Inhibition of CDK1 as a potential therapy for tumors over-expressing MYC. *Nat. Med.* 13, 820–827. doi: 10.1038/nm1606

Gopinathan, L., Tan, S. L., Padmakumar, V. C., Coppola, V., Tessarollo, L., and Kaldis, P. (2014). Loss of Cdk2 and cyclin A2 impairs cell proliferation and tumorigenesis. *Cancer Res.* 74, 3870–3879. doi: 10.1158/0008-5472.Can-13-3440

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hartmann, W., Küchler, J., Koch, A., Friedrichs, N., Waha, A., Endl, E., et al. (2009). Activation of phosphatidylinositol-3′-kinase/AKT signaling is essential in hepatoblastoma survival. *Clin. Cancer Res.* 15, 4538–4545. doi: 10.1158/1078-0432.Ccr-08-2878

He, J., Guo, X., Sun, L., Wang, N., and Bao, J. (2016). Regulatory network analysis of genes and microRNAs in human hepatoblastoma. *Oncol. Lett.* 12, 4099–4106. doi: 10.3892/ol.2016.5196

Hiyama, E. (2014). Pediatric hepatoblastoma: diagnosis and treatment. *Transl. Pediatr.* 3, 293–299. doi: 10.3978/j.issn.2224-4336.2014.09.01

Hung, Y. H., Huang, H. L., Chen, W. C., Yen, M. C., Cho, C. Y., Weng, T. Y., et al. (2017). Argininosuccinate lyase interacts with cyclin A2 in cytoplasm and modulates growth of liver tumor cells. *Oncol. Rep.* 37, 969–978. doi: 10.3892/or.2016.5334

Kanawa, M., Hiyama, E., Kawashima, K., Hiyama, K., Ikeda, K., Morihara, N., et al. (2019). Gene expression profiling in hepatoblastoma cases of the Japanese study group for pediatric liver tumors-2 (JPLT-2) trial. *Eur. J. Mol. Cancer* 1, 1–8. doi: 10.31487/J.EJMC.2018.01.003

Kidokoro, T., Tanikawa, C., Furukawa, Y., Katagiri, T., Nakamura, Y., and Matsuda, K. (2008). CDC20, a potential cancer therapeutic target, is

negatively regulated by p53. *Oncogene* 27, 1562–1571. doi: 10.1038/sj.onc.1210799

Kim, J. M., Sohn, H. Y., Yoon, S. Y., Oh, J. H., Yang, J. O., Kim, J. H., et al. (2005). Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. *Clin. Cancer Res.* 11, 473–482.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Leek, J., Johnson, W., Parker, H., Jaffe, A., and Storey, J. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Linabery, A. M., and Ross, J. A. (2008). Trends in childhood cancer incidence in the U.S. (1992–2004). *Cancer* 112, 416–432. doi: 10.1002/cncr.23169

Liu, W. T., Wang, Y., Zhang, J., Ye, F., Huang, X. H., Li, B., et al. (2018). A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma. *Cancer Lett.* 425, 43–53. doi: 10.1016/j.canlet.2018.03.043

Mondal, G., Sengupta, S., Panda, C. K., Gollin, S. M., Saunders, W. S., and Roychoudhury, S. (2007). Overexpression of Cdc20 leads to impairment of the spindle assembly checkpoint and aneuploidization in oral cancer. *Carcinogenesis* 28, 81–92. doi: 10.1093/carcin/bgl100

Murakami, Y., Tripathi, L., Prathipati, P., and Mizuguchi, K. (2017). Network analysis and in silico prediction of protein-protein interactions with applications in drug discovery. *Curr. Opin. Struct. Biol.* 44, 134–142. doi: 10.1016/j.sbi.2017.02.005

Norbury, C., and Nurse, P. (1992). Animal cell cycles and their control. *Annu. Rev. Biochem.* 61, 441–470. doi: 10.1146/annurev.bi.61.070192.002301

Perilongo, G., Shafford, E., Maibach, R., Aronson, D., Brugières, L., Brock, P., et al. (2004). Risk-adapted treatment for childhood hepatoblastoma. final report of the second study of the International Society of Paediatric Oncology–SIOPEL 2. *Eur. J. Cancer* 40, 411–421. doi: 10.1016/j.ejca.2003.06.003

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

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharp, D. J., Rogers, G. C., and Scholey, J. M. (2000). Microtubule motors in mitosis. *Nature* 407, 41–47. doi: 10.1038/35024000

Shi, S., Qin, X., Wang, S., Wang, W., Zhu, Y., Lin, Y., et al. (2020). Identification of potential novel differentially-expressed genes and their role in invasion and migration in renal cell carcinoma. *Aging* 12, 9205–9223. doi: 10.18632/aging.103192

Shin, E., Lee, K. B., Park, S. Y., Kim, S. H., Ryu, H. S., Park, Y. N., et al. (2011). Gene expression profiling of human hepatoblastoma using archived formalin-fixed and paraffin-embedded tissues. *Virchows Arch.* 458, 453–465. doi: 10.1007/s00428-011-1043-8

Spector, L., and Birch, J. (2012). The epidemiology of hepatoblastoma. *Pediatr. Blood Cancer* 59, 776–779. doi: 10.1002/pbc.24215

Sumazin, P., Chen, Y., Treviño, L., Sarabia, S., Hampton, O., Patel, K., et al. (2017). Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups. *Hepatology* 65, 104–121. doi: 10.1002/hep.28888

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tan, J., Xu, W., Lei, L., Liu, H., Wang, H., Cao, X., et al. (2020). Inhibition of aurora kinase a by alisertib reduces cell proliferation and induces apoptosis and autophagy in huh-6 human hepatoblastoma cells. *Oncol. Targets Ther.* 13, 3953–3963. doi: 10.2147/ott.S228656

Tulla, M., Berthold, F., Graf, N., Rutkowski, S., von Schweinitz, D., Spix, C., et al. (2015). Incidence, trends, and survival of children with embryonal tumors. *Pediatrics* 136, e623–e632. doi: 10.1542/peds.2015-0224

Ubersax, J. A., Woodbury, E. L., Quang, P. N., Paraz, M., Blethrow, J. D., Shah, K., et al. (2003). Targets of the cyclin-dependent kinase Cdk1. *Nature* 425, 859–864. doi: 10.1038/nature02062

Wang, M., Wang, J., Liu, J., Zhu, L., Ma, H., Zou, J., et al. (2020). Systematic prediction of key genes for ovarian cancer by co-expression network analysis. *J. Cell. Mol. Med.* 24, 6298–6307. doi: 10.1111/jcmm.15271

Yam, C. H., Fung, T. K., and Poon, R. Y. (2002). Cyclin A in cell cycle control and cancer. *Cell. Mol. Life Sci.* 59, 1317–1326. doi: 10.1007/s00018-002-8510-y

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

Yuan, L., Chen, L., Qian, K., Qian, G., Wu, C., Wang, X., et al. (2017). Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC). *Genom. Data* 14, 132–140. doi: 10.1016/j.gdata.2017.10.006

Zhang, D., Zhang, B., Zhou, L. X., Zhao, J., Yan, Y. Y., Li, Y. L., et al. (2016). Deacetylisovaltratum disrupts microtubule dynamics and causes G(2)/M-phase arrest in human gastric cancer cells in vitro. *Acta Pharmacol. Sin.* 37, 1597–1605. doi: 10.1038/aps.2016.91

Zhang, Y., Zhao, Y., Wu, J., Liangpunsakul, S., Niu, J., and Wang, L. (2018). MicroRNA-26-5p functions as a new inhibitor of hepatoblastoma by repressing lin-28 homolog B and aurora kinase a expression. *Hepatol. Commun.* 2, 861–871. doi: 10.1002/hep4.1185

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 March 2021
doi: 10.3389/fgene.2021.625741

Loss of Long Distance Co-Expression in Lung Cancer

Sergio Daniel Andonegui-Elguera[1], *José María Zamora-Fuentes*[1], *Jesús Espinal-Enríquez*[1,2]* and *Enrique Hernández-Lemus*[1,2]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

Lung cancer is one of the deadliest, most aggressive cancers. Abrupt changes in gene expression represent an important challenge to understand and fight the disease. Gene co-expression networks (GCNs) have been widely used to study the genomic regulatory landscape of human cancer. Here, based on 1,143 RNA-Seq experiments from the TCGA collaboration, we constructed GCN for the most common types of lung tumors: adenocarcinoma (TAD) and squamous cells (TSCs) as well as their respective control networks (NAD and NSC). We compared the number of intra-chromosome (*cis*-) and inter-chromosome (*trans*-) co-expression interactions in normal and cancer GCNs. We compared the number of shared interactions between TAD and TSC, as well as in NAD and NSC, to observe which phenotypes were more alike. By means of an over-representation analysis, we associated network topology features with biological functions. We found that TAD and TSC present mostly *cis*- small disconnected components, whereas in control GCNs, both types have a giant *trans*- component. In both cancer networks, we observed *cis*- components in which genes not only belong to the same chromosome but to the same cytoband or to neighboring cytobands. This supports the hypothesis that in lung cancer, gene co-expression is constrained to small neighboring regions. Despite this loss of distant co-expression observed in TAD and TSC, there are some remaining *trans*- clusters. These clusters seem to play relevant roles in the carcinogenic processes. For instance, some clusters in TAD and TSC are associated with the immune system, response to virus, or control of gene expression. Additionally, other non-enriched *trans*- clusters are composed of one gene and several associated pseudo-genes, as in the case of the FTH1 gene. The appearance of those common *trans*- clusters reflects that the gene co-expression program in lung cancer conserves some aspects for cell maintenance. Unexpectedly, 0.48% of the edges are shared between control networks; conversely, 35% is shared between lung cancer GCNs, a 73-fold larger intersection. This suggests that in lung cancer a process of de-differentiation may be occurring. To further investigate the implications of the loss of distant co-expression, it will become necessary to broaden the investigation with other omic-based approaches. However, the present approach provides a basis for future work toward an integrative perspective of abnormal transcriptional regulatory programs in lung cancer.

Keywords: lung adenocarcinoma, squamous lung cancer, gene co-expression networks, differentiation processes in cancer, loss of distant co-expression

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer, University of Oslo, Norway

Reviewed by:
Mukesh Bansal, Psychogenics, United States
Nurcan Tuncbag, Middle East Technical University, Turkey

***Correspondence:**
Jesús Espinal-Enríquez
jespinal@inmegen.gob.mx
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

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

Received: *03 November 2020*
Accepted: *29 January 2021*
Published: *10 March 2021*

Citation:
Andonegui-Elguera SD, Zamora-Fuentes JM, Espinal-Enriquez 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

INTRODUCTION

Lung cancer is one of the most deadly types of cancer nowadays. The survival range for lung cancer barely reaches 5.8%, quite below that of other malignant tumors (Torre et al., 2015). The World Health Organization places malignant tumors of the trachea, bronchi, and lung as the sixth leading cause of death globally (Marciniuk et al., 2017). Lung cancer occupies the first place in incidence and worldwide mortality among malignant tumors. Each year there are about 1.8 million new cases and 1.59 million deaths worldwide.

Currently, based on the type of tissue, lung cancer can be classified into two main categories: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). They represent around 80 and 20% of cases, respectively. NSCLC tumors are subclassified into squamous cells (TSC), adenocarcinoma (TAD), and large cell (LC) carcinoma. TSC occurs more frequently in the central area of the lungs, while TAD is found in peripheral areas, arising from bronchial glands and bronchial epithelium (Travis et al., 2015).

Treatment largely depends on histological diagnosis and tumor status. Detection is performed via chest X-ray and low-dose spiral tomography. Currently, only one-third of patients—diagnosed at an early stage—may be candidates for a surgical resection. However, recurrence after surgery reaches 30–60% even with adjuvant chemotherapy. For advanced states, the first line of treatment is chemotherapy with an average response between 30–40%.

Molecular Biology of Lung Cancer

Several mechanisms of genomic alterations have been found in lung cancer. For instance, DNA-repair pathways are triggered by exposure to tobacco-derived carcinogenic chemicals. Several single nucleotide polymorphisms (SNPs) have been identified in these pathways. The helicase ERCC2/XPD involved in DNA repair, the PHACTR2 protein that regulates the cytoskeleton, the DUSP1 protein that negatively regulates the MAP-kinase pathway are examples where SNPs have been identified (≈25% of cases) in lung adenocarcinoma (Spinola et al., 2007).

In terms of epigenetic marks, alterations have been reported via sputum analysis. In smokers, 14 genes with altered methylation patterns were identified (p16INK4a, DAPK, RASSF1A, PAX5, MGMT, GATA5, among others). These genes were associated with an increase of 50% in the risk of developing lung cancer. On the other hand, the p16 region has been found hypermethylated in 25–74% of lung cancer patients in different studies (Suzuki et al., 2014).

Alternative splicing events have been reported generating gene fusion in lung adenocarcinoma. Tyrosine kinase domain fusions have been identified by sequencing, including dimerization domains, such as EML4-ALK, KIF5B-RET, and CD74-ROS1, among other combinations. Additionally, some of these alterations have been observed to be involved in drug resistance. Patients with the EML4-ALK fusion treated with an ALK tyrosine kinase inhibitors have shown better results than traditional chemotherapy (Campbell et al., 2016). In nonsmoking women from Korea with adenocarcinoma mutations, gene fusions, among other alterations, were identified in c-Ret kinase as well as genes involved in mitotic progression and G2/M transition pathways (Campbell et al., 2016).

Two molecular pathways have been identified as relevant for lung cancer in recent years: the epidermal growth factor receptor (EGFR) and the anaplastic lymphoma kinase (ALK), respectively. These pathways can be affected by mutations in the kinase domain, amplification of the copy number or translocations, thus inducing new transcriptional control. Clinical trials have shown that patients whose malignant tumor is strongly related to EGFR or ALK can be treated with drugs targeting the kinase activities of these proteins, obtaining a 60% favorable response range (Suzuki et al., 2014).

Copy number alterations have also been identified in lung cancer. Chromosomal amplification of region 14q13.3 has been frequently found in tumor adenocarcinoma (TAD). One of the altered genes in copy number is NKX2-1, a transcription factor related to the differentiation and epithelial morphogenesis of the lung.

Several mutations have been reported in crucial genes associated with carcinogenic processes of the lung. KRAS, HER2, BRAF, EGFRvIII, and PIK3CA, among others, are frequently mutated in patients with NSCLC. Mutated KRAS is present in 15–25% of adenocarcinoma cases. There is no directed treatment targeting KRAS, but the subsequent effector route, RAS/ RAF/MEK, possesses inhibitors which may be effective in patients with diagnosed NSCLC and mutant KRAS (Shames and Wistuba, 2014). These are just some examples of the multiplicity of mutations and functional events related to abnormal regulation in lung cancer and its consequences. The purpose of this work is to further contribute to the understanding of these complex phenomena.

Large-Scale Studies on Abnormal Gene Regulation in Lung Cancer

Several efforts involving next-generation sequencing techniques have been developed by international groups. The objective is to provide a better understanding of the molecular changes that cells and tissues suffer during cancer progression. Endeavors such as The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) (Consortium et al., 2010) represent world-wide referents that have broadened our knowledge of cancer.

Collaborations like the ones mentioned above have helped to establish the relevance of cancer genomics and provided large amounts of data that have contributed to improve not only our basic knowledge of cancer biology but also oncological treatment and clinical practice. Data generated by such consortia are public and available to develop new knowledge based on such state-of-the-art experiments for thousands of samples.

A useful and powerful type of data to implement -omic analysis is the one generated by RNA-Seq technology. In the case of lung cancer, TCGA RNA-Seq-based gene expression databases include more than 1,100 samples for patients with adenocarcinoma (533), squamous cell carcinoma (502), as well as their adjacent-to-tumor healthy counterpart samples (101).

This kind of information allows researchers to explore in-depth the molecular mechanisms behind each cancerous genotype and at the same time to explore the functional implications of the concomitant phenotypes.

Gene expression data are one of the most used types of genomic information. However, gene expression analysis alone is not always sufficient to fully characterize and differentiate one type of cancer from another, even in the same tissue. Recently (Zamora-Fuentes et al., 2020, published in this research topic), we showed that, for clear cell renal carcinoma, gene expression signatures do not change during cancer progression. However, what remarkably differs between stages is the co-expression signature.

Gene co-expression networks are a helpful tool to analyze not only network parameters to distinguish global features, such as node degree or betweenness centrality between cases, but also functional implications based on the network structure for each phenotype (Amar et al., 2013; Alcalá-Corona et al., 2016, 2017, 2018; Drago-García et al., 2017; van Dam et al., 2018; Fionda, 2019; Tieri et al., 2019).

Despite several efforts to dissect the molecular mechanisms behind lung cancer origins and development, unsolved issues regarding the effect of gene co-expression and the relationship between co-expression patterns and phenotypic manifestations are still missing.

In this work, based on 1,143 gene-expression profiles of NSCLC patients, we constructed, inferred, and analyzed gene co-expression networks of lung cancer, as well as their healthy counterparts. To construct the networks, we separated cancer samples in adenocarcinoma tumors (TAD) and squamous carcinoma tumors (TSC).

We investigated how similar are both types of lung cancer at the expression and co-expression levels. We compared the resulting probabilistic co-expression networks in terms of shared interactions between lung cancer networks (TAD and TSC) and between the healthy ones (NAD and NSC). Finally, based on the gene co-expression signatures for both cancer networks, we performed over-representation analysis to observe those biological processes in which key genes participate.

MATERIALS AND METHODS

Data Acquisition

RNA-Seq files were obtained from the Genomic Data Commons database https://portal.gdc.cancer.gov/ for the two most common subtypes of lung cancer (TAD and TSC) as well as for adjacent-to-tumor normal lung tissue.

Files were downloaded using the following filters: Primary site = lung, sample type = primary tumor or solid normal tissue, experimental strategy = RNA-Seq, and workflow type = HTSeq-Counts. Data files consisted of 502 TAD samples, 49 adjacent-to-TAD normal samples (NAD); 533 TSC samples and 59 adjacent-to-TSC normal samples (NSC).

Data were annotated and harmonized for subsequent analysis using the latest genomic reference (GRCh38). Genomic information for gene stable ID, chromosome/scaffold name, gene start (bp), gene end (bp), gene% GC content, and gene type was mapped using BioMart database (version GRCh38.p12). This data pre-processing pipeline has been previously implemented to analyze RNA-Seq data from breast cancer (Drago-García et al., 2017; Espinal-Enriquez et al., 2017; Dorantes-Gilardi et al., 2020; García-Cortés et al., 2020; Serrano-Carbajal et al., 2020) and clear cell renal carcinoma (Zamora-Fuentes et al., 2020).

Data Pre-Processing

Quality control was performed using "Biotype detection" and "Sequencing depth" functions from `NOISeq` package (Tarazona et al., 2011). The most frequent sources of biases in RNAseq sequencing are associated with GC content, transcript length, and RNA composition (Tarazona et al., 2015). These biases were removed using full quantile normalization for GC content and length and TMM (Trimmed Mean of M) for RNA composition, all functions from NOISeq package. In addition, structural noise like batch effects were removed using `ARSyN` (Nueda et al., 2012) package. Finally, genes with *countspermillion* $<$ 10 were removed. Data pre-processing was carried out using R version 3.6.0.

For data normalization, we used the `DESeq2` R package (Love et al., 2014). After normalization of the four matrices, we preserved only those transcripts that were conserved in all four matrices. The number of resulting transcripts was 20,101. This number included protein coding genes, long noncoding RNA, microRNAs, pseudogenes, and other types of RNA species. The whole pre-processing and normalization code can be accessed and/or downloaded from https://github.com/CSB-IG/regulaciontrans-pipeline.

Differential Expression

`Limma` (Ritchie et al., 2015) is a Bioconductor component package based on a linear model to compare gene expression between two different gene sets. It can be used to analyze both types of data: microarrays or RNA-Seq. With this tool, we obtained the information about average expression, as well as the differential expression in terms of Log_2 fold change for TAD vs. NAD, and TSC vs. NSC samples. An absolute difference of fold change $\geq$ 1.5 and a Benjamini & Hochberg corrected *p-value* $<$ 0.01 were set as thresholds.

Gene Co-Expression Network (GCN) Inference

For inferring our four GCNs (NAD, TAD, NSC, and TSC), we used mutual information (MI) as the measure to determine gene co-expression. `ARACNe` (Margolin et al., 2006) is a standard method to calculate the MI between two data series. This algorithm was applied to the four gene expression profiles to establish correlations between pairs of genes. We used the serial `C++` version without Adaptive Partitioning Inference.

To improve the performance of this method, we developed a multicore version based on the aforementioned algorithm. This interface accelerates MI calculation depending on the number of available cores. For this work, we inferred a GCN of $\approx$ $200,000,000$ ($20,000^2/2$, corresponding to the total of genes in the matrix) of 100 sample expression matrix in 30 min using an 80-core server. This interface is available on github

(https://github.com/josemaz/aracne-multicore). We decided to analyze and conserve the top-10,000 interactions for the four GCNs in order to have the same size of the four graphs, as well as being able to compare them. Additionally, this network size has been previously observed to be significant to analyze them in terms of structural and functional characteristics (Alcalá-Corona et al., 2016, 2017; Velazquez-Caldelas et al., 2019; Zamora-Fuentes et al., 2020).

cis-/*trans-* Proportion Calculations

Previously, we observed in breast cancer GCNs (Espinal-Enriquez et al., 2017; de Anda-Jáuregui et al., 2019a,b,c; García-Cortés et al., 2020) that gene co-expression interactions occur in a preferential manner between genes from the same chromosome, and inter-chromosome interactions appear more frequently in noncancer breast tissue networks. We decided to observe whether or not that effect is also found in lung cancer networks. For that purpose, we separated gene co-expression interactions between intra-chromosome (*cis-*) and inter-chromosome (*trans-*).

For these analyses, we also used the top-10,000 interactions. However, in order to corroborate that any result generated by the analysis with 10,000 edges network was not related to the network size, we also performed calculations for a range of three orders of magnitude in terms of edges, i.e., we analyzed the *cis-*/*trans-* proportion in GCNs from 1,000 to 100,000 interactions. Finally, network visualizations and topological analyses were performed using `Cytoscape` v3.8.1.

We mentioned that the number of cancer samples is much larger than healthy samples. To assure that the obtained results for cancer GCNs were not due to the sample size, we developed a method to select 100 random cancer samples from the cancer expression matrix (table with samples and gene expression). For this work, we generated 10 randomized expression matrices with 100 samples for adenocarcinoma samples and other 10 matrices for squamous cancer data. The networks obtained using this method were pruned to 10,000 interactions, and compared with their healthy counterpart in terms of *cis-*/*trans-* proportion.

Functional Enrichment

Genes that presented a relevant network topology were in turn mapped into Gene Ontology categories to observe those processes that are allegedly enriched. For that purpose, we used `g:Profiler` web interface tool (Raudvere et al., 2019). We used the *g:SCS* option for multiple testing correction. The significance threshold was set to 10^{-5}. In **Figure 1**, the workflow presented in this paper is depicted.

RESULTS AND DISCUSSION

Gene Co-Expression Is Chromosome Dependent in Lung Cancer

Figure 2 shows the lung carcinoma (TSC and TAD) GCNs, compared with their healthy counterpart (NSC and NAD). The difference between both networks in terms of the component sizes is remarkable. The giant component of the healthy GCNs covers more than the half of the total size of the networks. Meanwhile, for the tumor-derived GCNs, there is no giant component; the larger one contains 123 genes and 336 edges (for TSC).

Aside from topological differences in network structure, in the tumor GCNs, components are formed mostly by genes from the same chromosome, which indicates that the majority of interactions are intra-chromosome or *cis-* interactions. Conversely, in the healthy networks genes co-express with other genes with no particular bias or trends in terms of the chromosomal location. The difference in *cis-* and *trans-* interactions between tumor and normal GCNs is observed in all chromosomes ($p-_{val} < 10^{-8}$ in both cases). In **Supplementary Material 1A**, we show all *cis-* interactions per chromosome for the four GCNs.

Furthermore, in the TAD and TSC GCNs, genes are correlated with other genes appearing in the same chromosome, but co-expressed genes are also physically close (in terms of chromosomal location) among them. This phenomenon can be observed in **Figure 3**. There, we depicted all interactions appearing in chromosome 19 for NAD and TAD GCNs. Genes are placed according to its gene start position. Turquoise interactions represent long-range *cis-* interactions, meanwhile purple edges show close co-expression relationships (both genes belong to the same cytoband).

Potential De-Differentiation of the Gene Co-Expression Program in Lung Cancer

Since both healthy and both cancer networks at first sight seem to be topologically similar, we decided to compare them in terms of shared genes an interactions: NAD vs. NSC and TAD vs. TSC. This was made out with the aims of observing the percentage of similarity between phenotypes.

In **Figure 4**, we observe the intersection of interactions for healthy and cancer GCNs. For the healthy networks, the number of shared genes is high, but they only share 0.48% of their edges. On the other hand, the TAD and TSC networks share 35% of their interactions. The intersection between cancer GCNs is then 73-fold larger.

The organizational principles that determine the structure in cancer GCNs are more similar than control networks. The observed co-expression program may indicate that the cancer cell suffers a process of de-differentiation, since cancer networks become more alike than the different lung cell types of origin.

The idea that TAD and TSC networks are suffering a de-differentiation process, and can be appreciated from the increase of intersected edges between cancer networks with respect to the normal counterpart, is based on the following premises:

- The gene co-expression program, in particular, the set of higher co-expression interactions, represents a reliable and significant example of the cellular state of a given phenotype.
- The gene co-expression program can be represented by a network, where nodes correspond to genes, and the edges

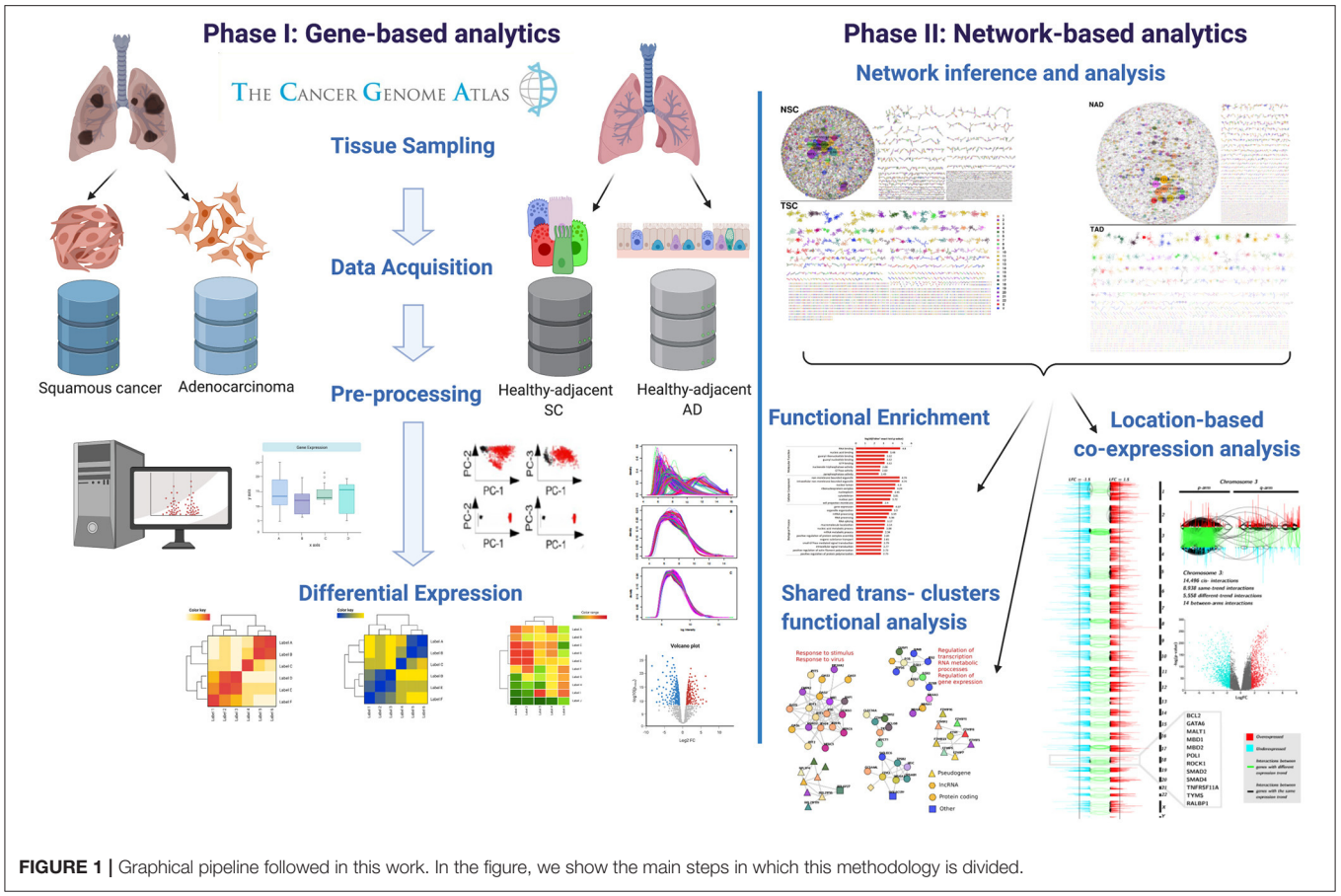

FIGURE 1 | Graphical pipeline followed in this work. In the figure, we show the main steps in which this methodology is divided.

connecting nodes represent a kind of interaction between any couple of genes.

- A gene co-expression interaction can be defined by a certain type of correlation observed between any two genes. In this case, the measurement used to define an interaction is MI.
- The similarity between two networks can be used, to a certain extent, as a proxy to assess the similarity between two gene co-expression phenotypes.
- The TSC cancer network came from the same cells that give form to the normal tissue-derived NSC network. Analogously, the TAD network comes from normal tissue-derived NAD network.
- NSC and NAD networks came from different cell types.

The similarity between tumor GCNs may be explained (at least partially) by a process of cellular de-differentiation. The NSC and NAD networks share little connectivity, but half of the genes are shared. This implies that although they express at least half of the same genes, they do not co-express in the same way; this is probably because they are well-differentiated cells with specialized tasks.

On the other hand, TSC and TAD networks share 76% of genes and 35% of the co-expression pattern. Tumor cells have a lower degree of differentiation and a higher proliferative power. Two tumors of different origin may be more similar to each other than two samples of specialized normal tissue.

trans- Clusters May Play a Crucial Role in the Carcinogenic Process

Components With *cis*- Co-Expression Belong to Neighboring Karyobands

Most of the components that form tumor GCNs contain genes from the same chromosome. The genes from each component, in addition to being from the same chromosome, are located in neighboring regions of the chromosomes. Co-expressed genes are usually within the same karyotype band in all chromosomes (p-val$< 10^{-8}$ for TAD, and p-val$< 10^{-6}$ for TSC, **Supplementary Material 1B**). In other words, the co-expression of neighboring genes is stronger than between distant genes, even within the same chromosome. These *cis*- components are not, however, significantly associated with biological processes in enrichment analysis.

A plausible explanation regarding the mechanisms for which we observed such a decrease in long-range gene couples, and a concomitant elevation of close gene co-expression interactions, could be chromosomal aberrations or the aforementioned copy number alterations (CNAs). This latter could be partially answered by an analysis of copy number alteration data and contrasted that with our network data. Preliminary results in breast cancer have shown that copy number alteration events are not highly correlated with clusters of physically close genes with high co-expression interactions. The complete analysis of

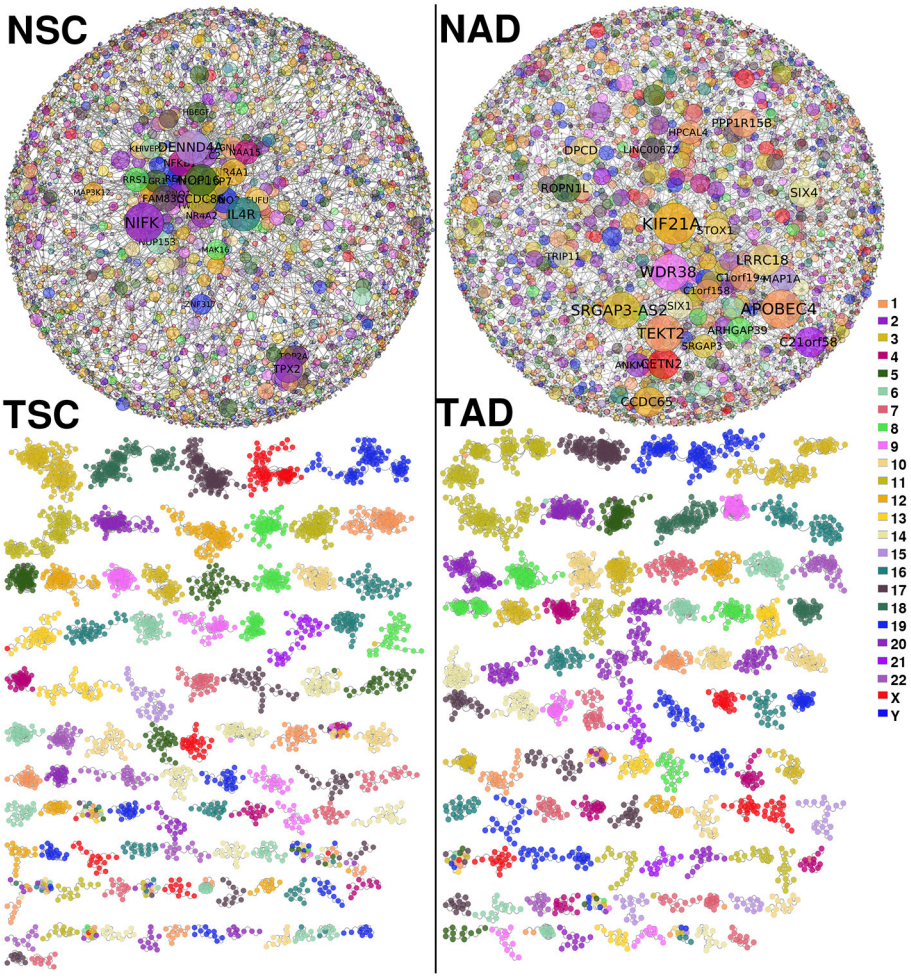

FIGURE 2 | Non-small cell lung cancer (NSCLC) and normal tissue-derived gene co-expression networks. **(Top-left)** Largest component of normal tissue-derived network for NSC. **(Top-right)** Correspond to the giant component of NAD gene co-expression network (GCN). **(Bottom-left)** Squamous carcinoma-derived gene co-expression network. **(Bottom-right)** Tumor adenocarcinoma network. In both tumor GCNs, components with more than 10 genes are depicted. Genes are colored according to the chromosome location. In healthy GCNs, gene size is proportional to the gene connectivity.

CNAs implication in the lung cancer co-expression program is under development.

Shared *trans*- Components Are Significantly Associated With Biological Functions

Despite the fact that the large majority of gene co-expression interactions in cancer GCNs are *cis*-, a small subset of *trans*-edges appears in both cancer networks. In fact, some *trans*-clusters are also shared between cancer phenotypes. **Figure 5** shows the shared *trans*- clusters between the two lung cancer GCNs. Additionally, two of those components are significantly associated with biological processes.

One of them, composed of OAS1, or IFIT genes, resulted enriched in 26 terms (**Supplementary Material 2**). They are related to processes such as response to virus and response to stimulus. The second enriched *trans*- component (with EGR, FGR, FOSB, and JUNB genes) is associated with the regulation of gene expression, regulation of transcription, and metabolic processes. Thirty-four GO categories resulted enriched for this geneset (**Supplementary Material 3**).

We previously reported (Alcalá-Corona et al., 2018) a gene co-expression network for HER2+ subtype breast cancer, which contained a component, very similar to the one with IFIT and OAS genes. This component was also associated with viral response. In Alcalá-Corona et al. (2018), additional to the association with virus-related processes, these genes were mostly overexpressed. Here, these genes in TSC network are mostly underexpressed. Moreover, in TAD network, this gene subset is not biased to a particular differential expression trend. The differential expression of all genes in this analysis can be found in **Supplementary Material 4**.

It is worth to notice that the HER2+ breast cancer network considered there was constructed based on microarray data, and this one is an RNA-Seq-based analysis. Despite technologies are different and also the primary organ in which this gene subset was found, the co-expression associations are the same in a very small

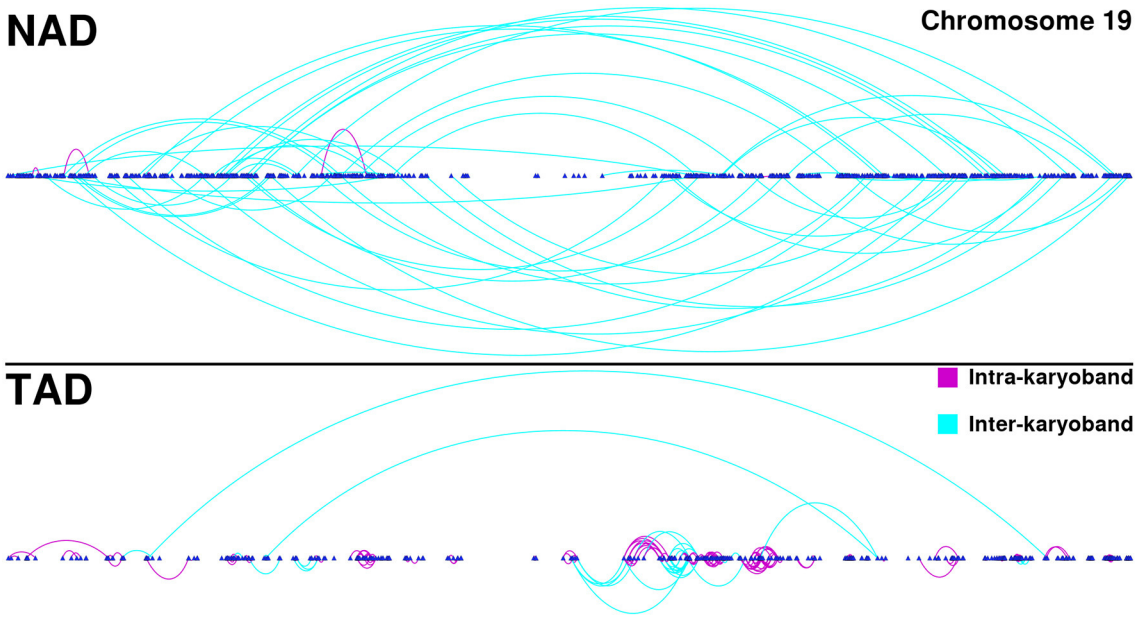

FIGURE 3 | Chromosome 19 gene co-expression networks (GCNs) for NAD and TAD. In this figure, *cis*- interactions for NAD (top) and TAD (bottom) are depicted. Genes are placed according to its starting base pair. Turquoise links join genes from different cytobands, meanwhile purple interactions take account for intra-cytoband relationships.

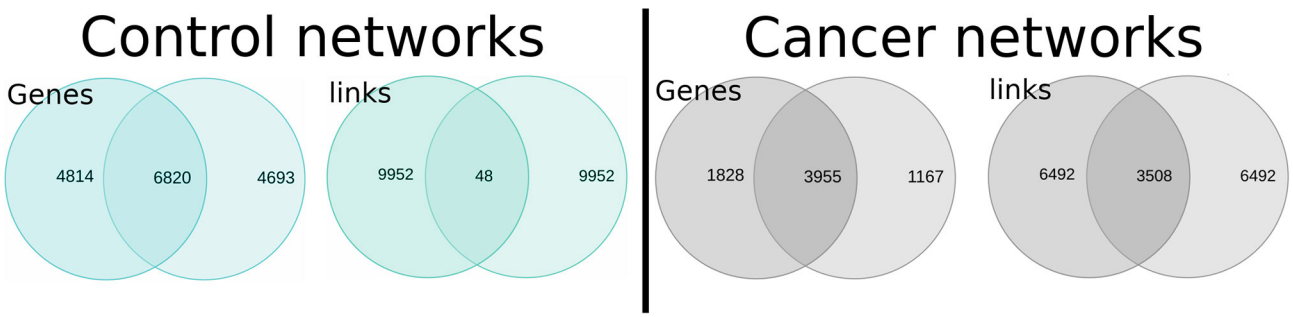

FIGURE 4 | Intersections between normal **(Left)** and lung cancer **(Right)** gene co-expression networks (GCNs). The upper section of the figure shows (in the form of Venn diagrams) the number of shared genes and interactions between NAD and NSC networks (blue diagrams); the right part represents the intersection between TAD and TSC GCNs.

group of genes. It is more remarkable that both cases present opposite differential expression trend. This could be another instance in which co-expression features are more robust that gene expression itself.

Cancer Networks Edges Are Biased to Genes With the Same Differential Expression Trend

Within the Top-10,000 GCN, we observed 5,783 genes for TAD and 5,122 for TSC. Hence, the GCNs do not contain the sufficient number of genes to analyze their whole genome differential expression patterns. To overcome this, we decided to analyze larger GCNs. For that purpose, we conserved gene interactions with a $p - value < 10^{-8}$ for both cancerous phenotypes.

In the case of TAD, the resulted GCN contains 170,190 interactions and 14,073 genes, which means that almost all genes in the genome participate in the structure of that network. By setting the Log_2FC threshold in ± 1.5, the number of significant DEGs was 1,056 for overexpressed and 1,304 showed underexpression.

In **Figures 6A,B**, only *cis*- interactions are depicted. Green links join co-expressed genes with an opposite differential expression trend, i.e., one gene presents positive Log_2FC and the other one has a negative Log_2FC. Black interactions join *cis*- genes that have the same differential expression trend: both genes are over- or underexpressed. There are

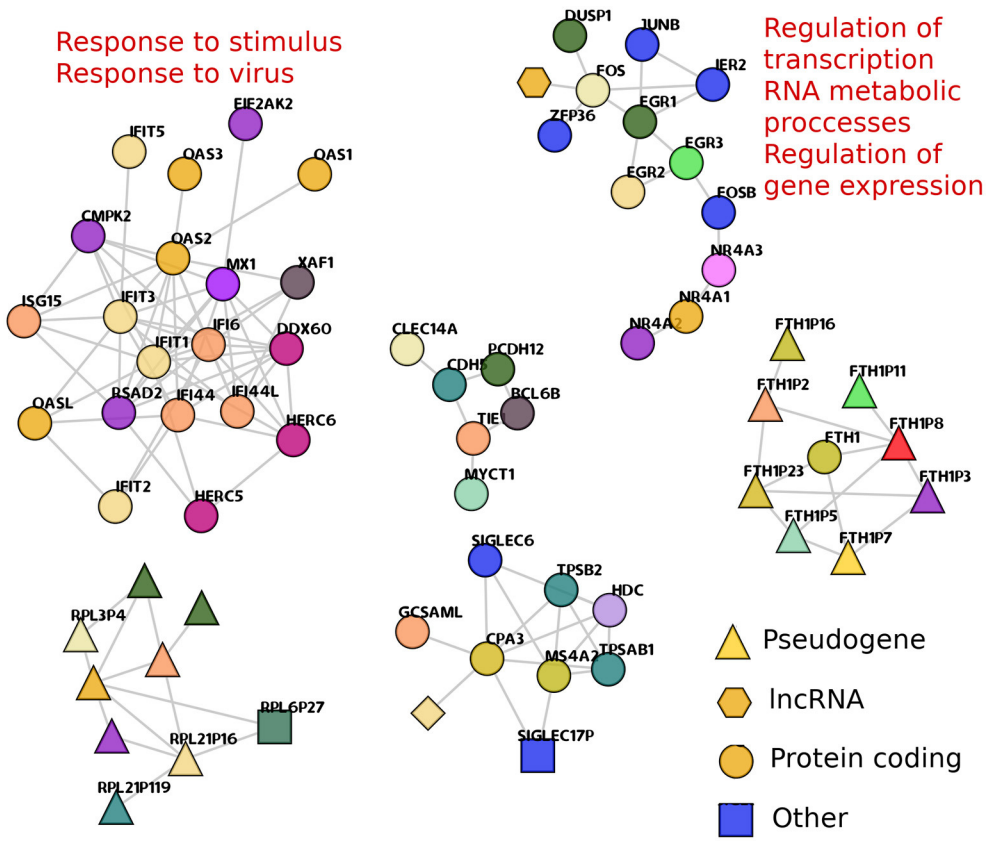

FIGURE 5 | Shared *trans*- components between TAD and TSC networks. *trans*- clusters that are found in both TAD and TSC are depicted. The shape of nodes represents the transcriptional species of the nodes. Enriched processes that are associated with a particular component are presented. The FTH1 component (bottom right) shows one only gene, that codifies for ferritin1, and all other nodes in the component are ferritin pseudogenes, showing that in this case, co-expression favor sequence similarities over gene physical closeness.

more underexpressed genes than overexpressed ones (1,304 vs. 1,056, **Figure 6C**). Additionally, underexpressed genes are more broadly differentially expressed than the overexpressed ones (**Supplementary Material 4**).

Regarding interactions, there are more black edges (joining same-expression-trend genes) than green ones in all chromosomes (119,574 vs. 40,445, p-val$< 10^{-8}$, **Supplementary Material 5**). Moreover, the large majority of same-trend interactions occurs between genes with positiveLog_2FC (110,714) than those with negative differential expression (8,860) in all chromosomes ($p - value < 10^{-10}$, **Supplementary Material 5**).

In the case of interactions between negative Log_2FC genes, chromosomes 3, 8, and 18, have the majority of intra-trend links. The p-arm of chromosome 3 has dense interactions hotspots in both intra- and inter-trend genes. There is a common deletion in Chr3p in lung cancer (Lerman et al., 2000; Kou et al., 2020). It is known that several tumor suppressor genes are located at 3p (Varella-Garcia, 2010). Partial deletion of 3p occurs in almost all lung carcinomas (Kou et al., 2020). This deletion includes tumor suppressor genes, such as RASSF1 (3p21.3) or TUSC2 (FUS1, 3p21.3) (Kok et al., 1997). These genes are found in the TAD network and both are downregulated.

Another zone with a high number of intra-trend edges is the q-arm of chromosome 10. A deletion in Chr10q24-26 in small cell lung carcinoma has been reported (Petersen et al., 1997; Kim et al., 1998). PTEN gene is located on 10q23.3 and it is also present in the network, downregulated but non-significantly underexpressed. Alterations in PTEN have been reported in around 20% of SCLC (Yokomizo et al., 1998). Despite this analysis was performed on nonsmall cell lung carcinomas, the intra-trend interactions hotspot observed in Chr10 could be associated with chromosomal-level events in NSCLC.

cis- interactions between genes that belong to different arms are also scarce. In the top right part of **Figure 6**, the zoom in of Chr3 shows that from almost 15,000 *cis*- Chr3 gene co-expression relationships, only 14 appear between genes from different arms, and none of them are between different expression trend genes.

All of these results appear to indicate that in NSCLC, the co-expression landscape is dominated by physically close genes. These genes, in turn, share other characteristics, such as the differential expression pattern. At this point, we do not know the functional causes behind this phenomenon.

In the case of TSC, the difference between same-arm co-expression interactions as compared to those in different-arm ones is even larger. The total number of significant interactions for TSC is 232,355. Intra-arm *cis*- interactions are 222,839,

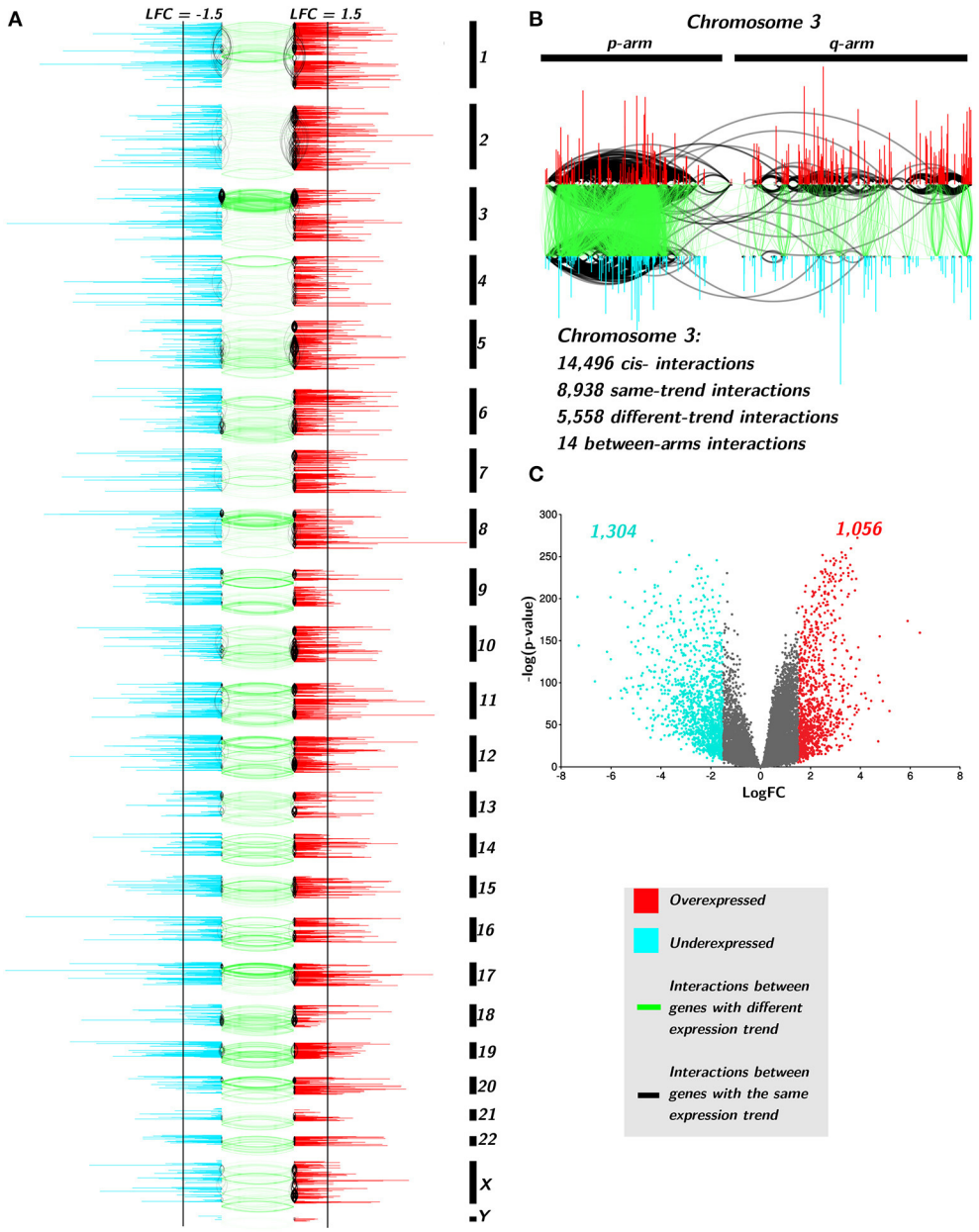

FIGURE 6 | Differential expression trend influences the interactions of TAD network. **(A)** Genes are placed according to its starting base per chromosome. Color of genes takes account for the differential expression trend: red for positive differential expression and turquoise for negative ones. Black vertical lines indicate the threshold point for differential expression (±1.5). Black edges join genes with the same differential expression trend, meanwhile green links represent interactions between different trend genes. **(B)** Zoom-in to Chr3. **(C)** volcano plot for TAD genes with the aforementioned threshold.

i.e., the 95.9% of all interactions. The *trans-* interactions cover 9,081, the 3.9% of all interactions. The inter-arm *cis-* interactions are only 435. The fact that for TSC network, we observed 20 times fewer interactions between inter-arm *cis-* relationships than *trans-* interactions, which was unexpected. The latter may suggest that some *trans-* interactions are crucial to maintain certain processes in the tumor cell. In **Supplementary Material 6**, we provide the Cytoscape session .cys file containing all networks used in this work.

Loss of *trans-* Co-Expression in Cancer Does Not Depend on the Network Size or the Number of Samples

As mentioned in section 2, we analyzed the GCNs with the top 10,000 interactions. To assess the validity of the results shown here, we decided to carry out calculations for a broader range of interactions, from the top 1,000 edges to the top 100,000, i.e., three orders of magnitude, to evaluate whether or not, the differences in the *cis-*/*trans-* proportion were maintained.

Supplementary Material 7 shows the proportion of *cis*-interactions of the total of edges at different cutoff values. As it is noted, the proportion of this imbalance in lung cancer networks is essentially preserved independent of the cutoff value. This confirms our assertion that the fundamental phenomenon we are observing, regarding structural features of GCNs, is indeed maintained over a fairly wide range of interaction cutoffs.

We commented above that the number of cancer samples is much larger than healthy ones ($\sim$ 1,000 vs. 100). To assure that the obtained results for cancer networks were not due to the sample size, we generated 10 randomized expression matrices with 100 samples for adenocarcinoma tumors and other 10 matrices for squamous cancer data. The GCNs obtained with this method were pruned to 10,000 interactions and calculated its *cis/trans* proportion.

Supplementary Material 8 contains a Cytoscape session file, including the 20 different realizations of randomized networks, 10 for TAD and 10 for TSC. There it can be observed that with 100 random samples, the effect of loss of *trans*- co-expression prevails in all instances.

CONCLUSIONS

As a summary of findings, in this work we have shown that:

- gene co-expression networks in lung cancer suffer a dramatic loss of distant interactions;
- adenocarcinoma and squamous cell lung cancer GCNs are much more alike (in terms of gene interactions) than the networks formed by adjacent-to-tumor normal-derived tissue;
- the co-expression interactions in lung cancer are biased to appear in genes that are in the same chromosome;
- in lung cancer, interactions occur preferably between genes from the same cytoband;
- top gene interactions in lung cancer occur often between genes with the same differential expression trends, in special between upregulated genes;
- shared *trans*- (inter-chromosome) connected components are strongly associated with important biological functions such as immune response and regulation of gene transcription;
- these features has been observed for the first time in lung tissue-derived GCNs.

We have observed an important intersection between genes and links in lung cancer networks, which is opposed to the observed in healthy lung tissue-derived networks. This finding leads us to suggest that a de-differentiation mechanism appears during lung carcinogenesis.

The networks used in this work were inferred from lung cancer samples with no other filter than the type of lung cancer (adenocarcinoma and squamous cells). Further investigation in this line of research must be focused on constructing and infer networks based on progression stages of these types of cancer to observe whether or not later stages are more similar than the early ones.

We strongly believe that the current knowledge regarding gene co-expression and the concomitant functional regulation of the transcriptional program in cancer phenotypes will be improved and better understood by aggregating other omic layers to these systems. Furthermore, the effect of loosing the long-range co-expression observed in more than one cancer tissue (breast, kidney, and now lung) may be an instance of a more complicated phenomenon that could be behind of a novel—not yet described—hallmark of cancer.

In any case, the present results contribute to advancing our knowledge of the deep intricacies behind transcriptional regulation in cancer. This, in turn, will be helpful not only to establish better the basic foundations of cancer biology but also to devise ways in which this knowledge may be translated into diagnostics, prognostics, and therapies for lung cancer patients.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SA-E performed calculations, analyzed the data, and wrote the paper. JZ-F performed calculations and wrote the code. JE-E co-conceived the project, analyzed the data, made the figures, and wrote the paper. EH-L co-conceived the project, coordinated the theoretical sections, analyzed the data, and wrote the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], 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. EH-L is a recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences. JE-E is a recipient of a 2017 Miguel Alemán Fellowship on the Medical Sciences.

ACKNOWLEDGMENTS

Authors want to thank to Diana García-Cortés for her valuable support during the implementation of the computational workflow. The authors also want to thank Gabriela Graham for her support with language editing and proof-reading of this manuscript. **Figure 1** was generated with Biorender (biorender.com).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.625741/full#supplementary-material

Supplementary Material 1A | Number of *cis*- interactions for the four networks.

Supplementary Material 1B | Number of interactions within genes from the same karyotype band (intra-karyoband) and different karyoband (inter-karyoband) per chromosome.

Supplementary Material 2 | Over-representation analysis of shared *trans*-cluster composed of OASL1 and IFIT, among other genes. The output of this file correspond to the standard output of g:ProfileR web interface.

Supplementary Material 3 | Over-representation analysis of shared *trans*-cluster composed of FOSB and JUN, among other genes.

Supplementary Material 4 | Differential expression values for TAD and TSC samples.

Supplementary Material 5 | Table with the number of same trend positive interactions, same trend negative interactions, and different trend interactions in all chromosomes.

Supplementary Material 6 | Cytoscape .cys file containing all networks inferred for this study.

Supplementary Material 7 | Loss of *trans*- co-expression is not dependent of the network size. In this plot, we provide the proof that network size does not influence the effect of loss of inter-chromosome interactions proportion in lung carcinoma. Different network cutoffs were calculated for this purpose. Note that 1,000 to 100,000 top interactions given on X-axis. On Y-axis, the *cis*- proportion is represented, i.e., the number of *cis*- interactions over the total interactions in said cutoff. Green and yellow dots represent tumor networks, whereas blue and orange ones take account for normal networks.

Supplementary Material 8 | Cys file with cancer networks obtained by random selection of 100 samples.

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. Physiol.* 8:915. doi: 10.3389/fphys.2017.00915

Alcalá-Corona, S. A., Espinal-Enríquez, J., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2018). The hierarchical modular structure of her2+ breast cancer network. *Front. Physiol.* 9:1423. doi: 10.3389/fphys.2018.01423

Alcalá-Corona, S. A., Velázquez-Caldelas, T. E., Espinal-Enríquez, J., and Hernández-Lemus, E. (2016). Community structure reveals biologically functional modules in MEF2C transcriptional regulatory network. *Front. Physiol.* 7:184. doi: 10.3389/fphys.2016.00184

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.* 9:e1002955. doi: 10.1371/journal.pcbi.1002955

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., et al. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* 48:607. doi: 10.1038/ng.3564

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

de Anda-Jáuregui, G., Alcalá-Corona, S. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019a). Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Appl. Netw. Sci.* 4:22. doi: 10.1007/s41109-019-0129-0

de Anda-Jáuregui, G., Espinal-Enriquez, J., and Hernández-Lemus, E. (2019b). Spatial organization of the gene regulatory program: an information theoretical approach to breast cancer transcriptomics. *Entropy* 21:195. doi: 10.3390/e21020195

de Anda-Jáuregui, G., Fresno, C., Garcíia-Cortés, D., Enríquez, J. E., and Hernández-Lemus, E. (2019c). Intrachromosomal regulation decay in breast cancer. *Appl. Math. Nonlin. Sci.* 4, 223–230. doi: 10.2478/AMNS.2019.1.00020

Dorantes-Gilardi, R., Garcíia-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, 1–23. doi: 10.1007/s41109-020-00291-1

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of EMT and met micro-RNA regulation in breast cancer. *Sci. Rep.* 7, 1–17. doi: 10.1038/s41598-017-13903-1

Espinal-Enriquez, 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, 1–19. doi: 10.1038/s41598-017-01314-1

Fionda, V. (2019). "Networks in biology," in *Encyclopedia of Bioinformatics and Computational Biology*, eds S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Academic Press), 915–921. doi: 10.1016/B978-0-12-809633-8.20420-2

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernandez-Lemus, E., and Espinal-Enriquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol.* 10:1232. doi: 10.3389/fonc.2020.01232

Kim, S. K., Ro, J. Y., Kemp, B. L., Lee, J. S., Kwon, T. J., Hong, W. K., et al. (1998). Identification of two distinct tumor-suppressor loci on the long arm of chromosome 10 in small cell lung cancer. *Oncogene* 17, 1749–1753. doi: 10.1038/sj.onc.1202073

Kok, K., Naylor, S. L., and Buys, C. H. (1997). "Deletions of the short arm of chromosome 3 in solid tumors and the search for suppressor genes," in *Advances in Cancer Research, Vol. 71* (Elsevier), 27–92. doi: 10.1016/S0065-230X(08)60096-2

Kou, F., Wu, L., Ren, X., and Yang, L. (2020). Chromosome abnormalities: new insights into their clinical significance in cancer. *Mol. Ther. Oncolyt.* doi: 10.1016/j.omto.2020.05.010

Lerman, M. I., and Minna, J. D. (2000). The 630-kb lung cancer homozygous deletion region on human chromosome 3p21. 3: identification and evaluation of the resident candidate tumor suppressor genes. *Cancer Res.* 60, 6116–6133.

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. doi: 10.1186/s13059-014-0550-8

Marciniuk, D., Schraufnagel, D., and Society, E. R. (2017). *The Global Impact of Respiratory Disease*. European Respiratory Society.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7:S7. doi: 10.1186/1471-2105-7-S1-S7

Nueda, M. j., Ferrer, A., and Conesa, A. (2012). ARSYN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics* 13, 553–566. doi: 10.1093/biostatistics/kxr042

Petersen, I., Langreck, H., Wolf, G. E., Schwendel, A., Psille, R., Vogt, P., et al. (1997). Small-cell lung cancer is characterized by a high incidence of deletions on chromosomes 3p, 4q, 5q, 10q, 13q and 17p. *Br. J. Cancer* 75, 79–86. doi: 10.1038/bjc.1997.13

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucl. Acids Res.* 47, W191–W198. doi: 10.1093/nar/gkz369

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

Serrano-Carbajal, E. A., Espinal-Enríquez, J., and Hernàndez-Lemus, E. (2020). Targeting metabolic deregulation landscapes in breast cancer subtypes. *Front. Oncol.* 10:97. doi: 10.3389/fonc.2020.00097

Shames, D. S., and Wistuba, I. I. (2014). The evolving genomic classification of lung cancer. *J. Pathol.* 232, 121–133. doi: 10.1002/path.4275

Spinola, M., Leoni, V. P., Galvan, A., Korsching, E., Conti, B., Pastorino, U., et al. (2007). Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. *Cancer Lett.* 251, 311–316. doi: 10.1016/j.canlet.2006.11.029

Suzuki, A., Makinoshima, H., Wakaguri, H., Esumi, H., Sugano, S., Kohno, T., et al. (2014). Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucl. Acids Res.* 42, 13557–13572. doi: 10.1093/nar/gku885

Tarazona, S., Furió-Tarí, P., Turrá, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with noiseq R/Bioc package. *Nucl. Acids Res.* 43:e140. doi: 10.1093/nar/gkv711

Tarazona, S., García, F., Ferrer, A., Dopazo, J., and Conesa, A. (2011). Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet J.* 17, 18–19. doi: 10.14806/ej.17.B.265

Tieri, P., Farina, L., Petti, M., Astolfi, L., Paci, P., and Castiglione, F. (2019). "Network inference and reconstruction in bioinformatics," in *Reference Module in Life Sciences* doi: 10.1016/B978-0-12-809633-8.20290-2

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *Cancer J. Clin.* 65, 87–108. doi: 10.3322/caac.21262

Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H., Beasley, M. B., et al. (2015). The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J. Thorac. Oncol.* 10, 1243–1260. doi: 10.1097/JTO.0000000000000630

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

Varella-Garcia, M. (2010). Chromosomal and genomic changes in lung cancer. *Cell Adhes. Migrat.* 4, 100–106. doi: 10.4161/cam.4.1.10884

Velazquez-Caldelas, T. E., Alcalá-Corona, S. A., Espinal-Enríquez, J., and Hernandez-Lemus, E. (2019). Unveiling the link between inflammation and adaptive immunity in breast cancer. *Front. Immunol.* 10:56. doi: 10.3389/fimmu.2019.00056

Yokomizo, A., Tindall, D. J., Drabkin, H., Gemmill, R., Franklin, W., Yang, P., et al. (1998). PTEN/MMAC1 mutations identified in small cell, but not in non-small cell lung cancers. *Oncogene* 17, 475–479. doi: 10.1038/sj.onc.1201956

Zamora-Fuentes, J. M., Hernandez-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.* 11:1232. doi: 10.3389/fgene.2020.578679

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 March 2021
doi: 10.3389/fgene.2021.617512

An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation

Soledad Ochoa[1], *Guillermo de Anda-Jáuregui*[1,2,3]* *and Enrique Hernández-Lemus*[1,2]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Conacyt Research Chairs, National Council on Science and Technology, Mexico City, Mexico

Breast cancer is a complex, highly heterogeneous disease at multiple levels ranging from its genetic origins and molecular processes to clinical manifestations. This heterogeneity has given rise to the so-called intrinsic or molecular breast cancer subtypes. Aside from classification, these subtypes have set a basis for differential prognosis and treatment. Multiple regulatory mechanisms—involving a variety of biomolecular entities—suffer from alterations leading to the diseased phenotypes. Information theoretical approaches have been found to be useful in the description of these complex regulatory programs. In this work, we identified the interactions occurring between three main mechanisms of regulation of the gene expression program: transcription factor regulation, regulation via noncoding RNA, and epigenetic regulation through DNA methylation. Using data from The Cancer Genome Atlas, we inferred probabilistic multilayer networks, identifying key regulatory circuits able to (partially) explain the alterations that lead from a healthy phenotype to different manifestations of breast cancer, as captured by its molecular subtype classification. We also found some general trends in the topology of the multi-omic regulatory networks: Tumor subtype networks present longer shortest paths than their normal tissue counterpart; epigenomic regulation has frequently focused on genes enriched for certain biological processes; CpG methylation and miRNA interactions are often part of a regulatory core of conserved interactions. The use of probabilistic measures to infer information regarding theoretical-derived multilayer networks based on multi-omic high-throughput data is hence presented as a useful methodological approach to capture some of the molecular heterogeneity behind regulatory phenomena in breast cancer, and potentially other diseases.

Keywords: breast cancer, probabilistic multilayer networks, information theory, co-expression networks, multiomics analysis

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer, University of Oslo, Norway

Reviewed by:
Giuseppe Jurman, Bruno Kessler Foundation, Italy
Tatiana Belova, University of Oslo, Norway

***Correspondence:**
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx
Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx

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

Received: *14 October 2020*
Accepted: *05 February 2021*
Published: *18 March 2021*

Citation:
Ochoa S, de Anda-Jáuregui G and Hernández-Lemus E (2021) An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation. Front. Genet. 12:617512. doi: 10.3389/fgene.2021.617512

1. INTRODUCTION

Cancer is a collection of complex diseases characterized by uncontrolled proliferation (GM., 2000). The complexity of cancer comes, among other sources, from the interaction of different molecular layers and the environment and results in both intra- and inter-tumor heterogeneity (Tian et al., 2011; Burrell et al., 2013; Turashvili and Brogi, 2017). In the case of breast cancer, this heterogeneity has been intended to be captured by tumor sub-classification. Breast cancer has been

thus classified into subtypes with specific molecular signatures and treatment options (Prat et al., 2015), though each altered molecular layer groups differently (Cancer Genome Atlas Network, 2012). Some of these layers, such as gene expression and DNA methylation, have been intensively studied, while others like chromatin accessibility are still gaining attention (Liu, 2020). However, all these layers are interrelated (Wang et al., 2014) and the study of their collective effect calls for multi-omic approaches.

Multi-omic approaches have become possible only recently due to their more stringent methodological requirements. A (relatively large) minimal number of samples are required to find significant patterns, and the needed sample size increases with the noise added per each additional omic. Measurements must refer to the same set of samples, with sustained quality, no matter the differences in data type and range (Kristensen et al., 2014; Bersanelli et al., 2016; Tarazona et al., 2020).

The ability to model heterogeneous and high-dimensional data has made networks a promising tool for multi-omics integration (Vaske et al., 2010; Kim et al., 2012; Wang et al., 2014). For instance, mutual information (MI) networks combining miRNA and gene expressions have been built to gain insight on the regulatory mechanisms behind breast cancer (Drago-García et al., 2017). Such networks pinpointed miR-200 and miR-199 as regulators of the acquisition of epithelial and mesenchymal traits. Another example is the coupling of promoter methylation, transcription factors (TFs), and gene expression in several cancers proposed by Liu et al. Based on those networks, they fitted per target regression models that suggest key cancer processes are jointly regulated by TFs and CpG sites, not by either one alone. Those processes turned out to be different than the processes dominated by copy number variants (Liu et al., 2019).

Gene co-expression networks have been extensively studied in the context of breast cancer subtypes, both from our group (de Anda-Jáuregui et al., 2016; de Anda-Jáuregui et al., 2019; Espinal-Enriquez et al., 2017; Dorantes-Gilardi et al., 2020; García-Cortés et al., 2020; Ochoa et al., 2020) and others (Tang et al., 2018; Bhuva et al., 2019). Here, we are presenting the results on the incorporation of CpG methylation in addition to the study of coding transcripts (for both TFs and other genes) and miRNA expression analyzed in each breast cancer subtype. The goal is to identify CpG sites, TF transcripts (referred to as TF-genes from here on) and miRNAs associated with the biological processes differentially activated in breast cancer, since these may perform potential roles as regulators of the phenotype. Integrated analyses may thus provide us with additional hints toward the possible discovery of synergistic or cooperative effects of these different regulators.

2. MATERIALS AND METHODS

2.1. Data Acquisition

Concurrent-sample measurements of DNA methylation, transcript abundance, and miRNA expression were downloaded from the GDC (https://portal.gdc.cancer.gov/repository) in May 2019. Samples quantified with the Illumina Human Methylation 27 BeadChip, which covers a smaller portion of the genome, were discarded. Instead, we used data obtained with the Infinium HumanMethylation450 BeadChip, which covers 99% of RefSeq genes, at both transcription repressive sites around promoters and transcription favorable sites on the body of genes (Dedeurwaerder et al., 2011). Since these measurements pertain to three distinct techniques: methylation beadchip, RNAseq, and miRNAseq; we treat them as separate omics, here on identified as CpG sites, transcripts, and miRNAs. By including the whole set of features, we wanted to recover the highest possible number of interactions. Subtype classification was also downloaded from the GDC metadata using the `TCGABiolinks` R package (Colaprico et al., 2016).

Each omic was pre-processed independently according to Aryee et al. (2014), Tarazona et al. (2015), and Tam et al. (2015) by using `biomaRt` v95. Preprocessing included filtering of transcripts and miRNAs with low counts, TMM normalization and batch effect correction with ARSyN. Low count thresholds are less than 10 counts per million for transcripts and, less than 5 counts for 25% or more of the samples for every subtype, in the case of miRNAs. Transcripts were also normalized for length and GC content via full method. Annotation was downloaded to tag transcripts coding for TFs (TF-genes).

For methylation data, we discarded sites with over 75% missing values, nonmapped or located within sexual chromosomes or SNPs. Remaining missing values were imputed via nearest neighbors. Resulting beta value matrices were transformed into *M*-value matrices. This way, values of 384,575 methylation probes, 16,475 coding transcripts, and 433 miRNA precursors were obtained for 45 unique samples belonging to the Her2+ subtype, 395 of LumA, 128 of LumB, and 125 of Basal subtypes, plus 75 samples of nontumor (normal adjacent) tissue. All samples correspond to women, ranging in age at diagnosis between 26 and 91 years, and further details can be found in the **Supplementary File 1**.

2.2. Inference of MI Networks

Normalized data matrices for methylation data, coding transcripts, and miRNA expression were merged by sample and used as input to the MI-based `ARACNE` network deconvolution algorithm (Margolin et al., 2006).

`ARACNE` calculates mutual information between every pair of features and returns values above a threshold, set either as an MI value or as a permutation *p*-value. There is no restriction on the features that get paired by MI calculation, and it was not required for CpG sites to be on the same chromosome than targets, nor that target promoters carry some TF motif. The only restriction made was for CpG-CpG interactions, which were not calculated due to the space needed to save all possible combinatoria. In a nutshell, pairwise mutual information calculations were performed for the expression patterns for all genes and miRNAs, as well as the beta values for genomewide CpG methylation. Co-expression networks on the different layers were built from the most significant interactions as follows:

Since MI distribution has been shown to change depending on the type of molecules (Drago-García et al., 2017), a unique threshold cannot be set. A unique MI threshold has the risk of discarding significant interactions between molecules whose values simply fall in a lower range or accepting nonsignificant interaction between molecules exhibiting values on a higher than the threshold range. A threshold based on *p*-values induces a

similar problem because MI and *p*-values are roughly inversely proportional. For example, it is possible to see that setting the threshold value to 0.1 in **Figure 2C** would discard most miRNA to transcript interactions while retaining all the interactions among transcripts, and that such pruning of edges would affect differently the distinct networks, producing disparate results due to methodology. Mutual information distributions and their respective threshold values have a direct impact on the topology of the underlying networks and in particular in the degree distributions. So, by choosing MI cutoffs one is indeed imposing an associated network topology.

To overcome this issue, top 10,000 interaction for each type of molecules paired were selected, that is, the 10,000 interactions with the highest MI values linking CpG sites and transcripts (both genes and TF-genes), CpG sites and miRNAs, transcripts (both genes and TF-genes) and miRNAs, and interactions within these two last groups. This way, the topology resulting from such a set of interactions is comparable among cancer subtypes and normal tissue. Thus, we take the focus from the varying MI distributions to a defined topology size. This strategy has been previously validated and used by our group for the reconstruction of biologically relevant networks from high-throughput data (de Anda-Jáuregui et al., 2016).

Fixed bandwidth ARACNE calculations ran with kernel width parameter (h) of 0.165024 for Basal data, 0.211612 for Her2+, 0.12527 for LumA, 0.16567 for LumB, and 0.18679 for normal tissue. To check the significance of the interactions in these networks, maximal MI for each pair of molecules was registered for different *p*-value thresholds. Thresholds with MI values larger than those observed in a network contain the network's interactions. The *p*-value upper limits for the final networks are reported in **Supplementary Table 1**. Finally, MI distributions were compared via Kolmogorov-Smirnov tests with False Discovery Rate (FDR) correction.

Kernel width variation between subtypes can be attributed to the size of the datasets. We estimated z-scores with subsets of the data to evaluate how size differences are affecting the networks. To this end, 100 subsamples of size 45 were taken from luminal and Basal subtypes, and from the normal tissue data. The subsample size was set to 45 for direct comparison with Her2-associated networks. MI was calculated using these subsets and resulting distributions served for z-score calculation. Results can be observed in **Supplementary Table 2**.

By keeping the same number of links in each layer, we are able to directly compare network parameters between layers. However, it should be noted that since the number of possible links increases (quadratically) with the number of nodes, there may be differences in the statistical significance. However, all our networks have an equivalent *p*-value of less than 1E-6 (corresponding to the CpG layer in Her2+ samples, i.e., the layer with more features analyzed for the subtype with the lowest number of samples).

2.3. Functional Enrichment

Independently of network construction, differential expression vs. normal tissue was calculated for every subtype using `limma`'s treat (McCarthy and Smyth, 2009) function with null fold change equal to 1.5. Afterwards, the complete rank of differential expression *t*-values was used as input for a GSEA on each subtype, as implemented in the R package `fgsea` (Sergushichev, 2016), vs. the biological process gene ontology.

Processes with Benjamini and Hochberg adjusted *p*-value lesser than 0.01 were subject to over-representation analysis on the corresponding subtype network. Processes with Benjamini and Hochberg adjusted *p*-value over 0.05 were regarded as nonrepresented in the network. The rest was examined for CpG sites, miRNAs, and TF-genes associated via their MI value with the functionally annotated transcripts, since these serve as potential regulators of the function. For the normal tissue, all the processes significant for a subtype were submitted to the over-representation analysis. There are processes present in a subtype network, but absent from the normal tissue network. This results in a total of 176 processes over-represented in at least one subtype network, from which only 128 have a match in the normal tissue network. In this step, a mean of 59.05% nodes was removed from the MI networks, a breakdown of which can be found in **Supplementary Table 3**.

Resulting networks were visualized using `Cytoscape` (Shannon et al., 2003) with a prefuse force directed layout. Nodes were added to account for the enriched functions in order to find out which biological processes were potentially regulated. Hereafter, these networks are denominated as *final networks* or *functionally enriched networks* to distinguish them from the purely probabilistically inferred networks. These focus on the processes whose expression is the most associated with the subtype, and that rely on interactions with the highest MI; these functions are potentially relevant for the subtypes and so it may be useful to elucidate the associated regulatory patterns.

2.4. Validation of MI Interactions

To check for additional support for the interactions in the final networks, regulator-target databases were reviewed per omic. CpG annotation was taken from Illumina's manifest file, and the genes affected by each site are considered as *validated*. CpG sites on the same chromosome than the target gene are considered as plausible regulators and regarded when adding predictions. These are distinguished from one another as *mapped* and *same chromosome* sites in **Supplementary Table 1**.

Transcription factor targets were downloaded via `tftargets` https://github.com/slowkow/tftargets, a package that queries TRED, ITFP, ENCODE, and TRRUST databases, and the lists compiled by (Neph et al., 2012; Marbach et al., 2016). Only TRRUST TF-targets are considered as validated, since those were manually curated from PubMed articles. The associations between transcripts and miRNAs were sought on DIANA-microT-CDS, ElMMo, MicroCosm, miRanda, miRDB, PicTar, PITA, TargetScan, miRecords, miRTarBase, and TarBase via multiMiR (Ru et al., 2014).

Targets for both TF and miRNA were searched in the tables obtained from each package. The only tuning needed for TF's search was to track ENTREZ gene IDs, HGNC symbols, and Ensembl IDs; this was done according to biomaRt data. Since GDC measurements are identified by precursor miRNA IDs,

TABLE 1 | Networks description.

Edges	Basal	Her2+	LumA	LumB	Normal
CpG–mRNA	2,456	3,847	1,932	4,334	4,732
	(554)	(88)	(536)	(708)	(28)
TF-genes–mRNA	2,735	2,498	1,686	2,746	2,544
	(5)	(2)	(5)	(1)	(14)
miRNA–mRNA	3,483	3,889	2,065	4,074	4,953
	(167)	(226)	(111)	(201)	(284)
mRNA–mRNA	4,189	4,523	2,276	4,709	5,088
Nodes					
Biological processes	109	119	34	123	128
CpG sites	2,254	3,769	1,553	3,638	3,863
Transcripts	4,567	6,356	2,834	5,235	4,733
TF-genes	658	748	375	618	684
miRNAs	433	432	408	433	14

Validated interactions appear between parentheses. Edges correspond to significant statistical dependencies inferred via MI calculations.

while databases use mature miRNA tags, this search requires translation from one to the other using mirBase records.

2.5. Characterization of the Potential Regulators

Looking for differences between subtypes, total regulators of each type were added for every process. Retrieved counts were compared between each subtype and the normal tissue via Fisher tests with FDR correction. Enrichment is only considered if the process has associated regulators of any type, in both the normal tissue and the subtype under evaluation. Statistical tests were one-tailed. Null hypothesis is set to be opposite to expected trends, that is, "greater" for the CpG nodes and "less" for both TF-genes and miRNAs.

To weight the abundance of each regulatory layer, counts per regulator type were divided by the total number of regulators associated with the process, obtaining the percentages displayed in **Supplementary File 2**.

Node topological parameters were calculated over the MI networks, that is, ignoring the *biological processes* nodes, which have to be excluded given the different nature of their associated edges: probabilistically inferred or database curated. Distributions were compared via Wilcoxon rank sum test with continuity correction and p-values were FDR corrected.

2.6. Potential Regulators Comparison

Both intra and inter-subtype comparisons were made. To this end, Jaccard index was calculated for each pair of processes from the same subtype for the intra-subtype comparison and for the same process in different subtypes for the inter-subtype comparison. Inter-subtype contrasts count edges instead of nodes, because in this case, the interest is on conserved regulatory interactions. Obtained distributions were evaluated via Kolmogorov–Smirnov tests with FDR correction.

The number of potential regulators either shared or exclusive between processes of the same subtype was evaluated via Fisher tests with the corresponding alternative hypothesis set "greater" for the CpG sites, and "less" for TF-genes and miRNAs, as previously stated.

All the code used for the described analysis is available at https://github.com/CSB-IG/MI-MultiOmics.git.

3. RESULTS

MI networks were constructed for each breast cancer subtype and for normal tissue combining three different omics: CpG methylation, transcript gene expression, and miRNA expression. The second omic includes two layers of information, regulated genes, and TF-genes. No restriction was made on the features that can get paired by MI calculation, CpG sites can get linked to targets on a different chromosome, and TFs may associate with targets without the akin binding motifs. Let us recall that mutual information does not assume any a priori mechanism and relies instead on statistical dependencies. **Table 1** presents all the different networks of MI-inferred potential gene regulators (CpG-mRNA, TF-gene-mRNA, miRNA-mRNA, mRNA-mRNA) plus the biological processes associated with them.

MI networks went through two pruning steps, first by edge significance (see section 2.2) and then by functional annotation of the nodes (see section 2.3). The first one retains only the most significant interactions, i.e., those with the largest MI. For the second pruning, biological processes with significant GSEA enrichment scores were mapped to the networks, keeping only the nodes involved in an enriched process and their first neighbors. For the normal tissue, all the processes significant for a subtype were subjected to over-representation analysis. This way, only nodes linked to transcripts involved in a process altered in the subtype are kept. Then, final networks carry only CpG–transcript, miRNA–transcript, and transcript–transcript interactions with the highest MI. The hypothesis is that nodes with gene expression regulatory roles may regulate the associated biological process. This would be partially explained, if regulators co-vary (even in a nonlinear fashion) with their targets, thus becoming detectable as MI statistical dependencies. It is relevant

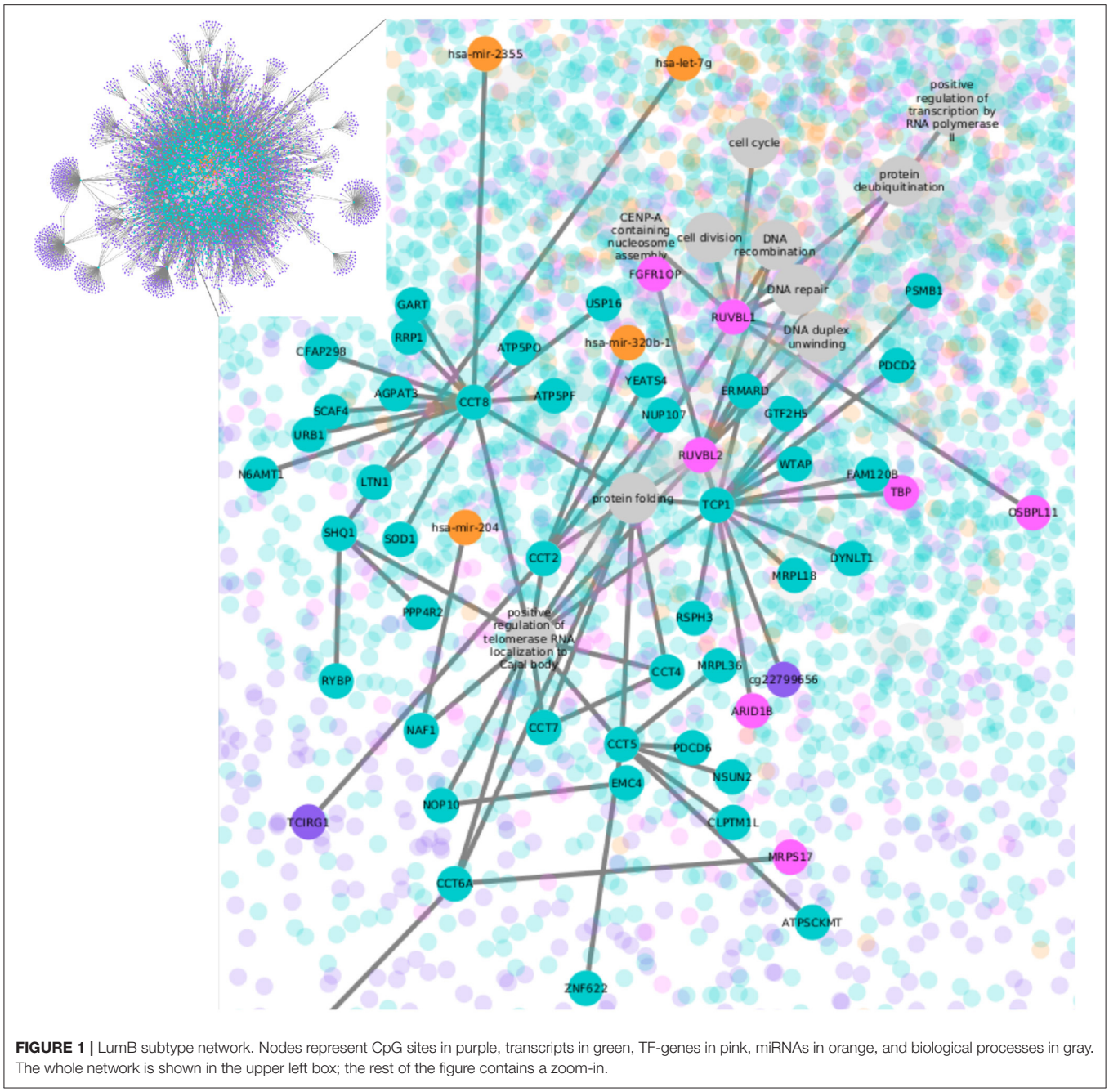

FIGURE 1 | LumB subtype network. Nodes represent CpG sites in purple, transcripts in green, TF-genes in pink, miRNAs in orange, and biological processes in gray. The whole network is shown in the upper left box; the rest of the figure contains a zoom-in.

to recall, however, that regulatory mechanisms are proxied here by the information given by the omics under study. Other regulatory mechanisms—including those of (explicit) chromatin remodeling, as well as post-transcriptional and post-translational modifications among others—may not be fully accounted by the statistical dependencies structures just outlined.

To assess the contribution of linear correlation measures, we are including further calculations in **Supplementary Figures 1–5** to show how many of the MI edges would be lost if the criterion was instead an FDR-corrected Pearson correlation with an associated $p < 0.05$.

To identify unequivocally the functions linked to each transcript, nodes representing the biological processes were added, resulting in multipartite graphs as the one shown in **Figure 1**. The multipartite nature of the network comes from the three different molecules (CpG sites, transcripts, and miRNAs) associated with the biological process nodes. There are also two kinds of edges: (1) MI edges, which indicate molecule covariation, and (2) functional annotation edges, which make explicit the link of a transcript and a process. All the five networks, four for the breast cancer subtypes and one for the normal tissue, consist of one giant single connected component. As expected,

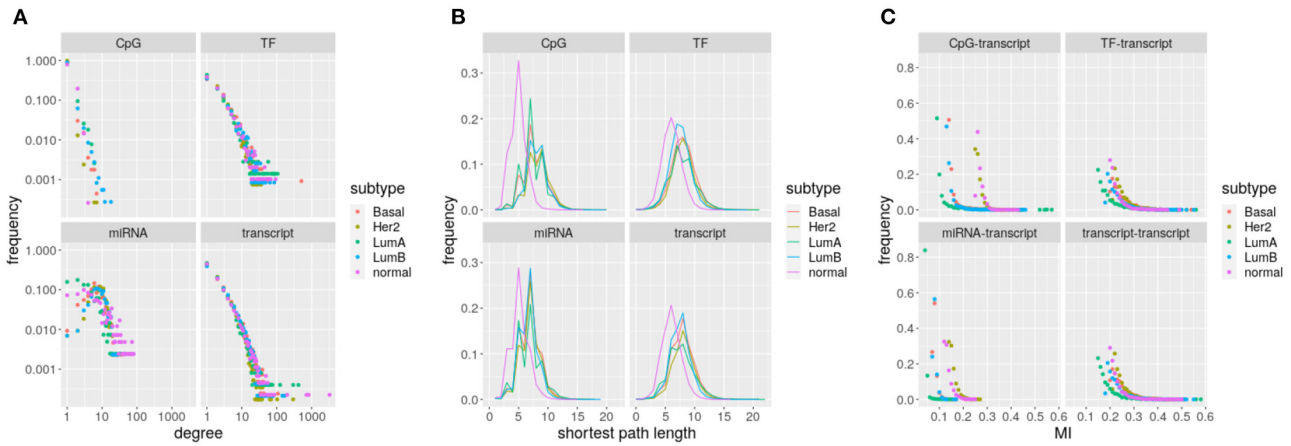

FIGURE 2 | Networks parameters: **(A)** degree, **(B)** shortest path, and **(C)** edges with mutual information. The subtype depicted by each network is indicated by the color code.

CpG methylation, which has the largest number of features, is the most represented omic in the networks.

By contrasting the molecules paired with databases on regulator-target, we can see how many of the found interactions were already known. Interactions absent from the databases can be new, previously unknown relationships, or simply indirect associations caused by the statistical co-variation of the molecules. Between 1.67 and 11.47% of the interactions linking a transcript with a potential regulator, that is a CpG, a TF-gene, or a miRNA, have been validated. The number of validated edges per subtype is shown in **Table 1**. If predictions are included (see section 2.4), 8.26–23.52% of the interactions have additional support. The effect on the networks of considering only some of the potential regulatory CpGs can be seen in **Supplementary Figure 6**. A large number of TF target predictions are based on ChIP-seq experiments, not necessarily performed on breast tissue, which may lower such matches.

Having described the general features of the five networks (one for each tumor subtype plus the one for normal tissue), we proceeded to search for differences between the behavior of the different omics among subtypes. Focus was made on differences on the potential regulators, since this could translate to regulatory features behind the subtypes.

3.1. Network Parameters Vary Between Omics

As stated earlier, there are two types of edges in the networks, edges that account for co-expression (i.e., significant statistical dependency) with a given value of MI, and edges that record functional annotation as presented in curated databases. Given the difference of meaning, interactions need to be analyzed separately.

Focusing only on MI edges, the number of components grows from 1 to hundreds. Average degree is around 3 for all the networks, but distributions vary between omics (Wilcoxon rank sum test q-value $\leq$ 1.666712e-22, **Figure 2A**). Though TF-genes and gene transcripts are measured by the same omic, distributions are significantly different (Wilcoxon rank sum test q-value $\leq$ 0.0237) for the five networks. The case of miRNAs stands out because distributions are not scale-free like. CpG sites show the lowest degrees, with an average of 89.42% nodes connected only with another node. Thus, most CpG sites do not contribute to network communication as they do not interlink paths.

The constrained (bounded) degree distribution of CpGs translates into a large portion of unreachable target nodes, an average of 32.23% of targets cannot be reached from some CpGs. Consistently, miRNAs have an average of 19.71% of unreachable targets, which is the lowest frequency. Despite range similarity, distributions change significantly across omics and between tumor subtypes and normal tissue (Wilcoxon rank sum test q-value $\simeq$ 0). Again, distributions for TF-genes and gene transcripts are significantly different (see **Figure 2B**, Wilcoxon rank sum test q-value $\leq$ 0.0002). The shift in the position of the peak in breast cancer subtypes relative to normal tissue suggests a loss of communication.

Edges also differ depending on the omics involved. Differences on mutual information distributions between omics and subtypes are significant (Kolmogorov–Smirnov q-value $\leq$ 5.53264e-06). TF-genes and gene transcripts follow the same distribution on each network. It is noticeable how small is the range of miRNA interactions and how CpG distributions segregate.

In **Table 1** and **Figure 2**, we have characterized the interactions occurring within and between different omics in each molecular subtype of breast cancer. We may appreciate that both intra-layer and inter-layer interaction sets are specific to each biological condition. In what follows, we will now leverage both the monolayer and multilayer interactions to further elucidate biological functions associated with each molecular subtype.

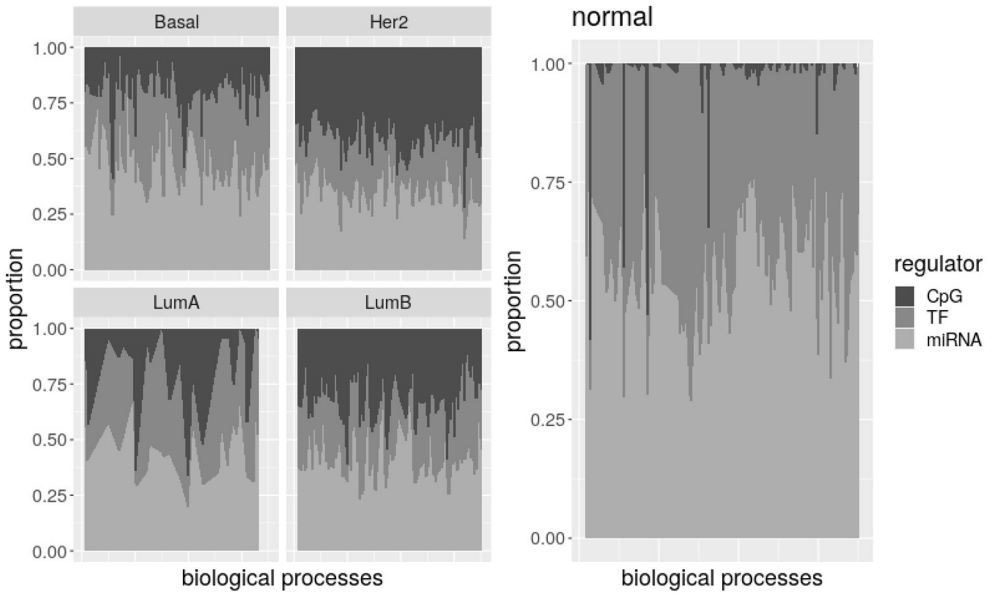

FIGURE 3 | Potential regulators per biological process. There is an area plot per network. Each column is a process. The proportion of regulators of each type associated with the process is denoted by the grayscale. All the processes together show how common are the three potential regulators in the subtype.

3.2. Representation of Potential Regulators Changes With the Subtype

To further explore the differences among potential regulators, its abundance per biological processes was calculated. To this end, total number of CpG, TF-genes, and miRNA nodes were obtained for each biological process. The proportion of regulators of each type is shown in **Figure 3** as a simple measure of the impact a regulatory layer has in a given subtype. A version of this figure with labels for biological processes and the corresponding table are available as **Supplementary Material**.

Despite variability, it is evident that the number of CpG nodes increases on breast cancer subtypes relative to normal tissue, while TF-genes and miRNA numbers of nodes are lower. The plot for Luminal A subtype is less noisy because this subtype has less processes on its network. Nevertheless, by comparing processes represented in each subtype and normal tissue, we found most processes are significantly enriched of CpG nodes in the Basal, Her2+, and LumB subtypes. Simultaneously, TF-genes and miRNAs are significantly under-represented on more than half of the processes in the Her2+ and LumB networks. Additionally, between 20 and 33% of the Basal- and LumA-associated processes show under-representation of TF-genes and miRNAs, and almost half of LumA processes are enriched of CpG nodes.

If potential regulators are actually regulating their associated processes, this may indicate transcriptional and post-transcriptional regulations are subdue in breast cancer subtypes while epigenetic regulation gains strength. By considering the combined effect across layers (inter-layer regulation) as well as the effects on a single type of molecular interaction, as given by each omic dataset (intra-layer regulation), it is possible to develop a deeper understanding of cross-regulatory effects. This will be considered in the next subsections in the context of the different tumor subtypes.

3.3. Normal Interactions With Potential Regulators Are Almost Absent in Breast Cancer Networks

Having seen that the abundance of complete regulatory layers is not maintained across subtypes, we wondered what happens to specific regulatory interactions. With this in mind, we calculated the extent to which interactions with potential regulators are shared among networks by calculating their associated Jaccard indices. The Jaccard index weights the size of the intersection between two sets with the size of their union. In other words, it counts what fraction of the elements is shared from the total. This way, sets of different extensions are assigned values between 0 and 1, and can be objectively compared.

From the total of 176 biological processes enriched in any network, 86.36% appear in at least two subtypes and also are able to share edges. Interactions with miRNAs are poorly shared, while TF-genes and CpG-edges reach a similar maximum but following different distributions (Kolmogorov–Smirnov test q-value $\leq$ 2.498002e-16). Links with any regulator are almost not shared between the breast cancer subtypes and the normal tissue (Kolmogorov–Smirnov test q-value $\leq$ 1.541449e-06), but TF-genes are visibly more shared. The five biological processes with the highest Jaccard index are shown in **Figure 4**.

Localization of telomerase RNA (hTR) to the Cajal body has the highest index for miRNAs for the sharing among Basal and Her2+ networks. This process is also the fifth for TF-genes, but pairing Her2+ and LumB. **Figure 5** shows that the elevated Jaccard indices are driven by only few shared interactions among sets of small size. Although potential regulation changes, the

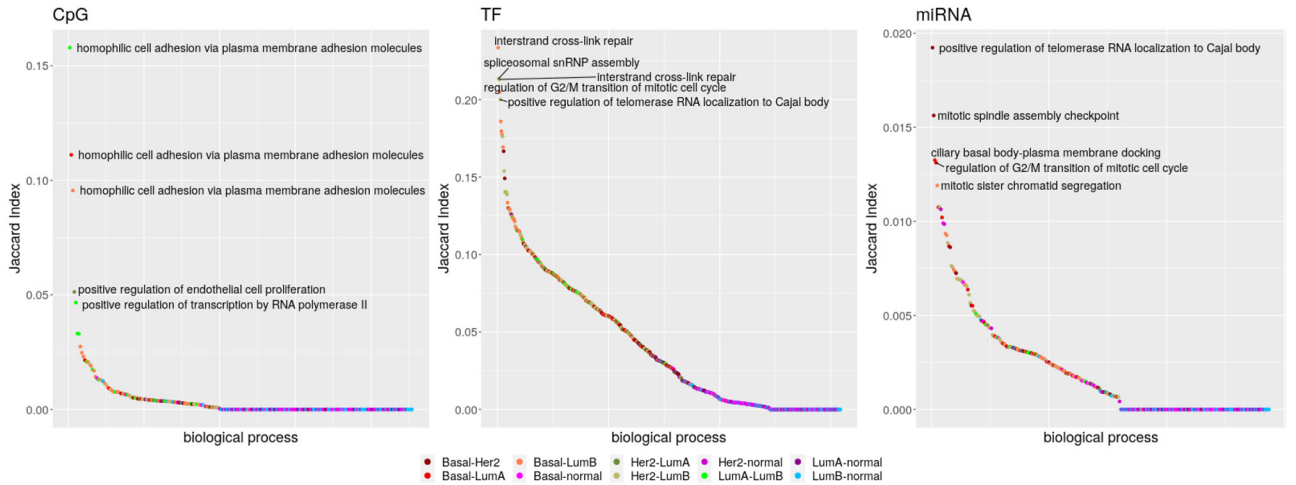

FIGURE 4 | Inter-subtype sharing. Biological processes are symbolized by dots, ordered from those sharing more potentially regulatory interactions to those sharing less. The color code indicates which networks are contrasted. Comparisons with the Basal network are in reddish colors, and those with Her2+ are in dry greens; indexes involving the normal tissue go from pink to blue and those between the luminal subtypes are in bright green.

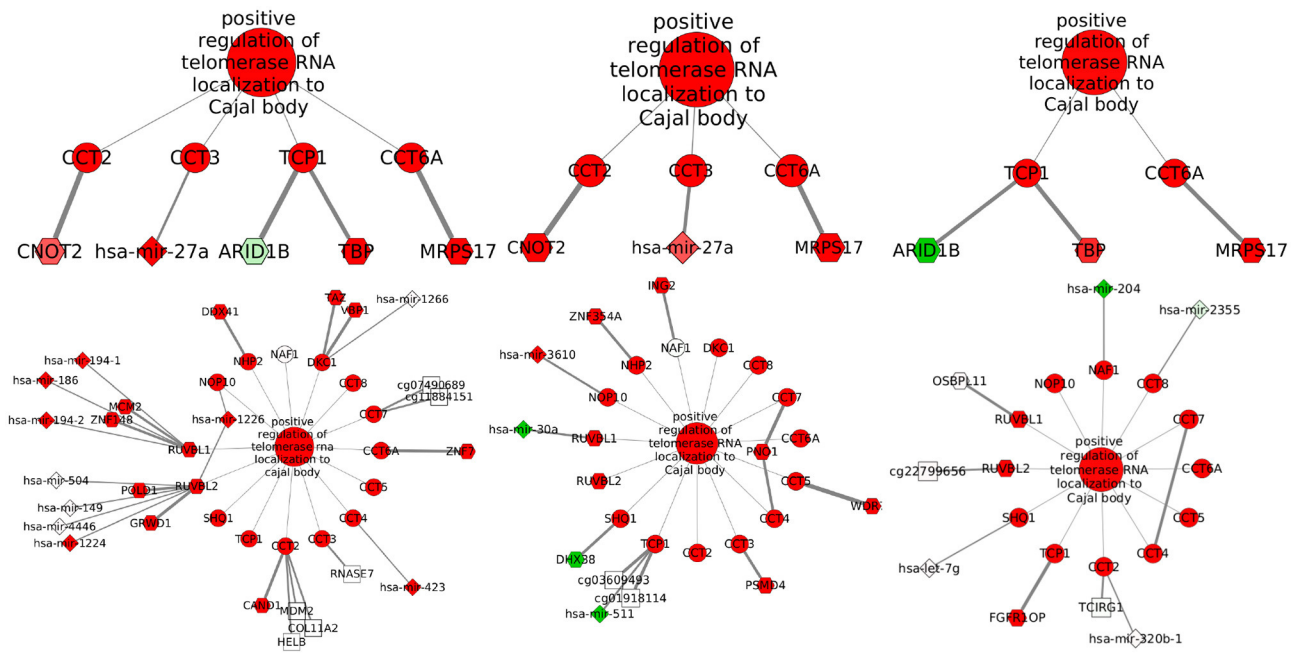

FIGURE 5 | Potential regulators associated with telomerase localization to Cajal body. From left to right, subgraphs with shared **(top)** and exclusive **(bottom)** interactions are shown for Basal, Her2+, and LumB subtypes. Nodes are colored in red if differential expression values or GSEA normalized enrichment scores are positive or in green if values are negative. Node transparency represents statistical significance. Nodes for biological processes and transcripts are circles, TF-genes are hexagons, miRNAs are diamonds, and CpG sites are squares. When the altered gene is known, its name is on the CpG label. Edge thickness represents MI values.

process is equivalently activated in these three subtypes. The interaction linking Chaperonin Containing TCP1 Subunit 6A (CCT6) with Mitochondrial Ribosomal Protein S17 (MRPS17) is shared across these three subtypes, but may be an artifact of the physical proximity of the genes.

3.4. Within Subtypes, CpG Nodes Are Exclusive of Processes, but miRNAs Do Not

For a complementary perspective, we checked if regulators are shared between the distinct biological processes enriched in a

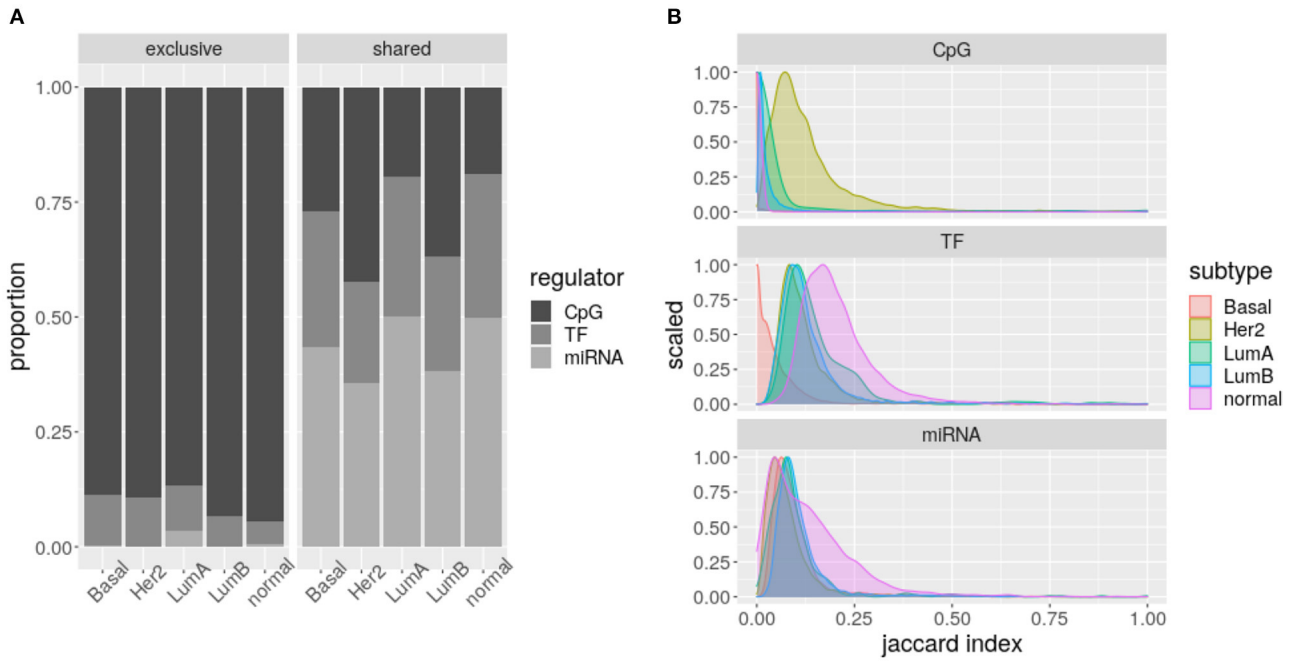

FIGURE 6 | Intra-subtype sharing: **(A)** Proportion of regulators shared or exclusive, across biological processes per subtype. Each column represents a different network. The grayscale indicates the potential regulator. **(B)** Jaccard Index distributions per regulator. The color code indicates the network represented.

single subtype. Degree distributions suggest that CpG sites are exclusive, while miRNAs and TF-genes are shared.

Figure 6A shows how CpG sites are mostly exclusive of one biological process (Fisher test q-value $\leq$ 1.949349e-67), while TF-genes and specially miRNAs are shared between various processes (Fisher test q-value $\leq$ 1.411310e-11). That is, miRNA expression seems to connect different biological processes, while for CpG methylation this effect is much lower.

When calculating the Jaccard index of the biological processes enriched for each subtype and regulator, significantly different distributions are obtained (Kolmogorov–Smirnov test q-value $\leq$ 0.0221, **Figure 6B**). Consistently, as presented in **Figure 6A**, these distributions show CpG sites are less shared, but TF-genes seem to be more shared than miRNAs. The CpG sites of Her2+ and the TF-genes of Basal subtypes call for attention.

4. DISCUSSION

With the aim of exploring potential regulatory patterns of breast cancer subtype expression, we reconstructed via mutual information, multi-omics networks, functionally enriched in GO biological processes. The hypothesis is that there may be a transitive property between the regulators of a transcript and the function associated with the transcript.

This way, potential regulators emerging from the networks are associated with the biological processes significantly enriched. Potential regulators separate domains topologically from non-regulatory transcripts and from each other. Degree distributions are coherent with the pattern of exclusivity and sharing across processes, observed later for CpG sites and TF-genes–miRNAs, respectively. Both results coincide with what is known for the molecule types. Namely, CpG sites have a rather local effect (Li and Zhang, 2014), while TF-genes and miRNAs are *promiscuous*, spanning through a much wider chromosome range (Cho, 2007).

Given the pattern of sharing/exclusivity across processes, one could expect that targeting DNA methylation may drive focused changes, while miRNAs and TF-genes targeting may show pleiotropy. However, current modulators of DNA methylation act over the whole genome, making impossible to change sites related to specific processes. On the contrary, CpG sites linked to specific processes may have potential as predictors of process alteration. Such potential is promising given the early timing of methylation alterations in other cancer types (Vrba and Futscher, 2019). For example, there are 19 CpG sites associated with DNA damage checkpoint in Her2+ subtype, suggesting a possible monitoring mechanism. Nevertheless, it would be necessary to have a whole new project to test the predictability of such sites. The value of the multilayer networks presented here is to propose this kind of hypothesis among all possible combinations, though they need further testing.

To verify that CpG exclusivity per process is not induced by the omission of CpG–miRNA and miRNA–miRNA interactions, non-functionally enriched networks were revisited (**Supplementary Figure 7**). Distributions still change per omic (Wilcoxon rank sum test q-value $\leq$ 4.657478e-16), while the percentage of CpG nodes with degree equal to one is maintained above 90%, indicating that observations made for the first neighbors are relevant when considering farther neighbors. By considering the top 10,000 MI interactions per paired molecules,

we observed that CpG sites do not significantly participate in the regulatory circuitry flow but are often endpoints.

Shortest-paths distributions point out to a decrease in communication independently of the omic observed. This is in line with the under-representation of TF-genes and miRNAs detected specially in Her2+ and Luminal B associated processes. To reconcile communication reduction with over-representation of CpG sites on the subtypes, it is necessary to remember that most CpG nodes do not participate in network connection. These layer level patterns consistently match literature reports on alteration of CpG methylation (Cancer Genome Atlas Network, 2012; Berger et al., 2018), and miRNA expression in breast cancer (O'Day and Lal, 2010; Bertoli et al., 2015; Klinge, 2018).

Two subtype-specific patterns attracted our attention, elevated sharing of CpG nodes between the processes enriched for the subtype Her2+, and decreased sharing of Basal TF-genes. The 2,112 CpG sites shared by Her2+ processes are all over the genome, with a slight increase in chromosomes 1 and 17. While chromosome 1 has been reported as severely affected by differential methylation (Lindqvist et al., 2014), the characteristic amplification of chromosome 17 cannot be fully accounted for the excess sharing. Only 76 from the 1576 genes affected by shared CpG sites co-amplify with the *Her2* gene. Similarly, only 22.91% of affected genes have evidence of AR regulation, a TF postulated to crosstalk with Her2 amplification (Daemen and Manning, 2018).

The other pattern that caught our attention is the decrease in TF-genes linking any two processes in the network for the Basal subtype. This is not caused by a decrease in TF-genes, since the quantity of TF-gene nodes associated with the processes is equivalent for all the networks. Uniqueness of biological processes in the Basal network are neither responsible, seeing that only 6 processes are exclusive for this subtype. Instead, we speculate the pattern is related to promoter accessibility because of ATAC-seq data groups tumors in Basal and non-basal networks (Corces et al., 2018). Further characterization finds a pro-metastasis open-chromatin signature elevated in the Basal subtype (Cai et al., 2020). By its side, protein level measures integrated with copy number normalized gene expression suggest TF-genes as relevant drivers of this subtype (Koh et al., 2019).

Only one edge level pattern was found, but it is a remarkable one. Interactions with regulatory potential are poorly shared among all networks, but the edges of the normal tissue network are almost endemic, especially in the case of CpG sites and miRNAs. If we conform to the idea that DNA methylation preserves cell type identity (Szyf, 2012), our results advert mammary gland defining methylation has been lost in processes like T-cell receptor signaling pathway and inflammatory response.

Localization of hTR to the Cajal body is a biological process linked with cancer cell's unlimited division, given that these organelles have been implicated in the biogenesis of telomerase (Tomlinson et al., 2008). Associated subgraphs exhibit how few edges are shared across subtypes and suggest a convergence of different regulatory schemes to a single outcome. The relative uniformity of enrichment scores across subtypes (**Supplementary Figure 8**) indicates this could be common. Such pattern is important because the way a tumor gains an expression signature might create different vulnerabilities. An example is given by tumors compatible with Her2-enriched expression, but lacking the mutation that makes tumors sensitive to targeted treatment (Godoy-Ortiz et al., 2019).

We must, however, stress that one limitation of the current approach resides on the relatively small sample size. This is a constraint due to lack of availability of a larger dataset comprising the same types of multi-omic data. Limited availability of additional independent datasets also precluded us to validate our findings on an independent cohort. To partially alleviate this, we have resorted to subsampling procedures and null models. The effect of data size differences can be seen in **Supplementary Figure 9** and **Supplementary Table 2**. **Supplementary Table 2** and **Supplementary Figure 9** show the dispersion between MI values estimated with the whole set of samples as well as values obtained through subsampling, for the interactions with the lowest, most varying significance, those between miRNAs and transcripts. Though subsampling repetition is low (100), it catches a tendency toward small z-scores and noisier low subsampled MI values. This means higher z-scores are not necessarily bad, since the large difference between complete and subsampled values maintains points at the top of the range. Altogether, subsampling suggests adding samples would reach higher MI values, but would not alter the ranking dramatically, which supports the (cautious) usage of datasets such as the one used for Her2+. Nevertheless, our analysis could only take advantage of an increase of the number available samples.

As with other areas of molecular biology, one driving force behind the development of multi-omics is the expectation that the results from these technologies may lead to novel pharmacological interventions (de Anda-Jáuregui and Hernández-Lemus, 2020). Nevertheless, the translation from the identification of a perturbation to clinical implementation is not straightforward (Silverman et al., 2020). In this regard, pharmaceutical interventions in each of the analyzed layers are unevenly distributed: drugs that have effects on epigenetic modifications such as methylation have not attained the efficacy that was expected (Buocikova et al., 2020), although they remain an important research area. Meanwhile, gene expression has been able to identify biomarkers as well as drug repurposing opportunities (Mejía-Pedroza et al., 2018; Koudijs et al., 2019). In this context, the type of analyses that we present here provides the opportunity to identify not only the deregulation features in each regulatory layer but also the way it connects to other molecular elements. As such, the opportunity to modulate virtually undruggable targets through the control of its neighbors may help unblock therapeutic opportunities. However, as we mentioned previously, the path from these initial data analyses toward a translational and eventually a clinical setting is long and not necessarily direct.

4.1. Summary of Findings

In brief, the main findings that have been derived from our analysis may be summarized as follows:

- For networks associated with tumor subtypes:
  - Shortest paths are longer for the four subtypes than for the normal tissue.
  - Most biological processes (over 85%) are enriched for CpG nodes in Basal, Her2+, and LumB. Only 41.38% of the processes in LumA are enriched for CpG nodes.
  - Most biological processes (over 50%) are under-represented of TF-gene and miRNA nodes in Her2+ and LumB.
  - Interactions with CpGs and miRNAs found in normal tissue network are near endemic.
  - Her2+ CpG nodes are more shared between processes than expected.
  - Basal TF-gene nodes are less shared between processes than expected.
- For differences in the representation of different omics:
  - CpG nodes tend to show degree = 1, which translates into exclusivity for each process.
  - TF-genes have fewer nodes with degree = 1, and miRNAs have even less. Consistently, these nodes are more shared between processes thus participating in concerted network communication.
  - miRNAs degree distribution shape is remarkably different.
- For shared interactions:
  - Those with CpGs and miRNAs are less maintained than those with TF-genes.

5. CONCLUSIONS

Together, the observations made from multi-omic mutual information networks for the different breast cancer subtypes build a landscape of the differential influence the distinct regulatory layers may exert over the phenotypes. This expands our understanding of breast cancer associated regulatory phenomena and poses possible treatment alternatives to be further explored. For example, now that there is evidence that CpG methylation coordinates with the expression of Her2-associated genes involved in most biological processes more than in any other subtype, experiments with de-methylation agents on this specific subtype seem relevant to analyze.

So far, the interaction between regulatory layers has been overlooked due to the paucity of data and inadequacy of methods. Yet, mutual information calculations and the available algorithms just presented have no formal restriction to handle different omics, unlike other correlation measures MI allows to handle variables with disparate dynamic ranges as it relies in the probability distributions, and has proven capable to retrieve single omics regulatory interactions. Results obtained with the multi-omic setting are encouraging, though refinement of post-MI analysis is needed and is indeed a further avenue of research within our group.

In order to capture CpG methylation and miRNAs linked to biological processes via the interaction with one another, a more sophisticated method would be needed. For example, a computationally expensive recovery of all the paths between transcripts associated with functions. Another possible improvement would be the implementation of a multi-omics data processing inequality (DPI). DPI states that the edge with the smaller MI in a triangle can be filtered out as indirect. However, MI distribution changes for every type of omics paired complicating MI comparisons. Perhaps a better alternative will be to resort to tensor representations of probabilistic multilayer networks (Hernández-Lemus, 2020).

It is also pertinent to recall that higher mutual information does not translate into causal interactions. The so-called *potential regulators* may simply co-vary with transcript expression, or causality may be dependent on an intermediate node. Even if linked CpGs sites regulated gene expression, omics that are not included like copy number variation may also play relevant roles. To identify the potential regulators whose patterns are most related to transcripts expression, there are other strategies available (Lê Cao et al., 2009), which may benefit from MI interaction scores (Koh et al., 2019). There are however more insights to be extracted from the multi-omics networks yet.

With the set of potential regulators associated with a biological process, we aspire to multi-layer regulatory models that include examples like the one described for miRNA processing enzymes Drosha and Dicer (Rupaimoole et al., 2014). Here, we present general results, but particular cases can be further examined within this general approach. When the focus is on particular models, the distinct regulators connected to single gene allow the proposal of hypothesis about synergy and antagonism among regulation layers. Nevertheless, this approach calls for a much more detailed scrutiny.

All in all, due to the relative simplicity and generalizability of the approach, the use of combined probabilistic modeling and knowledge discovery in databases presented here allows for the inference of regulatory models that may be refined by resorting to more specialized techniques, both experimental and computational.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SO organized data, developed code, performed calculations, analyzed data, and drafted the manuscript. GA-J contributed to the methodological approach, analyzed data, discussed results, and co-supervised the project. EH-L envisioned the project, devised the methodological strategy, designed the study, contributed to the methodological approach, analyzed data, discussed results, reviewed the manuscript, and supervised the project. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], 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. EH-L is recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

The authors want to thank Gabriela Graham for her support with language editing and proofreading of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617512/full#supplementary-material

REFERENCES

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369. doi: 10.1093/bioinformatics/btu049

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* 33, 690–705.e9.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 17(Suppl. 2):15. doi: 10.1186/s12859-015-0857-9

Bertoli, G., Cava, C., and Castiglioni, I. (2015). MicroRNAs: New biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer. *Theranostics* 5, 1122–1143. doi: 10.7150/thno.11543

Bhuva, D. D., Cursons, J., Smyth, G. K., and Davis, M. J. (2019). Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome Biol.* 20, 1–21. doi: 10.1186/s13059-019-1851-8

Buocikova, V., Rios-Mondragon, I., Pilalis, E., Chatziioannou, A., Miklikova, S., Mego, M., et al. (2020). Epigenetics in breast cancer therapy-new strategies and future nanomedicine perspectives. *Cancers* 12:3622. doi: 10.3390/cancers12123622

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Cai, W. L., Greer, C. B., Chen, J. F., Arnal-Estapé, A., Cao, J., Yan, Q., et al. (2020). Specific chromatin landscapes and transcription factors couple breast cancer subtype with metastatic relapse to lung or brain. *BMC Med. Genomics* 13:33. doi: 10.1186/s12920-020-0695-0

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Cho, W. C. S. (2007). Oncomirs: the discovery and progress of microRNAs in cancers. *Mol. Cancer* 6:60. doi: 10.1186/1476-4598-6-60

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. *Nucl. Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., et al. (2018). The chromatin accessibility landscape of primary human cancers. *Science* 362:eaav1898. doi: 10.1126/science.aav1898

Daemen, A., and Manning, G. (2018). Her2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in AR-driven breast tumors. *Breast Cancer Res.* 20:8. doi: 10.1186/s13058-018-0933-y

de Anda-Jáuregui, G., Alcalá-Corona, S. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019). Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Appl. Netw. Sci.* 4, 1–13. doi: 10.1007/s41109-019-0129-0

de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Computational oncology in the multi-omics era: state of the art. *Front. Oncol.* 10:423. doi: 10.3389/fonc.2020.00423

de Anda-Jáuregui, G., Velázquez-Caldelas, T. E., Espinal-Enríquez, J., and Hernández-Lemus, E. (2016). Transcriptional network architecture of breast cancer molecular subtypes. *Front. Physiol.* 7:568. doi: 10.3389/fphys.2016.00568

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the infinium methylation 450k technology. *Epigenomics* 3, 771–784. doi: 10.2217/epi.11.105

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, 1–23. doi: 10.1007/s41109-020-00291-1

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of EMT and MET micro-RNA regulation in breast cancer. *Sci. Rep.* 7:13534. doi: 10.1038/s41598-017-13903-1

Espinal-Enriquez, 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, 1–19. doi: 10.1038/s41598-017-01314-1

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernandez-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

GM., C. (2000). *The Development and Causes of Cancer. The Cell: A Molecular Approach, 2nd Edn*. Sunderland: Sinauer Associates.

Godoy-Ortiz, A., Sanchez-Muñoz, A., Chica Parrado, M. R., Álvarez, M., Ribelles, N., Rueda Dominguez, A., et al. (2019). Deciphering her2 breast cancer disease: biological and clinical implications. *Front. Oncol.* 9:1124. doi: 10.3389/fonc.2019.01124

Hernández-Lemus, E. (2020). On a class of tensor Markov fields. *Entropy* 22:451. doi: 10.3390/e22040451

Kim, D., Shin, H., Song, Y. S., and Kim, J. H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Inform.* 45, 1191–1198. doi: 10.1016/j.jbi.2012.07.008

Klinge, C. M. (2018). Non-coding RNAs: long non-coding RNAs and microRNAs in endocrine-related cancers. *Endocr. Relat. Cancer* 25, R259–R282. doi: 10.1530/ERC-17-0548

Koh, H. W. L., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iomicspass: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst. Biol. Appl.* 5:22. doi: 10.1038/s41540-019-0099-y

Koudijs, K. K. M., Terwisscha van Scheltinga, A. G. T., Böhringer, S., Schimmel, K. J. M., and Guchelaar, H.-J. (2019). Transcriptome signature reversion as a method to reposition drugs against cancer for precision oncology. *Cancer J.* 25, 116–120. doi: 10.1097/PPO.0000000000000370

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Lê Cao, K.-A., Martin, P. G. P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10:34. doi: 10.1186/1471-2105-10-34

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

Lindqvist, B. M., Wingren, S., Motlagh, P. B., and Nilsson, T. K. (2014). Whole genome dna methylation signature of Her2-positive breast cancer. *Epigenetics* 9, 1149–1162. doi: 10.4161/epi.29632

Liu, Y. (2020). Clinical implications of chromatin accessibility in human cancers. *Oncotarget* 11, 1666–1678. doi: 10.18632/oncotarget.27584

Liu, Y., Liu, Y., Huang, R., Song, W., Wang, J., Xiao, Z., et al. (2019). Dependency of the cancer-specific transcriptional regulation circuitry on the promoter DNA methylome. *Cell Rep*. 26, 3461–3474.e5. doi: 10.1016/j.celrep.2019.02.084

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366–370. doi: 10.1038/nmeth.3799

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

McCarthy, D. J., and Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a treat. *Bioinformatics* 25, 765–771. doi: 10.1093/bioinformatics/btp053

Mejía-Pedroza, R. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2018). Pathway-based drug repositioning for breast cancer molecular subtypes. *Front. Pharmacol*. 9:905. doi: 10.3389/fphar.2018.00905

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286. doi: 10.1016/j.cell.2012.04.040

Ochoa, S., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Multi-omic regulation of the pam50 gene signature in breast cancer molecular subtypes. *Front. Oncol*. 10:845. doi: 10.3389/fonc.2020.00845

O'Day, E., and Lal, A. (2010). Micrornas and their target gene networks in breast cancer. *Breast Cancer Res*. 12:201. doi: 10.1186/bcr2484

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., et al. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* 24, S26–S35. doi: 10.1016/j.breast.2015.07.008

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., et al. (2014). The multimir R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucl. Acids Res*. 42:e133. doi: 10.1093/nar/gku631

Rupaimoole, R., Wu, S. Y., Pradeep, S., Ivan, C., Pecot, C. V., Gharpure, K. M., et al. (2014). Hypoxia-mediated downregulation of miRNA biogenesis promotes tumour progression. *Nat. Commun*. 5:5202. doi: 10.1038/ncomms6202

Sergushichev, A. A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*. 1–40. doi: 10.1101/060012

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Silverman, E. K., Schmidt, H. H. H. W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., et al. (2020). Molecular networks in network medicine: development and applications. *Wiley Interdisc. Rev*. 12:e1489. doi: 10.1002/wsbm.1489

Szyf, M. (2012). Dna methylation signatures for breast cancer classification and prognosis. *Genome Med*. 4:26. doi: 10.1186/gm325

Tam, S., Tsao, M.-S., and McPherson, J. D. (2015). Optimization of miRNA-seq data preprocessing. *Brief. Bioinformatics* 16, 950–963. doi: 10.1093/bib/bbv019

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

Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D., Schmidt, A., Imhof, A., Hankemeier, T., et al. (2020). Harmonization of quality metrics and power calculation in multi-omic studies. *Nat. Commun*. 11:3092. doi: 10.1038/s41467-020-16937-8

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with noiseq R/bioc package. *Nucl. Acids Res*. 43:e140. doi: 10.1093/nar/gkv711

Tian, T., Olson, S., Whitacre, J. M., and Harding, A. (2011). The origins of cancer robustness and evolvability. *Integr. Biol*. 3, 17–30. doi: 10.1039/C0IB00046A

Tomlinson, R. L., Abreu, E. B., Ziegler, T., Ly, H., Counter, C. M., Terns, R. M., et al. (2008). Telomerase reverse transcriptase is required for the localization of telomerase RNA to cajal bodies and telomeres in human cancer cells. *Mol. Biol. Cell* 19, 3793–3800. doi: 10.1091/mbc.e08-02-0184

Turashvili, G., and Brogi, E. (2017). Tumor heterogeneity in breast cancer. *Front. Med*. 4:227. doi: 10.3389/fmed.2017.00227

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* 26, i237–i245. doi: 10.1093/bioinformatics/btq182

Vrba, L., and Futscher, B. W. (2019). Dna methylation changes in biomarker loci occur early in cancer progression. *F1000Research* 8:2106. doi: 10.12688/f1000research.21584.1

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 March 2021
doi: 10.3389/fgene.2021.617282

Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms

*Yoonjee Kang, Denis Thieffry and Laura Cantini**

Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR 8197, INSERM U1024, Ecole Normale Supérieure, Paris Sciences et Lettres Research University, Paris, France

Networks are powerful tools to represent and investigate biological systems. The development of algorithms inferring regulatory interactions from functional genomics data has been an active area of research. With the advent of single-cell RNA-seq data (scRNA-seq), numerous methods specifically designed to take advantage of single-cell datasets have been proposed. However, published benchmarks on single-cell network inference are mostly based on simulated data. Once applied to real data, these benchmarks take into account only a small set of genes and only compare the inferred networks with an imposed ground-truth. Here, we benchmark six single-cell network inference methods based on their reproducibility, i.e., their ability to infer similar networks when applied to two independent datasets for the same biological condition. We tested each of these methods on real data from three biological conditions: human retina, T-cells in colorectal cancer, and human hematopoiesis. Once taking into account networks with up to 100,000 links, GENIE3 results to be the most reproducible algorithm and, together with GRNBoost2, show higher intersection with ground-truth biological interactions. These results are independent from the single-cell sequencing platform, the cell type annotation system and the number of cells constituting the dataset. Finally, GRNBoost2 and CLR show more reproducible performance once a more stringent thresholding is applied to the networks (1,000–100 links). In order to ensure the reproducibility and ease extensions of this benchmark study, we implemented all the analyses in scNET, a Jupyter notebook available at https://github.com/ComputationalSystemsBiology/scNET.

Keywords: biological networks, scRNA-seq, single-cell, transcriptome, network inference, network theory, reproducibility

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer, Centre for Molecular Medicine Norway, Faculty of Medicine, University of Oslo, Norway

Reviewed by:
Van Anh Huynh-Thu, University of Liège, Belgium
Rudiyanto Gunawan, University at Buffalo, United States

***Correspondence:**
Laura Cantini
cantini@bio.ens.psl.eu

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

Received: *14 October 2020*
Accepted: *24 February 2021*
Published: *22 March 2021*

Citation:
Kang Y, Thieffry D and Cantini L (2021) Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. Front. Genet. 12:617282. doi: 10.3389/fgene.2021.617282

INTRODUCTION

Biological systems are inherently complex, in particular because of the emergent phenotypic properties arising from the interaction of their numerous molecular components. Characterizing genotype to phenotype connections and pathological deregulations thus requires to identify the biological macromolecules involved (e.g., genes, mRNAs, proteins), but also how these interact in a huge diversity of cellular pathways and networks (Barabási and Oltvai, 2004).

In the post-genomic era, biological networks have been extensively exploited to investigate such complex interactions among biological macromolecules (Barabási et al., 2011;

Sonawane et al., 2019; Silverman et al., 2020). Network-based studies brought crucial insights into cell functioning and diseases (Basso et al., 2005; Margolin et al., 2006; Ideker and Sharan, 2008). A network is a graph-based representation of a biological system, where the nodes represent objects of interest (e.g., genes, mRNAs, proteins), while the edges represent relations between these objects (e.g., gene co-expression, or binding between two proteins). Different approaches can be used to reconstruct biological networks. Here, we focus on data-driven methods, which infer networks from gene expression data with the help of reverse engineering techniques (Sonawane et al., 2019).

Network inference algorithms were first proposed to extract information from bulk gene expression data, and their development has been an active area of research for more than 20 years (Barabási et al., 2011; The DREAM5 Consortium et al., 2012; Verny et al., 2017; Sonawane et al., 2019; Silverman et al., 2020). With the advent of single-cell RNA sequencing (scRNA-seq), we started to gather transcriptomic data from individual cells, enabling proper studies of their heterogeneity. However, the analysis of scRNA-seq data comes with a variety of computational challenges (e.g., small number of sequencing reads, systematic noise due to the stochasticity of gene expression at single-cell level, dropouts) that distinguish this data type from its bulk counterpart. For this reason, network inference methods originally developed for bulk gene expression data may not be suitable for data generated from single cells. The development of network inference algorithms has thus recently undergone a strong shift towards the design of methods targeting single-cell data (Fiers et al., 2018).

Two benchmarks of single-cell network inference methods have been published (Chen and Mar, 2018; Pratapa et al., 2020). Both works evaluate network inference algorithms by comparing the inferred network with a ground-truth. These works are also mostly focused on simulated data and they apply a strong filtering on genes (leaving only 100–1,000 genes for network inference). Chen and Mar (2018) considered five methods targeting bulk data and three methods specifically designed for single-cell data. More recently, Pratapa et al. (2020) focused on 12 methods designed for single-cell data. Both benchmarks concluded that the overall performances of all methods were quite disappointing, and that network inference remains a challenging problem.

Here, we evaluate network inference algorithms based on their reproducibility, i.e., their ability to infer similar networks once applied to two independent datasets for the same biological condition (e.g., two independent scRNA-seq datasets obtained from colorectal tumors). The rationale behind this comparison is that, if the two independent datasets are profiled from the same biological condition (e.g., colorectal cancer, CRC) involving the same cell types, we can expect that the regulatory programs underlying them should strongly overlap. As a consequence, a good network inference algorithm should infer highly overlapping networks when applied to single-cell datasets profiled from the same biological condition. We selected six algorithms spanning the main network inference formulations that do not require an ordering of the cells according to pseudo-time, and we tested the reproducibility of the inferred networks in three biological systems: human retina, T-cells in CRC and human hematopoiesis. Differently from previous benchmarks, we only applied a soft filtering on genes, thus testing the algorithms based on their performances to infer networks involving from 6,000 to 12,000 nodes/genes.

From our benchmark, once an high number of links is taken into account (100,000), GEne Network Inference with Ensemble of Trees (GENIE3) results to be the most reproducible algorithm and, together with GRNBoost2, show the highest intersection with ground-truth biological interactions. GRNBoost2 and Context Likelihood of Relatedness (CLR) have instead better performances for low link numbers (1,000–100). In order to ensure the reproducibility and ease extensions of this benchmark study, we implemented all the analyses in a Jupyter notebook, called scNET and available at https://github.com/ComputationalSystemsBiology/scNET.

MATERIALS AND METHODS

Benchmarked Single-Cell Network Inference Algorithms

Starting from the exhaustive collection of single-cell network inference algorithms presented in Chen and Mar (2018) and Pratapa et al. (2020), two main categories of methods can be distinguished. Some methods interpret scRNA-seq as time-course expression data, where the pseudo-time corresponds to the time information. These methods are frequently based on Ordinary Differential Equations (ODEs) and are relevant for biological systems undergoing dynamic transcriptional changes (e.g., scRNA-Seq performed on differentiating cells) (Matsumoto et al., 2017). In contrast, other methods do not use pseudo-time information to infer networks. These methods generally use statistical measures (e.g., correlation, mutual information) to infer regulatory connections and are thus better suited for transcriptomic data not affected by strong dynamical processes (e.g., retina cells in normal state).

Testing reproducibility strictly requires the availability of two independent scRNA-seq datasets reflecting the same biological condition and presenting as few as possible technical variations. Indeed, the presence of technical variations due to the sequencing or experimental procedures could drastically impact the outcome of our comparison. In this respect, finding independent scRNA-seq datasets reflecting dynamic transcriptional changes, generated with the same experimental procedure, is really challenging. We thus decided to focus our benchmark study on network inference methods that do not use the pseudo-time information. In addition, only algorithms provided in R or Python code are here taken into account. Six single-cell network inference methods are thus considered in this evaluation: GENIE3 (Huynh-Thu et al., 2010), GRNBoost2 (Moerman et al., 2019), PPCOR (Kim, 2015), Partial Information Decomposition and Context (PIDC; Chan et al., 2017), CLR (Faith et al., 2007), and GeneNet (Opgen-Rhein and Strimmer, 2007). All the methods selected for this benchmark were originally designed for bulk data and they span the main mathematical formulations of network inference, as described

in The DREAM5 Consortium et al. (2012). Of note, GENIE3, GRNBoost2 and PIDC are also the best performing in the single-cell benchmark of Pratapa et al. (2020).

GEne Network Inference with Ensemble of Trees (Huynh-Thu et al., 2010) is a tree-based network inference method. For each gene g_i in the expression dataset, GENIE3 solves a regression problem, determining the subset of genes whose expression is the most predictive of the expression of g_i. This method was the best performing algorithm in the DREAM4 In Silico Multifactorial challenge (Greenfield et al., 2010). GENIE3 requires in input the scRNA-seq expression matrix and a list of Transcription Factors (TFs). In our tests the list of human TFs provided in input corresponds to the intersection between the expressed genes and those annotated as encoding TFs by Chawla et al. (2013). The output of GENIE3 is a weighted network linking TFs with predicted target genes. The weight associated with each link corresponds to its Importance Measure (IM), which represents the weight that the TF has in the prediction of the level of expression of the target gene. We run GENIE3 from the Arboreto library (Moerman et al., 2019) using default parameters.

GRNBoost2 (Moerman et al., 2019) has been developed as a faster alternative to GENIE3. It is thus based on a regression model, using a stochastic gradient boosting machine regression. The inputs and outputs of GRNBoost2 have the same structure of those of GENIE3. Both GRNBoost2 and GENIE3 are part of the SCENIC workflow (Aibar et al., 2017). We run GRNBoost2 from the Arboreto library (Moerman et al., 2019) using default parameters.

PPCOR (Kim, 2015) infers the presence of a regulatory interaction between two genes by computing the correlation of their expression patterns. To control for possible indirect effects, partial correlation is used instead of a simple correlation, where partial correlation is a measure of the relationship between two variables while controlling for the effect of other variables. The only input of PPCOR is the expression matrix. The output of PPCOR is a weighted network, where all links are weighted based on the partial correlation between the expression values of the linked nodes/genes.

Partial Information Decomposition and Context (Chan et al., 2017) is based on concepts from information theory and uses partial information decomposition (PID) to identify potential regulatory relationships between genes. The only input of PIDC is the expression matrix and its output is a weighted gene-gene network.

Context Likelihood of Relatedness (Faith et al., 2007) is another commonly used approach based on concepts from information theory. The measure used by CLR to infer links in between genes is Mutual Information (MI). In contrast with other algorithms also based on MI, such as ARACNE (Margolin et al., 2006), CLR adjusts the link weights for the background distribution of the MI values to control for false positives interactions.

GeneNet (Opgen-Rhein and Strimmer, 2007) is a method for statistical learning of a high-dimensional causal network. The method first converts a correlation network into a partial correlation graph. Subsequently, a partial ordering of the nodes is established by multiple testing of the log-ratio of standardized partial variances.

To make the different network inference algorithms comparable, we applied the same thresholding to all of them, by keeping only the top K links (K = 100,000). For GeneNet, inferring less than 100,000 links, no filtering has been applied.

Data Acquisition and Preprocessing

Fourteen public scRNA-seq datasets have been used for this benchmark (**Table 1**): Lukowski et al. (2019) and Menon et al. (2019) obtained by profiling human retina cells; Li et al. (2017) and Zhang et al. (2019) profiling T-cells in CRC; Hay et al. (2018) and Setty et al. (2019) profiling human hematopoiesis cells. See **Table 1** for a complete description of these datasets. The hematopoiesis datasets were split according to their cell type of origin. Only those cell types reported in both studies by Hay et al. (2018) and Setty et al. (2019) were considered. We thus obtained a total of 10 scRNA-seq datasets in hematopoiesis spanning five cell types: HSC, CLP, Monocyte, Erythroblast, and Dendritic Cell.

After downloading the data, we filtered the genes based on their total count number ($<3 \times 0.01 \times$ number of cells), as well as on the number of cells in which they are detected ($>0.01 \times$ number of cells), as described in Aibar et al. (2017). The gene filtering is performed on each dataset independently. Then, for each biological condition (CRC T-cells, retina, and hematopoiesis), only the genes retained for both datasets were selected for network inference. The number of genes retained after filtering are reported in the last column of **Table 1**. Finally, the data were log2-normalized before applying the different network inference algorithms.

Indexes Employed to Measure the Reproducibility of the Network Inference Algorithms

Percentage of intersection (perINT) and Weighted Jaccard Similarity (WJS) have been employed here to assess the reproducibility of the network inference algorithms. The percentage of intersection is used to detect the presence of links shared between two compared networks, while WJS takes into account the similarity of the weights associated with the links shared between the compared networks.

Given two networks N_1 and N_2 inferred respectively from scRNAseq datasets D_1 and D_2, and indicating as $|N|$ the number of links in the network N, the perINT is computed as:

$$perINT\left(N_1, N_2\right) = \frac{|N_1 \cap N_2|}{min\left(|N_1|, |N_2|\right)},$$

while the WJS (Tantardini et al., 2019), is defined as

$$WJS\left(N_1, N_2\right) = \frac{\sum_{i=1}^{|N|} min\left(w_i^1, w_j^2\right)}{\sum_{i=1}^{|N|} max\left(w_i^1, w_j^2\right)},$$

where w^1, w^2 are the vectors of weights associated with the links in common between N_1 and N_2.

In addition, to compare the inferred links to a ground-truth, we considered two additional scores: RcisTarget and

TABLE 1 | Datasets employed in this benchmark.

Data Name	Biological context	Sequencing technology	Number of cells	Cell type annotation strategy	Associated publication	Number of genes after preprocessing
Menon	Human retina	10X Genomics	20,091	Manually curated marker genes	Menon et al., 2019	6,212
Lukowski	Human retina	10X Genomics	20,009	No annotation	Lukowski et al., 2019	6,212
Zhang	CRC T-cells	Smart-Seq2,	10,805	FACS sorted	Zhang et al., 2019	11,242
Li	CRC T-cells	HiSeq 2000 Illumina	375 cells (of which 35 T-cells)	Manually curated marker genes	Li et al., 2017	11,242
Hay	human hematopoiesis	10X Genomics	101,935	MarkerFinder ICGS	Hay et al., 2018	7,038
Setty	human hematopoiesis	10X Genomics	12,046	Sorted bulk hematopoietic populations	Setty et al., 2019	7,038

Regulatory Circuit scores. We derived the RcisTarget score from the application of the RcisTarget tool (Aerts et al., 2010; Aibar et al., 2017). Given a network of TF-gene interaction, RcisTarget predicts candidate target genes of a TF by looking at the DNA motifs that are significantly over-represented in the surroundings of the Transcription Start Site (TSS) of all the genes that are linked to the TF. We here consider the links validated by RcisTarget as ground-truth and we compare them with the inferred networks, by computing:

$$RcisTargetScore\,(N_1) = \frac{NumberLinks \in N_1 \cap ValidatedByRcisTarget}{|N_1|}$$

In the case of the methods inferring links between all genes, a selection of links connecting TFs with possible target genes is performed before computing the RcisTaget score.

The Regulatory Circuits score instead is obtained by computing the intersection between an inferred network and tissue-specific regulatory circuits from http://www.regulatorycircuits.org (Marbach et al., 2016). The regulatory circuits considered are the following: adult retina for retina, lymphocytes for CRC T-cells and CD34 stem cell derived for hematopoiesis. We here computed the Regulatory Circuits score for a network N_1 as:

$$RegulatoryCircuitScore\,(N_1) = \frac{\left|N_1 \cap AssociatedRegulatoryCircuits\right|}{|N_1|}$$

RESULTS

Based on previous works (Chen and Mar, 2018; Pratapa et al., 2020), we selected the six single-cell network inference algorithms that do not require an ordering of the cells according to pseudo-time (GENIE3, GRNBoost2, PPCOR, PIDC, CLR and GeneNet see section "Materials and Methods") and we evaluated them based on their reproducibility, i.e., their ability to infer similar networks once applied to two independent datasets from the same biological condition (e.g., two independent scRNA-seq datasets of CRC). The reproducibility is measured based on the perINT and WJS indexes (see section "Materials and Methods"). In addition, we computed the intersection with two instances of ground-truth, based on the RcisTarget and on Regulatory Circuits scores (see section "Materials and Methods"). The evaluation is repeated across three biological conditions: human retina, T-cells in CRC and human hematopoiesis, for a total of 14 independent scRNAseq datasets. See **Figure 1** for an overview of the benchmark workflow.

While in previous benchmarks (Chen and Mar, 2018; Pratapa et al., 2020), a low number of highly variable genes had been taken into account (100–1,000 genes), we here tested the ability of the algorithms to infer networks involving all expressed genes (see section "Materials and Methods" for details on the procedure used to filter genes). Indeed, filtering only the top 100–1,000 varying genes is a strong limitation. Restricting the nodes of the inferred network to a low number of genes is reasonable when a manually curated list of relevant genes is available (for example marker genes identified by wet-lab experiments). However, when such a list is not available, working only with the top 100–1,000 varying genes may overlook genes and interactions playing a key role in the regulatory programs of the biological system. We thus tested the various network inference algorithms once applied to scRNAseq datasets containing 6,000–11,000 genes.

In our test cases, PIDC failed to reconstruct networks for two main reasons: (i) the algorithms was slow, especially in the discretization step required to infer a network and (ii) the use of multivariate information measures impose to have a number of genes much lower than the number of cells, thus requiring to drastically filter out the starting set of genes. Overall, PIDC thus resulted to be more adequate to infer small networks (100–1,000 nodes/genes), which are not the focus of this work.

Reproducibility in Human Retina

We applied GENIE3, GRNBoost2, PPCOR, CLR, and GeneNet to two independent scRNA-seq datasets of human retina, reported in Menon et al. (2019) and Lukowski et al. (2019) (see section "Materials and Methods"). After filtering, the two datasets span 6,212 common genes across a comparable number of cells: 20,091 in Menon versus 20,009 in Lukowski.

We thus inferred a total of ten networks. Details on the number of links before and after thresholding are provided in the **Supplementary Table 1**. We then evaluated the reproducibility of each algorithm by computing the perINT and the WJS between the networks inferred independently from the two datasets.

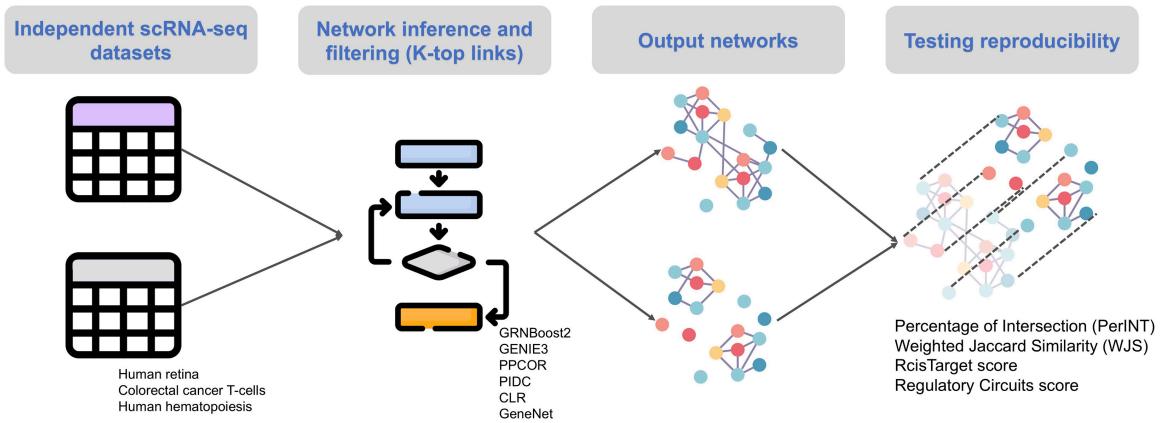

FIGURE 1 | Summary of the workflow followed in this benchmark.

While perINT is intended to test the amount of common links between the two networks, the WJS takes also into account the similarity of the weights associated with the common links.

As shown in **Figure 2A**, GENIE3 (45.9% perINT and 0.28 WJS) and GRNBoost2 (41.1% perINT and 0.25 WJS) are the algorithms showing the highest reproducibility, with GENIE3 performing slightly better. At the same time, in agreement with the results of the previous benchmarks, the intersection with the ground truth considered remains rather low, but higher for GRNBoost2 (1% RcisTarget score and 4.2% Regulatory Circuits score). Similar performances apply also for the other network inference methods.

Reproducibility in Colorectal Cancer T-Cells

We further tested the performances of GENIE3, GRNBoost2, PPCOR, CLR, and GeneNet in CRC T-cells. The two datasets used in this case are taken from Zhang et al. (2019) and Li et al. (2017) (see section "Materials and Methods"), restricting the last dataset to only T-cells (see section "Materials and Methods"). After filtering, we obtained datasets composed of 11,242 common genes and a widely varying number of cells: 10,805 for Zhang, and 35 for Li.

We applied GENIE3, GRNBoost2, PPCOR, CLR and GeneNet independently to the two datasets (for details on the number of links before and after thresholding, refer to **Supplementary Table 1**). Of note, PPCOR has been excluded from this comparison, as it produced partial correlation values outside the range [-1;1] for the Li et al. dataset.

After computation of the perINT and WJS (**Figure 2B**), GENIE3 (3% perINT and 0.008 WJS) and GRNBoost2 (3.4% perINT and 0.007 WJS) emerged as the best performing methods. The reproducibility indexes are quite low in this test case, probably due to the low number of cells present in the Li dataset (35 cells). The RcisTarget and Regulatory Circuits scores reflecting the intersection with a ground-truth are also quite low for all algorithms, with GRNBoost2 showing better performances (4% RcisTarget score and 14.6% Regulatory Circuits score).

Reproducibility in Human Hematopoiesis

Human hematopoiesis has been used as the third biological context for the comparison of GENIE3, GRNBoost2, PPCOR, CLR, and GeneNet. The hematopoiesis datasets were split according to the different cell types profiled: HSC, CLP, Monocyte, Erythroblast, and Dendritic Cell, obtaining a total of 10 scRNA-seq datasets. Networks were thus inferred on each cell type independently with GENIE3, GRNBoost2, PPCOR, CLR, and GeneNet, resulting in a total of 50 networks. Details on the number of links before and after thresholding are available in **Supplementary Table 1**. As for CRC T-cells, PPCOR produced networks composed of links with partial correlation higher than 1 and/or lower than -1 for some CLPs, and Monocytes. For this reason, we did not consider PPCOR in the reproducibility evaluation for these cell types.

The reproducibility was then tested for each cell type using the perINT and WJS indexes (**Figures 2C,D**). GENIE3 displayed the best performances with percentages of intersection of 26–56% and WJS at 0.13–0.37. Consistently with previous observations, the RcisTarget and Regulatory Circuits scores remain low for all cell types and all methods, with GRNBoost2 having slightly better performances than GENIE3 (approx. 2–4.2% and 4–7.6%, respectively) (**Figures 2E,F**).

Stability With Respect to Link Thresholding in the Inferred Networks

In the previous experiments, the 100,000 top-ranked links have been taken into account for all methods, except GeneNet having less than 100,000 links (see section "Materials and Methods," **Supplementary Table 1**). Here we test to which extent our conclusions, regarding the reproducibility of the benchmarked methods, are stable with respect to the number (K) of links retained in each network. We thus apply a more stringent filtering, considering an identical number (K) of top-ranked links of 10,000, 1,000, and 100 for all compared methods. GeneNet has been excluded from this analysis, as the number of its inferred links is lower than 1,000 in most of the cases. After thresholding, the intersection between the networks inferred from independent

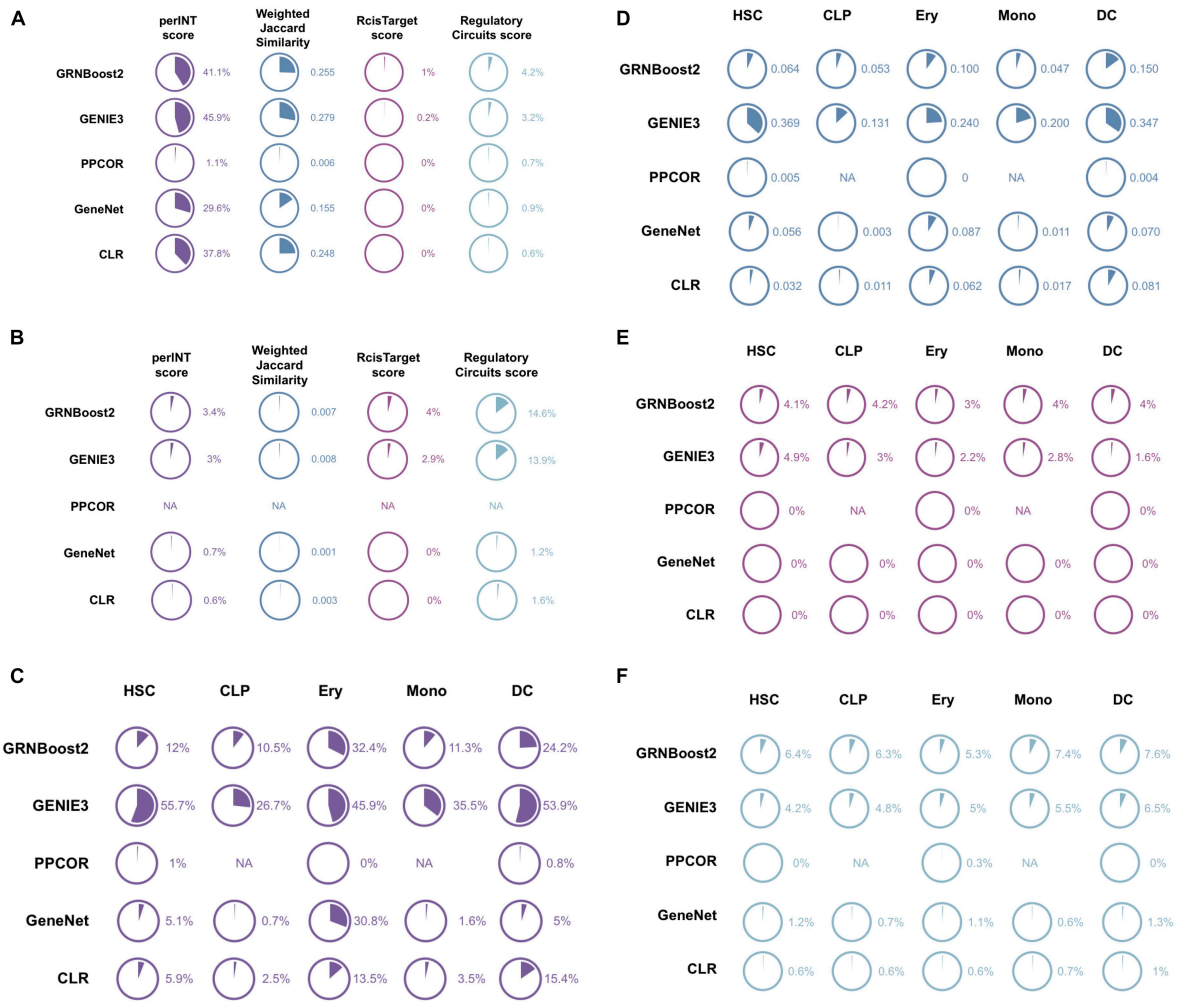

FIGURE 2 | Reproducibility performances of the various network inference algorithms across the three biological contexts: human retina, colorectal cancer T-cells and human hematopoiesis. Panels **(A,B)** report the Percentage of intersection (perINT), Weighted Jaccard Similarity (WJS), RcisTarget score and Regulatory Circuits score obtained by each the benchmarked algorithm (GRNBoost2, GENIE3, PPCOR, CLR, and GeneNet) in human retina and colorectal cancer T-cells, respectively. Panels **(C–F)** summarize the performances of the same algorithms on the hematopoiesis datasets, with perINT **(C)**, WJS **(D)**, RcisTarget score **(E)**, and Regulatory Circuits score **(F)**.

datasets from the same biological condition were evaluated, using the percINT and WJS as above.

As shown in **Figure 3**, the performances of the various algorithms are quite heterogeneous once different thresholds (K) are considered. As observed in the previous sections, GENIE3 tends to have better performances for high K. However, for low numbers of links (K = 1,000 and 100), GRNBoost2 and CLR tend to predominate in most of the cases.

Stability With Respect to Technical Variations in the Input Data: Number of Profiled Cells, Sequencing Platform, and Cell Type Annotation

In the experiments performed above, we tested the reproducibility of the network inference algorithms by using two independent datasets for each biological condition (e.g., human retina). A limitation of this approach comes from the technical differences between the protocols followed to generate these datasets: different sequencing platforms, different procedures used for the annotation of the cell types, and different number of cells. All these technical differences could impact our results.

To evaluate the stability of the results against technical variations, we used the largest dataset, from Menon et al. (2019), encompassing 20,091 cells. We splitted this dataset into two subsets, keeping the proportions of the various cell types constant. We then applied the five network inference algorithms independently to the two subsets, and we evaluated the reproducibility of the algorithms using perINT and WJS, as in the previous tests. To further assess the effect of the number of cells on network inference, we split the same scRNAseq dataset generated by Menon et al. (2019) three times to obtain

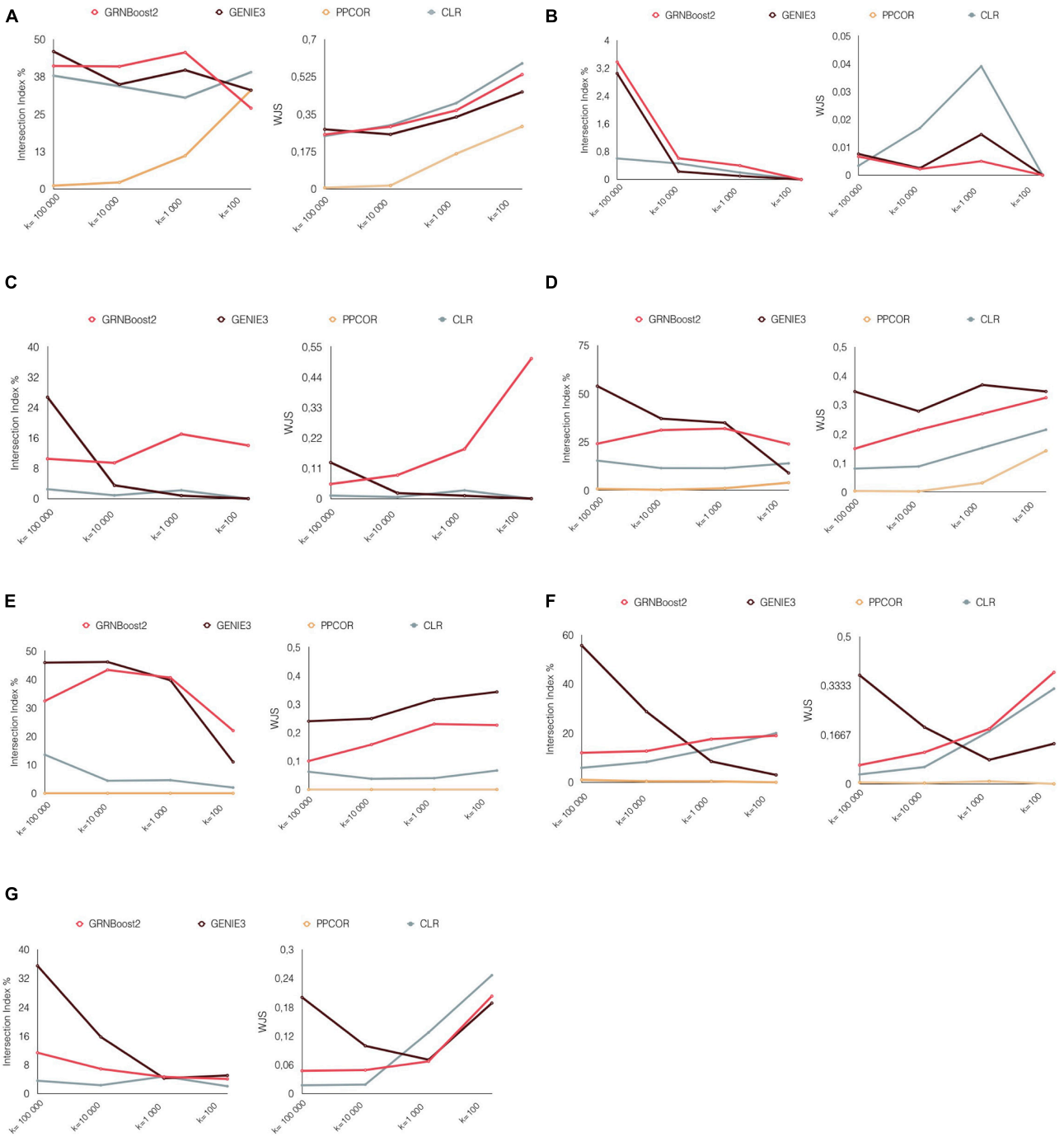

FIGURE 3 | perINT and WJS according to different network thresholding. The perINT and WJS are reported for varying values of the threshold K on the network links: 100,000, 10,000, 1,000, and 100. The results are reported for all the tested datasets **(A)** retina, **(B)** CRC T-cells, **(C)** CLPs, **(D)** Dendritic cells, **(E)** Erythrocytes, **(F)** HSCs, and **(G)** Monocytes.

couples of datasets encompassing decreasing numbers of cells: 10,000, 1,000, and 100. Note that for all these comparisons, the sequencing platform and/or the method/technique used to annotate the cells are identical for all subsets. PPCOR inferred networks for 10,000 and 1,000 cells, but failed at 100 cells (see **Supplementary Table 2**). Details on the number of links before and after thresholding (K = 100,000) are provided in the **Supplementary Table 2**.

Overall, as shown in **Figure 4**, GENIE 3 emerged again as the best performing method in all cases. Of note, for low number of cells, a general decrease in reproducibility is observed for all network inference methods, which can be justified by a

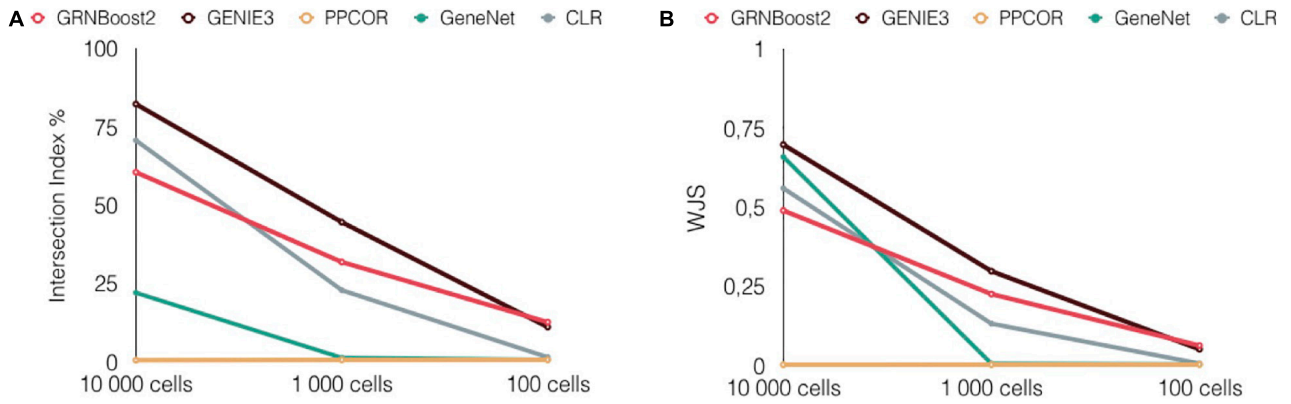

FIGURE 4 | Stability of the network inference performances with respect to technical variations in the input data. Reproducibility scores of GRENBoost2 (red), GENIE3 (black) PPCOR (yellow), CLR (gray) and GeneNet (green) across different splittings of the Menon retina dataset. Panels **(A,B)** report the percentage of intersection (perINT) and Weighted Jaccard Similarity (WJS), respectively.

lower accuracy in the link estimation due to the low number of observations (cells).

The scNET Jupyter Notebook

To foster the reproducibility of all the results and figures presented in this study, we implemented the corresponding code in a Jupyter notebook, available on GitHub, at the url https://github.com/ComputationalSystemsBiology/scNET, together with a Conda package containing all the required libraries. Importantly, scNET can be used to benchmark new network inference algorithms based on their reproducibility, or further test GENIE3, PPCOR, GRNBoost2, CLR, and GeneNet on user-provided datasets.

DISCUSSION

Starting from the benchmark of Pratapa et al. (2020), we evaluated the network inference algorithms from a complementary perspective by assessing their reproducibility. We were interested in assessing weither the algorithms would infer similar networks when applied to pairs of independent datasets from the same biological condition (e.g., T-cells in CRC). Our benchmark focused on real patient-derived data spanning three biological contexts: human retina, T-cells in CRC, and human hematopoiesis cells. We thus considered highly different biological contexts, going from cancer tissue, to isolated healthy immune cells, and to a mixture of normal retina cells combined in a single dataset. Importantly, we aimed at inferring networks involving a much higher number of genes compared to previous works.

In agreement with previous benchmarks, all network inference algorithms generated networks having low intersections with ground-truth. Of note, the ground-truth considered here, based on RcisTarget and regulatory circuits, is different and complementary to those used in previous benchmarks. This disappointing result might arise for different reasons, potentially adding up. Limitations can be present in the input data, as scRNAseq may not provide sufficient resolution for reliable network inference, and technical and experimental factors present in the input data might affect information content. Turning to the inference algorithm, limitations may arise from underlying statistical assumptions and the documented lack of uniqueness in the solution of the network inference problem. Finally, the ground-truth network considered here and in previous benchmarks may not be sufficiently comprehensive.

PPCOR provided weights outside the normal range of correlation values ([-1,1]) for datasets having less than 1,000 cells. Such inconsistencies are likely due to numerical problems arising when the input dataset encompasses many more genes than cells. PIDC was the algorithm that suffered the most when applied to high numbers of genes. Overall, for high link numbers (K = 100,000), GENIE3 consistently generated the most reproducible results across all the three biological contexts considered. Furthermore, its performances proved to be stable with respect to the single-cell sequencing platform, the cell type annotation system and the number of cells considered. Once a more stringent filtering is considered (K = 1,000 or 100), CLR and GRNBoost2 show better performances. However, even the best performing methods show reproducibility scores that are less than ideal (26–54% perINT and 0.1–0.3 WJS), indicating that further improvements are still needed in the design of network inference methods for scRNA-seq data.

We considered network inference methods that are highly heterogeneous. Some algorithms, as PPCOR and GeneNet, infer links between all possible couples of genes, while others, as GENIE3 and GRNBoost2, only infer links between TFs and possible target genes. We tried to make the inferred networks comparable by fixing the number of links in all networks to a certain value K, thus obtaining networks with the same density. However, in principle, methods inferring only TF-target links should have higher chances to be reproducible

in our comparison. At the same time, once the links of PPCOR and GeneNet are restricted to only TF-target links, the dimension of the networks drastically decreases (sometimes empty networks are obtained).

The main limitation of this benchmark is the number of considered network inference algorithms. Future extensions of this study could include pseudotime-based network inference methods, once adequate datasets will become available. To date, available independent datasets relevant for pseudotime-based network inference algorithms (e.g., cells profiled during development stimulation) present too many experimental variations to be employed for a reliable evaluation of reproducibility. Of note, such extensions will be greatly facilitated by taking advantage of the Jupyter notebook (scNET) provided as **Supplementary Material**.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The datasets for this study can be accessed from their associated publications (see **Table 1**). All the analyses are reproducible using the scNET Jupyter notebook available at https://github.com/ComputationalSystemsBiology/scNET.

AUTHOR CONTRIBUTIONS

LC designed the analysis. YK performed the analysis. LC and DT co-supervised the study. All authors contributed to the manuscript and approved the submitted version.

FUNDING

The project leading to this publication has received funding from the Agence Nationale de la Recherche (ANR) – project scMOmix.

ACKNOWLEDGMENTS

We thank the bioinformatics platform of IBENS for the computational support. We thank Michael Mason, Anaïs Baudot, and Sabine Tejpar for the scientific feedbacks on the work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617282/full#supplementary-material

REFERENCES

Aerts, S., Quan, X.-J., Claeys, A., Naval Sanchez, M., Tate, P., Yan, J., et al. (2010). Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in *Drosophila* uncovers a regulatory basis for sensory specification. *PLoS Biol.* 8:e1000435. doi: 10.1371/journal.pbio.1000435

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463

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

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390. doi: 10.1038/ng1532

Chan, T. E., Stumpf, M. P. H., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267.e3. doi: 10.1016/j.cels.2017.08.014

Chawla, K., Tripathi, S., Thommesen, L., Lægreid, A., and Kuiper, M. (2013). TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors. *Bioinformatics* 29, 2519–2520. doi: 10.1093/bioinformatics/btt432

Chen, S., and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* 19:232. doi: 10.1186/s12859-018-2217-z

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-Scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5:e8. doi: 10.1371/journal.pbio.0050008

Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genom.* 17, 246–254. doi: 10.1093/bfgp/elx046

Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. (2010). DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS One* 5:e13397. doi: 10.1371/journal.pone.0013397

Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L., and Salomonis, N. (2018). The human cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* 68, 51–61. doi: 10.1016/j.exphem.2018.09.004

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e12776. doi: 10.1371/journal.pone.0012776

Ideker, T., and Sharan, R. (2008). Protein networks in disease. *Genome Res.* 18, 644–652. doi: 10.1101/gr.071852.107

Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods* 22, 665–674. doi: 10.5351/CSAM.2015.22.6.665

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49, 708–718. doi: 10.1038/ng.3818

Lukowski, S. W., Lo, C. Y., Sharov, A. A., Nguyen, Q., Fang, L., Hung, S. S., et al. (2019). A single−cell transcriptome atlas of the adult human retina. *EMBO J.* 38:e100811. doi: 10.15252/embj.2018100811

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366–370. doi: 10.1038/nmeth.3799

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., et al. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* 33, 2314–2321. doi: 10.1093/bioinformatics/btx194

Menon, M., Mohammadi, S., Davila-Velderrain, J., Goods, B. A., Cadwell, T. D., Xing, Y., et al. (2019). Single-cell transcriptomic atlas of the human retina

identifies cell types associated with age-related macular degeneration. *Nat. Commun.* 10:4902. doi: 10.1038/s41467-019-12780-8

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., et al. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics (Oxford, England)* 35, 2159–2161. doi: 10.1093/bioinformatics/bty916

Opgen-Rhein, R., and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* 1:37. doi: 10.1186/1752-0509-1-37

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17, 147–154. doi: 10.1038/s41592-019-0690-6

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* 37, 451–460. doi: 10.1038/s41587-019-0068-4

Silverman, E. K., Schmidt, H. H. H. W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., et al. (2020). Molecular networks in network medicine: development and applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 12:e1489. doi: 10.1002/wsbm.1489

Sonawane, A. R., Weiss, S. T., Glass, K., and Sharma, A. (2019). Network medicine in the age of biomedical big data. *Front. Genet.* 10:294. doi: 10.3389/fgene.2019.00294

Tantardini, M., Ieva, F., Tajoli, L., and Piccardi, C. (2019). Comparing methods for comparing networks. *Sci. Rep.* 9:17557. doi: 10.1038/s41598-019-53708-y

The DREAM5 Consortium, Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Verny, L., Sella, N., Affeldt, S., Singh, P. P., and Isambert, H. (2017). Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* 13:e1005662. doi: 10.1371/journal.pcbi.1005662

Zhang, Y., Zheng, L., Zhang, L., Hu, X., Ren, X., and Zhang, Z. (2019). Deep single-cell RNA sequencing data of individual T cells from treatment-naïve colorectal cancer patients. *Sci. Data* 6:131. doi: 10.1038/s41597-019-0131-5

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 April 2021
doi: 10.3389/fgene.2021.625246

System-Level Analysis of Alzheimer's Disease Prioritizes Candidate Genes for Neurodegeneration

Jeffrey L. Brabec[1], Montana Kay Lara[1], Anna L. Tyler[2] and J. Matthew Mahoney[1,2] for the Alzheimer's Disease Neuroimaging Initiative[†]*

[1] Department of Neurological Sciences, University of Vermont, Burlington, VT, United States, [2] The Jackson Laboratory, Bar Harbor, ME, United States

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer,
University of Oslo, Norway

Reviewed by:
Annikka Polster,
University of Oslo, Norway
Deborah Weighill,
Harvard University, United States

***Correspondence:**
J. Matthew Mahoney
Matt.Mahoney@jax.org

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

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

Received: 02 November 2020
Accepted: 22 February 2021
Published: 06 April 2021

Citation:
Brabec JL, Lara MK, Tyler AL and Mahoney JM (2021) System-Level Analysis of Alzheimer's Disease Prioritizes Candidate Genes for Neurodegeneration. Front. Genet. 12:625246. doi: 10.3389/fgene.2021.625246

Alzheimer's disease (AD) is a debilitating neurodegenerative disorder. Since the advent of the genome-wide association study (GWAS) we have come to understand much about the genes involved in AD heritability and pathophysiology. Large case-control meta-GWAS studies have increased our ability to prioritize weaker effect alleles, while the recent development of *network-based functional prediction* has provided a mechanism by which we can use machine learning to reprioritize GWAS hits in the functional context of relevant brain tissues like the hippocampus and amygdala. In parallel with these developments, groups like the Alzheimer's Disease Neuroimaging Initiative (ADNI) have compiled rich compendia of AD patient data including genotype and biomarker information, including derived volume measures for relevant structures like the hippocampus and the amygdala. In this study we wanted to identify genes involved in AD-related atrophy of these two structures, which are often critically impaired over the course of the disease. To do this we developed a combined score prioritization method which uses the cumulative distribution function of a gene's functional and positional score, to prioritize top genes that not only segregate with disease status, but also with hippocampal and amygdalar atrophy. Our method identified a mix of genes that had previously been identified in AD GWAS including *APOE*, *TOMM40*, and *NECTIN2*(*PVRL2*) and several others that have not been identified in AD genetic studies, but play integral roles in AD-effected functional pathways including *IQSEC1*, *PFN1*, and *PAK2*. Our findings support the viability of our novel combined score as a method for prioritizing region- and even cell-specific AD risk genes.

Keywords: gene prioritization, machine learning, GWAS, Alzheimer's disease (AD), network-based functional prediction, Alzheimer's Disease Neuroimaging Initiative (ADNI)

INTRODUCTION

The central goal of genome-wide association studies (GWAS) in Alzheimer's disease (AD) is to identify novel candidate genes influencing risk for developing AD. Like other complex disorders, AD has highly polygenic risk, where hundreds or even thousands of small-effect alleles modify the probability of developing AD (Lee et al., 2013; Carmona et al., 2018). Fundamentally, this genetic complexity arises from the underlying biological complexity of AD, where all the major

cell types of the brain and multiple highly differentiated brain structures have established roles in pathogenesis or symptom severity (Calderon-Garcidueñas and Duyckaerts, 2017; Jaroudi et al., 2017). To fully capture this biological complexity for genetic mapping, the international community has undertaken multiple strategies, including *case-control GWAS* and *imaging GWAS*, that capture distinct components of the genetic risk for AD. In particular, case-control GWAS is well powered to detect risk alleles but cannot ascribe these effects to specific brain pathologies. On the other hand, imaging GWAS can localize the effect of alleles, but these studies have limited sample size and, therefore, limited statistical power. In this study, we apply a *network-based gene reprioritization* (NGR) strategy that leverages mature functional prioritization methods to integrate AD risk-gene networks from case-control GWAS with imaging GWAS data to predict genes that specifically influence hippocampal and amygdalar atrophy.

The spectrum of AD risk alleles is well studied, particularly in European populations (Hu et al., 2017; Solomon et al., 2018; Jansen et al., 2019; Rajan et al., 2019; Andrews et al., 2020). Using gold-standard cognitive exams that provide robust *premortem* diagnoses of AD, modern case-control GWAS are powered to detect small-effect alleles using large cohorts. These efforts have culminated most recently in a meta-analysis of AD GWAS assessing the effect of 9,862,738 SNPs in 71,880 cases and 383,378 controls (Jansen et al., 2019). With such large-scale studies, it has been possible to detect 2,357 variants and 29 genes with genome-level significant associations to AD (Jansen et al., 2019). However, increasing population size has diminishing marginal returns. Newly resolved effects are ever weaker. Moreover, the functional role of these alleles cannot be localized to any of the relevant cellular or regional drivers of AD pathology based on case-control status alone. Nevertheless, with a valid AD diagnosis as an endpoint, the alleles mapped in case-control GWAS can be confidently attributed to AD risk.

As an alternative to large case-control studies, the Alzheimer's Disease Neuroimaging Initiative (ADNI) uses structural magnetic resonance imaging (MRI) as a phenotype for GWAS (Wyman et al., 2013). In contrast to cognitive exams, which measure the complex emergent functions of distributed neural circuits, neuroimaging localizes particular structural pathologies. In principle, alleles that have a small overall effect on disease risk could have a comparatively stronger effect on critical pathologies, including hippocampal and amygdalar atrophy, that mediate the genetic risk factors for developing AD. However, MRI is expensive and time-consuming, so the ADNI sample size is limited to the thousands, not hundreds of thousands, of subjects. To date, 2272 patients have been recruited, a subset of 556 of which have both imaging and genotype data (ADNI-1 cohort) (Weiner et al., 2015). This dramatically limits statistical power relative to case-control GWAS. Moreover, while some longitudinal data have been gathered (Bhagwat et al., 2018), it is currently impossible to dissociate background developmental differences in brain structures from pathogenic changes due to AD. Thus, for example, alleles influencing the growth of the hippocampus cannot be distinguished from alleles that exacerbate hippocampal atrophy.

To leverage the independent strengths of case-control and imaging GWAS, we performed an integrative analysis. Using NGR with the well-powered case-control meta-GWAS (Jansen et al., 2019), we identified hippocampus- and amygdala-specific functional networks that were enriched for AD risk genes. We then used a novel approach to combine these functional results with imaging GWAS results for low hippocampal and amygdalar volume in patients with AD. By combining AD specificity from NGR with genetic influences on low hippocampal and amygdalar volume, we can prioritize high-confidence genes for AD-induced hippocampal and amygdalar atrophy.

The key insight to NGR is that the tail of low *p*-values from a GWAS is typically highly enriched for genes in disease-relevant biological processes, independent of whether most of those genes achieve genome-wide significance (Greene et al., 2015). For any choice of statistical cutoff there is a tradeoff between (*a priori* unknown) false positives and false negatives. In particular, genome-wide significance is a conservative threshold that has many false negatives. With a more liberal threshold, one captures more true positives at the cost of more false positives, with no way to discriminate one from the other using GWAS data alone. In order to distinguish likely true positives from false positives, NGR augments the GWAS statistical signals with functional gene-gene interactions. The essential idea of NGR is that true positive genes, by virtue of being functionally related to the disease, are likely to be functionally related to each other. By identifying subnetworks that are enriched for interactions among nominally significant GWAS genes, we can distinguish the likely true positives from spurious associations. Several approaches to NGR have been recently developed, including strategies based on support vector machines (SVM) (Greene et al., 2015), network diffusion (Li and Li, 2012), and Bayesian data integration (Wu et al., 2017). All methods return a *functional score* for every gene in the genome (a *reprioritization*) that measures how strongly each gene interacts with the nominally significant GWAS hits. Using NGR, many groups have shown significant improvements in disease gene prediction (Greene et al., 2015; Wu et al., 2017), including in AD (Song et al., 2016; Yao et al., 2017).

In this study, following Guan et al. (2010), we used an ensemble of SVMs to reprioritize AD risk genes from case-control GWAS using hippocampus- and amygdala-specific functional networks. We then integrated these tissue-specific functional scores with imaging GWAS *p*-values for hippocampal and amygdalar volume. Using a combined score based on the joint cumulative density function of functional scores and imaging GWAS *p*-values, we prioritized candidate genes for hippocampal and amygdalar atrophy in AD and defined the putative AD gene networks in which these candidate genes function.

MATERIALS AND METHODS

Data

We used two distinct GWAS data sets and processed them through separate pipelines (**Figure 1**). The first data set is from the ADNI database and includes genotype and

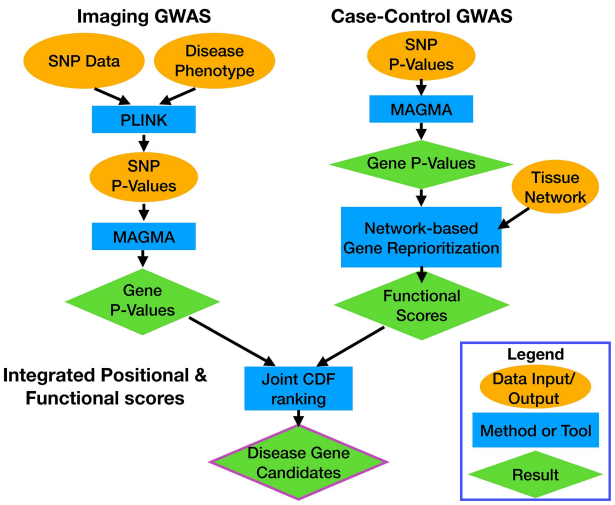

FIGURE 1 | Workflow overview. Each data stream, and the calculation of the integrated score, are indicated by the bolded labels. Each section shows how data (orange ovals) were processed by computational tools (blue rectangles) to get results (green diamonds). Arrows indicate flow of information through the pipeline.

structural MRI imaging data[1]. The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Identifying novel biomarkers of AD will help aid clinicians and researchers develop effective treatments and interventions.

Alzheimer's Disease Neuroimaging Initiative is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols, in addition to the ongoing ADNI-3, have recruited over 2200 adults, ages 55–90, to participate in the research, consisting of control, non-AD (CN) older individuals, people with early or late MCI (EMCI or LMCI), and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. Thousands of longitudinal imaging scans (Jack et al., 2008; Jagust et al., 2010), performance on neuropsychological and clinical assessments (Petersen et al., 2010) and biological samples (Shaw et al., 2009) were collected at baseline and at follow-up visits for all or a subset of participants. Genome-wide genotyping data (Saykin et al., 2010) are available on the full ADNI sample. For up to-date information, see www.adni-info.org.

[1]adni.loni.usc.edu

Freesurfer version 5.1 was used to extract hippocampal volume and amygdalar volume measures from the 1.5 T baseline MRI scans of the ADNI-1 participants as described previously (Risacher et al., 2013). The measurements were retrieved from the ADNI data archive.

Genotype data of all participants from ADNI-1 were downloaded, quality controlled, and imputed to get full coverage beyond the initial 600,000 SNPs available on the Illumina 610Quad platform. Initial QC was performed using PLINK 1.9[2] (Chang et al., 2015). Genotype data were processed as follows: (1) Samples missing more than 10% of their genotype calls were removed (one person removed), (2) SNPs with a minor allele frequency (MAF) greater than 0.05 were filtered for samples missing greater than 5% of the genotype calls and those with an MAF less than 0.05 were filtered for samples missing greater than 1% of genotype calls (48,026 variants removed), (3) duplicated samples were removed (14,238 variants removed), (4) samples that failed Hardy–Weinberg Equilibrium (HWE) ($p < 10^{-7}$) were filtered out (434 variants removed). After QC, we performed genotype imputation using BEAGLE 5.1[3] (Browning et al., 2018). Briefly, genotype data were split by chromosome and each chromosome was mapped onto the appropriate reference genome (hg37) and imputed to the CEU 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) reference panel. Imputed chromosomes were recombined using PLINK 1.9 and underwent an additional round of QC following the procedures listed above (433 variants removed for not meeting HWE). After imputation, 14,403,717 variants and 683 samples passed QC. Hippocampal and amygdalar volumes were used as the phenotypes in two separate GWAS analyses. A total of 556 individuals had both genotyping data and imaging phenotype data (n = 120 AD, n = 261 MCI, n = 175 CN). Genome scans were performed using PLINK 1.9 using a linear regression model with covariates for age, sex, education, and intracranial volume (ICV), following the GWAS protocol of a recent ADNI study using a related network-based gene reprioritization approach (Song et al., 2016).

SNP-level p-values were mapped to gene level p-values using MAGMA[4] (de Leeuw et al., 2015). SNPs were annotated to genes using the hg37 genetic reference and a 10 kb annotation window on either side of the gene. The window size was chosen to match that used for gene mapping the AD meta-GWAS study (Jansen et al., 2019). Of the 14,403,717 SNPs contained within the ADNI genotype data, a total of 6,989,349 SNPs mapped to 18,385 genes. The HV GWAS yielded 338 nominally significant genes and three genes that reached a Bonferroni–Holm corrected, genome-wide significant p-value (**Supplementary File 1**). The AV GWAS yielded 276 nominally significant genes and 1 gene that reached a Bonferroni–Holm corrected genome-wide significant p-value (**Supplementary File 2**).

The second data set we analyzed was the AD meta-GWAS study conducted previously (Jansen et al., 2019). In that study, Jansen et al. (2019) performed a meta-analysis on case-control

[2]https://www.cog-genomics.org/plink2/

[3]http://faculty.washington.edu/browning/beagle/beagle.html

[4]https://ctg.cncr.nl/software/magma

AD data from four major studies including the Alzheimer's disease working group of the Psychiatric Genomics Consortium (PGC-ALZ), the International Genomics of Alzheimer's Project (IGAP), the Alzheimer's Disease Sequencing Project (ADSP), and UK Biobank (UKB). This analysis resulted in 71,880 AD cases and 383,378 non-AD controls and 9,862,738 SNPs passing quality control. SNP associations were calculated by regression as follows:

(1) Logistic regression was used to calculate SNP association with case control phenotypes from ADSP, PGC-ALZ, and IGAP.
(2) Linear regression was used to calculate associations for a continuous phenotype from UKB (calculated as the number of parents with AD).
(3) Associations were adjusted for sex as well as age. However, the ADSP study did not use age as a covariate as the study group was highly enriched for older patients and inclusion of age as a covariate in that study eliminated true AD associations (see Methods: Data Analysis in Jansen et al., 2019).
(4) The first four ancestry principal components (PCs) were also used to adjust statistical associations. A total of 20 were calculated and more were used if they showed a strong association with the phenotype.
(5) For UKB 12 PCs, age, sex, genotyping array, and testing center were all used as covariates.

SNP summary statistics were downloaded from the Center for Neurogenomics and Cognitive Research website: https://ctg.cncr.nl/software/summary_statistics. We used MAGMA to compute gene-level *p*-values as above. Of the 13,367,299 SNPs contained within the meta-GWAS summary statistics, 6,536,525 mapped to a total of 18,456 genes. At a nominal level of significant ($p < 0.01$) the meta-GWAS had 735 significant genes, while a Bonferroni–Holm corrected *p*-value yielded 28 genome-wide significant genes (**Supplementary File 3**).

Network-Based Gene Repositioning

To functionally score every gene in the genome for relevance to AD, we performed NGR. NGR requires two inputs: a set of positive examples of *disease-associated genes*, and a *functional network* encoding gene-gene interactions (*cf.* Greene et al., 2015). From these data, NGR uses the network to propagate the "disease-associated" annotation to genes that are well connected to the disease-associated gene set. In this study, we used nominally significant AD-GWAS genes ($p < 0.01$) from the MAGMA analysis of the meta-GWAS as disease-associated genes. For functional networks, we used the hippocampus and amygdala tissue-specific functional networks freely available for download at HumanBase[5] ('hippocampus_top' and 'amygdala_top') (Wong et al., 2018). Briefly, these networks were generated using a regularized Bayesian knowledge integration based on tissue ontology and a combination of gene expression datasets from the Gene Expression Omnibus (Barrett et al., 2013) representing 20,868 conditions (Greene et al., 2015). Each functional network is a weighted network, where each pair of genes (g_i, g_j) is linked with a weight, $W_{g_ig_j}$, encoding the predicted probability that those genes functionally interact in that tissue. We define a *feature vector*, f_g, for each gene, g, in the genome as the vector of weights connecting g to the n AD-GWAS genes, $p_1, \ldots, p_n$ (i.e., positive examples),

$$f_g = [W_{gp_1}, \ldots, W_{gp_n}].$$

Using these feature vectors, we trained an ensemble of 100 (linear) SVM classifiers to distinguish between AD-GWAS genes and the rest of the genes in the genome. Formally, this problem is an instance of *positive-unlabeled (PU) learning* (PU), as we only have positive examples of AD-relevant genes (i.e., GWAS hits), but the status of all other genes is unknown. In the PU learning setting, we can treat all unlabeled examples as negatives for the sake of training the model, with the understanding that many unlabeled examples are likely AD-associated genes (Elkan and Noto, 2008). For each of the 100 SVMs, we trained using all positive examples and a random, balanced set of unlabeled examples as putative negatives. Each SVM was cross-validated to optimize its cost hyperparameter, C, over a grid, as described previously (Tyler et al., 2019). Each model M_i assigns each gene, g_j, a model-based, real-valued prediction score $M_i(g_j)$, where large positive scores correspond to high confidence that the gene is a positive example and negative scores correspond to low confidence. To normalize prediction scores across models prior to aggregation, we computed an *unlabeled-predicted-positive rate* (UPPR) for each model, M_i, and gene, g_j, as,

$$\mathrm{UPPR_{ij}} = \frac{\#\{g \in Unlabeled \mid M_i(g) > M_i(g_j)\}}{\#\{g \in Unlabeled\} \mid M_i(g) > M_i(g_j)\} + \#\{g \in Unlabeled \mid \mathrm{M_i}(g_j) > M_i(g)\}}$$

where '#' denotes the cardinality of a finite set. The UPPR is the PU-learning equivalent of the false positive rate, where lower values indicate higher confidence that a gene is functionally associated with the AD GWAS genes. We averaged UPPR over all models and took the negative logarithm to obtain a final *functional score*, $FS(g_j)$

$$FS(g_j) = -\log_{10}\left(\frac{1}{100}\sum_{i=1}^{100} UPPR_{ij}\right).$$

The functional score ranges from zero to infinity, with higher values indicating greater confidence. Models were trained using the *e1071* R package (Meyer et al., 2019).

Integrating Functional and Positional Scores

To integrate functional scores for AD-specificity with imaging GWAS *p*-values, we computed a novel *combined score* based on the empirical joint cumulative density function (CDF) of the two scores. Specifically, every gene, g, had a functional score $FS(g)$, and a positional score $PS(g) = -\log_{10}(p_g)$,

[5]https://hb.flatironinstitute.org/download

where p_g is the MAGMA p-value for g in the imaging GWAS. To quantify how highly ranked a gene, g_j, is along both measures simultaneously, we used the value of the empirical joint CDF as a combined score, $CS(g_j)$,

$$\mathrm{CS}\left(\mathrm{g_j}\right) = \frac{\#\{\mathrm{g} \in \mathrm{Genome} |\ \mathrm{FS}\left(\mathrm{g}\right) < \mathrm{FS}\left(\mathrm{g_j}\right)\ \ \&\ \mathrm{PS}\left(\mathrm{g}\right) < \mathrm{PS}\left(\mathrm{g_j}\right)\}}{\mathrm{N}},$$

where N is the number of genes in the genome. Note that this is equivalent to the probabilistic definition using the empirical joint distribution of the two scores. Thus, the combined score represents the probability that a randomly chosen gene in the genome will score lower on both measures than g_j.

Functional Enrichment Analysis

To compare the functional enrichments of ADNI imaging genetics p-values versus the combined scores, we used the g:GOSt tool in the *gprofiler2* R package to identify significantly enriched Gene Ontology terms (Kolberg et al., 2020). Specifically, we ranked all genes by either p-value or combined score and tested the significance of all Gene Ontology Biological Process (GO:BP) terms (Ashburner et al., 2000; Carbon et al., 2018). We then summarized the enriched term lists into high-level annotations using the REVIGO online ontology analysis tool (Supek et al., 2011). Finally, we plotted high-level annotations as pie charts using *ggplot2* (Wickham, 2016).

Modularity and Gene Enrichment Analysis of Functional Networks

To visualize and interpret the outputs of our SVM predictions, we plotted sub-networks of high-ranking genes and performed enrichment analyses of network modules. For both the hippocampal and amygdalar networks, we extracted the sub-networks of genes with functional scores greater than two (i.e., average UPPR < 0.01). We visualized these sub-networks using force-directed layout (Jacomy et al., 2014) in Gephi[6] (Bastian et al., 2009). We identified modules in this sub-network using maximum modularity as implemented in Gephi (Blondel et al., 2008). The list of genes in each module was then sorted by functional score and input to g:GOSt (Raudvere et al., 2019), resulting in significantly enriched Gene Ontology (Ashburner et al., 2000; Carbon et al., 2018), KEGG (Kanehisa and Goto, 2000), and Reactome (Jassal et al., 2019) terms. Network modules were annotated by manually curating a set of representative functional terms, and the full output g:GOSt can be viewed in **Supplementary Files 4, 5**.

Code Availability

To ensure rigor and reproducibility of our results, all analysis code used in this study is freely available at https://github.com/MahoneyLabGroup/AD_NBFP.

[6]https://gephi.org

RESULTS

Hippocampal Volume, Amygdalar Volume, and AD Diagnosis Captured Distinct Genetic Signals

The ADNI-1 dataset contains measures of hippocampal volume (HV) and amygdalar volume (AV) of patients and controls derived from structural MRI, as well as multiple relevant covariates: sex, age, educational attainment, and total ICV. Both of these brain volume measures correlated strongly with a patient's clinical cognitive status (**Figure 2A**). Regional volumes were highest in control, non-AD (CN) subjects, lower in late mild cognitive impairment (LMCI) subjects, and lowest in patients with AD (**Figure 2A**). While there was overlap between the subgroups in HV and AV, the average size of each structure was significantly different between each clinical group (**Figure 2A**), as has been previously shown in prior ADNI work (Schuff et al., 2009; Whitwell et al., 2012).

The hippocampus and amygdala take part in overlapping limbic system neural pathways and are physically close to one another in the temporal lobe, suggesting that atrophy of each of these structures in AD could be highly correlated (Cavedo et al., 2011; Wang et al., 2016). To assess this, we corrected HV and AV for diagnosis at baseline, ICV, years of education, age, and sex using a linear model and computed the correlation of the residuals (**Figure 2B**). The residuals were significantly correlated (R^2 = 0.41, p = 3.2e-66), indicating a significant, but moderate, correlation between the sizes of the two structures. The moderate correlation indicates that there are likely overlapping processes driving the size of these structures, but also biological processes that are unique to each. It is interesting to note that, after controlling for covariates, the distributions of HV and AV are unimodal and do not have any obvious subgroupings. Thus, for the remainder of the study, we treated HV and AV as quantitative traits.

To identify genetic drivers HV and AV in patients with AD, we used PLINK 1.9 (Chang et al., 2015) to statistically associate SNPs to HV and AV, and used MAGMA (de Leeuw et al., 2015) to integrate SNP-level association to gene-level associations (**Figure 1**). Overall, three genes—*APOC1, TOMM40,* and *APOE*—were significant after correcting for multiple comparisons for HV, and one gene—*APOC1*—was significant for AV. Furthermore, 338 and 276 genes were nominally significant at the p = 0.01 level for HV and AV, respectively. The top-ranked genes by p-value for both HV and AV were *APOC1*, *TOMM40*, and *APOE*, which all have well-established associations to AD (Zhou et al., 2014; Chiba-Falek et al., 2018; Zhao et al., 2018). Examining the nominally significant genes, we found that HV and AV independently associated with a unique subset of genes (**Figure 2C**). For example, the gene *GRIN2B*, which plays a role in brain development and is a candidate gene for temporal lobe epilepsy and autism spectrum disorder due to its effects on the hippocampus (Parrish et al., 2013; Varghese et al., 2017), was nominally significant for HV but not AV. Conversely, the gene *EDN1*, which is a candidate gene antagonist for multiple system atrophy (Gu et al., 2018), was

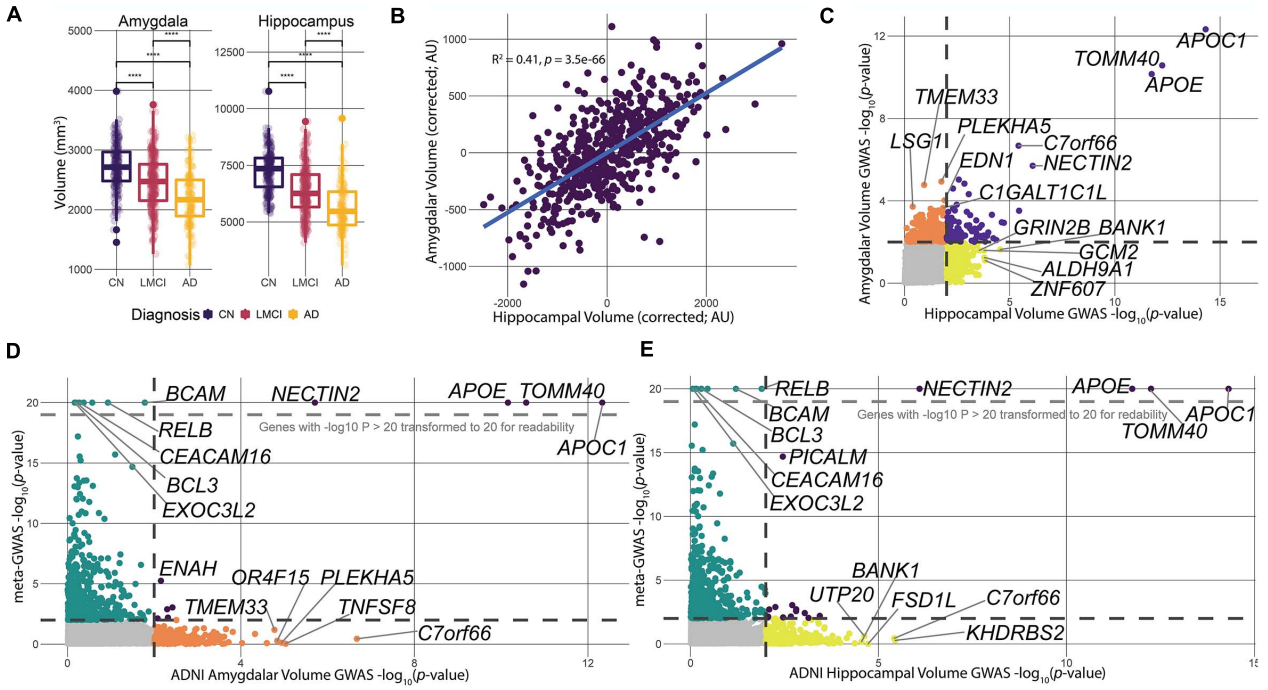

FIGURE 2 | Data processing and comparison of hippocampal and amygdalar volume GWAS and meta-GWAS summary statistics. **(A)** We calculated the correlation between each of the three clinical diagnoses for ADNI-1 and the respective tissue volume measures. All three diagnoses showed significant ($p < 2e\text{-}16$) differences in average volume size across tissues. **(B)** Linear models predicting HV and AV were calculated to determine the correlation in size of the two structures. There was a significant, moderate correlation ($R^2 = 0.41$, $p = 3.6e\text{-}66$) between the volumes of the two structures controlling for diagnosis at baseline, education, sex, ICV, and age. **(C)** Comparison of *p*-value distributions for the two GWAS volume measures. Dotted lines indicate a nominal significance cutoff of 0.01. **(D,E)** Comparison of *p*-value distributions between the meta-GWAS data and the respective volume GWAS data. Genes with log transformed *p*-values greater than 20 were transformed to 20. Black dotted lines indicate a nominal significance cutoff of 0.01.

nominally significant for AV but not HV. These results suggest that large-effect genes may have pleiotropic effects on HV and AV, but also that separate pathways may be driving atrophy in particular structures.

The virtue of endophenotypic measures such as HV and AV is they can potentially resolve biologically specific components of a disease that are otherwise too convoluted with other disease mechanisms when considering disease status alone. However, because the ADNI data are cross-sectional, it is not clear *a priori* whether genetic effects on HV or AV relate to genetic differences in brain developmental or to AD-induced atrophy. To assess the concordance between gene associations for HV and AV associations with AD risk *per se*, we compared gene-level *p*-values for HV and AV to corresponding *p*-values from the AD meta-GWAS study recently published (Jansen et al., 2019) (**Figures 2D,E**). The Jansen et al. (2019) study is the largest AD meta-GWAS to date, and provides the most robust data set to identify any HV- or AV-specific hits influencing AD risk. Like the comparison between HV and AV *p*-values, the meta-GWAS shares several genome-wide significant genes with HV and AV (**Figures 2D,E**). Furthermore, the meta-GWAS shares some nominally significant genes with imaging GWAS, for example, *ENAH* with AV and *PICALM* for HV (**Figures 2D,E**). These overlapping hits, at a nominal significance level, suggest that at least some of the variation in HV and AV is potentially driven by factors influencing genetic AD risk.

NGR Identified Distinct Hippocampal and Amygdalar Functional Gene Networks Connecting AD Risk Genes

As major components of AD pathology, genetic risk factors for AD-induced hippocampal and amygdalar atrophy are expected to be a subset of all AD risk factors. However, differences in sample size (i.e., statistical power) and study population between the case-control and imaging GWAS limit our ability to detect these overlapping associations. Nevertheless, we expect that, beyond specific shared gene associations between HV and AV and disease risk, risk genes for imaging endophenotypes should lie in AD risk gene pathways. To identify the hippocampal and amygdalar pathways involved in AD pathogenesis, we performed NGR using hippocampus- and amygdala-specific functional genomic networks (Wong et al., 2018) to rank every gene in the genome by how well they connect to AD-GWAS genes. Briefly, we trained an ensemble of SVM classifiers to distinguish between AD-GWAS genes and the rest of the genome using connection weights to AD-GWAS genes in the tissue networks as features (see section "Materials and Methods"). The output of this analysis was a ranked list of genes with each gene receiving a *functional score*

(formally, the negative logarithm of the unlabeled-predicted-positive rate) that quantifies how well connected a gene is to AD-GWAS genes. As positive examples we used all genes that reached a nominal level of significance (p = 0.01) in the meta-GWAS dataset (n = 735 genes). The remaining genes were treated as unlabeled.

To aid the interpretation of top functional hits, we visualized the sub-networks of genes that had functional scores greater than 2 for the hippocampus and amygdala networks. We performed modularity analysis in Gephi (Bastian et al., 2009) and identified four modules in both sub-networks (**Figure 3**). We assigned functional annotations to the genes from each network module using g:GOSt (Raudvere et al., 2019). While the number of modules were the same for both tissue sub-networks, the functional annotations underscored distinct pathways. The hippocampus sub-network modules were enriched for genes taking part in *endothelial cell migration* (GO:0043542), *regulation of cell adhesion* (GO:0030155), *Rho/RAS mediated GTPase activity* (GO:0007266, GO:0046578), and *regulation of macroautophagy* (GO:0016241) (**Figure 3A**). The amygdala sub-network modules were enriched for genes involved in *regulation of the ERK signaling cascade and protein ubiquitination* (GO:0070372, GO:0030433), *cytoskeletal and organelle organization* (GO:0051493, GO:0033043), *chromatin and chromosome organization* (GO:0006325), and *apoptotic signaling and cell death* (GO:2001233, GO:0010941) (**Figure 3B**). These enrichments covered a diverse range of processes, some of which overlapped between tissues (e.g., *regulation of macroautophagy* and *apoptotic signaling and cell death*), while others appeared to be tissue-specific (e.g., *endothelial cell migration* in the hippocampus).

Integration of Functional Scores With Imaging GWAS *p*-Values Predicted Risk Genes for AD-Induced Hippocampal and Amygdalar Atrophy

The HV and AV measurements are cross-sectional and cannot resolve whether a genetic association is due to AD-driven atrophy or a genetically encoded difference in brain development. Thus, the genes that associate with HV and AV need not necessarily associate with disease status. In order to identify genes that were simultaneously associated with HV or AV and functionally connected to AD disease risk, we computed a combined score using the joint cumulative density function of the imaging GWAS p-values and the functional scores from NGR. The resulting scores ranged continuously from zero to one, with values closer to one indicating a higher rank on both genetic and functional metrics. Plotting the functional score *vs.* the negative logarithm of the imaging GWAS p-value with a color gradient indicating each gene's combined score, we see that some genes in the upper-right quadrant of the point cloud scored

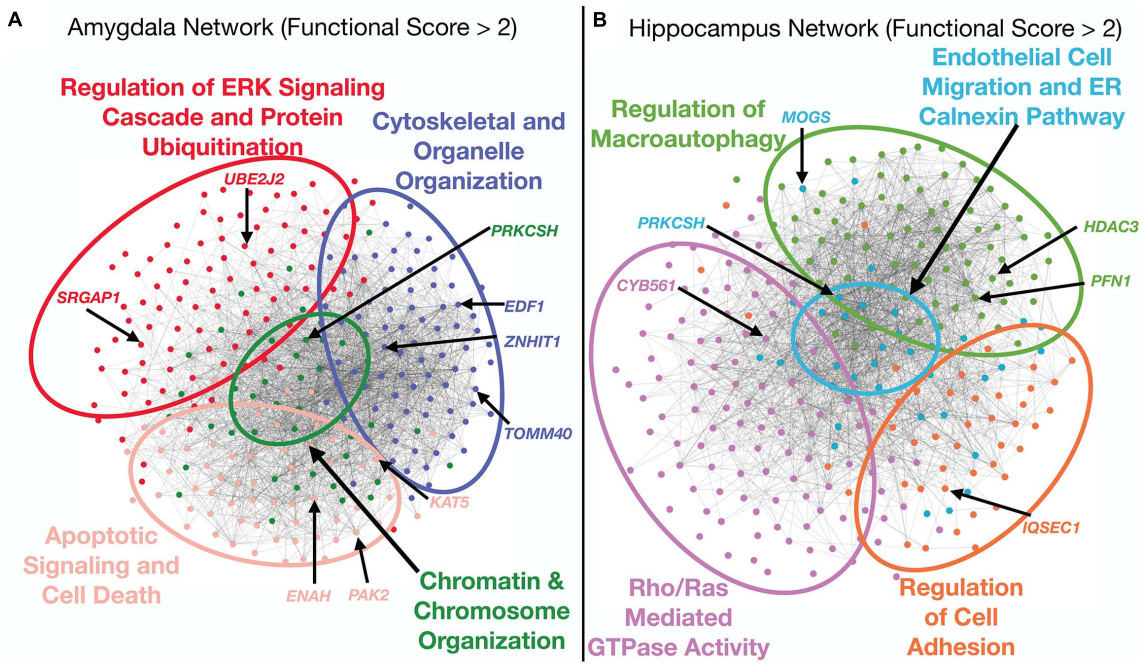

FIGURE 3 | Analysis of top functional subnetworks. The functional subnetwork of genes with a functional scores greater than 2 were extracted from both tissue networks, run through a modularity algorithm, and the modules were functionally annotated using functional enrichment analysis by g:OSt. Several of the top hit genes from the combined ranking appeared in a diverse array of functional classifications. Each gene is colored by the functional module in which it is a member. Network edges were filtered to only include weights greater than 0.25 for visual clarity. **(A)** Amygdala sub-network analysis. The top functional sub-network for this tissue was enriched for genes in pathways that regulate apoptosis and cell death, cytoskeletal and organelle organization and chromosomal organization. **(B)** Hippocampus sub-network analysis. This top functional sub-network was enriched for genes involved in immune signaling as well as cell adhesion and ER regulation.

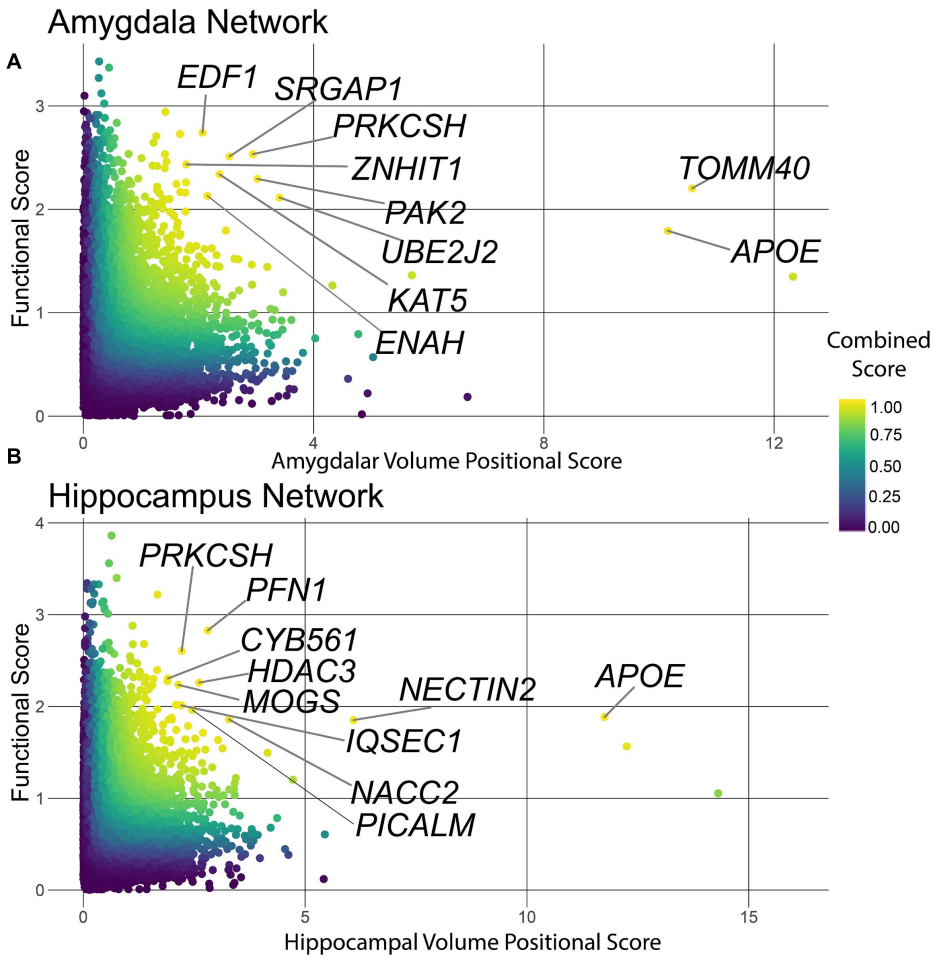

FIGURE 4 | Combined score ranking. Points are colored on a gradient by combined score with yellow points scoring highest, and blue points scoring lowest. **(A)** The combined score plot for the amygdala tissue. Several of the top ranked genes were involved in the regulation gene transcription (*EDF1*) or the maintenance of organelles (*PRKCSH, UBE2J2*) and integrity of the synapse (*PAK2, ENAH*). **(B)** The combined score plot for the hippocampus tissue. Several genes were involved in processes required for the maintenance of the synapse (*PFN1, IQSEC1*) regulation of gene transcription (*HDAC3*) and proper ER regulation (*PRKCSH, MOGS*).

better than 95% of the genes in the genome on both axes (**Figures 4A,B**).

The purpose of the combined score was to prioritize AD-specific genes and distinguish them from genes influencing HV and AV through developmental pathways. To establish a specific enrichment for AD-relevant pathways, we compared functional enrichments between ranking genes by *p*-value (ascending) and by combined score (descending). To summarize the large lists of enriched terms, we used REVIGO to compress the enrichments into representative high-level terms (Supek et al., 2011). For the hippocampus and amygdala (**Figure 5**), the *p*-value analyses revealed an enrichment for genes involved in cholesterol metabolism and cell adhesion. On the other hand, the combined score in the hippocampus was enriched for terms involved in the regulation of the immune response and cellular stress related to the endoplasmic reticulum (ER). Similarly, for the amygdala, the combined score was enriched for pathways involved in ER stress and neuron growth. These results demonstrate that the combined score prioritizes genes involved in AD-relevant functional pathways, distinct from those regulated by *APOE* (e.g., cholesterol metabolism) (Schliebs and Arendt, 2011; Heneka et al., 2015; Gerakis and Hetz, 2018).

Notably, while the combined score ranked genes involved in AD-relevant pathways highly, many of the top-10 genes have not been previously annotated to the disease (**Tables 1**, **2**). High-scoring hippocampal genes are involved in actin regulation (*PFN1*, *IQSEC1*, *PAK2*), protein regulation in the ER (*MOGS* and *PRKCSH*), and transcriptional regulation (*HDAC3*). Highly ranked genes in the amygdala are involved in a wide range of processes, including regulation of proteins in the ER (*PRKCSH* and *UBE2J2*), transcription modification or cell cycle modulation (*KAT5*, *EDF1*, and *ZNHIT1*), and the maintenance and development of healthy synapses (*SRGAP1* and *PAK2*). The top 10 genes in both the hippocampus and amygdala were distributed throughout the NGR functional networks and were present in all functional modules (**Figure 3**).

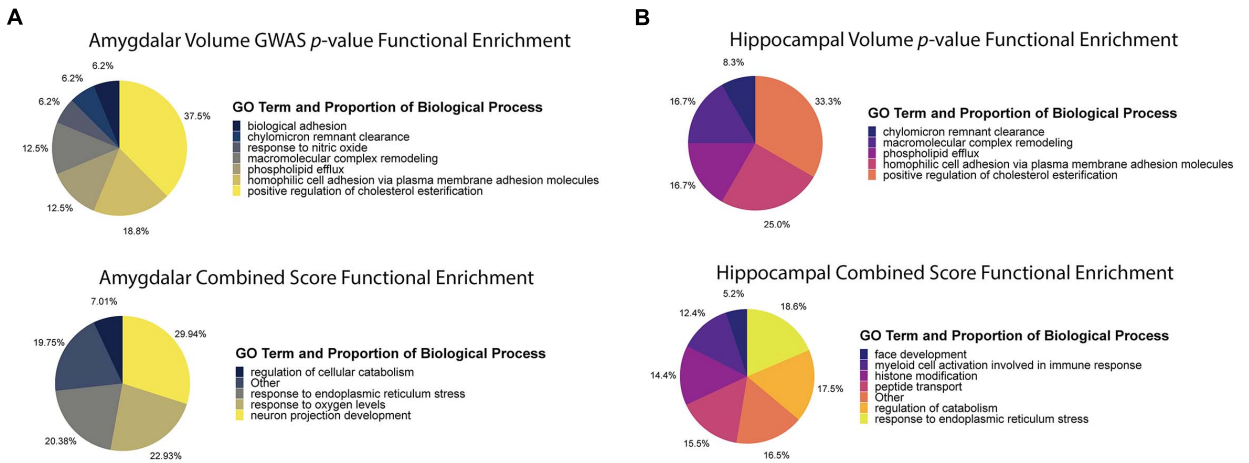

FIGURE 5 | Functional enrichment analysis using *p*-value and combined score. **(A)** Functional enrichment analysis of amygdala volume GWAS *p*-value ranking and combined score rankings. The *p*-value functional enrichment analysis revealed terms like biological cell adhesion and macromolecular complex remodeling. By comparison, the functional enrichment analysis for the combined score revealed terms relating to ER stress, neuron projection development, and response to oxygen levels which are all pathways affected by AD pathophysiology. **(B)** Functional enrichment analysis of hippocampal volume GWAS *p*-value ranking and combined score rankings. The *p*-value functional enrichment analysis saw similar enrichments as the amygdala *p*-value functional enrichment analysis. The combined score functional enrichment analysis on the other hand, saw enrichment of pathways involved in development, ER stress, and immune response regulation.

DISCUSSION

As a complex disease, the genetic risk for AD is distributed over a wide variety of cellular and molecular pathways. Thus, the genetic architecture of AD is expected to be dominated by thousands of small-effect variants that each slightly perturb brain physiology toward a more AD-susceptible state, rather than a small set of highly penetrant mutations. Indeed, even the well-studied *APOE-E4* risk allele has an odds ratio of only 11.8 in the Caucasian population, which is by no means a certainty for any carrier (Jia et al., 2020). The value of genetic network analysis to the study of the architecture of complex disease, therefore, is to aggregate these many small perturbations into a pathway- and process-level description of the full disease. To this end, our results clearly implicate common mutations in many genes as perturbations of pathways that react to the aberrant accumulation of Aβ in the brain (**Figure 6**; discussed below). Far from being statistical noise, genes with nominally significant *p*-value from the imaging GWAS are enriched for AD-specific biology. Interestingly, the gene-level *p*-values largely did not replicate between imaging GWAS and the case-control meta-GWAS. It was only after identifying the relevant tissue-specific functional sub-networks with NGR that we could resolve the likely AD-specific genes for HV and AV. Validating any of these high-ranking genes as specifically influencing hippocampal or amygdalar atrophy is beyond the scope of this study, but many top hits have strong connections to well-established AD biology.

The pathognomonic signature of AD is the aggregation of amyloid β (Aβ) peptide into amyloid plaques in the brain. Beyond aggregating into plaques, however, Aβ is associated with a number of pathological processes, including loss of synaptic integrity (Rönicke et al., 2011; Parsons and Raymond, 2014; Wang and Reddy, 2016; Singh et al., 2017; Kang and Woo, 2019; Schaeverbeke et al., 2019) and dysregulating neuronal and astrocytic calcium channels (Yu et al., 2005; Rönicke et al., 2011; Parsons and Raymond, 2014; Lim et al., 2016; Wang and Reddy, 2016; Verkhratsky et al., 2017; Liu et al., 2019). At the astrocyte, Aβ has been shown to bind Alpha-7 nicotinic acetylcholine receptors (α7 nAChRs), causing an influx of calcium to the astrocyte and glutamate release into the synapse (Pirttimaki et al., 2013). At the synapse, Aβ has been shown to bind to *N*-methyl-D-aspartate receptors (NMDARs) preventing glutamate from activating the channel to allow an influx of calcium ions (Liu et al., 2019). Loss of current through NMDARs drives depression of synaptic strength at that synapse, as lower levels of calcium initially drive the endocytosis of α-amino-3-hydroxy-5-methyl-4-isoxasolepropionic acid receptors (AMPARs) and later NMDARs in the postsynaptic neuron (Tigaret et al., 2006; Yu et al., 2010). Loss of synaptic efficacy is a critical signal for synaptic pruning (Lüscher and Malenka, 2012), and an accumulated loss of synapses is one possible mechanism for loss of network function. Beyond synaptic pruning, Aβ is associated with a loss of synaptic integrity, where the neurotransmitters, such as glutamate, can leak out of the synapse and activate extra-synaptic receptors (Hardingham and Bading, 2010; Parsons and Raymond, 2014). It has been hypothesized that the high level of glutamate release by astrocytes leads to an increase in extra-synaptic glutamate signaling and excitotoxicity (Sattler et al., 2000; Hardingham and Bading, 2010; Parsons and Raymond, 2014; Wang and Reddy, 2016), which is hypothesized to both induce ER stress (Sokka et al., 2007; Concannon et al., 2008) and activate pro-apoptotic pathways (Hardingham et al., 2002), while antagonizing pro-survival pathways, particularly brain-derived neurotrophic factor (BDNF) signaling, leading to neuron death (Hardingham et al., 2002; Hardingham and Bading, 2010; Parsons and Raymond, 2014; Wang and Reddy, 2016). Thus, the accumulation of Aβ

TABLE 1 | Brief descriptions of the top genes according to the combined score for the hippocampus.

Gene	Functional Score	*p*-value	Role (with PMID)
PFN1	0.00148	1.56E-03	Increased actin depolymerization in hippocampus of APP/PS1 mice indicates impaired synaptic plasticity (PMID: 31472195). Actin remodeling mediated by SGK1, a gene involved in spatial memory formation and consolidation (PMID: 31981651). Critical for proper PNS myelination, organization, and development (PMID: 24598164).
HDAC3	0.00549	2.44E-03	Nuclear HDAC3 is significantly increased in the hippocampus of 6- and 9-month-old APP/PS1 mice compared with age-matched wild-type C57BL/6 mice. Inhibition of HDAC3 in the hippocampus attenuated spatial memory deficits, and decreased amyloid plaque load and ABeta levels. Dendiritic spine density increased while microglial activation alleviated after HDAC3 inhibition. Over expression led to an increase in hipppocampal feels of Abeta, activation of microglia, and decreased dendritic spine density (PMID: 28771976).
PRKCSH	0.00249	6.06E-03	Colocalizes with IP3Rs which mediate calcium release from the ER, specifically in hippocampal neurons. Additionally, *PRKCSH* enhances IP3-induced calcium release and has been found to regulate ATP-induced CA2(+ (PMID: 18990696).
APOE (29107063)	0.0130	1.78E-12	Lipid transporter that binds to cell-surface receptors to aid in cholesterol transport and membrane homeostasis. It is present in a broad range of functional pathways within the CNS including synaptic plasticity, mitochondrial function, and neuroinflammation. Its epsilon 4 allele is one of the biggest risk factors for AD (PMID: 28434655).
MOGS	0.00581	7.21E-03	Located in the lumen of the ER where it performs N-linked glycosylation. Several mutations within the gene can lead to congenital diseases of glycosylation which can lead to major structural malformations within the brain, liver, lungs, and many other higher-order tissues and organs (PMID: 30587846).
NECTIN2 (29107063)	0.0141	8.12E-07	Also known as *PVRL2,* this gene is a component protein of adherens junctions between cells. Has wide ranging roles in cell signaling to natural killer cells to leukocyte transport in endothelial cells (PMID: 28062492).
PICALM (19734902)	0.0109	3.52E-03	Involved in clathrin assembly. Two SNPs 5′ to the gene are associated with Reduced LOAD Risk (PMID: 19734902; 24162737; 19734903), but their functions have not yet been determined. It colocalizes with APP and over-expression of PICALM *in vivo* increases plaque deposition in AD transgenic mice (PMID: 22539346). Binds to autophagosomes, suggesting a role in autophagy mediated Abeta clearance (PMID: 24067654).
NACC2	0.0139	5.19E-04	Transcription repressor within the p53 pathway: inhibits the expression of MDM2 which stabilizes the expression of p53 an important tumor suppressor (PMID: 22926524).
IQSEC1	0.00974	6.14E-03	Loss of function affects a wide variety of actin-dependent cellular processes, including AMPA and NMDA receptor trafficking at synapses (PMID: 20547133). Mutations have led to intellectual disability and developmental delays in those affected (PMID: 31607425).
CYB561	0.00496	1.24E-02	An electron transporter critical for the conversion of dopamine to epinephrine and norepinephrine. A mutation in this gene, which disrupts the final production of norepinephrine, has been observed in families with severe orthostatic hypotension (PMID: 29343526).

Genes in bold have been previously found in AD GWAS. PMIDs from supporting papers are included in parentheses next to bolded gene names.

acts through multiple complex pathways—at the synapse, at the ER, and through transcriptional regulation—to cause atrophy of neural tissue. Importantly, our top-ranking genes in both the hippocampus and the amygdala act in these Aβ-response pathways.

Multiple High-Ranking Genes Influence Synaptic Structure Through the Cytoskeleton

Altered synaptic structure and function are well-established in AD (Spires-Jones and Knafo, 2012; Pozueta et al., 2013; Chabrier et al., 2014; Price et al., 2014; Mango et al., 2019; Koller and Chakrabarty, 2020). The highest-ranking hippocampal gene, *PFN1* (**Figure 6A** and **Table 1**), encodes an actin-monomer binding protein that is known to regulate the cytoskeleton of neurites (Murk et al., 2012), but has also been shown to support the highly mobile *F*-actin in astrocytic projections that surround synaptic clefts (Schweinhuber et al., 2015). It has been associated with impaired synaptic plasticity and spatial memory in the *APP*/*PS1* mouse model of AD (Sun et al., 2019; Lian et al., 2020). Alterations to the function of *PFN1* due to AD risk mutations could account for alterations in synaptic maintenance, leading to increased glutamate signaling to extra-synaptic NMDARs. *PFN1* activity is promoted by BDNF, which is hypothesized to be inhibited by extrasynaptic glutamate signaling, and loss of that signal could stop proper formation of actin at neurite outgrowths and potentially in astrocytic processes supporting synaptic clefts

TABLE 2 | Brief descriptions of the top genes according to the combined score for the amygdala.

Gene	Functional Score	*p*-value	Role (with PMID)
PRKCSH	0.0293	1.13E-03	Colocalizes with IP3Rs which mediate calcium release from the ER, specifically in hippocampal neurons. Additionally, *PRKCSH* enhances IP3-induced calcium release and has been found to regulate ATP-induced CA2+ (PMID: 18990696).
TOMM40 (29107063)	0.00626	2.66E-11	Mitochondrial membrane protein critical for transport of protein precursors into the mitochondria and is associated with mitochondrial dysfunction in AD. Further, it has recently been found to be associated with functional connectivity of brain regions via fMRI (PMID: 31568198). It is in LD with APOE.
PAK2	0.00508	9.50E-04	Haploinsufficiency of PAK2 has been observed to decrease synapse density, impair LTP, and drive autism related behaviors in mice (PMID: 30134165). Strong regulator of cellular senescence and organismal aging through gene-expression and the H3.3 nucleosome assembly (PMID: 31209047).
SRGAP1	0.00308	2.89E-03	A GTPase activator that works with CDC42 to negatively regulate neuronal migration. Interacts with ROBO1 to inactivate CDC42 (PMID: 11672528).
UBE2J2	0.00771	2.88E-04	Ubiquitination by this protein is a potential mechanism for endoplasmic reticulum-associated depredation (ERAD) (PMID: 19951915; 25083800).
KAT5	0.00459	4.31E-03	A histone acetyl transferase (HAT) that plays a role in DNA repair and apoptosis as well as signal transduction. Complexes with the intracellular domain of the cleaved APP products to form nuclear spheres which seem to have a role in cell-cycle regulation, but are not well understood (PMID: 27644079).
EDF1	0.00181	8.50E-03	Transcriptional regulator of PPAR-gamma which has a wide array of roles in combatting AD pathophysiology including amyloid clearance and metabolic regulation (PMID: 22109891, 24838579).
ENAH	0.00740	6.99E-03	Complexes with FE65 and that association may have an effect on APP biogenesis (PMID: 9407065). Also involved in actin polymerization and cell motility (PMID: 10069337, 10892743).
ZNHIT1	0.00368	1.63E-02	Induces arrest of cell cycle at G1 and CDK6 was strongly down-regulated by Znhit1 through transcriptional repression (PMID: 19501046). CDK6 is unregulated in patients in AD compared to non-AD controls (PMID: 26766955).
APOE (29107063)	0.0162	7.00E-11	Lipid transporter that binds to cell-surface receptors to aid in cholesterol transport and membrane homeostasis. It is present in a broad range of functional pathways within the CNS including synaptic plasticity, mitochondrial function, and neuroinflammation. Its epsilon 4 allele is one of the biggest risk factors for AD (PMID: 28434655).

Genes in bold have been previously found in AD GWAS. PMIDs from supporting papers are included in parentheses next to bolded gene names.

(Murk et al., 2012; Parsons and Raymond, 2014; Schweinhuber et al., 2015).

Another high-ranking hippocampal gene was *IQSEC1* (also known as *BRAG2*), which encodes a guanine nucleotide exchange factor, ARF-GEF$_{100}$, that is critical for the proper maintenance of excitatory synapses through AMPA and NMDA receptor trafficking, and regulating synaptic long-term depression (Ottis et al., 2013; Elagabani et al., 2016; Um, 2017; Ansar et al., 2019) (**Figure 6B** and **Table 1**). Loss of function mutations in *IQSEC1* have been associated with intellectual disability (Elagabani et al., 2016) and a biallelic variant mutation has been observed in two families exhibiting intellectual disability and developmental delays (Ansar et al., 2019). A recent study in Wistar rats found that *BRAG2* is a member of a small network of proteins that are dysregulated in response to age-induced changes in proteostasis (Ottis et al., 2013). Significantly, changes in this protein network lead to impaired learning and memory performance (Ottis et al., 2013). Thus, common variants in *IQSEC1* could play a role in synaptic reorganization in response to aging and Aβ burden in AD.

The highest scoring amygdala gene was *PAK2*, which supports actin formation and the promotion of dendritic spine formation (Bokoch, 2003; Shin et al., 2009; Wang et al., 2018) (**Figure 6C** and **Table 2**). Mutations in *PAK2* are associated with other neurological disorders, including autism spectrum disorder and a 3q29 microdeletion syndrome with a range of neurological phenotypes including intellectual disability and autism (Wang et al., 2018). PAK family proteins have been associated with impaired dendritic spine formation in *in vitro* AD models (Ma et al., 2008), and PAK2 has been shown to be cleaved by caspase resulting in cell death (Marlin et al., 2011). Recent work has also shown that LIMK1, a downstream signaling molecule from *PAK2*, is involved in a ROCK2 actin regulatory pathway which mediates Aβ42-induced spine degeneration as well as neuronal hyperexcitability in hAPP mice (Henderson et al., 2019). *PAK2* activity is regulated by the Slit/roundabout (ROBO) signaling pathway (Dubrac et al., 2016; Xu et al., 2018), which is primarily involved in modulating axonal guidance and neuronal migration (Dickson and Gilestro, 2006; Mastick et al., 2010; Slováková et al., 2012), via the CDC42 GTPase (Wong et al., 2001; Xu et al., 2018; Huang et al., 2020). Another high-ranking amygdala gene, *SRGAP1*, suppresses the activity of *PAK2* through the Slit/ROBO signaling pathway (Dubrac et al., 2016; Xu et al., 2018) (**Figure 6C** and **Table 2**). Slit binds to ROBO and activates the SRGAP1

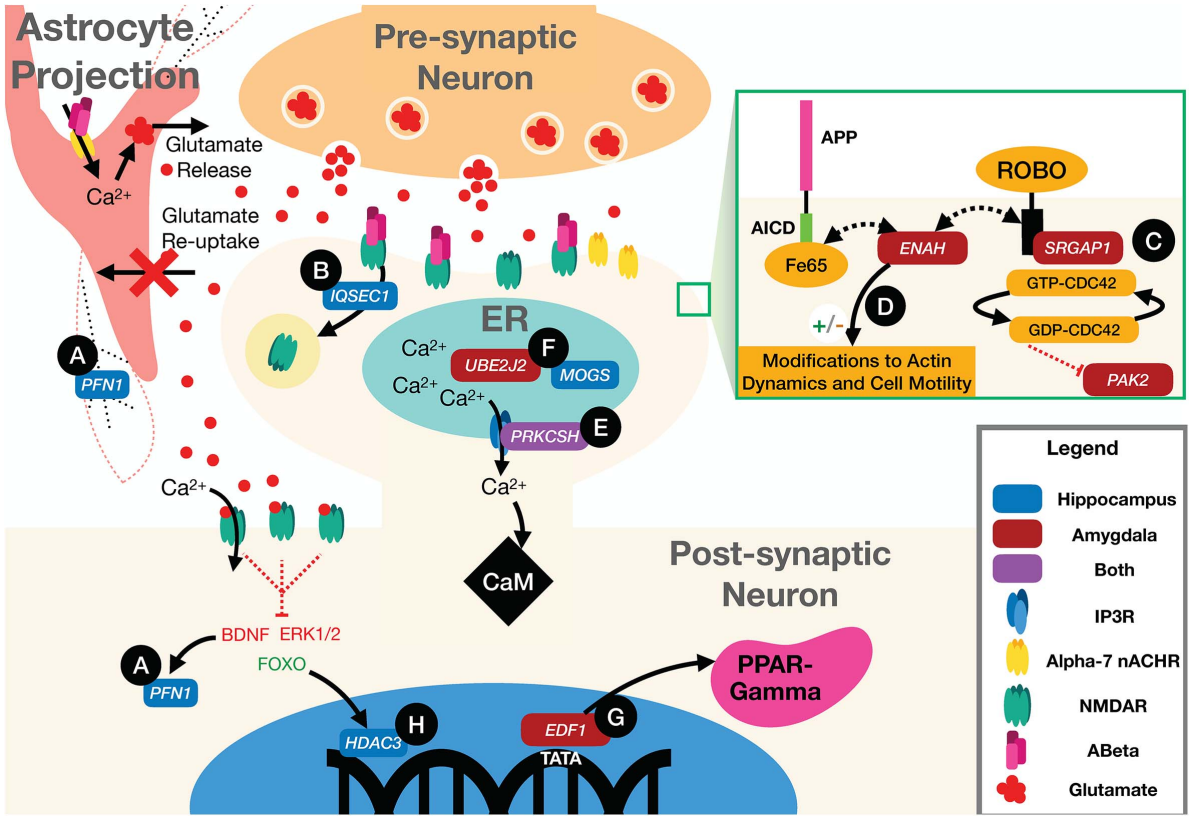

FIGURE 6 | Illustration of neuronal and astrocytic pathways implicated in the disease according to the top scoring genes according to our integrated ranking. **(A)** Loss of signal from BDNF to *PFN1* downregulates the activity of the gene, impairing F-actin structure in the neuron and astrocyte. **(B)** *IQSEC1* mediates AMPAR and NMDAR receptor internalization. **(C)** *PAK2* activity is partially regulated by the CDC42 GTPase which is activated by binding of the SLIT protein to ROBO at the membrane. This activity can be interrupted by the activity of *SRGAP1* which inhibits ROBO signaling. **(D)** *ENAH* can bind to both ROBO and Fe65. When bound to ROBO it acts to inhibit actin polymerization and motility, but while bound to Fe65 and the AICD the ROBO pathway functions normally, promoting cell motility, and actin polymerization. **(E)** *PRKCSH* co-localizes with IP3Rs increasing calcium current through the channels, increasing cytosolic levels of Ca^{2+}. **(F)** *UBE2J2* and *MOGS* are both involved in the proper regulation of intra-ER processes. Impairment of the proper activity of post-translational modification by *MOGS* could drive ERAD mediated by ubiquitination by *UBE2J2*. **(G)** *EDF1* is also a transcriptional regulator for PPAR-gamma. Activation of PPAR-gamma helps to regulate disturbed metabolic states and A plaque clearance. **(H)** *HDAC3* regulates transcription of many genes that have an effect on Aβ burden, microglial activation, and dendritic spine density.

protein which triggers the hydrolysis of GTP by the CDC42 GTPase, which attenuates *PAK2* activity (Dubrac et al., 2016; Feng et al., 2016). Thus, common variants that modify the activity of *PAK2* or its upstream regulator, *SRGAP1*, could lead to alterations in synaptic morphology and axonal migration, and possibly to cleaved PAK2 signaling for neuronal death.

A final cytoskeletal protein among the top-rankings genes was *ENAH* in the amygdala. The ENAH protein has been found to form a complex with Fe65, a transcriptional activator and protein involved in neurite outgrowth and binding partner of amyloid precursor protein (APP) (Sabo et al., 2001; Li et al., 2018) (**Figure 6D** and **Table 2**). ENAH also binds to ROBO and profilin (PFN), acting as an inhibitor of motility and regulator of actin dynamics, respectively (Gertler et al., 1996; Lanier et al., 1999; Bear et al., 2000; Lanier and Gertler, 2000). Greater association of ENAH with the Fe65-APP complex supports neurite outgrowth and motility, whereas binding to ROBO inhibits that activity (Sabo et al., 2001). Common variants in *ENAH*, therefore, could influence synaptic plasticity through its association with the major AD risk factor *APP* (Trillaud-Doppia and Boehm, 2018).

PRKCSH Potentially Regulates Excitotoxicity in AD

Loss of synaptic integrity coupled with impaired glutamate clearance by astrocytes caused by Aβ leads to high levels of extracellular glutamate, which binds to NMDARs increasing intracellular calcium levels (Parsons and Raymond, 2014; Liu et al., 2019). Under physiological conditions, the ER and other organelles act as calcium sinks that modulate intracellular ion levels.

Excitotoxicity occurs when intracellular calcium levels exceed the buffering capacity of the cell. The only top-ten gene shared by both tissues, aside from *APOE*, was *PRKCSH* (**Tables 1**, **2**), which encodes the protein kinase C substrate 80K-H (80K-H), a glucosidase enzyme in the ER. 80K-H is known to colocalize with the inositol triphosphate receptor (IP3R), an ER-resident calcium

channel that facilitates calcium currents in the ER (Kawaai et al., 2009) (**Figure 6E**). Common variants in *PRKCSH* could modify neuronal responses to excitotoxic levels of calcium, potentially exacerbating tissue atrophy in the hippocampus and amygdala.

ER Stress and Misfolded Protein Response Genes Could Contribute to Apoptotic Signaling

Several other high-ranking genes are integral to the proper folding of proteins in the ER. ER stress occurs when the ability of the ER to properly fold proteins becomes saturated (Lin et al., 2007). The hippocampal gene *MOGS* encodes a glycosylation enzyme that aids in protein folding (Sadat et al., 2014; Li et al., 2019) (**Figure 6F** and **Table 1**). Common variants in *MOGS* could modify the rate at which ER stress occurs and exacerbate AD-related hippocampal atrophy.

When the ER reaches a critical state of misfolded proteins, ER-associated degradation (ERAD) can be triggered. ERAD is a process by which misfolded proteins are ubiquitinated and then proteolyzed to prevent the misfolded polymers from causing cellular damage. The amygdalar gene *UBE2J2* encodes a ubiquitin conjugating enzyme that marks misfolded proteins for degradation (Wang et al., 2009; Glaeser et al., 2018) (**Figure 6F** and **Table 2**). In some cases, ERAD can be triggered as part of apoptosis, and ubiquitination enzymes, including UBE2J2, are recruited to ubiquitinate misfolded proteins (Glaeser et al., 2018). Common variants in *UBE2J2* could affect the misfolded protein response and exacerbate cellular damage due to misfolded proteins.

High Ranking Transcriptional Regulators Could Have Pleiotropic Effects on AD

A final set of high-ranking genes was broadly involved in transcriptional regulation. The high-ranking amygdala gene *EDF1* encodes a factor that acts as a transcriptional coactivator of peroxisome proliferator-activated receptor-gamma (PPARγ) (**Figure 6G** and **Table 2**). PPARγ has multiple functions, including regulating metabolism (Pipatpiboon et al., 2012), supporting vascular endothelial cells (Cazzaniga et al., 2018), and promoting BDNF expression (d'Angelo et al., 2019). It has been hypothesized that PPARγ counteracts insulin resistance and metabolic dysfunction in AD (Hoyer and Lannert, 1999; Pipatpiboon et al., 2012). It potentially also plays a role in modifying extracellular Aβ levels by facilitating increased uptake of Aβ by neurons and glia (Mandrekar-Colucci et al., 2012). PPARγ also downregulates the pro-inflammatory mechanisms of AD pathology (Combs et al., 2000; Govindarajulu et al., 2018). Common variants within the *EDF1* gene could have pleiotropic effects on cellular function through the regulation of PPARγ.

The hippocampal gene *HDAC3* encodes a histone deacetylase enzyme that epigenetically regulates gene expression (McQuown and Wood, 2011; Nott et al., 2016) (**Figure 6H** and **Table 1**). Extra-synaptic glutamate signaling drives pro-apoptotic gene expression, in part through the FOXO transcription factor, which is upregulated by extra-synaptic signaling (Parsons and Raymond, 2014). FOXO forms a complex with HDAC3, the protein product of *HDAC3*, and suppresses gene transcription (Nott et al., 2016). Thus, common variants in *HDAC3* could influence pro-apoptotic gene expression, exacerbating hippocampal atrophy.

HDAC3, and other members of the HDAC family, also negatively regulate long-term memory formation (McQuown and Wood, 2011; Zhu et al., 2017), via the "molecular brake pad hypothesis" (McQuown and Wood, 2011). The molecular brake pad hypothesis posits that the tight binding of HDACs to the promoters of genes that drive memory formation requires high-levels of activity-dependent signaling to dissociate them and enable protein synthesis-dependent long-term memory formation (McQuown and Wood, 2011). Notably, *HDAC3* has also been found to affect dendritic spine density, amyloid burden, microglial activation, and spatial memory in the *APP/PS1* AD mouse model (Zhu et al., 2017). Furthermore, in the 3xTG-AD mouse model, inhibition of HDAC3 reversed AD-related pathologies (Janczura et al., 2018), and in cultured rat hippocampal neurons, inhibition of HDAC3 reversed Aβ-induced plasticity deficits (Krishna et al., 2016). Interestingly, another histone deacetylase inhibitor, HDAC2, is emerging as a potential drug target in AD (Choubey and Jeyakanthan, 2018). Together, these results suggest pleiotropic roles for *HDAC3* as a gene influencing hippocampal atrophy in AD.

In summary, the genes prioritized by our integrative method are robustly related to AD by prior research and have clear pathways connecting them to neuron death, and therefore, to the imaging signals of low HV and AV.

The present study was potentially limited by a number of important factors. First, by treating HV and AV independently as quantitative traits, we potentially miss important population substructure (e.g., discrete patient subgroups with extreme neuropathology). While we do not see obvious subgroups in the HV/AV data (**Figure 2**), it is possible that by paring MRI with other phenotypic measures, such groups could appear. Future multi-trait analyses could have greater power to detect risk factors for patient subgroups, such as those that have been detected in gene expression data (Mukherjee et al., 2020). In particular, with emerging longitudinal data, it may become possible to identify subgroups that have distinct disease trajectories. Second, we have applied an NGR method that has been extensively tested, applied, and validated (Guan et al., 2010; Gorenshteyn et al., 2015; Goya et al., 2015; Greene et al., 2015; Krishnan et al., 2016; Song et al., 2016; Yao et al., 2018; Tyler et al., 2019). However, NGR methods are under active development, with new variants using different machine learning strategies or molecular networks. Future work can benchmark different NGR strategies prior to our integrative prioritization to identify the most robust combination of molecular network and learning algorithm for AD GWAS. Third, the present study focused on the genomic data alone. Neither the meta-GWAS or the ADNI-1 data in this study have gene expression for the study participants. However, gene expression data from patients with AD exist in other data sets, such as the Religious Orders Study (Bennett et al., 2012). Future work could integrate gene expression data into a prioritization pipeline, which has been done in other fields, such as cancer (Ritchie et al., 2013). Finally, we have not validated any of our gene candidates experimentally, and the proposed mechanisms for our highly ranked genes are speculative.

Despite the above limitations, however, the integrative approach we have taken has strongly implicated cytoskeletal dynamics, ER stress, and transcriptional dysregulation as major cellular processes driving neural atrophy. While it is beyond the scope of the present study to validate any of our candidates, by highlighting specific cellular processes and genes taking part in those processes, we can design robust *in vivo* and *in vitro* experiments to test them. For example, recent results in cultured neurons implicate impaired dendritic dynamics as a hallmark of AD (Froula et al., 2018; Boros et al., 2019; Henderson et al., 2019; Walker and Herskowitz, 2020; Walker et al., 2021). Such culture systems could be used for follow up experiments in which our candidate genes could feasibly be tested at scale.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: http://adni.loni.usc.edu/data-samples/access-data/, https://ctg.cncr.nl/documents/p1651/AD_sumstats_Jansenetal_2019sept.txt.gz. We recognize that some may not be able to gain access to the ADNI data so we have made the gene-level summary statistics for the ADNI and MetaGWAS datasets available on the GitHub repository for this paper, along with all the code required to replicate this analysis: https://github.com/MahoneyLabGroup/AD_NBFP.

AUTHOR CONTRIBUTIONS

JB and JM designed the study. JB obtained the data and analyzed it using the pipeline designed by AT and JM. JB, ML, AT, and JM drafted and revised this manuscript. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Library of Medicine (5R21LM012615-02) and the National Institute of General Medical Sciences (5P20GM130454-02).

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.625246/full#supplementary-material

Supplementary Data Sheet 1 | List of Hippocampus Volume Significant Genes.

Supplementary Data Sheet 2 | List of Amygdala Volume Significatn Genes.

Supplementary Data Sheet 3 | MetaGWAS Significant Genes.

Supplementary Data Sheet 4 | Top Hippocampus Network Module GO Terms.

Supplementary Data Sheet 5 | Top Amygdala Network Module GO Terms.

Supplementary Data Sheet 6 | Final Hippocampus Scores.

Supplementary Data Sheet 7 | Final Amygdala Scores.

REFERENCES

1000 Genomes Project Consortium, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Andrews, S. J., Fulton-Howard, B., and Goate, A. (2020). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol.* 19, 326–335. doi: 10.1016/s1474-4422(19)30435-1

Ansar, M., Chung, H. L., Al-Otaibi, A., Elagabani, M. N., Ravenscroft, T. A., Paracha, S. A., et al. (2019). Bi-allelic variants in IQSEC1 cause intellectual disability, developmental delay, and short stature. *Am. J. Hum. Genetics* 105, 907–920. doi: 10.1016/j.ajhg.2019.09.013

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

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: an open source software for exploring and manipulating networks," in *Proceedings of the 3rd international AAAI conference on weblogs and social media, ICWSM 2009*, San Jose, CA. doi: 10.13140/2.1.1341.1520

Bear, J. E., Loureiro, J. J., Libova, I., Fässler, R., Wehland, J., and Gertler, F. B. (2000). Negative regulation of fibroblast motility by Ena/VASP proteins. *Cell* 101, 717–728. doi: 10.1016/s0092-8674(00)80884-3

Bennett, D. A., Schneider, J. A., Arvanitakis, Z., and Wilson, R. S. (2012). Overview and findings from the religious orders study. *Curr. Alzheimer Res.* 9, 628–645. doi: 10.2174/156720512801322573

Bhagwat, N., Viviano, J. D., Voineskos, A. N., Chakravarty, M. M., and Initiative, A. D. N. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput. Biol.* 14:e1006376. doi: 10.1371/journal.pcbi.1006376

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Arxiv* [Preprint]. doi: 10.1088/1742-5468/2008/10/p10008

Bokoch, G. M. (2003). Biology of the P21-activated kinases. *Annu. Rev. Biochem.* 72, 743–781. doi: 10.1146/annurev.biochem.72.121801.161742

Boros, B. D., Greathouse, K. M., Gearing, M., and Herskowitz, J. H. (2019). Dendritic spine remodeling accompanies Alzheimer's disease pathology and genetic susceptibility in cognitively normal aging. *Neurobiol. Aging* 73, 92–103. doi: 10.1016/j.neurobiolaging.2018.09.003

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Calderon-Garcidueñas, A. L., and Duyckaerts, C. (2017). Chapter 23 Alzheimer disease. *Handb. Clin. Neurol.* 145, 325–337. doi: 10.1016/b978-0-12-802395-2.00023-7

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., et al. (2018). The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* 47:gky1055. doi: 10.1093/nar/gky1055

Carmona, S., Hardy, J., and Guerreiro, R. (2018). Chapter 26 the genetic landscape of Alzheimer disease. *Handb. Clin. Neurol.* 148, 395–408. doi: 10.1016/b978-0-444-64076-5.00026-0

Cavedo, E., Boccardi, M., Ganzola, R., Canu, E., Beltramello, A., Caltagirone, C., et al. (2011). Local amygdala structural differences with 3T MRI in patients with Alzheimer disease. *Neurology* 76, 727–733. doi: 10.1212/wnl.0b013e31820d62d9

Cazzaniga, A., Locatelli, L., Castiglioni, S., and Maier, J. (2018). The contribution of EDF1 to PPARγ transcriptional activation in VEGF-treated human endothelial cells. *Int. J. Mol. Sci.* 19:1830. doi: 10.3390/ijms19071830

Chabrier, M. A., Cheng, D., Castello, N. A., Green, K. N., and LaFerla, F. M. (2014). Synergistic effects of amyloid-beta and wild-type human tau on dendritic spine loss in a floxed double transgenic model of Alzheimer's disease. *Neurobiol. Dis.* 64, 107–117. doi: 10.1016/j.nbd.2014.01.007

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 1–16. doi: 10.1186/s13742-015-0047-8

Chiba-Falek, O., Gottschalk, W. K., and Lutz, M. W. (2018). The effects of the TOMM40 poly-T alleles on Alzheimer's disease phenotypes. *Alzheimers Dement.* 14, 692–698. doi: 10.1016/j.jalz.2018.01.015

Choubey, S. K., and Jeyakanthan, J. (2018). Molecular dynamics and quantum chemistry-based approaches to identify isoform selective HDAC2 inhibitor – a novel target to prevent Alzheimer's disease. *J. Recept. Signal. Transduct. Res.* 38, 1–13. doi: 10.1080/10799893.2018.1476541

Combs, C. K., Johnson, D. E., Karlo, J. C., Cannady, S. B., and Landreth, G. E. (2000). Inflammatory mechanisms in Alzheimer's disease: inhibition of β-amyloid-stimulated proinflammatory responses and neurotoxicity by PPARγ agonists. *J. Neurosci.* 20, 558–567. doi: 10.1523/jneurosci.20-02-00558.2000

Concannon, C. G., Ward, M. W., Bonner, H. P., Kuroki, K., Tuffy, L. P., Bonner, C. T., et al. (2008). NMDA receptor—mediated excitotoxic neuronal apoptosis in vitro and in vivo occurs in an ER stress and PUMA independent manner. *J. Neurochem.* 105, 891–903. doi: 10.1111/j.1471-4159.2007.05187.x

d'Angelo, M., Castelli, V., Catanesi, M., Antonosante, A., Dominguez-Benot, R., Ippoliti, R., et al. (2019). PPARγ and cognitive performance. *Int. J. Mol. Sci.* 20:5068. doi: 10.3390/ijms20205068

de Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* 11:e1004219. doi: 10.1371/journal.pcbi.1004219

Dickson, B. J., and Gilestro, G. F. (2006). Regulation of commissural axon pathfinding by slit and its robo receptors. *Annu. Rev. Cell Dev. Biol.* 22, 651–675. doi: 10.1146/annurev.cellbio.21.090704.151234

Dubrac, A., Genet, G., Ola, R., Zhang, F., Pibouin-Fragner, L., Han, J., et al. (2016). Targeting NCK-mediated endothelial cell front-rear polarity inhibits neovascularization. *Circulation* 133, 409–421. doi: 10.1161/circulationaha.115.017537

Elagabani, M. N., Briševac, D., Kintscher, M., Pohle, J., Köhr, G., Schmitz, D., et al. (2016). Subunit-selective N-Methyl-d-aspartate (n.d.) receptor signaling through brefeldin A-resistant Arf guanine nucleotide exchange factors BRAG1 and BRAG2 during synapse maturation. *J. Biol. Chem.* 291, 9105–9118. doi: 10.1074/jbc.m115.691717

Elkan, C., and Noto, K. (2008). "Learning classifiers from only positive and unlabeled data," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, 213–220. doi: 10.1145/1401890.1401920

Feng, Y., Feng, L., Yu, D., Zou, J., and Huang, Z. (2016). srGAP1 mediates the migration inhibition effect of Slit2-Robo1 in colorectal cancer. *J. Exp. Clin. Canc Res.* 35:191. doi: 10.1186/s13046-016-0469-x

Froula, J. M., Henderson, B. W., Gonzalez, J. C., Vaden, J. H., Mclean, J. W., Wu, Y., et al. (2018). α-Synuclein fibril-induced paradoxical structural and functional defects in hippocampal neurons. *Acta Neuropathol. Commun.* 6:35. doi: 10.1186/s40478-018-0537-x

Gerakis, Y., and Hetz, C. (2018). Emerging roles of ER stress in the etiology and pathogenesis of Alzheimer's disease. *FEBS J.* 285, 995–1011. doi: 10.1111/febs.14332

Gertler, F. B., Niebuhr, K., Reinhard, M., Wehland, J., and Soriano, P. (1996). Mena, a relative of VASP and *Drosophila* enabled, is implicated in the control of microfilament dynamics. *Cell* 87, 227–239. doi: 10.1016/s0092-8674(00)81341-0

Glaeser, K., Urban, M., Fenech, E., Voloshanenko, O., Kranz, D., Lari, F., et al. (2018). ERAD—dependent control of the Wnt secretory factor Evi. *EMBO J.* 37:e97311. doi: 10.15252/embj.201797311

Gorenshteyn, D., Zaslavsky, E., Fribourg, M., Park, C. Y., Wong, A. K., Tadych, A., et al. (2015). Interactive big data resource to elucidate human immune pathways and diseases. *Immunity* 43, 605–614. doi: 10.1016/j.immuni.2015.08.014

Govindarajulu, M., Pinky, P. D., Bloemer, J., Ghanei, N., Suppiramaniam, V., and Amin, R. (2018). Signaling mechanisms of selective PPARγ modulators in Alzheimer's disease. *PPAR Res.* 2018, 1–20. doi: 10.1155/2018/2010675

Goya, J., Wong, A. K., Yao, V., Krishnan, A., Homilius, M., and Troyanskaya, O. G. (2015). FNTM: a server for predicting functional networks of tissues in mouse. *Nucleic Acids Res.* 43, W182–W187. doi: 10.1093/nar/gkv443

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. doi: 10.1038/ng.3259

Gu, X., Chen, Y., Zhou, Q., Lu, Y.-C., Cao, B., Zhang, L., et al. (2018). Analysis of GWAS-linked variants in multiple system atrophy. *Neurobiol. Aging* 67, 201.e1–201.e4. doi: 10.1016/j.neurobiolaging.2018.03.018

Guan, Y., Ackert-Bicknell, C. L., Kell, B., Troyanskaya, O. G., and Hibbs, M. A. (2010). Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.* 6:e1000991. doi: 10.1371/journal.pcbi.1000991

Hardingham, G. E., and Bading, H. (2010). Synaptic versus extrasynaptic NMDA receptor signalling: implications for neurodegenerative disorders. *Nat. Rev. Neurosci.* 11, 682–696. doi: 10.1038/nrn2911

Hardingham, G. E., Fukunaga, Y., and Bading, H. (2002). Extrasynaptic NMDARs oppose synaptic NMDARs by triggering CREB shut-off and cell death pathways. *Nat. Neurosci.* 5, 405–414. doi: 10.1038/nn835

Henderson, B. W., Greathouse, K. M., Ramdas, R., Walker, C. K., Rao, T. C., Bach, S. V., et al. (2019). Pharmacologic inhibition of LIMK1 provides dendritic spine resilience against β-amyloid. *Sci. Signal.* 12:eaaw9318. doi: 10.1126/scisignal.aaw9318

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

Hoyer, S., and Lannert, H. (1999). Inhibition of the neuronal insulin receptor causes Alzheimer—like disturbances in oxidative/energy brain metabolism and in behavior in adult rats. *Ann. N. Y. Acad. Sci.* 893, 301–303. doi: 10.1111/j.1749-6632.1999.tb07842.x

Hu, Y., Zheng, L., Cheng, L., Zhang, Y., Bai, W., Zhou, W., et al. (2017). GAB2 rs2373115 variant contributes to Alzheimer's disease risk specifically in European population. *J. Neurol. Sci.* 375, 18–22. doi: 10.1016/j.jns.2017.01.030

Huang, J., Huang, A., Poplawski, A., DiPino, F., Traugh, J. A., and Ling, J. (2020). PAK2 activated by Cdc42 and caspase 3 mediates different cellular responses to oxidative stress-induced apoptosis. *Biochim. Biophys. Acta BBA Mol. Cell Res.* 1867:118645. doi: 10.1016/j.bbamcr.2020.118645

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS One* 9:e98679. doi: 10.1371/journal.pone.0098679

Jagust, W. J., Bandy, D., Chen, K., Foster, N. L., Landau, S. M., Mathis, C. A., et al. (2010). The Alzheimer's disease neuroimaging initiative positron emission tomography core. *Alzheimers Dement.* 6, 221–229. doi: 10.1016/j.jalz.2010.03.003

Janczura, K. J., Volmar, C.-H., Sartor, G. C., Rao, S. J., Ricciardi, N. R., Lambert, G., et al. (2018). Inhibition of HDAC3 reverses Alzheimer's disease-related pathologies in vitro and in the 3xTg-AD mouse model. *Proc. Natl. Acad. Sci. U.S.A.* 115, 201805436. doi: 10.1073/pnas.1805436115

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* 51, 404–413. doi: 10.1038/s41588-018-0311-9

Jaroudi, W., Garami, J., Garrido, S., Hornberger, M., Keri, S., and Moustafa, A. A. (2017). Factors underlying cognitive decline in old age and Alzheimer's disease: the role of the hippocampus. *Rev. Neurosci.* 28, 705–714. doi: 10.1515/revneuro-2016-0086

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2019). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi: 10.1093/nar/gkz1031

Jia, L., Xu, H., Chen, S., Wang, X., Yang, J., Gong, M., et al. (2020). The APOE ε4 exerts differential effects on familial and other subtypes of Alzheimer's disease. *Alzheimers Dement.* 16, 1613–1623. doi: 10.1002/alz.12153

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kang, D. E., and Woo, J. A. (2019). Cofilin, a master node regulating cytoskeletal pathogenesis in Alzheimer's disease. *J. Alzheimers Dis.* 72, S131–S144. doi: 10.3233/jad-190585

Kawaai, K., Hisatsune, C., Kuroda, Y., Mizutani, A., Tashiro, T., and Mikoshiba, K. (2009). 80K-H interacts with inositol 1,4,5-trisphosphate (IP3) receptors and regulates IP3-induced calcium release activity. *J. Biol. Chem.* 284, 372–380. doi: 10.1074/jbc.m805828200

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020). gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000research* 9:ELIXIR–709. doi: 10.12688/f1000research.24956.2

Koller, E. J., and Chakrabarty, P. (2020). Tau-mediated dysregulation of neuroplasticity and glial plasticity. *Front. Mol. Neurosci.* 13:151. doi: 10.3389/fnmol.2020.00151

Krishna, K., Behnisch, T., and Sajikumar, S. (2016). Inhibition of histone deacetylase 3 restores amyloid-β oligomer-induced plasticity deficit in hippocampal CA1 pyramidal neurons. *J. Alzheimers Dis.* 51, 783–791. doi: 10.3233/jad-150838

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., et al. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* 19, 1454–1462. doi: 10.1038/nn.4353

Lanier, L. M., Gates, M. A., Witke, W., Menzies, A. S., Wehman, A. M., Macklis, J. D., et al. (1999). Mena is required for neurulation and commissure formation. *Neuron* 22, 313–325. doi: 10.1016/s0896-6273(00)81092-2

Lanier, L. M., and Gertler, F. B. (2000). From Abl to actin: Abl tyrosine kinase and associated proteins in growth cone motility. *Curr. Opin. Neurobiol.* 10, 80–87. doi: 10.1016/s0959-4388(99)00058-6

Lee, S. H., Harold, D., Nyholt, D. R., ANZGene Consortium, International Endogene Consortium, Genetic and Environmental Risk for Alzheimer's disease Consortium, et al. (2013). Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum. Mol. Genet.* 22, 832–841. doi: 10.1093/hmg/dds491

Li, M., Xu, Y., Wang, Y., Yang, X.-A., and Jin, D. (2019). Compound heterozygous variants in MOGS inducing congenital disorders of glycosylation (CDG) IIb. *J. Hum. Genet.* 64, 265–268. doi: 10.1038/s10038-018-0552-6

Li, W., Tam, K. M. V., Chan, W. W. R., Koon, A. C., Ngo, J. C. K., Chan, H. Y. E., et al. (2018). Neuronal adaptor FE65 stimulates Rac1-mediated neurite outgrowth by recruiting and activating ELMO1. *J. Biol. Chem.* 293, 7674–7688. doi: 10.1074/jbc.ra117.000505

Li, Y., and Li, J. (2012). Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC Genomics* 13:S27. doi: 10.1186/1471-2164-13-s7-s27

Lian, B., liu, M., Lan, Z., Sun, T., Meng, Z., Chang, Q., et al. (2020). Hippocampal overexpression of SGK1 ameliorates spatial memory, rescues Aβ pathology and actin cytoskeleton polymerization in middle-aged APP/PS1 mice. *Behav. Brain Res.* 383:112503. doi: 10.1016/j.bbr.2020.112503

Lim, D., Rodríguez-Arellano, J. J., Parpura, V., Zorec, R., Zeidán-Chuliá, F., Genazzani, A. A., et al. (2016). Calcium signalling toolkits in astrocytes and spatio-temporal progression of Alzheimer's disease. *Curr. Alzheimer Res.* 13, 359–369. doi: 10.2174/1567205013666151116130104

Lin, J. H., Walter, P., and Yen, T. S. B. (2007). Endoplasmic reticulum stress in disease pathogenesis. *Annu. Rev. Pathol. Mech. Dis.* 3, 399–425. doi: 10.1146/annurev.pathmechdis.3.121806.151434

Liu, J., Chang, L., Song, Y., Li, H., and Wu, Y. (2019). The role of NMDA receptors in Alzheimer's disease. *Front. Neurosci. Switz.* 13:43. doi: 10.3389/fnins.2019.00043

Lüscher, C., and Malenka, R. C. (2012). NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). *CSH Perspect. Biol.* 4:a005710. doi: 10.1101/cshperspect.a005710

Ma, Q.-L., Yang, F., Calon, F., Ubeda, O. J., Hansen, J. E., Weisbart, R. H., et al. (2008). p21-activated kinase-aberrant activation and translocation in Alzheimer disease pathogenesis. *J. Biol. Chem.* 283, 14132–14143. doi: 10.1074/jbc.m708034200

Mandrekar-Colucci, S., Karlo, J. C., and Landreth, G. E. (2012). Mechanisms underlying the rapid peroxisome proliferator-activated receptor-γ-mediated amyloid clearance and reversal of cognitive deficits in a murine model of Alzheimer's disease. *J. Neurosci.* 32, 10117–10128. doi: 10.1523/jneurosci.5268-11.2012

Mango, D., Saidi, A., Cisale, G. Y., Feligioni, M., Corbo, M., and Nisticò, R. (2019). Targeting synaptic plasticity in experimental models of Alzheimer's disease. *Front. Pharmacol.* 10:778. doi: 10.3389/fphar.2019.00778

Marlin, J. W., Chang, Y.-W. E., Ober, M., Handy, A., Xu, W., and Jakobi, R. (2011). Functional PAK-2 knockout and replacement with a caspase cleavage-deficient mutant in mice reveals differential requirements of full-length PAK-2 and caspase-activated PAK-2p34. *Mamm. Genome* 22, 306–317. doi: 10.1007/s00335-011-9326-6

Mastick, G. S., Farmer, W. T., Altick, A. L., Nural, H. F., Dugan, J. P., Kidd, T., et al. (2010). Longitudinal axons are guided by Slit/Robo signals from the floor plate. *Cell Adhes. Migr.* 4, 337–341. doi: 10.4161/cam.4.3.11219

McQuown, S. C., and Wood, M. A. (2011). HDAC3 and the molecular brake pad hypothesis. *Neurobiol. Learn. Mem.* 96, 27–34. doi: 10.1016/j.nlm.2011.04.005

Meyer, D., Dimitriadou, E., Hornik, K., Maintainer, A., and Leisch, F. (2019). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). TU Wien*. R package version 1.7-3.

Mukherjee, S., Heath, L., Preuss, C., Jayadev, S., Garden, G. A., Greenwood, A. K., et al. (2020). Molecular estimation of neurodegeneration pseudotime in older brains. *Nat. Commun.* 11:5781. doi: 10.1038/s41467-020-19622-y

Murk, K., Wittenmayer, N., Michaelsen-Preusse, K., Dresbach, T., Schoenenberger, C.-A., Korte, M., et al. (2012). Neuronal profilin isoforms are addressed by different signalling pathways. *PLoS One* 7:e34167. doi: 10.1371/journal.pone.0034167

Nott, A., Cheng, J., Gao, F., Lin, Y.-T., Gjoneska, E., Ko, T., et al. (2016). Histone deacetylase 3 associates with MeCP2 to regulate FOXO and social behavior. *Nat. Neurosci.* 19, 1497–1505. doi: 10.1038/nn.4347

Ottis, P., Topic, B., Loos, M., Li, K. W., de Souza, A., Schulz, D., et al. (2013). Aging-induced proteostatic changes in the rat hippocampus identify ARP3, NEB2 and BRAG2 as a molecular circuitry for cognitive impairment. *PLoS One* 8:e75112. doi: 10.1371/journal.pone.0075112

Parrish, R. R., Albertson, A. J., Buckingham, S. C., Hablitz, J. J., Mascia, K. L., Haselden, W. D., et al. (2013). Status epilepticus triggers early and late alterations in brain-derived neurotrophic factor and NMDA glutamate receptor Grin2b DNA methylation levels in the hippocampus. *Neuroscience* 248, 602–619. doi: 10.1016/j.neuroscience.2013.06.029

Parsons, M. P., and Raymond, L. A. (2014). Extrasynaptic NMDA receptor involvement in central nervous system disorders. *Neuron* 82, 279–293. doi: 10.1016/j.neuron.2014.03.030

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., et al. (2010). Alzheimer's disease neuroimaging initiative (ADNI) clinical Characterization. *Neurology* 74, 201–209. doi: 10.1212/wnl.0b013e3181cb3e25

Pipatpiboon, N., Pratchayasakul, W., Chattipakorn, N., and Chattipakorn, S. C. (2012). PPARγ agonist improves neuronal insulin receptor function in hippocampus and brain mitochondria function in rats with insulin resistance induced by long term high-fat diets. *Endocrinology* 153, 329–338. doi: 10.1210/en.2011-1502

Pirttimaki, T. M., Codadu, N. K., Awni, A., Pratik, P., Nagel, D. A., Hill, E. J., et al. (2013). α7 nicotinic receptor-mediated astrocytic gliotransmitter release: Aβ effects in a preclinical Alzheimer's mouse model. *PLoS One* 8:e81828. doi: 10.1371/journal.pone.0081828

Pozueta, J., Lefort, R., and Shelanski, M. L. (2013). Synaptic changes in Alzheimer's disease and its models. *Neuroscience* 251, 51–65. doi: 10.1016/j.neuroscience.2012.05.050

Price, K. A., Varghese, M., Sowa, A., Yuk, F., Brautigam, H., Ehrlich, M. E., et al. (2014). Altered synaptic structure in the hippocampus in a mouse model of Alzheimer's disease with soluble amyloid-β oligomers and no plaque pathology. *Mol. Neurodegener.* 9:41. doi: 10.1186/1750-1326-9-41

Rajan, K. B., Weuve, J., Barnes, L. L., Wilson, R. S., and Evans, D. A. (2019). Prevalence and incidence of clinically diagnosed Alzheimer's disease dementia from 1994 to 2012 in a population study. *Alzheimers Dement.* 15, 1–7. doi: 10.1016/j.jalz.2018.07.216

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. doi: 10.1093/nar/gkz369

Risacher, S. L., Kim, S., Shen, L., Nho, K., Foroud, T., Green, R. C., et al. (2013). The role of apolipoprotein E (APOE) genotype in early mild cognitive impairment (E-MCI). *Front. Aging Neurosci.* 5:11. doi: 10.3389/fnagi.2013.00011

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2013). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16, 85–97. doi: 10.1038/nrg3868

Rönicke, R., Mikhaylova, M., Rönicke, S., Meinhardt, J., Schröder, U. H., Fändrich, M., et al. (2011). Early neuronal dysfunction by amyloid β oligomers depends on activation of NR2B-containing NMDA receptors. *Neurobiol. Aging* 32, 2219–2228. doi: 10.1016/j.neurobiolaging.2010.01.011

Sabo, S. L., Ikin, A. F., Buxbaum, J. D., and Greengard, P. (2001). The Alzheimer amyloid precursor protein (APP) and Fe65, an APP-binding protein, regulate cell movement. *J. Cell Biol.* 153, 1403–1414. doi: 10.1083/jcb.153.7.1403

Sadat, M. A., Moir, S., Chun, T.-W., Lusso, P., Kaplan, G., Wolfe, L., et al. (2014). Glycosylation, hypogammaglobulinemia, and resistance to viral infections. *New Engl. J. Med.* 370, 1615–1625. doi: 10.1056/nejmoa1302846

Sattler, R., Xiong, Z., Lu, W.-Y., MacDonald, J. F., and Tymianski, M. (2000). Distinct roles of synaptic and extrasynaptic NMDA receptors in excitotoxicity. *J. Neurosci.* 20, 22–33. doi: 10.1523/jneurosci.20-01-00022.2000

Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., et al. (2010). Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement.* 6, 265–273. doi: 10.1016/j.jalz.2010.03.013

Schaeverbeke, J., Gille, B., Adamczuk, K., Vanderstichele, H., Chassaing, E., Bruffaerts, R., et al. (2019). Cerebrospinal fluid levels of synaptic and neuronal integrity correlate with gray matter volume and amyloid load in the precuneus of cognitively intact older adults. *J. Neurochem.* 149, 139–157. doi: 10.1111/jnc.14680

Schliebs, R., and Arendt, T. (2011). The cholinergic system in aging and neuronal degeneration. *Behav. Brain Res.* 221, 555–563. doi: 10.1016/j.bbr.2010.11.058

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L. M., Trojanowski, J. Q., et al. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain* 132, 1067–1077. doi: 10.1093/brain/awp007

Schweinhuber, S. K., Meßerschmidt, T., Hänsch, R., Korte, M., and Rothkegel, M. (2015). Profilin isoforms modulate astrocytic morphology and the motility of astrocytic processes. *PLoS One* 10:e0117244. doi: 10.1371/journal.pone.0117244

Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., et al. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann. Neurol.* 65, 403–413. doi: 10.1002/ana.21610

Shin, E.-Y., Shim, E.-S., Lee, C.-S., Kim, H. K., and Kim, E.-G. (2009). Phosphorylation of RhoGDI1 by p21-activated kinase 2 mediates basic fibroblast growth factor-stimulated neurite outgrowth in PC12 cells. *Biochem. Biophys. Res. Commun.* 379, 384–389. doi: 10.1016/j.bbrc.2008.12.066

Singh, A. K., Kashyap, M. P., Tripathi, V. K., Singh, S., Garg, G., and Rizvi, S. I. (2017). Neuroprotection through rapamycin-induced activation of autophagy and PI3K/Akt1/mTOR/CREB signaling against amyloid-β-induced oxidative stress, synaptic/neurotransmission dysfunction, and neurodegeneration in adult rats. *Mol. Neurobiol.* 54, 5815–5828. doi: 10.1007/s12035-016-0129-3

Slováková, J., Speicher, S., Sánchez-Soriano, N., Prokop, A., and Carmena, A. (2012). The actin-binding protein Canoe/AF-6 forms a complex with robo and is required for slit-robo signaling during Axon pathfinding at the CNS midline. *J. Neurosci.* 32, 10035–10044. doi: 10.1523/jneurosci.6342-11.2012

Sokka, A.-L., Putkonen, N., Mudo, G., Pryazhnikov, E., Reijonen, S., Khiroug, L., et al. (2007). Endoplasmic reticulum stress inhibition protects against excitotoxic neuronal injury in the rat brain. *J. Neurosci.* 27, 901–908. doi: 10.1523/jneurosci.4289-06.2007

Solomon, A., Kivipelto, M., Molinuevo, J. L., Tom, B., Ritchie, C. W., and Consortium, E. (2018). European prevention of Alzheimer's dementia longitudinal cohort study (EPAD LCS): study protocol. *BMJ Open* 8:e021017. doi: 10.1136/bmjopen-2017-021017

Song, A., Yan, J., Kim, S., Risacher, S. L., Wong, A. K., Saykin, A. J., et al. (2016). Network-based analysis of genetic variants associated with hippocampal volume in Alzheimer's disease: a study of ADNI cohorts. *BioData Min.* 9:3. doi: 10.1186/s13040-016-0082-8

Spires-Jones, T., and Knafo, S. (2012). Spines, plasticity, and cognition in Alzheimer's model mice. *Neural. Plast.* 2012:319836. doi: 10.1155/2012/319836

Sun, H., Liu, M., Sun, T., Chen, Y., Lan, Z., Lian, B., et al. (2019). Age-related changes in hippocampal AD pathology, actin remodeling proteins and spatial memory behavior of male APP/PS1 mice. *Behav. Brain Res.*, 376:112182. doi: 10.1016/j.bbr.2019.112182

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

Tigaret, C. M., Thalhammer, A., Rast, G. F., Specht, C. G., Auberson, Y. P., Stewart, M. G., et al. (2006). Subunit dependencies of N-Methyl-d-aspartate (n.d.) receptor-induced α-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor internalization. *Mol. Pharmacol.* 69, 1251–1259. doi: 10.1124/mol.105.018580

Trillaud-Doppia, E., and Boehm, J. (2018). The amyloid precursor protein intracellular domain is an effector molecule of metaplasticity. *Biol. Psychiatry* 83, 406–415. doi: 10.1016/j.biopsych.2016.12.015

Tyler, A. L., Raza, A., Krementsov, D. N., Case, L. K., Huang, R., Ma, R. Z., et al. (2019). Network-based functional prediction augments genetic association to predict candidate genes for histamine hypersensitivity in mice. *G3 Gene. Genom. Genet.* 9, 4223–4233. doi: 10.1534/g3.119.400740

Um, J. W. (2017). Synaptic functions of the IQSEC family of ADP-ribosylation factor guanine nucleotide exchange factors. *Neurosci. Res.* 116, 54–59. doi: 10.1016/j.neures.2016.06.007

Varghese, M., Keshav, N., Jacot-Descombes, S., Warda, T., Wicinski, B., Dickstein, D. L., et al. (2017). Autism spectrum disorder: neuropathology and animal models. *Acta Neuropathol.* 134, 537–566. doi: 10.1007/s00401-017-1736-4

Verkhratsky, A., Rodríguez-Arellano, J. J., Parpura, V., and Zorec, R. (2017). Astroglial calcium signalling in Alzheimer's disease. *Biochem. Biophys. Res. Commun.* 483, 1005–1012. doi: 10.1016/j.bbrc.2016.08.088

Walker, C. K., Greathouse, K. M., Boros, B. D., Poovey, E. H., Clearman, K. R., Ramdas, R., et al. (2021). Dendritic spine remodeling and synaptic tau levels in

PS19 tauopathy mice. *Neuroscience* 455, 195–211. doi: 10.1016/j.neuroscience.2020.12.006

Walker, C. K., and Herskowitz, J. H. (2020). Dendritic spines: mediators of cognitive resilience in aging and Alzheimer's disease. *Neurosci* 107385842094596. doi: 10.1177/1073858420945964

Wang, R., and Reddy, P. H. (2016). Role of glutamate and NMDA receptors in Alzheimer's disease. *J. Alzheimers Dis.* 57, 1041–1048. doi: 10.3233/jad-160763

Wang, X., Herr, R. A., Rabelink, M., Hoeben, R. C., Wiertz, E. J. H. J., and Hansen, T. H. (2009). Ube2j2 ubiquitinates hydroxylated amino acids on ER-associated degradation substrates. *J. Cell Biol.* 187, 655–668. doi: 10.1083/jcb.200908036

Wang, Y., Zeng, C., Li, J., Zhou, Z., Ju, X., Xia, S., et al. (2018). PAK2 haploinsufficiency results in synaptic cytoskeleton impairment and autism-related behavior. *Cell Rep.* 24, 2029–2041. doi: 10.1016/j.celrep.2018.07.061

Wang, Z., Zhang, M., Han, Y., Song, H., Guo, R., and Li, K. (2016). Differentially disrupted functional connectivity of the subregions of the amygdala in Alzheimer's disease. *J. Xray Sci. Technol.* 24, 329–342. doi: 10.3233/xst-160556

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., et al. (2015). 2014 Update of the Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimers Dement.* 11, e1–e120. doi: 10.1016/j.jalz.2014.11.001

Whitwell, J. L., Wiste, H. J., Weigand, S. D., Rocca, W. A., Knopman, D. S., Roberts, R. O., et al. (2012). Comparison of imaging biomarkers in the Alzheimer disease neuroimaging initiative and the mayo clinic study of aging. *Arch. Neurol. Chicago* 69, 614–622. doi: 10.1001/archneurol.2011.3029

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wong, A. K., Krishnan, A., and Troyanskaya, O. G. (2018). GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Res.* 46, W65–W70. doi: 10.1093/nar/gky408

Wong, K., Ren, X.-R., Huang, Y.-Z., Xie, Y., Liu, G., Saito, H., et al. (2001). Signal transduction in neuronal migration roles of GTPase activating proteins and the small GTPase Cdc42 in the slit-robo pathway. *Cell* 107, 209–221. doi: 10.1016/s0092-8674(01)00530-x

Wu, M., Lin, Z., Ma, S., Chen, T., Jiang, R., and Wong, W. H. (2017). Simultaneous inference of phenotype-associated genes and relevant tissues from GWAS data via Bayesian integration of multiple tissue-specific gene networks. *J. Mol. Cell Biol.* 9, 436–452. doi: 10.1093/jmcb/mjx059

Wyman, B. T., Harvey, D. J., Crawford, K., Bernstein, M. A., Carmichael, O., Cole, P. E., et al. (2013). Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimers Dement.* 9, 332–337. doi: 10.1016/j.jalz.2012.06.004

Xu, R., Qin, N., Xu, X., Sun, X., Chen, X., and Zhao, J. (2018). Inhibitory effect of SLIT2 on granulosa cell proliferation mediated by the CDC42-PAKs-ERK1/2 MAPK pathway in the prehierarchical follicles of the chicken ovary. *Sci. Rep.* 8:9168. doi: 10.1038/s41598-018-27601-z

Yao, V., Kaletsky, R., Keyes, W., Mor, D. E., Wong, A. K., Sohrabi, S., et al. (2018). An integrative tissue-network approach to identify and test human disease genes. *Nat. Biotechnol.* 36, 1091–1099. doi: 10.1038/nbt.4246

Yao, X., Jingwen, Y., Kefei, L., Sungeun, K., Kwangsik, N., Risacher, S. L., et al. (2017). Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules. *Bioinformatics* 33, 3250–3257. doi: 10.1093/bioinformatics/btx344

Yu, S. Y., Wu, D. C., and Zhan, R. Z. (2010). GluN2B subunits of the NMDA receptor contribute to the AMPA receptor internalization during long-term depression in the lateral amygdala of juvenile rats. *Neuroscience* 171, 1102–1108. doi: 10.1016/j.neuroscience.2010.09.038

Yu, W.-F., Guan, Z.-Z., Bogdanovic, N., and Nordberg, A. (2005). High selective expression of α7 nicotinic receptors on astrocytes in the brains of patients with sporadic Alzheimer's disease and patients carrying Swedish APP 670/671 mutation: a possible association with neuritic plaques. *Exp. Neurol.* 192, 215–225. doi: 10.1016/j.expneurol.2004.12.015

Zhao, N., Liu, C.-C., Qiao, W., and Bu, G. (2018). Apolipoprotein E, receptors, and modulation of Alzheimer's disease. *Biol. Psychiatry* 83, 347–357. doi: 10.1016/j.biopsych.2017.03.003

Zhou, Q., Zhao, F., Lv, Z., Zheng, C., Zheng, W., Sun, L., et al. (2014). Association between APOC1 Polymorphism and Alzheimer's disease: a case-control study and meta-analysis. *PLoS One* 9:e87017. doi: 10.1371/journal.pone.0087017

Zhu, X., Wang, S., Yu, L., Jin, J., Ye, X., Liu, Y., et al. (2017). HDAC3 negatively regulates spatial memory in a mouse model of Alzheimer's disease. *Aging Cell* 16, 1073–1082. doi: 10.1111/acel.12642

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 April 2021
doi: 10.3389/fgene.2021.629475

Luminal A Breast Cancer Co-expression Network: Structural and Functional Alterations

*Diana García-Cortés [1,2], Enrique Hernández-Lemus [1,3] and Jesús Espinal-Enríquez [1,3]**

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

Luminal A is the most common breast cancer molecular subtype in women worldwide. These tumors have characteristic yet heterogeneous alterations at the genomic and transcriptomic level. Gene co-expression networks (GCNs) have contributed to better characterize the cancerous phenotype. We have previously shown an imbalance in the proportion of intra-chromosomal (*cis*-) over inter-chromosomal (*trans*-) interactions when comparing cancer and healthy tissue GCNs. In particular, for breast cancer molecular subtypes (Luminal A included), the majority of high co-expression interactions connect gene-pairs in the same chromosome, a phenomenon that we have called loss of *trans*- co-expression. Despite this phenomenon has been described, the functional implication of this specific network topology has not been studied yet. To understand the biological role that communities of co-expressed genes may have, we constructed GCNs for healthy and Luminal A phenotypes. Network modules were obtained based on their connectivity patterns and they were classified according to their chromosomal homophily (proportion of *cis*-/*trans*- interactions). A functional overrepresentation analysis was performed on communities in both networks to observe the significantly enriched processes for each community. We also investigated possible mechanisms for which the loss of *trans*- co-expression emerges in cancer GCN. To this end we evaluated transcription factor binding sites, CTCF binding sites, differential gene expression and copy number alterations (CNAs) in the cancer GCN. We found that *trans*- communities in Luminal A present more significantly enriched categories than *cis*- ones. Processes, such as angiogenesis, cell proliferation, or cell adhesion were found in *trans*- modules. The differential expression analysis showed that FOXM1, CENPA, and CIITA transcription factors, exert a major regulatory role on their communities by regulating expression of their target genes in other chromosomes. Finally, identification of CNAs, displayed a high enrichment of deletion peaks in *cis*- communities. With this approach, we demonstrate that network topology determine, to at certain extent, the function in Luminal A breast cancer network. Furthermore, several mechanisms seem to be acting together to avoid *trans*- co-expression. Since this phenomenon has been observed in other cancer tissues, a remaining question is whether the loss of long distance co-expression is a novel hallmark of cancer.

Keywords: loss of long range co-expression, gene co-expression networks, Luminal A breast cancer, breast cancer, transcription factor analysis, CTCF binding site analysis

OPEN ACCESS

Edited by:
Kimberly Glass, Brigham and Women's Hospital and Harvard Medical School, United States

Reviewed by:
Rebekka Burkholz, Harvard University, United States
Alexander Lachmann, Icahn School of Medicine at Mount Sinai, United States

***Correspondence:**
Jesús Espinal-Enríquez
jespinal@inmegen.gob.mx

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

Received: *14 November 2020*
Accepted: *17 March 2021*
Published: *20 April 2021*

Citation:
García-Cortés D, Hernández-Lemus E and Espinal-Enríquez J (2021) Luminal A Breast Cancer Co-expression Network: Structural and Functional Alterations. Front. Genet. 12:629475. doi: 10.3389/fgene.2021.629475

1. BACKGROUND

Gene co-expression networks (GCN) enable the study of interactions of highly correlated genes in a transcriptional program, capturing global and local connectivity properties emerging from those interactions (Sonawane et al., 2019). These type of networks are built from gene expression profiles, a measurable output of transcription. Therefore, they outline the contribution of the regulatory elements operating at different levels of the transcription process to ensure the expression of specific sets of genes. In this sense, GCNs might provide insights about shared regulatory mechanisms and their alterations in a disease, such as cancer (Emmert-Streib et al., 2014; Yang et al., 2014; Wu et al., 2019; Liao et al., 2020). Those alterations in cancer disrupt the transcriptional process and lead to altered gene expression and the promotion of tumor progression (Garraway and Lander, 2013; Lee and Young, 2013).

There are multiple studies where GCNs are constructed and important aspects of the connectivity structure are analyzed to identify genes prognosis markers (Hsu et al., 2019), metabolic deregulation (Serrano-Carbajal et al., 2020), and differences in transcriptional profiles (van Dam et al., 2018).

In breast cancer GCNs, there is an imbalance in the proportion of intra-chromosomal (*cis*-) over inter-chromosomal (*trans*-) gene co-expression interactions, meaning that the majority of high co-expression links connect gene-pairs in the same chromosome (Espinal-Enríquez et al., 2017; de Anda-Jáuregui et al., 2019a; Dorantes-Gilardi et al., 2020). This phenomenon has been called loss of long distance co-expression. Furthermore, a highly localized co-expression pattern associated with chromosome cytobands has been observed (García-Cortés et al., 2020). These features are not present in the healthy tissue GCN. In the entire set of co-expression interactions, the loss of long distance co-expression in breast cancer (measured in base pairs) subtypes is displayed as a decay in the *cis*- co-expression values dependent on gene physical distance (de Anda-Jáuregui et al., 2019b; García-Cortés et al., 2020).

The structural characteristics evaluated in the co-expression networks are different for each breast cancer molecular subtype, displaying another instance of their emblematic heterogeneity (Alcalá-Corona et al., 2017, 2018a). The four breast cancer molecular subtypes, Luminal A, Luminal B, HER2+ and Basal-like, are classified according to their gene expression profiles and they represent different cancer manifestations, with distinct molecular traits, genomic alterations, and prognosis (Perou et al., 2000; Prat and Perou, 2011; Berger et al., 2018). Hormone status, evaluated through the expression of estrogen and progesterone receptors (ER and PR correspondingly), and the presence of human epidermal growth factor receptor 2 (HER2), play a major role for breast cancer molecular subtypes characterization and the election of therapeutic strategies (Zhang et al., 2014).

Luminal A is the most frequent breast cancer molecular subtype. Almost a half of the total cases of breast cancer correspond to this phenotype (Fan et al., 2006). These tumors are often positive to estrogen receptor (ER) and negative to ERBB2 receptor, and they also present overexpression on the ER-regulated genes. This subtype is associated with highest median survival, best prognosis (Hu et al., 2006), and lower recurrence rates (Arvold et al., 2011; Metzger-Filho et al., 2013).

Nevertheless, clinical and molecular heterogeneity is present within Luminal A tumors, where differences in genomic alterations have been potentially associated with resistance to endocrine therapy (Ciriello et al., 2013).

Additionally, the Luminal A GCN presents the least dissimilar structure compared with the healthy GCN (García-Cortés et al., 2020). A relevant measure to analyze differences in cancer GCNs, is the size of connected components. In the case of healthy GCN, as well as in the case of Luminal A GCN, they present a giant component (a set of connected genes that contains more than the half of the total amount of nodes in the networks). The other breast cancer subtype GCNs have only small intra-chromosomal connected components. Furthermore, Luminal A GCN is the one with the highest number of inter-chromosomal (*trans*-) interactions.

The structure of a GNC is often organized into *communities* or modules (Alcalá-Corona et al., 2016), this is, subsets of connected genes so that the density of within-connections is higher than that of between-connections (Girvan and Newman, 2002; Porter et al., 2009; Fortunato and Hric, 2016; Alcalá-Corona et al., 2018a). In the case of GCNs, communities may correspond to a co-regulated set of genes (Wilkinson and Huberman, 2004; Zhu et al., 2008; Cantini et al., 2015). The structure of said modules may capture the phenomenology behind biological processes (Alcalá-Corona et al., 2017, 2018a,b).

Being the subtype with the best prognosis, the most similar co-expression network, and taking into account that community structure in GCN may be implicated in the functional regulation of a cancerous phenotype, in this work we analyzed the structure of communities of the Luminal A GCN, in order to determine the relevance of the loss of long distance co-expression in the biological functions associated to that network. Additionally, we evaluated possible mechanisms for which we observe the preference for *cis*- interactions in this breast cancer subtype. We analyzed the influence of differential gene expression, transcription factor binding sites, copy number alterations, and CTCF binding sites, in order to understand the regulatory mechanisms underlying the appearance of the loss of long distance interactions in cancer GCNs.

2. RESULTS AND DISCUSSION

2.1. Community Structure Displays Loss of *trans*- Co-expression

Figure 1A displays GCNs built from the 20,217 (see Methods section) most significant mutual information interactions in the Luminal A and the Healthy co-expression profiles. Genes are colored according to the chromosome where they are located. As previously reported, the Healthy GCN has a giant component with interactions linking genes from different chromosomes. The Luminal A network also has a giant component but the

Abbreviations: CNA, copy number alteration; GCN, gene co-expression network; GTRD, gene transcription regulation database; LFC, Log2 fold change.

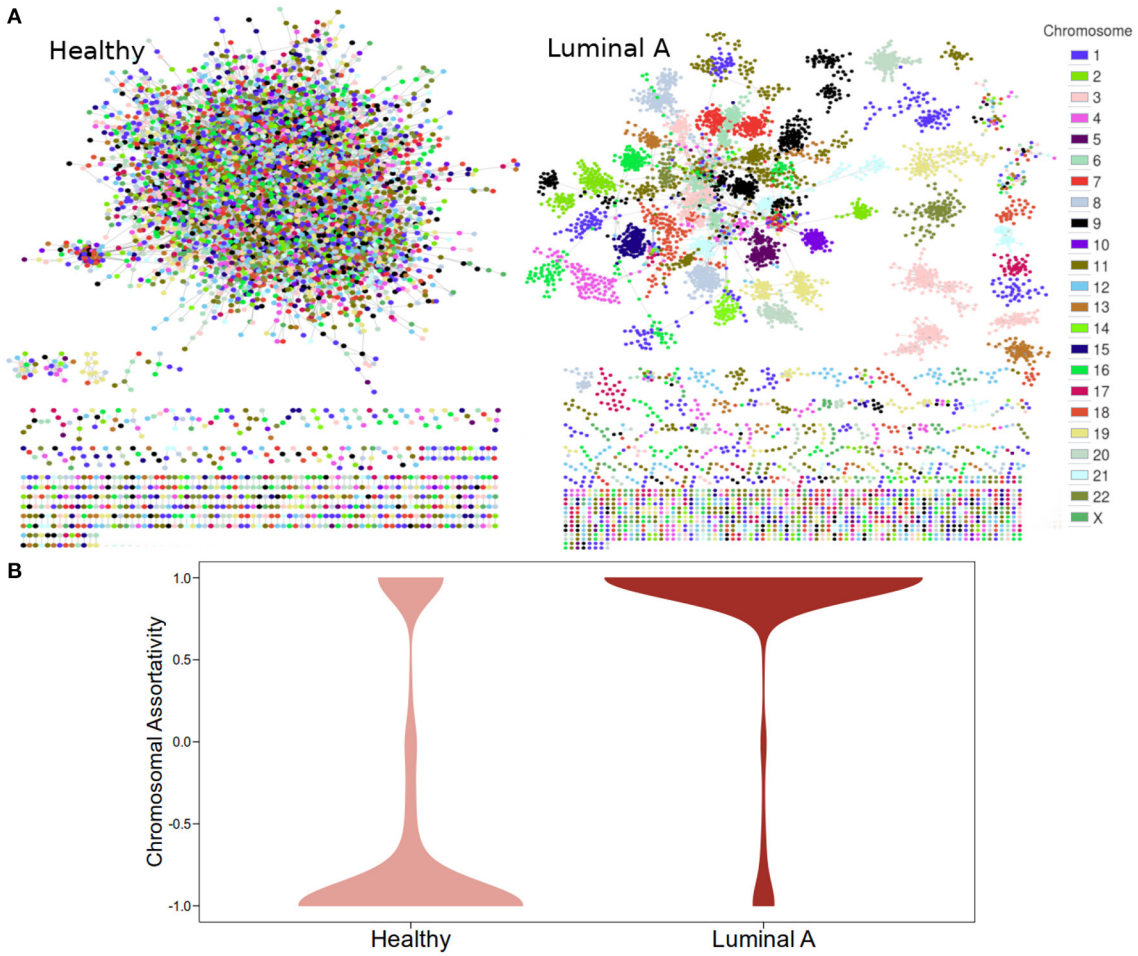

FIGURE 1 | Co-expression networks for Healthy and Luminal A tissue. **(A)** GCNs built from the 20,217 most significant gene pair mutual information values for both phenotypes. Node colors are assigned according to the chromosome where each gene is located. **(B)** Distribution of chromosomal assortativity in network communities.

layout suggests that genes from the same chromosome are preferentially linked.

To evaluate the previous observation, communities were detected in both networks using four algorithms for weighted networks implemented in the `igraph` package: Fast Greedy, Infomap, Leading Eigenvector, and Louvain. **Supplementary Material 1** presents results for all algorithms. Jaccard indexes where calculated among communities detected by the four algorithms. More than 95% of the total number of communities detected by Fast Greedy, Leading Eigenvector, and Louvain have a Jaccard Index equal to 1, while Infomap displays more dissimilar results. Given that Louvain presents the highest modularity values, results for this algorithm are presented in the main text. **Table 1** contains the number of communities and modularity values for the four algorithms applied to the Healthy and the Luminal A network.

Chromosomal assortativity, ASS_{chr} was calculated by taking the number of intra-chromosomal links minus the number of inter-chromosomal links divided by the total number of links in a community. **Figure 1B** displays the distribution of the ASS_{chr} in both networks in the form of violin plots. The differences in the distributions allow us to confirm the loss of *trans*- interactions in the Luminal A GCN.

2.2. Specific *trans*- Communities in the Luminal A GCN Are Highly Associated With Biological Processes

To identify the functional role of the highly co-expressed groups of genes identified by network communities, an overrepresentation analysis was performed, using the biological process category in Gene Ontology (GO). Results for all algorithms are presented in **Table 1**. *-cis* communities are the ones having ASS_{chr} equals to 1.

Half of the *-trans* communities with more than five nodes extracted by the Louvain algorithm in the Luminal A GCN were associated with biological processes. However, only 12% of the *-cis* communities where enriched. Despite having a larger

TABLE 1 | Features of *cis-* and *trans-* chromosomal communities in the Luminal A and the Healthy gene co-expression network.

Algorithm	Healthy							Luminal A						
	Modularity	Communities		Size ≥ 5		Enriched communities		Modularity	Communities		Size ≥ 5		Enriched communities	
		cis-	*trans-*	*cis-*	*trans-*	*cis-*	*trans-*		*cis-*	*trans-*	*cis-*	*trans-*	*cis-*	*trans-*
Fast Greedy	0.703	75	325	0	50	0	14	0.934	614	87	77	40	9	20
Infomap	0.674	83	768	1	386	1	47	0.907	826	93	194	39	16	20
Leading Eigenvector	0.696	71	283	1	32	1	18	0.892	594	84	58	37	9	20
Louvain	0.752	71	291	0	41	0	17	0.935	614	87	77	40	9	20

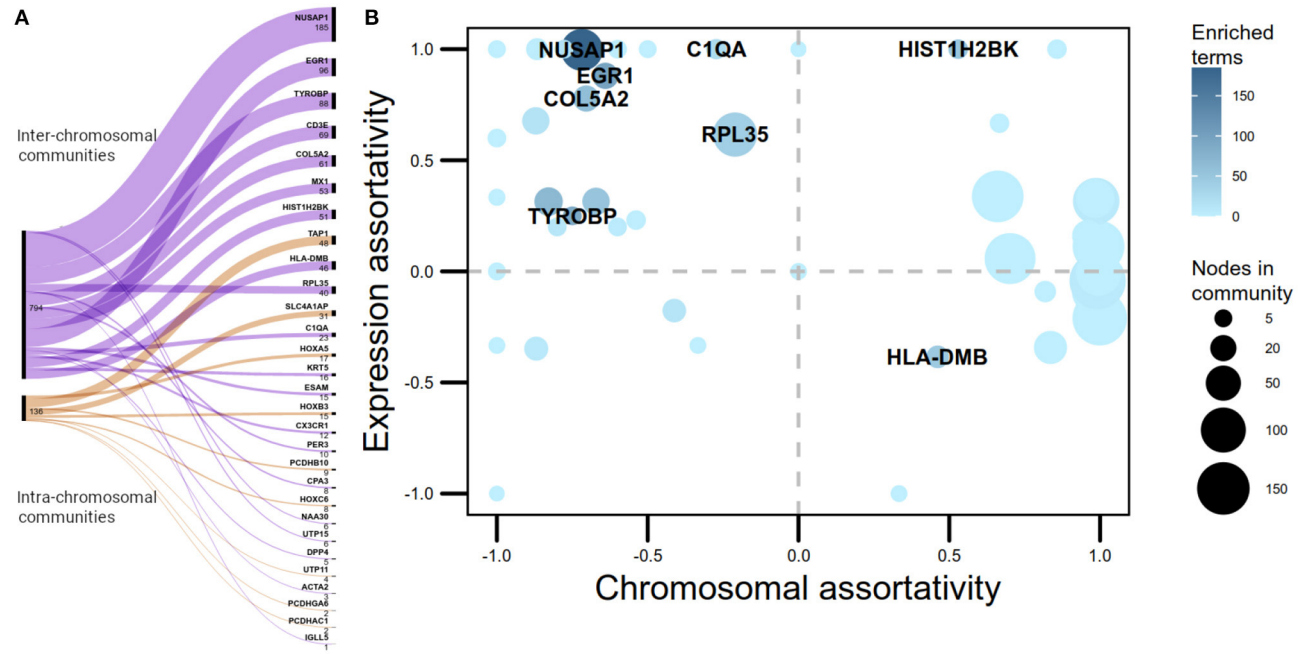

FIGURE 2 | *cis-* and *trans-* communities in the Luminal A network. **(A)** Alluvial graph displaying the proportion of overrepresented Gene Ontology biological processes per community in *trans-* (purple) and *cis-* (orange) communities. The name assigned for each community is the name of the gene with highest page rank value. **(B)** Communities plotted according to their chromosomal and expression assortativity values. Dot sizes correspond to the number of nodes in the community and node color, to the number of overrepresented GO terms. Communities with more than 20 terms are highlighted. Notice that the quadrant with more enriched communities is the one with high expression assortativity and low chromosomal assortativity.

number of intra-chromosomal *cis-* communities in the Luminal A network, the majority of communities with statistically significant biological processes associated are *trans-*. **Figure 2A** presents a visual representation in the form of an alluvial plot. There, the width of each line corresponds to the number of significantly enriched processes for a given community, named by the gene with highest page rank centrality. The difference in the amount of *cis-* and *trans-* communities with associated functions, may reflect that the set of biological processes annotated in GO do not tend to exhibit a bias toward an specific chromosome contrary to what it is observed in the Luminal A GCN communities.

There is a wide variety in the biological enriched processes in the Luminal A *trans-* communities. Processes associated with regulation of transcription, telomere maintenance, and regulation of cell division as well as gene silencing are found. **Supplementary Table 1** contains the entire set of significantly overrepresented processes for Luminal A and healthy GCNs, as well as the shared enriched terms between both networks.

On the other hand, the enriched Luminal A *cis-* communities are mainly composed of gene families located at the same regions in the genome. In this group we have the HOXA, HOXB, and HOXC genes, which are important for embryogenesis. They have been found to be expressed in normal and neoplastic breast tissue (Cantile et al., 2003), with altered patterns of expression levels in breast cancer molecular subtypes. In particular, HOXA genes in Luminal A subtype, have shown underexpression associated with the acquisition of repressive epigenetic marks, such as hypermethylation (Novak et al., 2006; Kamalakaran et al., 2011; Hur et al., 2014).

Protocadherins (PCDHA, PCDHB, and PCDHG genes) were also identified as three distinct *cis*- communities in the Luminal A network. Protocadherin genes were previously identified as the most densely connected component (almost a clique) in a breast cancer network (Espinal-Enríquez et al., 2017). There, it was also shown that all protocadherins resulted underexpressed. The observed underexpression of this cluster coincides with a reported hypermethylation of protocadherins in breast cancer (Novak et al., 2008).

In the Healthy network 41% of the *trans*- communities were associated with biological processes, and no *cis*- communities where enriched due to the fact that *cis*- communities identified in this network have <5 genes (the threshold set for the overrepresentation analysis, see Methods). The set of terms includes mostly metabolism-associated process, cell division, and mitochondrial functions.

The Healthy and the Luminal A GCN share 24 communities of only two nodes. Additionally, there is one community named HLA-DRB1 in the Healthy GCN, and HLA-DMB in the Luminal A GCN, with a Jaccard Index of 0.916. This community is associated with activation of the immune response, and it is composed by MHC class II HLA genes located on chromosome 6 region p21.32, plus CIITA (Class II Major Histocompatibility Complex Transactivator), on Chromosome 16, and CD74, located on chromosome 5, only in the Luminal A community.

One pair of communities named CPA3 in both networks share the set of associated processes, but displays a Jaccard index of 0.705 regarding their gene sets. Processes include peptide hormone processing and regulation of systemic arterial blood pressure. Members of this community, such as TPSAB1, CMA1, CTSG, CPA3, HDC, and MS4A2, are commonly found in Mast Cells expression, part of the immune response and usually recruited to breast tumors (Aponte-López et al., 2020). The presence of these immune-system associated communities as high co-expression sets in both networks might be an instance of multiple cell types present in the sample.

2.3. *trans*- Communities in the Luminal A Network Present Different Patterns of Differential Expression

Once we observed that biological processes were significantly associated with *trans*- communities, a differential expression analysis was performed to assess the influence of altered gene expression in *trans*- communities and their processes. **Supplementary Figure 1** presents the differential expression representation in the GCN and **Supplementary Table 2** contains the log2 fold change (LFC) values for each gene in the network.

The number of links joining genes with the same sign of LFC, minus the number of links between genes with different sign of LFC, over the total number of links, was computed per community as a measure of differential gene expression assortativity (ASS_{dge}). **Figure 2B** plots ASS_{dge} and ASS_{chr} for *trans*- communities, as well as the number of associated GO terms. Highly enriched communities (>20 GO terms) are highlighted. The majority of these communities are placed in the first quadrant of the plot, meaning that their genes tend to have similar differential expression but they are placed in different chromosomes. Moreover, those communities are not in the top-10 regarding size, hence functional association in *-trans* communities appears to be influenced by high ASS_{dge} and low ASS_{chr} values.

The community with the highest number of enriched GO terms is the NUSAP1 community which also contains highly overexpressed genes only (**Figure 3A**). Its enriched terms are associated with nuclear division, DNA replication, chromatid segregation, and cell cycle checkpoints, i.e., cell division processes. This community shares a Jaccard index of 0.5 regarding gene members and 0.718 regarding GO associated terms with the MKI67 community in the Healthy network.

NUSAP1 has already been identified as a hub gene in a network of ER positive breast cancer tumor tissues of patients treated with tamoxifen, and derived from a similar methodology but using micro-array data (Liu et al., 2015). In that study, five hub genes with high expression levels strongly associated with poor survival were identified, and four of them: CDK1, DLGAP5, NUSAP1 and RRM2, belong to this particular community.

High expression of several genes in this community, including NUSAP1, was also observed in patients with Luminal A breast cancer and obesity (Nuncia-Cantarero et al., 2018). Nuncia-Cantarero et al. reported 39 genes related with a poor outcome group for patients with both conditions and 26 are found in this community, including FOXM1 (Forkhead box proteinM1), a transcription factor that has been identified as a potential therapeutic target for breast cancer (Lu et al., 2018), highly associated with luminal tumors and ER expression (Millour et al., 2010; Carr et al., 2012).

Table 2 shows the 39 genes reported in Nuncia-Cantarero et al. (2018). The coincident genes found in our network community are bold and their corresponding log2 fold change values are displayed. Interestingly, none of the genes presented in Nuncia-Cantarero et al. (2018) are in the Luminal A GCN but those found in the NUSAP1 community.

From the highly enriched communities, RPL35 is the one with more genes. The majority of them are ribosomal proteins; therefore, among the enriched GO terms we find ribosome biogenesis, large and small ribosomal subunit assembly, as well as regulation of ubiquitin-protein transferase activity. Riboproteins in this community are mostly underexpressed (**Supplementary Figure 2**). Low levels of expression have been reported in breast cancer for RPL5 and RPL11, associated with a mechanism of apoptosis inhibition through P53 degradation (Tong et al., 2020), and induction of proliferation in MCF7 cells, a Luminal A-derived cell type (Fancello et al., 2017). It has been shown that riboproteins have high co-expression values in other gene co-expression networks (Prieto et al., 2008; Wang et al., 2020a,b). The finding of highly co-expressed cluster of riboproteins reported here, reinforces the fact that these GCNs are coherent and represent with some accuracy the actual co-expression landscape in Luminal A breast cancer.

To our knowledge, coordinated underexpression of ribosomal genes in a breast cancer subtype has not previously been described. On the contrary, an increased ribosomal content has been recently found to contribute to proliferative and

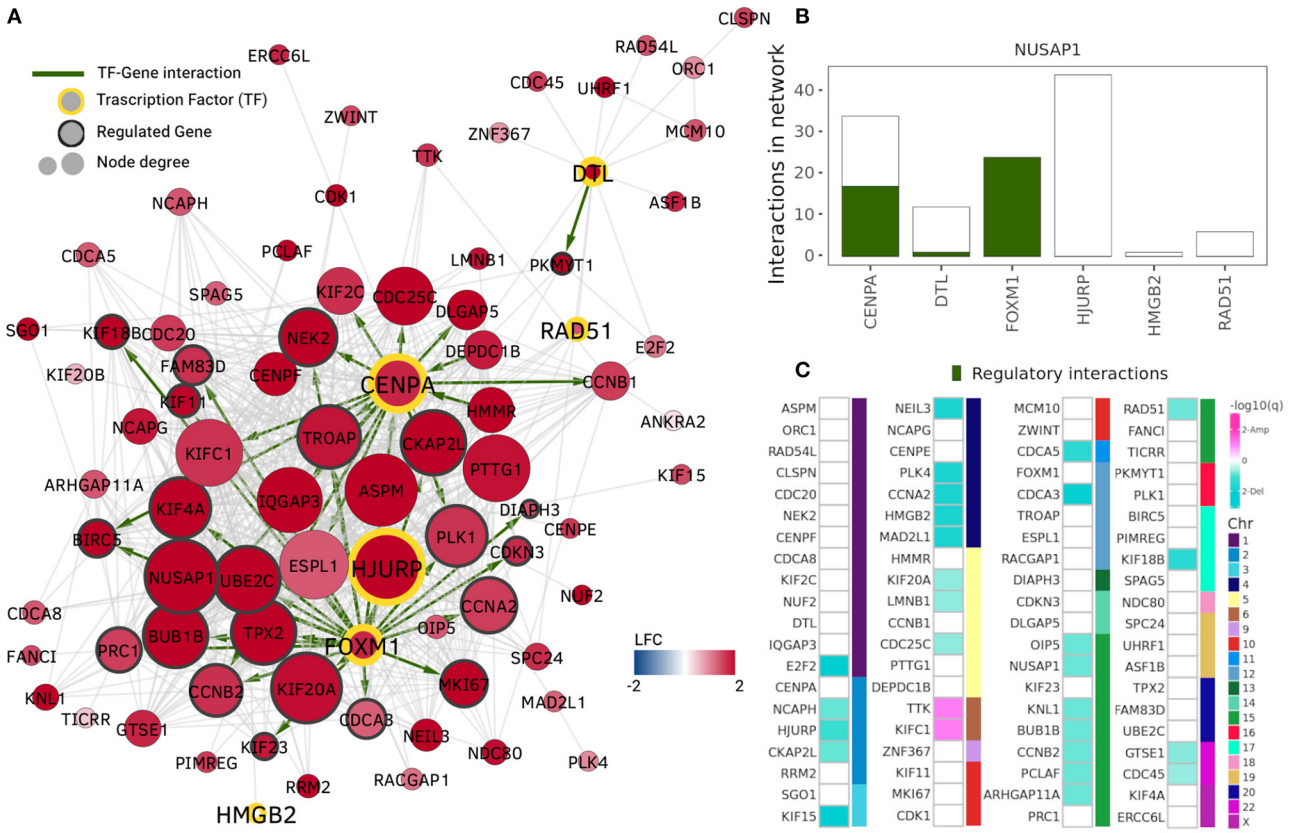

FIGURE 3 | NUSAP1 community. The NUSAP1 community is the community with the highest number of enriched terms and highest expression assortativity. **(A)** Nodes and edges in the NUSAP1 community. Node colors represent log2 fold change, thus, the entire community is overexpressed. Transcription factors (TF) are highlighted by a yellow border and their regulated genes (genes with at least one TF binding site in GTRD) are identified by a gray border. Green edges indicate regulatory interactions. **(B)** Transcription factors in the NUSAP1 community. Total number of network interactions for each TF. The number of interactions identified as regulatory is displayed in green. **(C)** Copy number alteration peaks in the NUSAP1 community. Squares represent each gene in the community and they are ordered according to the chromosome where they are located. Turquoise squares depict genes in deletion peaks while pink ones represent amplification events.

metastatic potential in breast cancer circulating tumor cells (Ebright et al., 2020). This discrepancy may be due to the fact that the overexpression of RPL transcripts, such as RPL15 observed in Ebright et al. (2020), was reported for circulating tumor cells. These tumor cells present additional alterations in their transcriptional profile, and they have acquired a highly proliferative capacity. Hence, the underexpression of ribosomal genes in the Luminal A network may be an indicative that the tumors are not as invasive as other subtypes. It is worth noticing again that Luminal A breast cancer subtype is the less aggressive, the one with the best prognostic and also the best in terms of response to therapy.

2.4. Effects of Transcription Factors and CNAs in *trans*- Communities

The general overexpression trend observed in the NUSAP1 community, and underexpression in the RPL35 module, suggested a contribution of altered mechanisms of transcriptional regulation promoting the formation of high co-expression clusters. To evaluate this, we analyzed the contribution of regulatory interactions from transcription factors (TFs) and the presence of deletion and amplification peaks in the Luminal A network communities.

TFs in the ten highlighted communities from **Figure 2B** were identified using data from the Gene Transcription Regulation Database (GTRD) (Yevshin et al., 2018). Five communities included at least one gene reported as TF in GTRD. The total number of interactions for these genes in the NUSAP1 community is presented in **Figure 3B**, where the number of genes having at least one binding site in the promoter region (1,000 bp upstream, 100 bp downstream from starting point) is shown in green. It can be observed that FOXM1 transcription factor has its entire set of adjacent links marked as regulatory interactions.

As stated in the previous section, the NUSAP1 community contains interactions that have been reported in luminal associated breast cancer phenotypes. Particularly, the FOXM1 transcriptional network was identified as the largest regulon by GPU-ARACNE, the accelerated parallel implementation of ARACNE, the algorithm used here to infer the gene co-expression networks (He et al., 2017). He et al. identified 121 FOXM1 interactions with 14 experimentally validated targets.

TABLE 2 | Previously reported genes in the NUSAP1 community.

Gene	Gene name	LFC
NEK2	**Serine/threonine-protein kinase Nek2**	3.564
KIF4A	**Kinesin Family Member 4**	3.098
ASPM	**Abnormal spindle-like microcephaly-associated protein**	2.567
CENPF	**Centromere protein F**	2.567
TPX2	**Protein TPX2**	2.567
KIF18B	**Kinesin Family Member 18B**	2.396
CDC25C	**M-phase inducer phosphatase**	2.316
DLGAP5	**Disks large-associated protein 5**	2.297
NUSAP1	**Nucleolar and spindle-associated protein**	2.223
MKI67	**Proliferation marker protein Ki-67**	2.191
UBE2C	**Ubiquitin-conjugating enzyme E2**	2.173
HMMR	**Hyaluronan mediated motility receptor**	2.162
BUB1B	**Mitotic checkpoint serine/threonine-protein kinase**	2.157
BIRC5	**Baculoviral IAP repeat-containing protein**	2.057
CDK1	**Cyclin-dependent kinase**	2.012
KIF11	**Kinesin Family Member 11**	1.963
RRM2	**Ribonucleoside-diphosphate reductase subunit M2**	1.961
KIF20A	**Kinesin Family Member 20**	1.898
ISG15	**Ubiquitin-like protein ISG15**	1.789
GTSE1	**G2 and S phase-expressed protein**	1.714
FOXM1	**Forkhead box protein M1**	1.699
CCNB2	**G2/mitotic-specific cyclin-B2**	1.621
CCNB1	**G2/mitotic-specific cyclin-B**	1.523
PRC1	**Protein regulator of cytokinesis**	1.504
KIF15	**Kinesin Family Member 15**	1.425
ZWINT	**ZW10 interactor**	1.416
OIP5	**Protein Mis18-beta**	1.299
BUB1	Mitotic checkpoint serine/threonine-protein kinase BUB1	
CEP55	Centrosomal protein of 55 kDa	
EZH2	Histone-lysine N-methyltransferase EZH2	
GDP-15	Growth/differentiation factor 15	
KIAA0101	PCNA-associated factor	
MELK	Maternal embryonic leucine zipper kinase	
MMP1	Matrix Metallopeptidase	
MYBL1	MYB Proto-Oncogene Like	
PBK	PDZ Binding Kinase	
RIPPLY3	Protein ripply3	
TOP2A	DNA topoisomerase 2-alpha	
TYMS	Thymidylate synthase	

39 Genes reported in Nuncia-Cantarero et al. (2018), related with poor outcome group for patients with obesity and Luminal A breast cancer. Highlighted genes are present in the NUSAP1 community. Their corresponding log2 fold change value is also displayed. Notice that all concordant genes are overexpressed.

In the NUSAP1 community, FOXM1 has 24 co-expression interactions with other genes in the module. All of these interacting genes contain a FOXM1 binding site in their promoter region according to the data gather by GTRD. From these 24 regulated genes, eight intersect with the experimentally validated targets reported in He et al. (2017).

Centromere protein A or CENPA, is another important transcription factor with overexpression in the NUSAP1 community. It regulates centromere integrity and chromosome segregation. This TF was identified in a mRNA signature correlated with lower survival ratio in Luminal A breast cancer (Xiao et al., 2018). One of its interacting proteins, HJURP, required for CENPA centromeric localization, is also a member of this community. HJURP mRNA expression level has been significantly associated with estrogen and progesterone receptor, and reported as clinically relevant for Luminal A breast cancer patients (Hu et al., 2010; Montes de Oca et al., 2015). Although HJURP is the transcription factor with more adjacent links in the NUSAP1 community, none of them was identified as a regulatory interaction; instead, HJURP was identified as regulated by FOXM1.

The remaining overexpressed TFs in the NUSAP1 community have also been found to play a role in the luminal breast cancer phenotype. Increased mRNA expression of RAD51, a gene in the double-strand breaks repair pathway, is associated with higher risk of tumor relapse and distant metastases in estrogen receptor positive breast cancer tumors (Barbano et al., 2011; Nieto-Jiménez et al., 2017). Overexpression of DTL and HMGB2 has also been associated with tumor progression in breast cancer (Perez-Peña et al., 2017; Fu et al., 2018), and resistance to endocrine therapies (Redmond et al., 2015). These results suggest a strong contribution of TFs, particularly from FOXM1 and CENPA, and their interactions found in the NUSAP1 community, to the process of tumorigenesis and progression in Luminal A breast cancer.

Gene copy number alteration (CNA) is a common trait of genomic instability in cancer and their presence has therapeutic relevance in breast cancer, specially for the Her2 enriched subtype (Andre et al., 2009; Inaki et al., 2014). Different levels of correlation have been identified between DNA amplification and deletion events, mRNA, and protein expression values in breast cancer, (Myhre et al., 2013), showing that it is not an homogeneous mechanism of altered expression. However, given the possible effect and importance for the breast cancer phenotype, amplification and deletion peaks may play a role in the formation of high co-expression clusters in the Luminal A network. For instance, in the case of breast cancer, correlation between CNVs and gene expression could reach until 25% (Lachmann, 2016).

Those gene expression alterations may influence importantly in the co-expression landscape. In Lachmann (2016), it was reported that CNVs may impact importantly the co-expression program, in particular for transcription factor targets.

To evaluate the role of CNVs in the Luminal A GCN, we obtained amplification and deletion peaks using the GISTIC2 algorithm (Mermel et al., 2011). **Figure 3C** presents the results for the NUSAP1 community. Turquoise squares represent genes in which a deletion has been observed, meanwhile amplifications are depicted in pink squares. Since the NUSAP1 community is *trans*-, the chromosome in which those genes are located is also depicted.

As observed, the majority of genes with copy number alterations correspond to deletions. Only two genes, TTK and KIFC1 (Chr6) present amplifications. However, 52 out of 80 genes do not present changes in copy number. This result shows

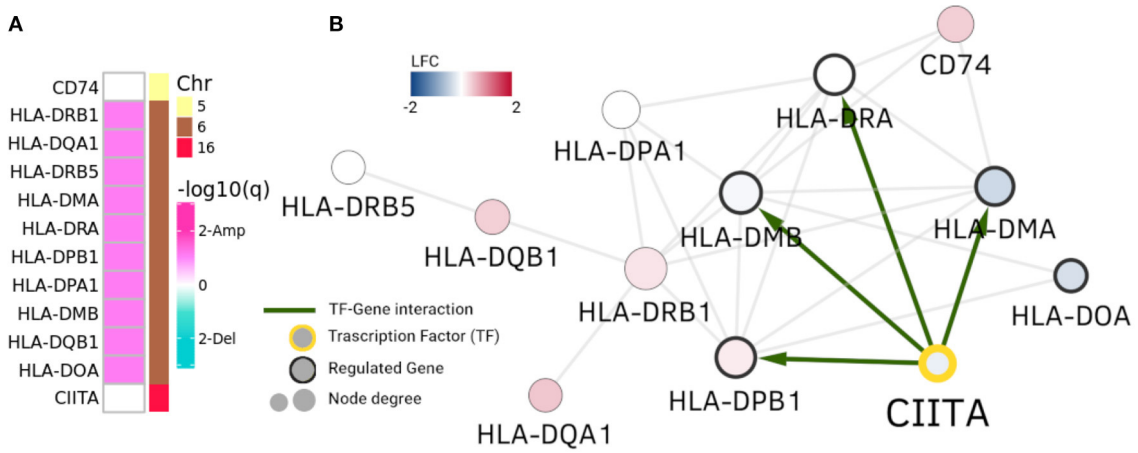

FIGURE 4 | HLA-DRB1 community. In this picture, analogue to **Figure 3**, HLA-DRB1 community is depicted. Panel **(A)** displays amplification/deletion peaks in genes in the community and panel **(B)** shows differential expression and regulatory interactions. Genes in Chr6 (MHC class II genes) present amplifications. However, their differential expression is neither uniform nor with the same sign.

that, at least for the NUSAP1 community, which is the one with the most differentially expressed genes, CNAs do not significantly influence neither expression nor co-expression patterns.

However, in the case of HLA-DRB1 community in **Figure 4**, we observe the opposite phenomenon: genes are not differentially expressed, but the ones that are placed in Chr6 belong to a clearly amplified region. This cluster is composed of MHC class II HLA genes. Interestingly, CIITA gene is a TF that regulates some of these human leukocyte antigen genes. As it can be observed in **Figure 4**, four of these genes have a CIITA binding site in their promoter region.

In this case CNAs and the CIITA regulation appear to exert a concomitant action with the observed copy number alterations to generate the community of MHC class II genes, independently of their differential expression. It is worth mentioning that CIITA (Class II Major Histocompatibility Complex Transactivator) is located at Chromosome 16, but clearly regulates the transcriptional and functional characteristics of HLA genes. The same representation for the RPL35 community is shown in **Supplementary Figure 2**. It is worth to stress that the HLA-DRB1 community in Luminal A GCN is almost identical to a community of the healthy GCN (Jaccard index = 0.916).

2.5. *cis*- Communities Are Enriched With Deletion Peaks

The presence of deletion and amplification peaks, and their effect in gene altered expression was also evaluated for *cis*- communities. **Figure 5** presents the results of an overrepresentation analysis where GISTIC2 peaks were analyzed. As it can be observed, communities are mostly enriched with deletion peaks, and their effect in the average log2 fold change in *cis*- communities varies. **Supplementary Figure 3** presents the entire set of alterations in these communities.

The pattern of amplification in the q arm of chromosome 1 and deletion in chromosome 16q, previously reported in a subset of Luminal A tumors (Ciriello et al., 2013) is also observed here. However, no other alteration matched that particular study. Luminal A tumors tend to have the lowest frequency of CNAs among breast cancer subtypes (Gatza et al., 2014), and as evaluated by our methodology, amplification and deletion peaks do not *a priori* determine the formation of *cis*- communities.

It is important to mention that copy number alterations are a key element affecting the gene expression of large sections of the genome (Freeman et al., 2006; Redon et al., 2006; McCarroll and Altshuler, 2007), specially in cancer (Shlien and Malkin, 2009; Lachmann, 2016; Shao et al., 2019). A large part of a chromosome being altered by a gain or loss of copy number, will trigger an equally abrupt change in several genes along that portion of the genome.

2.6. *cis*- Communities Are Not Bound by CTCF Binding Sites

The three-dimensional structure of DNA is another regulator of gene expression in eukaryotic cells. Regions with active transcription are characterized by open chromatin, whereas closed chromatin indicates regions of inactive or repressed transcription (Achinger-Kawecka et al., 2016; Corces and Corces, 2016). Furthermore, the regulatory effect of regions, such as enhancers and promoters, usually requires the formation of long distance chromatin loops that bring together distant genomic loci. These loops are maintained and regulated by architectural proteins, such as CTCF and cohesin, among others (Achinger-Kawecka and Clark, 2017; Pugacheva et al., 2020). Given the fact that CTCF proteins are able to modify the chromatin landscape, they may be underlying the appearance of a large amount of *cis*- communities in breast cancer.

To evaluate the role of CTCF in the appearance of *cis*- clusters of genes in the Luminal A breast cancer gene co-expression network, we calculated the number of CTCF binding sites at the boundaries of *cis*- communities. This was done using a previously

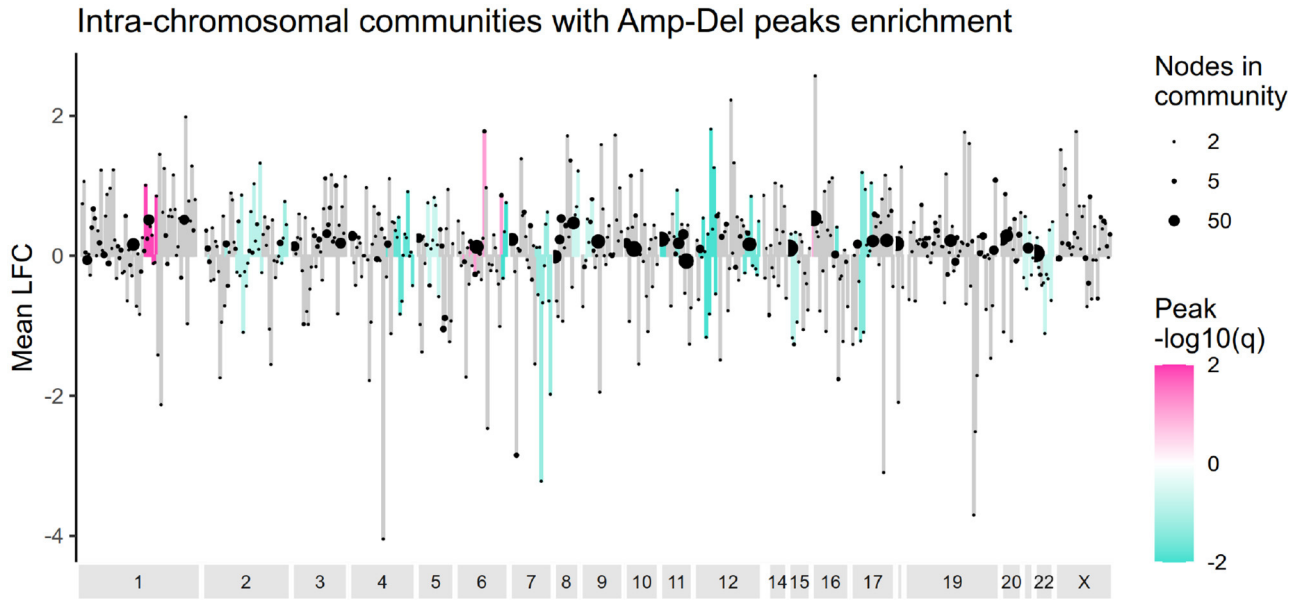

FIGURE 5 | Amplification and deletion peaks overrepresentation in *cis*- communities. Intra-chromosomal communities are ordered according to the start of their first gene and the mean log2 fold change value for each community is plotted in the *y* axis. Communities enriched with amplification or deletion peaks are colored in pink or turquoise, accordingly. Ending dots indicate size of each community.

reported dataset containing Chip-seq peaks in MCF7 cells, a Luminal A breast cancer cell line (Fiorito et al., 2016).

The number of binding sites in a window of 50k base pairs before the first gene and after the last one in a community was compared to the average number of binding sites in same size windows spanning the community region (see Methods). The distribution of these binding sites is shown in **Supplementary Figure 4**. No significant difference was found in the distribution of the number of binding sites in the boundaries and the middle sections of the communities. Actually, out of the 416 *cis*- communities with at least one CTCF binding site associated, only 197 had more binding sites at the boundaries than in middle regions.

2.7. Loss of Long-Distance Co-expression Does Not Depend on the Correlation Measure

We decided to construct GCNs for Luminal A and healthy phenotypes using Pearson correlation, to observe whether the phenomenon of loss of long-distance co-expression was maintained using other correlation measure. The results can be observed in the form of a heatmap in **Figure 6**. There, genes are placed according to its position in the chromosome. The color of the heatmap is proportional to the correlation value. The results show that, as observed with mutual information-inferred networks, the highest correlation values occur between genes from the same chromosome.

Additionally, it can also be appreciated that the Pearson correlation values are in general higher in the healthy matrix than in the Luminal A breast cancer one (except for those values close to the diagonal, which represent *cis*- interactions).

2.8. Loss of Long-Distance Co-expression Does Not Depend on the MI Threshold Value

Setting a threshold on the weight of edges so as to discard edges with strength less than a certain value is a well-known open problem in graph theory and network science. Determination of this threshold can be made by choosing among a number of methods. For instance, if an accurate measure of the signal-to-noise ratio in the correlations of the data under consideration can be obtained, one possible way to set the threshold is by allowing all edges valued above the noise-level. In most practical applications, however, this is not feasible.

To overcome this situation, we presented a comparison of *cis/trans* proportion in both networks. For this purpose, we constructed networks with different threshold values, ranging from the top-1,000 to the top-1,000,000 higher edges (**Figure 7**). As it can be appreciated in the figure, the proportion of *cis*-interactions is always higher in Luminal A network than in the healthy GCN.

Additionally, to assess the influence of the MI threshold value in the phenomenon of loss of long-distance co-expression in Luminal A breast cancer, we observed the distribution of MI values in both networks. We constructed (a) the histograms of all interactions (20,217) in both networks, (b) the histograms for only *cis*- interactions, and (c) the histogram for *trans*- edges in both phenotypes (**Figure 8**). There, it can be observed that

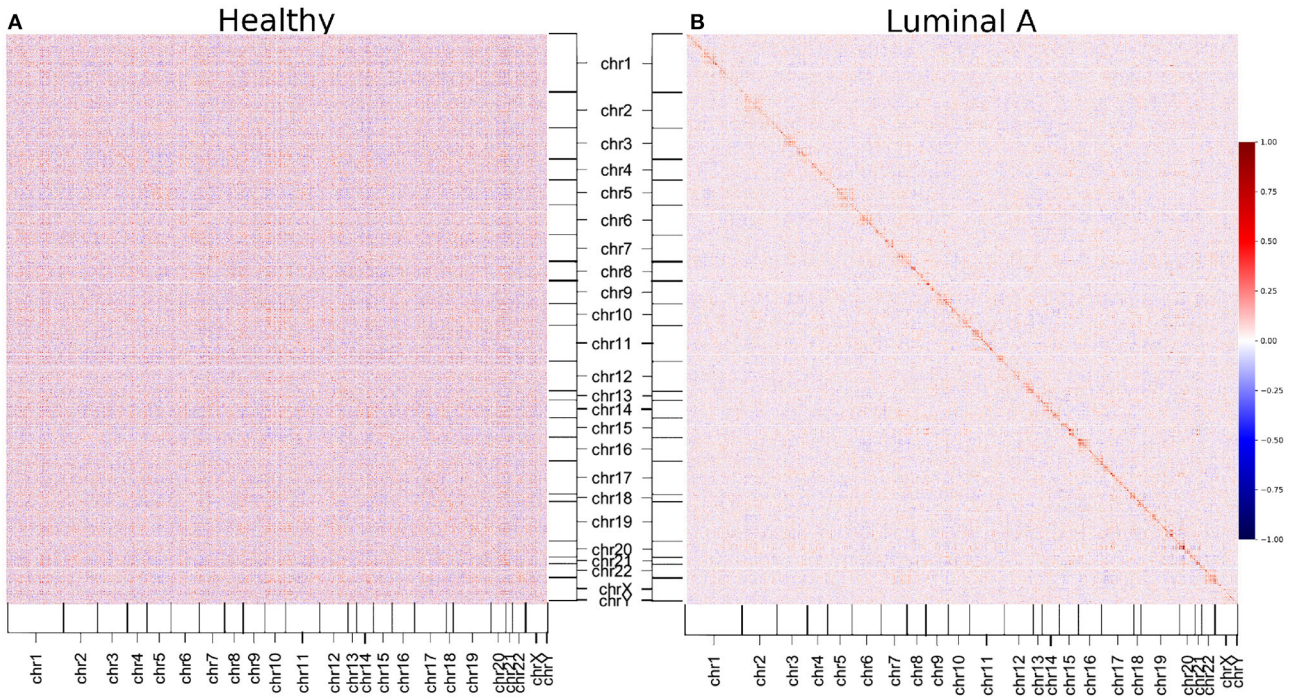

FIGURE 6 | Pearson correlation matrices of healthy and Luminal A phenotypes. Correlation between all gene couples of each phenotype are depicted. The color code corresponds to the correlation value. **(A)** Healthy matrix. **(B)** Luminal A matrix. As in the case of mutual information-derived networks, higher correlation values in Luminal A occur between genes from the same chromosome (close to the diagonal).

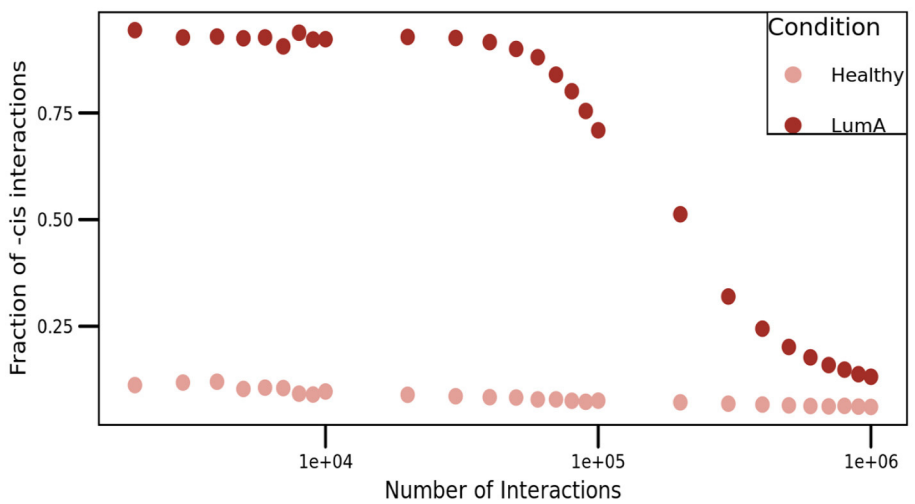

FIGURE 7 | Proportion of *cis*- interactions at different network sizes. This figure shows the fraction of intra-chromosome interactions (Y-axis) for healthy (pale pink), and Luminal A (brown) GCNs. X-axis represents the number of edges in each network, ranging from the top-1,000 to the top-1,000,000 links, i.e., three orders of magnitude.

independently of the threshold, healthy interactions have higher MI values.

The above mentioned result coincides with the one presented in the matrices of **Figure 6**. Correlation values (independent on the correlation measure), are in general higher in the healthy phenotype than in cancer, but for a subset intra-chromosome interactions.

Complementarily, in **Figure 8** we inserted a zoom of those histograms in the higher MI value region (0.3–0.7). There, it is shown that for *cis*- interactions, the Luminal A network has

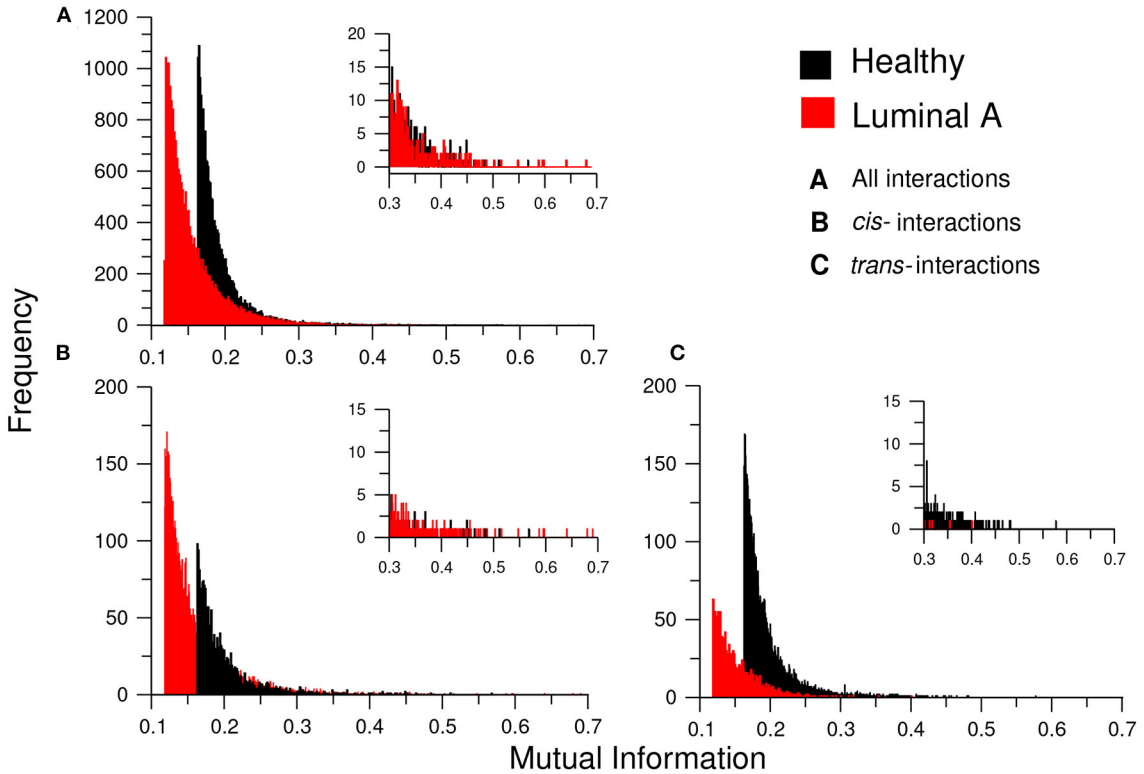

FIGURE 8 | Distribution of MI values in the GCNs. This plot shows the histograms for the MI values of the healthy (black) and Luminal A (red) GCNs. **(A)** The total of MI values. **(B)** Only *cis-* edges, **(C)** Only *trans-* interactions. Each histogram also contains an inset with a zoom of the highest interactions for each condition. Notice the absence of *trans-* interactions in the Luminal A case in the inserts of **(B,C)**; this reflects the loss of *trans-* co-expression in the cancer GCN.

more and higher interactions in the highest values; conversely, for the *trans-* interactions, the higher and more abundant links are observed in the healthy phenotype.

We have shown previously that the threshold value is not determinant to observe the loss of long-distance co-expression in other clear cell renal carcinoma (Zamora-Fuentes et al., 2020), as well as in lung cancer (Andonegui-Elguera et al., 2021). We have demonstrated for these cancer GCNs that the particular value of the threshold, affects the size and sparsity of the networks as expected. However, the proportion of inter- and intra-chromosomal links remains largely unchanged.

2.9. Implications of Network Topology in the Context of Luminal a Breast Cancer

We have shown that in Luminal A breast cancer, the already mentioned *loss of trans- co-expression* is not as strong as in other breast cancer subtype GCNs, but the effect is perceived. Actually, several *trans-* interactions appear in the top co-expressed pairs. Luminal A GNC topology allows us to:

- identify functional communities (mostly *trans-*)
- differentiate enriched functions between healthy and cancer GCNs
- observe mechanisms that may influence the appearance of this loss of long distance co-expression
- observe specific differential expression patterns depending on the community

The identification of significant biological processes, associated with particular sets of highly co-expressed genes is one of the most relevant improvements of using network topology to analyze the functional implications of RNA-Seq-based genome-wide multi-sample sets for a given phenotype. The use of network communities improves the specificity of the enrichment analysis over using the whole genome or using differentially expressed genes.

The number of enriched processes in *cis-* communities is significantly lower than the ones associated with *trans-* communities, given the total number of communities for each type. However, the functions that are significant for *cis-* communities, are also relevant for cell maintenance. For instance, HOXA community, whose genes are relevant for organism development. These genes are found together in chromosome 7p15.2, and they are all underexpressed. Analogously, the protocadherin cluster is found to be related to cell adhesion, which is one of the non-shared processes between Luminal A GCN and the healthy GCN (**Supplementary Material 1**).

From the alluvial diagram of **Figure 2** it can be observed that out of the 11 enriched *cis-* communities, 6 correspond to HOX and protocadherin clusters. This could be an indicative of the

importance of the conjugated action that these set of genes may have for the phenotype. Additionally, these clusters appear with the same differential expression trend.

3. CONCLUDING REMARKS

Based on the previous analysis, we may conclude that for the establishment of the regulatory program observed in the Luminal A subtype gene co-expression network, compared with the healthy GCN, several DNA modifications and regulatory elements must participate. DNA modifications (copy number alterations, transcription factor regulation, CTCF binding sites) should exert, to at certain extent, influence over the gene co-expression interactions. Additionally, differential gene expression is a relevant element to take into account, specially for *trans-* communities. We can establish that, for the manifestation of the *loss of trans- co-expression* in cancer it is not only necessary to observe separately differential gene expression, transcription factor regulation, CNAs, or CTCF binding sites, but to take them all into account.

Other regulatory elements should also participate in modifying the co-expression patterns between a healthy and a cancer co-expression network: micro-RNA regulation (Drago-García et al., 2017; de Anda-Jáuregui et al., 2018), topologically associated domains and their boundaries (Rafique et al., 2015; Achinger-Kawecka et al., 2020; Khoury et al., 2020), long non-coding RNAs (Hung et al., 2011; Zhang et al., 2019), the methylation profiles (Paz et al., 2003; Hernández-Lemus et al., 2019), among others, might delineate these imbalance between *cis-* and *trans-* genetic relationships.

More investigation regarding the aforementioned elements is also important in order to have an integral picture of the regulatory landscape in the cancer genome, and provide hypotheses that could explain the phenomenon of loss of long distance genetic interactions in cancer.

It is likely plausible that the *loss of trans- co-expression* observed in breast cancer (and breast cancer molecular subtypes) responds to a physical/mechanical principle in which the transcriptional machinery is somehow altered. Recently, we have observed the loss of long distance co-expression in clear cell renal carcinoma (Zamora-Fuentes et al., 2020), and in lung adenocarcinoma, as well as in squamous cell lung cancer (Andonegui-Elguera et al., 2021).

The ubiquity of this disruption of the *normal* transcriptional landscape led us to hypothesize that the physical principle behind this global alteration is the same in all of these cancer tissues. The consistency and relevance of this loss could be considered as a possible emergent hallmark of cancer. Further investigation toward this particular issue must be achieved beforehands, however, further investigation is required.

4. METHODS

4.1. Databases

Gene expression values for Luminal A and Healthy samples were retrieved from our previous publication (García-Cortés et al., 2020), with RNA-seq data obtained from The Cancer Genome Atlas (TCGA) breast invasive carcinoma dataset (Tomczak et al., 2015), downloaded from the Genomic Data Commons (GDC) Data Portal. The GDC Data portal case identifiers for Luminal A were use to download "Masked Copy Number Segment Files" for the GISTIC2 pipeline. The Chip-seq data was downloaded from the Gene Expression Omnibus dataset GSE85106 (Fiorito et al., 2016), and only the control sample for CTCF was used. The Homo sapiens genes promoter dataset from the Gene Transcription Regulation Database (GTRD) (Yevshin et al., 2018) was used to identify transcription factors and their regulatory interactions.

4.2. Data Processing

As detailed in García-Cortés et al. (2020), 113 samples for Healthy tissue and 1,102 cancer samples were acquired and pre-processed to log_2 normalized gene expression values. After applying the PAM50 algorithm using the Permutation-Based Confidence for Molecular Classification (Fresno et al., 2017) as implemented in the `pbcmc` R package (Fresno et al., 2016), and multidimensional noise reduction using `ARSyN` R implementation (Nueda et al., 2012), 217 samples for Luminal A breast cancer were identified.

The "Masked Copy Number Segment Files" were downloaded from GDC and integrated into one segmentation file to run `gistic2` (Mermel et al., 2011). The parameters suggested in the Copy Number Variation Analysis Pipeline from GDC and the GDC reference sequence, and markers file were used. The identified amplification and deletion regions in the lesions output file with 0.99 confidence were re-mapped to keep genes spanned entirely by peaks.

4.3. Network Construction

The `ARACNE` (Margolin et al., 2006) algorithm was used to calculate mutual information (MI) to quantify statistical dependence between pairs of genes. The method associates a significance value (p-value) to each MI value based on permutation analysis, as a function of the sample size. Only the highest interactions in terms of their statistical significance ($P \leq 1e^{-8}$) were kept for further analysis. The total number of interactions in the Luminal A and the Healthy network were reduced to 20,127, the number of significant interactions in the Healthy network.

4.4. Community Detection and Assortativity Calculation

Four community detection algorithms were evaluated: Fast Greedy (Clauset et al., 2004), Infomap (Rosvall and Bergstrom, 2008), Leading Eigenvector (Newman, 2006), and Louvain (Blondel et al., 2008; Rahiminejad et al., 2019). MI values were used as link weights. Their implementation in the `igraph` (Csardi and Nepusz, 2006) R package was used. Algorithm results were compared using the Jaccard index, a coefficient that measures similarity between two finite sets, defined as the size of their intersection divided by the size of their union. Genes in a community constitute a set and all communities identified by one algorithm were compared against communities identified by another one. The same approach was used to

TABLE 3 | CTCF binding sites location classification.

	Promoter	Gene body	Intergenic region
Dataset	868	8,047	11,438
In Luminal A network	177	1,343	887

compare the set of GO terms associated per community in the overrepresentation analysis.

$$J(C_1, C_2) = \frac{(C_1 \cap C_2)}{(C_1 \cup C_2)} \quad (1)$$

To calculate chromosomal assortativity, the chromosome location for each gene was used. For each community, the number of links joining genes in the same chromosome (-*cis* links) minus the number of links joining genes in different chromosomes (-*trans* links), was divided by the total number of links in the community. Expression assortativity was calculated in the same manner, using the log2 fold change sign to classify genes into overexpressed or underexpressed as the assortativity attribute.

$$ASS_{chr} = \frac{|\{\{x,y\} \mid x,y \in C_i \text{ and } x.\text{chr } = y.\text{chr }\}| - |\{\{x,y\} \mid x,y \in C_i \text{ and } x.\text{chr } \neq y.\text{chr }\}|}{|\{\{x,y\} \mid x,y \in C_i\}|}$$

C_i = community i in network.

4.5. Overrepresentation Analysis

The `enrichGO` function from the `clusterProfiler` (Yu et al., 2012) R package was used to identify over-represented or enriched terms in the Biological Process category in Gene Ontology (GO). Enrichment was performed for communities with five or more genes and GO terms with a minimum size of ten were retained. Genes in the original expression matrix defined the universe set. Terms with adjusted p-value below 0.005 using the Benjamini and Hochberg method for multiple testing were kept. The overrepresentation analysis for amplification and deletion peaks was conducted using the generic function `enricher` from the same package. The same universe set was used and no size threshold for communities or peaks was defined. An adjusted p-value of 0.05 was set as cutoff.

4.6. Differential Expression Analysis

Differential expression analysis was performed as described in (Espinal-Enríquez et al., 2017). The `limma` package (Ritchie et al., 2015) in `R` was used to determine overexpressed or underexpressed genes, by adjusting a gene based linear model. An absolute difference of log2 fold change $\geq$0.5 and a p-value < 0.05 was set as threshold.

4.7. Transcription Factors Identification

The entire set of gene promoters in the smallest region available, [−100, +10] base pairs from starting site was downloaded from the Gene Transcription Regulation Database (GTRD) (Yevshin et al., 2018). For the selected communities, gene members that matched transcription factors (TF) in GTRD were extracted and their neighboring genes were compared to the set of annotated genes that had at least one binding site from that TF in the ChIP-seq data.

4.8. CTCFs

We took the CTCFs in genes and promoters in the *cis*- Luminal A network communities that were not in other genes or promoters. For the Inter-regional CTCFs, we took the ones that were in a region <50k bps from the extreme of the promoter and the extreme of the gene.

Once filtered, the binding sites were classified according to their location. CTCFs in gene bodies, promoters (+1,000, −500 bps) and intergenic region were identifies. **Table 3** displays the classified binding sites for the complete dataset, as well as the binding sites present in genes comprising the Luminal A *trans*-communities. For the intergenic region, only CTCF binding sites in a window of 50k base pairs upstream the first gene and downstream the last one in *cis*- communities were kept.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Genomic Data Commons Data Portal https://portal.gdc.cancer.gov/, Gene Expression Omnibus dataset GSE85106 Gene Transcription Regulation Database, http://gtrd.biouml.org/, Relevant data used, and the scripts to generate the results and figures can be found in the following repository: https://github.com/ddiannae/luma.

AUTHOR CONTRIBUTIONS

DG-C performed the computational analyses, developed and implemented the programming code, performed the pre-processing and low-level data analysis, made the figures, drafted the manuscript. EH-L contributed to the theoretical analysis, co-supervised the project, contributed to the writing of the manuscript. JE-E conceived and designed the project, co-supervised the project, discussed the results, drafted the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by CONACYT (558985 student grant to DG-C, 285544/2016, and 2115/2018 to JE-E), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). Additional support has been granted by the National Laboratory of Complexity Sciences (232647/2014 CONACYT). JE-E was recipient of the 2018 Miguel Alemán Fellowship in Health Sciences. EH-L was a recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

DG-C was a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autnoma de México (UNAM). This work is part of her Ph.D. Thesis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.629475/full#supplementary-material

Supplementary Material 1 | Results of community detection algorithms evaluation.

Supplementary Table 1 | List of Gene Ontology biological processes overrepresented in the Healthy and the Luminal A gene co-expression networks (GCN), as well as the shared enriched terms between both networks.

Supplementary Table 2 | List of genes in the Luminal A GCN and the Healthy GCN with their chromosomal location, associated log2 fold change (LFC) value, and corresponding community for each algorithm.

Supplementary Figure 1 | Differential expression in the Luminal A GCN. The NUSAP1 community is highlighted.

Supplementary Figure 2 | RPL35 community. Left panel presents amplification and deletion peaks identified by GISTIC2, through pink and turquoise squares. Genes are ordered according to their corresponding chromosome. Right panel displays differential expression and regulatory interactions in genes in the community.

Supplementary Figure 3 | Amplification and deletion peaks in *cis*- communities. Entire set of copy number alterations identified in intra-chromosomal communities. Genes are displayed according to their starting site.

Supplementary Figure 4 | CTCF binding sites distribution over *cis*- communities. Biding sites at a distance of no more than 50,000 base pairs from a gene in the community are displayed.

REFERENCES

Achinger-Kawecka, J., and Clark, S. J. (2017). Disruption of the 3D cancer genome blueprint. *Epigenomics* 9, 47–55. doi: 10.2217/epi-2016-0111

Achinger-Kawecka, J., Taberlay, P. C., and Clark, S. J. (2016). Alterations in three-dimensional organization of the cancer genome and epigenome. *Cold Spring Harbor Symp. Quant. Biol.* 81, 41–51. doi: 10.1101/sqb.2016.81.031013

Achinger-Kawecka, J., Valdes-Mora, F., Luu, P.-L., Giles, K. A., Caldon, C. E., Qu, W., et al. (2020). Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nat. Commun.* 11, 1–17. doi: 10.1038/s41467-019-14098-x

Alcalá-Corona, S. A., de Anda-Jáuregui, G., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network modularity in breast cancer molecular subtypes. *Front. Physiol.* 8:915. doi: 10.3389/fphys.2017.00915

Alcalá-Corona, S. A., de Anda-Jáuregui, G., Espinal-Enriquez, J., Tovar, H., and Hernández-Lemus, E. (2018a). "Network modularity and hierarchical structure in breast cancer molecular subtypes," in *International Conference on Complex Systems* (Cham: Springer), 352–358. doi: 10.1007/978-3-319-96661-8_36

Alcalá-Corona, S. A., Espinal-Enríquez, J., De Anda Jáuregui, G., and Hernandez-Lemus, E. (2018b). The hierarchical modular structure of HER2+ breast cancer network. *Front. Physiol.* 9:1423. doi: 10.3389/fphys.2018.01423

Alcalá-Corona, S. A., Velázquez-Caldelas, T. E., Espinal-Enríquez, J., and Hernández-Lemus, E. (2016). Community structure reveals biologically functional modules in MEF2C transcriptional regulatory network. *Front. Physiol.* 7:184. doi: 10.3389/fphys.2016.00184

Andonegui-Elguera, S. D., Zamora-Fuentes, J., Espinal-Enríquez, J., and Hernandez-Lemus, E. (2021). Loss of long-distance co-expression in lung cancer. *Front. Genet.* 12:192. doi: 10.3389/fgene.2021.625741

Andre, F., Job, B., Dessen, P., Tordai, A., Michiels, S., Liedtke, C., et al. (2009). Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin. Cancer Res.* 15, 441–51. doi: 10.1158/1078-0432.CCR-08-1791

Aponte-López, A., Enciso, J., Muñoz-Cruz, S., and Fuentes-Pananá, E. M. (2020). An *in vitro* model of mast cell recruitment and activation by breast cancer cells supports anti-tumoral responses. *Int. J. Mol. Sci.* 21:5293. doi: 10.3390/ijms21155293

Arvold, N. D., Taghian, A. G., Niemierko, A., Abi Raad, R. F., Sreedhara, M., Nguyen, P. L., et al. (2011). Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. *J. Clin. Oncol.* 29:3885. doi: 10.1200/JCO.2011.36.1105

Barbano, R., Copetti, M., Perrone, G., Pazienza, V., Muscarella, L. A., Balsamo, T., et al. (2011). High RAD51 mRNA expression characterize estrogen receptor-positive/ progesteron receptor-negative breast cancer and is associated with patient's outcome. *Int. J. Cancer* 129, 536–545. doi: 10.1002/ijc.25736

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* 33, 690–705.e9. doi: 10.1016/j.ccell.2018.03.014

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Cantile, M., Pettinato, G., Procino, A., Feliciello, I., Cindolo, L., and Cillo, C. (2003). *In vivo* expression of the whole HOX gene network in human breast cancer. *Eur. J. Cancer* 39, 257–264. doi: 10.1016/S0959-8049(02)00599-3

Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep.* 5:17386. doi: 10.1038/srep17386

Carr, J. R., Kiefer, M. M., Park, H. J., Li, J., Wang, Z., Fontanarosa, J., et al. (2012). FoxM1 regulates mammary luminal cell fate. *Cell Rep.* 1, 715–29. doi: 10.1016/j.celrep.2012.05.005

Ciriello, G., Sinha, R., Hoadley, K. A., Jacobsen, A. S., Reva, B., Perou, C. M., et al. (2013). The molecular diversity of luminal A breast tumors. *Breast Cancer Res. Treat.* 141, 409–420. doi: 10.1007/s10549-013-2699-3

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* 70:066111. doi: 10.1103/PhysRevE.70.066111

Corces, M. R., and Corces, V. G. (2016). The three-dimensional cancer genome. *Curr. Opin. Genet. Dev.* 36, 1–7. doi: 10.1016/j.gde.2016.01.002

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJ. Complex Syst.* 5, 1–9.

de Anda-Jáuregui, G., Espinal-Enríquez, J., Drago-García, D., and Hernández-Lemus, E. (2018). Nonredundant, highly connected micrornas control functionality in breast cancer networks. *Int. J. Genomics* 2018:9585383. doi: 10.1155/2018/9585383

de Anda-Jáuregui, G., Espinal-Enriquez, J., and Hernández-Lemus, E. (2019a). Spatial organization of the gene regulatory program: an information theoretical approach to breast cancer transcriptomics. *Entropy* 21:195. doi: 10.3390/e21020195

de Anda-Jáuregui, G., Fresno, C., García-Cortés, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019b). Intrachromosomal regulation decay in breast cancer. *Appl. Math. Nonlin. Sci.* 4, 217–224. doi: 10.2478/AMNS.2019.1.00020

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, 1–23. doi: 10.1007/s41109-020-00291-1

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of emt and met micro-RNA regulation in breast cancer. *Sci. Rep.* 7:13534. doi: 10.1038/s41598-017-13903-1

Ebright, R. Y., Lee, S., Wittner, B. S., Niederhoffer, K. L., Nicholson, B. T., Bardia, A., et al. (2020). Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. *Science* 367, 1468–1473. doi: 10.1126/science.aay0939

Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.* 5:15. doi: 10.3389/fgene.2014.00015

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. doi: 10.1038/s41598-017-01314-1

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., et al. (2006). Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* 355, 560–569. doi: 10.1056/NEJMoa052933

Fancello, L., Kampen, K. R., Hofman, I. J., Verbeeck, J., and De Keersmaecker, K. (2017). The ribosomal protein gene RPL5 is a haploinsufficient tumor suppressor in multiple cancer types. *Oncotarget*. 8, 14462–14478. doi: 10.18632/oncotarget.14895

Fiorito, E., Sharma, Y., Gilfillan, S., Wang, S., Singh, S. K., Satheesh, S. V., et al. (2016). CTCF modulates estrogen receptor function through specific chromatin and nuclear matrix interactions. *Nucleic Acids Res.* 44, 10588–10602. doi: 10.1093/nar/gkw785

Fortunato, S., and Hric, D. (2016). Community detection in networks: a user guide. *Phys. Rep.* 659, 1–44. doi: 10.1016/j.physrep.2016.09.002

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., et al. (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949–961. doi: 10.1101/gr.3677206

Fresno, C., González, G. A., Llera, A. S., and Fernández, E. A. (2016). *PBCMC: Permutation-Based Confidence for Molecular Classification*. R package version 1.2.

Fresno, C., González, G. A., Merino, G. A., Flesia, A. G., Podhajcer, O. L., Llera, A. S., et al. (2017). A novel non-parametric method for uncertainty evaluation of correlation-based molecular signatures: its application on PAM50 algorithm. *Bioinformatics* 33, 693–700. doi: 10.1093/bioinformatics/btw704

Fu, D., Li, J., Wei, J., Zhang, Z., Luo, Y., Tan, H., et al. (2018). HMGB2 is associated with malignancy and regulates Warburg effect by targeting LDHB and FBP1 in breast cancer. *Cell Commun. Signal.* 16, 1–10. doi: 10.1186/s12964-018-0219-0

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

Garraway, L. A., and Lander, E. S. (2013). Lessons from the cancer genome. *Cell* 153, 17–37. doi: 10.1016/j.cell.2013.03.002

Gatza, M. L., Silva, G. O., Parker, J. S., Fan, C., and Perou, C. M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat. Genet.* 46, 1051–1059. doi: 10.1038/ng.3073

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

He, J., Zhou, Z., Reed, M., and Califano, A. (2017). Accelerated parallel algorithm for gene network reverse engineering. *BMC Syst. Biol.* 11:83. doi: 10.1186/s12918-017-0458-5

Hernández-Lemus, E., Reyes-Gopar, H., Espinal-Enríquez, J., and Ochoa, S. (2019). The many faces of gene regulation in cancer: a computational oncogenomics outlook. *Genes* 10:865. doi: 10.3390/genes10110865

Hsu, H. M., Chu, C. M., Chang, Y. J., Yu, J. C., Chen, C. T., Jian, C. E., et al. (2019). Six novel immunoglobulin genes as biomarkers for better prognosis in triple-negative breast cancer by gene co-expression network analysis. *Sci. Rep.* 9:4484. doi: 10.1038/s41598-019-40826-w

Hu, Z., Fan, C., Oh, D. S., Marron, J., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96. doi: 10.1186/1471-2164-7-96

Hu, Z., Huang, G., Sadanandam, A., Gu, S., Lenburg, M. E., Pai, M., et al. (2010). The expression level of HJURP has an independent prognostic impact and predicts the sensitivity to radiotherapy in breast cancer. *Breast Cancer Res.* 12, 1–15. doi: 10.1186/bcr2487

Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., et al. (2011). Extensive and coordinated transcription of noncoding rnas within cell-cycle promoters. *Nat. Genet.* 43, 621–629. doi: 10.1038/ng.848

Hur, H., Lee, J. Y., Yun, H. J., Park, B. W., and Kim, M. H. (2014). Analysis of HOX gene expression patterns in human breast cancer. *Mol. Biotechnol.* 56, 64–71. doi: 10.1007/s12033-013-9682-4

Inaki, K., Menghi, F., Woo, X. Y., Wagner, J. P., Jacques, P. É., Lee, Y. F., et al. (2014). Systems consequences of amplicon formation in human breast cancer. *Genome Res.* 24, 1559–1571. doi: 10.1101/gr.164871.113

Kamalakaran, S., Varadan, V., Giercksky Russnes, H. E., Levy, D., Kendall, J., Janevski, A., et al. (2011). DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Mol. Oncol.* 5, 77–92. doi: 10.1016/j.molonc.2010.11.002

Khoury, A., Achinger-Kawecka, J., Bert, S. A., Smith, G. C., French, H. J., Luu, P.-L., et al. (2020). Constitutively bound ctcf sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-019-13753-7

Lachmann, A. (2016). *Confounding effects in gene expression and their impact on downstream analysis* (Ph.D. thesis), Columbia University, New York, NY, United States.

Lee, T. I., and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251. doi: 10.1016/j.cell.2013.02.014

Liao, Y., Wang, Y., Cheng, M., Huang, C., and Fan, X. (2020). Weighted gene coexpression network analysis of features that control cancer stem cells reveals prognostic biomarkers in lung adenocarcinoma. *Front. Genet.* 11:311. doi: 10.3389/fgene.2020.00311

Liu, R., Guo, C. X., and Zhou, H. H. (2015). Network-based approach to identify prognostic biomarkers for estrogen receptor-positive breast cancer treatment with tamoxifen. *Cancer Biol. Ther.* 16, 317–324. doi: 10.1080/15384047.2014.1002360

Lu, X.-F., Zeng, D., Liang, W.-Q., Chen, C.-F., Sun, S.-M., and Lin, H.-Y. (2018). FoxM1 is a promising candidate target in the treatment of breast cancer. *Oncotarget* 9, 842–852. doi: 10.18632/oncotarget.23182

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006). Reverse engineering cellular networks. *Nat. Protoc.* 1, 662–671. doi: 10.1038/nprot.2006.106

McCarroll, S. A., and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* 39, S37–S42. doi: 10.1038/ng2080

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, 1–14. doi: 10.1186/gb-2011-12-4-r41

Metzger-Filho, O., Sun, Z., Viale, G., Price, K. N., Crivellari, D., Snyder, R. D., et al. (2013). Patterns of recurrence and outcome according to breast cancer subtypes in lymph node-negative disease: results from international breast cancer study group trials viii and ix. *J. Clin. Oncol.* 31:3083. doi: 10.1200/JCO.2012.46.1574

Millour, J., Constantinidou, D., Stavropoulou, A. V., Wilson, M. S., Myatt, S. S., Kwok, J. M., et al. (2010). FOXM1 is a transcriptional target of ERα and has a critical role in breast cancer endocrine sensitivity and resistance. *Oncogene* 29, 2983–2995. doi: 10.1038/onc.2010.47

Montes de Oca, R., Gurard-Levin, Z. A., Berger, F., Rehman, H., Martel, E., Corpet, A., et al. (2015). The histone chaperone HJURP is a new independent prognostic marker for luminal A breast carcinoma. *Mol. Oncol.* 9, 657–674. doi: 10.1016/j.molonc.2014.11.002

Myhre, S., Lingjærde, O. C., Hennessy, B. T., Aure, M. R., Carey, M. S., Alsner, J., et al. (2013). Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Mol. Oncol.* 7, 704–718. doi: 10.1016/j.molonc.2013.02.018

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74:036104. doi: 10.1103/PhysRevE.74.036104

Nieto-Jiménez, C., Alcaraz-Sanabria, A., Páez, R., Pérez-Peña, J., Corrales-Sánchez, V., Pandiella, A., et al. (2017). DNA-damage related genes and clinical outcome in hormone receptor positive breast cancer. *Oncotarget* 8, 62834–62841. doi: 10.18632/oncotarget.10886

Novak, P., Jensen, T., Oshiro, M. M., Watts, G. S., Kim, C. J., and Futscher, B. W. (2008). Agglomerative epigenetic aberrations are a common event in human breast cancer. *Cancer Res.* 68, 8616–8625. doi: 10.1158/0008-5472.CAN-08-1419

Novak, P., Jensen, T., Oshiro, M. M., Wozniak, R. J., Nouzova, M., Watts, G. S., et al. (2006). Epigenetic inactivation of the HOXA gene cluster in breast cancer. *Cancer Res.* 66, 10664–10670. doi: 10.1158/0008-5472.CAN-06-2761

Nueda, M. J., Ferrer, A., and Conesa, A. (2012). ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics* 13, 553–566. doi: 10.1093/biostatistics/kxr042

Nuncia-Cantarero, M., Martinez-Canales, S., Andrés-Pretel, F., Santpere, G., Ocaña, A., and Galan-Moya, E. M. (2018). Functional transcriptomic annotation and protein–protein interaction network analysis identify NEK2,

BIRC5, and TOP2A as potential targets in obese patients with luminal A breast cancer. *Breast Cancer Res. Treat.* 168, 613–623. doi: 10.1007/s10549-017-4652-3

Paz, M. F., Fraga, M. F., Avila, S., Guo, M., Pollan, M., Herman, J. G., et al. (2003). A systematic profile of DNA methylation in human cancer cell lines. *Cancer Res.* 63, 1114–1121.

Perez-Peña, J., Corrales-Sánchez, V., Amir, E., Pandiella, A., and Ocana, A. (2017). Ubiquitin-conjugating enzyme E2T (UBE2T) and denticleless protein homolog (DTL) are linked to poor outcome in breast and lung cancers. *Sci. Rep.* 7:17530. doi: 10.1038/s41598-017-17836-7

Perou, C. M., Sørile, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Ress, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Not. AMS* 56, 1082–1097.

Prat, A., and Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* 5, 5–23. doi: 10.1016/j.molonc.2010.11.003

Prieto, C., Risueño, A., Fontanillo, C., and De Las Rivas, J. (2008). Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS ONE* 3:e3911. doi: 10.1371/journal.pone.0003911

Pugacheva, E. M., Kubo, N., Loukinov, D., Tajmul, M., Kang, S., Kovalchuk, A. L., et al. (2020). CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc. Natl. Acad. Sci. U.S.A.* 117, 2020–2031. doi: 10.1073/pnas.1911708117

Rafique, S., Thomas, J. S., Sproul, D., and Bickmore, W. A. (2015). Estrogen-induced chromatin decondensation and nuclear re-organization linked to regional epigenetic regulation in breast cancer. *Genome Biol.* 16:145. doi: 10.1186/s13059-015-0719-9

Rahiminejad, S., Maurya, M. R., and Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics* 20:212. doi: 10.1186/s12859-019-2746-0

Redmond, A. M., Byrne, C., Bane, F. T., Brown, G. D., Tibbitts, P., O'Brien, K., et al. (2015). Genomic interaction between ER and HMGB2 identifies DDX18 as a novel driver of endocrine resistance in breast cancer cells. *Oncogene* 34, 3871–3880. doi: 10.1038/onc.2014.323

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454. doi: 10.1038/nature05329

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

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123. doi: 10.1073/pnas.0706851105

Serrano-Carbajal, E. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2020). Targeting metabolic deregulation landscapes in breast cancer subtypes. *Front. Oncol.* 10:97. doi: 10.3389/fonc.2020.00097

Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., et al. (2019). Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.* 20:175. doi: 10.1186/s12881-019-0909-5

Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. *Genome Med.* 1, 1–9. doi: 10.1186/gm62

Sonawane, A. R., Weiss, S. T., Glass, K., and Sharma, A. (2019). Network medicine in the age of biomedical big data. *Front. Genet.* 10:294. doi: 10.3389/fgene.2019.00294

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

Tong, D. D., Zhang, J., Wang, X. F., Li, Q., Liu, L. Y., Yang, J., et al. (2020). MeCP2 facilitates breast cancer growth via promoting ubiquitination-mediated P53 degradation by inhibiting RPL5/RPL11 transcription. *Oncogenesis* 9:56. doi: 10.1038/s41389-020-0239-7

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformatics* 19, 575–592. doi: 10.1093/bib/bbw139

Wang, J., Yi, Y., Chen, Y., Xiong, Y., and Zhang, W. (2020a). Potential mechanism of rrm2 for promoting cervical cancer based on weighted gene co-expression network analysis. *Int. J. Med. Sci.* 17:2362. doi: 10.7150/ijms.47356

Wang, J. C., Ramaswami, G., and Geschwind, D. H. (2020b). Gene co-expression network analysis in human spinal cord highlights mechanisms underlying amyotrophic lateral sclerosis susceptibility. *bioRxiv*. doi: 10.1101/2020.08.16.253377

Wilkinson, D. M., and Huberman, B. A. (2004). A method for finding communities of related genes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 5241–5248. doi: 10.1073/pnas.0307740100

Wu, Y., Luo, S., Yin, X., He, D., Liu, J., Yue, Z., et al. (2019). Co-expression of key gene modules and pathways of human breast cancer cell lines. *Biosci. Rep.* 39:BSR20181925. doi: 10.1042/BSR20181925

Xiao, B., Chen, L., Ke, Y., Hang, J., Cao, L., Zhang, R., et al. (2018). Identification of methylation sites and signature genes with prognostic value for luminal breast cancer. *BMC Cancer* 18:405. doi: 10.1186/s12885-018-4314-9

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* 5:3231. doi: 10.1038/ncomms4231

Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y., and Kolpakov, F. (2018). GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.* 47, D100–D105. doi: 10.1093/nar/gky1128

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

Zamora-Fuentes, J. M., Hernandez-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.* 11:1232. doi: 10.3389/fgene.2020.578679

Zhang, M. H., Man, H. T., Zhao, X. D., Dong, N., and Ma, S. L. (2014). Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials. *Biomed. Rep.* 2, 41–52. doi: 10.3892/br.2013.187

Zhang, T., Hu, H., Yan, G., Wu, T., Liu, S., Chen, W., et al. (2019). Long non-coding rna and breast cancer. *Technol. Cancer Res. Treat.* 18:1533033819843889. doi: 10.1177/1533033819843889

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* 40, 854–861. doi: 10.1038/ng.167

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

frontiers in Genetics

MINI REVIEW
published: 21 April 2021
doi: 10.3389/fgene.2021.649440

Community Detection in Large-Scale Bipartite Biological Networks

Genís Calderer[1] *and Marieke L. Kuijjer*[1,2]*

[1] *Centre for Molecular Medicine Norway, University of Oslo, Oslo, Norway,* [2] *Department of Pathology, Leiden University Medical Center, Leiden, Netherlands*

Networks are useful tools to represent and analyze interactions on a large, or genome-wide scale and have therefore been widely used in biology. Many biological networks—such as those that represent regulatory interactions, drug-gene, or gene-disease associations—are of a bipartite nature, meaning they consist of two different types of nodes, with connections only forming between the different node sets. Analysis of such networks requires methodologies that are specifically designed to handle their bipartite nature. Community structure detection is a method used to identify clusters of nodes in a network. This approach is especially helpful in large-scale biological network analysis, as it can find structure in networks that often resemble a "hairball" of interactions in visualizations. Often, the communities identified in biological networks are enriched for specific biological processes and thus allow one to assign drugs, regulatory molecules, or diseases to such processes. In addition, comparison of community structures between different biological conditions can help to identify how network rewiring may lead to tissue development or disease, for example. In this mini review, we give a theoretical basis of different methods that can be applied to detect communities in bipartite biological networks. We introduce and discuss different scores that can be used to assess the quality of these community structures. We then apply a wide range of methods to a drug-gene interaction network to highlight the strengths and weaknesses of these methods in their application to large-scale, bipartite biological networks.

Keywords: networks, genomic networks, community detection algorithms, community detection analysis, genomic data analysis, network analysis, biological network analysis, biological network clustering

OPEN ACCESS

Edited by:
Alfredo Pulvirenti, University of Catania, Italy

Reviewed by:
Pao-Yang Chen, Institute of Plant and Microbial Biology, Academia Sinica, Taiwan
Yuri Wolf, National Center for Biotechnology Information (NLM), United States

***Correspondence:**
Marieke L. Kuijjer
marieke.kuijjer@ncmm.uio.no

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

Received: *04 January 2021*
Accepted: *18 March 2021*
Published: *21 April 2021*

Citation:
Calderer G and Kuijjer ML (2021) Community Detection in Large-Scale Bipartite Biological Networks. Front. Genet. 12:649440. doi: 10.3389/fgene.2021.649440

1. INTRODUCTION

Many processes in biology are linked through complex patterns of physical and functional interactions, which can be represented in large-scale, genome-wide biological networks. Analysis of these networks can help our understanding of biology and medicine (Barabási et al., 2011). For example, a recent analysis of protein-protein interaction networks has helped to map cellular organization and genome function (Luck et al., 2020). Analysis of gene regulatory (Sonawane et al., 2017) and expression quantitative trait (eQTL) networks—where Single Nucleotide Polymorphisms (SNP) are connected to gene expression levels based on the strength of their association (Platig et al., 2016; Fagny et al., 2017)—have helped to highlight potential disease associations of genes and SNPs.

Most of the literature on genome-wide biological network analysis has focused on unipartite networks—networks with one type of node, where interactions can in principle form between all nodes. Examples of such networks are those that represent protein-protein interactions or gene-gene co-expression. However, many types of biological networks are naturally bipartite, meaning that there are two disjoint types of nodes, and interactions can only form between the different node types. Examples of genome-wide bipartite networks are gene regulatory networks (Emmert-Streib et al., 2014)—which include transcriptional, post-transcriptional, and post-translational regulatory networks (Koch, 2016; Statello et al., 2020; Guo and Amir, 2021)—eQTL networks, networks comprising gene-pathway associations (He et al., 2014), networks representing gene-disease (Goh et al., 2007; Halu et al., 2019) or non-coding RNA (ncRNA)-disease associations (Sumathipala et al., 2019), or drug-target interaction networks (Yildirim et al., 2007) (see Pavlopoulos et al., 2018 for an extensive overview of different types of bipartite biological networks).

Community detection is an approach to identify so-called "communities" or "modules"—sets of nodes that are densely connected internally (Newman, 2006). Community detection helps to define the higher-order structure of biological networks and allows researchers to extract and interpret biological signals (Pellegrini, 2019). For instance, in a network representing drug-gene associations, which we use as an example network in this mini review, one can apply community detection to identify groups of drugs that affect similar biological processes, thereby capturing potential new treatment strategies for patients who experience adverse effects to a specific drug. In eQTL networks, communities are often enriched for specific biological functions. SNPs in the center of these communities are enriched for regulatory elements and associated with disease phenotypes (Fagny et al., 2017). In regulatory networks,—which are often bipartite in nature, representing regulatory molecules and their targets as different types of nodes—community detection may help improve our understanding of the functions of specific regulatory molecules, as it places similar regulatory molecules in the context of their neighborhoods of targets (Sonawane et al., 2017). Community detection is particularly helpful in increasing our understanding of the biological processes that are targeted by relatively understudied regulatory molecules, for which specific functions are often unknown. These include, for example, ncRNAs (Kuijjer et al., 2020) or regulatory molecules that are not evolutionarily conserved. For a schematic overview of community detection in large-scale bipartite biological networks and their applications, please refer to **Figure 1**.

In this mini review, we discuss different community detection methods that can be applied to identify modules in large-scale bipartite biological networks. We start by giving a theoretical basis of bipartite networks and their community structures in general. We then discuss so-called "modularity" scores, which can be used to assess community structure quality. We show how calculating these modularity scores on bipartite networks differs from calculating them on unipartite networks. We then describe five widely used strategies for community detection that were specifically designed to be applied to bipartite networks. Finally, we assess the performance of these methods on a large-scale, near genome-wide, gene-drug interaction network and discuss the feasibility of applying these methods to genome-wide networks. We hope this overview will help shed light on the challenges with community detection in genome-wide networks in general, as well as on the advantages and disadvantages of applying some of the most widely-used community detection methods to large-scale bipartite genomic networks.

2. PROBLEM DEFINITION

We will first discuss the theoretical basis of some of the most widely used community detection methods that can be applied to networks in general (Diestel, 2005). We note that most of these methods were not initially designed for or tested on biological networks. However, they can be applied to biological networks and have been widely used in their analysis. We start by defining what a network is and, in particular, what a bipartite network represents. We also introduce the notation that we will use in the rest of this mini review.

Definition 1. A weighted network $G = (V, E, \omega)$ is a triple—a set of three elements—where V is a set of nodes, E is a set of edges between nodes in V, and ω is a function that assigns each edge $e \in E$ a weight. We denote n the number of nodes and $m = \sum_{e \in E} \omega(e)$ the sum of edge weights. If a network is unweighted, $\omega = 1$ and m is equal to the total number of edges. A network is said to be bipartite if V can be partitioned into two sets, V_1, V_2, such that every edge $e \in E$ is connected to a node in V_1 and to a node in V_2. From now on, we will use the term $G = (V_1 \cup V_2, E, \omega)$ to indicate a bipartite weighted network, unless otherwise stated.

For a unipartite network, the definition of a "community" is easy and intuitive: it is a set of nodes that are more connected within the same set compared to the rest of the network (Girvan and Newman, 2002). Given a bipartite network G, the problem of finding bipartite communities is more complex. We say that a *community structure* on G is a partition of $V_1 = \cup_{i=1}^{l} C_i$ and $V_2 = \cup_{j=1}^{k} D_j$, where C_i are pairwise disjoint subsets of V_1 and D_j are pairwise disjoint subsets of V_2, such that all nodes in a specific C_i are more connected to a particular subset of V_2 than the rest of nodes in V_1 are, and likewise for the partition of V_2.

As we discuss below, there are several precise definitions of what it means *to be more connected* in a network. Most of these are based on comparing the network structure to a null model, where the nodes are randomly connected, respecting the degree distribution (Barber, 2007; Murata, 2009). This allows an extension to weighted networks, since the degrees can be substituted by the sum of edge weights. We can then define scores, generally called modularities, that precisely measure how "good" a community structure is, in the sense of how much more connected the nodes are within communities compared to the random model. Most community finding strategies identify communities by maximizing such scores (Lancichinetti and Fortunato, 2011).

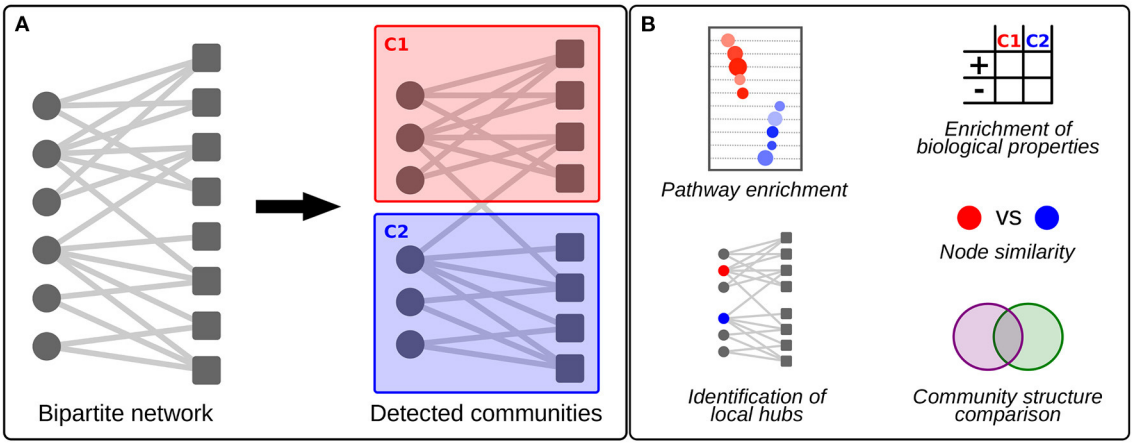

FIGURE 1 | Schematic visualization of bipartite community detection and its applications to large-scale biological networks. **(A)** An example of two communities (C1 and C2) detected in a bipartite network. **(B)** Possible applications of bipartite community detection in the analysis of large-scale biological networks. This includes pathway enrichment in communities, enrichment analysis of other biological properties by testing against external data, identification of "local hub" genes that are central to their community, node similarity detection, and community structure comparison between, for example,networks modeled on disease and control samples.

3. MODULARITY SCORES

The definition of bipartite modularity is an adapted version of the modularity for unipartite networks, which we will describe in the section below.

3.1. Unipartite Modularity

Let $G = (V, E, \omega)$ be a weighted unipartite network with n vertices and $m = \sum_{e \in E} \omega(e)$ edges and let this network be defined by its weighted adjacency matrix A. A is a matrix such that its ij entry is the weight of the edge that joins vertices i and j. In case of an unweighted network, $\omega = 1$. If each node i is assigned to a community g_i, we can define the modularity score (Newman, 2006) of this assignment as follows:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(g_i, g_j), \tag{1}$$

where P is a matrix with entries consisting of the expectation that i and j are connected in the null model, and δ is the Kronecker delta function. We denote $B = A - P$ the modularity matrix.

If the set of nodes in a given community C are more connected within the community itself than would be expected given a random network with same degree distribution, then, for nodes $i, j \in C$, their corresponding entry, B_{ij}, in the modularity matrix will be larger than zero. Per definition, $Q \in [-1, 1]$. When the given community assignment is not worse than a random partition of nodes, Q will be larger than or equal to zero. Such a community structure is said to be stronger when the modularity score Q is closer to 1.

3.2. Bipartite Modularity Scores

Extending the definition of modularity to adapt to the structure of bipartite networks is not completely straightforward and different approaches that do this exist. The most widely used methods are described below. Please note that these bipartite modularity scores were developed for general bipartite networks and can be calculated on any type of bipartite network, including large-scale bipartite biological networks. However, the performance of these scores has not been tested on large-scale biological networks and it is difficult to assess which method is the best. For an overview of how optimizing the different modularity scores might influence the detected community structure, please refer to (Xu et al., 2015).

3.2.1. Guimerà's Modularity

The first approach to a define modularity score for a given community structure on bipartite networks was designed by Guimerà (Guimerà et al., 2007). Guimerà's modularity is the cumulative deviation of the number of edges between nodes that are members of the same bipartite community from the random expectation. This score only takes into account nodes that are in one of the bipartite sets. Because of this, it is not used in any of the community finding methods that we will explore below and, thus, we will not discuss it in more detail.

3.2.2. Barber's Modularity

Barber's approach to defining bipartite modularity (Barber, 2007) is a direct adaptation of the unipartite version described in Equation (1). However, instead of working with the adjacency matrix, the biadjacency matrix $\tilde{A}$ is used. The biadjacency matrix is the non-zero block matrix in the adjacency matrix, if we order nodes first in V_1 and then in V_2. The bimodularity matrix is defined as $\tilde{B} = \tilde{A} - \tilde{P}$, with $\tilde{P}$ being a matrix of expectations corresponding to a null model where nodes are randomly connected, respecting the bipartite structure and degree distribution. This results in a modularity score for assigning nodes $i \in V_1$ to communities g_i and nodes $j \in V_2$ to

communities h_j, which is defined as

$$Q_B = \frac{1}{m}\sum_{i=1}^{p}\sum_{j=1}^{q}(\tilde{A}_{ij} - \tilde{P}_{ij})\delta(g_i, h_j), \quad (2)$$

where $p = |V_1|, q = |V_2|$.

Barber's modularity score takes into account the two node types and the bipartite structure of the network. However, it forces a one-to-one correspondence between the partition in V_1 and the partition in V_2. Thus, each set has to be partitioned into the same number of communities. This is an overly restrictive condition, as it limits the number of possible communities to $\min(p, q)$ (Murata, 2009).

3.2.3. Murata Modularity and Murata+ Modularity

Murata and Murata+ are two modularity scores that build on the previously defined ones. The Murata modularity score (Murata, 2009) was developed to overcome the restriction mentioned in the section above and thus does not force a one-to-one correspondence between the two partitions. It introduces the concept of a *co-cluster* of $C_i \subset V_1$, which is the community on V_2 that C_i shares the highest sum of edge weights with (or in the more intuitive, unweighted case, the largest number of edges).

Let $2M = \sum_e \omega(e)$ be the sum of edge weights. For communities $C \subset V_1$ and $D \subset V_2$, we define the normalized weight of their connection to be $e_{C,D} = e_{D,C} = \frac{1}{2M}\sum_e \omega(e)$, for e edges from $i \in C$ to $j \in D$. Each community contributes to $2M$ with a weight of $a_C = \frac{1}{2M}\sum_D e_{C,D}$. Moreover, we can define the *co-cluster* of a community C to be the community $D_C \subset V_2$ with the highest concentration of edges from C, that is $D_C = \arg\max_D(e_{C,D})$. With these definitions, Murata's modularity score for a given partition of V_1 and V_2 is

$$Q_M = \sum_{C\subset V_1}(e_{C,D_C} - a_C a_{D_C}) + \sum_{D\subset V_2}(e_{C_D,D} - a_{C_D}a_D). \quad (3)$$

This score pairs each community in V_1 to a community in V_2, its co-cluster, and computes the difference between intra-co-cluster edges and the expected edges in a randomly generated graph. This metric is less restrictive than Barber's modularity, because it assumes different community structures in each of the sets V_1 and V_2 that are related to one another by the co-cluster correspondences of each community in each of the sets.

In the *biLouvain* method (Pesantez-Cabrera and Kalyanaraman, 2016), which we describe in the next section, the definition of Murata's modularity is extended so that the co-cluster relationship is not necessarily symmetric. To do so, the choice of co-cluster is adapted to use the terms $a_C a_{D_C}$ and $a_{C_D}a_D$. This allows for even more flexibility, as the co-cluster $D \subset V_2$ of a community $C \subset V_1$ does not necessarily need to have C as its co-cluster. Thus, for a given partition, this new modularity score—which is called Murata+—has the same definition as in Equation 3, but the co-clusters are chosen as follows:

$$D_C = \arg\max_D(e_{C,D} - a_C a_D) \quad \text{and} \quad C_D = \arg\max_C(e_{C,D} - a_C a_D). \quad (4)$$

3.3. Resolution

Most community finding strategies rely on maximizing a modularity score (generally Barber's, see Equation 2). These approaches have been shown to retrieve true communities when applied to networks with a ground-truth community structure (Barber, 2007; Dao et al., 2017). However, there is a resolution limit when it comes to properly separating communities, which hampers community detection in large-scale networks. For unipartite networks, it was shown that communities with a number of internal edges $\leq O(\sqrt{m})$ may not be detected (Fortunato and Barthélemy, 2007). While this problem was highlighted with unipartite modularity, this also applies to bipartite networks with Barber's modularity.

This poses a problem when it comes to working with large-scale networks, such as genomic networks; certain small, tightly-knit communities might be too small to detect. This is particularly relevant in the analysis of biological networks, as this means that general processes can still be detected, but that the subtle differences that distinguish, for example, a disease network from a control network may be below the resolution limit and thus could be left undetected. This can be adjusted [in the case of Barber's modularity (Equation 2)] by introducing a resolution parameter $\lambda > 0$, such that

$$Q_B = \frac{1}{m}\sum_{i=1}^{p}\sum_{j=1}^{q}(\tilde{A}_{ij} - \lambda\tilde{P}_{ij})\delta(g_i, h_j). \quad (5)$$

Then if $\lambda > 1$, more, but smaller communities are detected and if $\lambda < 1$, fewer, but larger communities are found.

4. COMMUNITY DETECTION STRATEGIES

Most community finding methods, both in unipartite and bipartite networks, are based on optimizing a modularity function. There are several strategies to do this in a fast and optimal manner (Newman, 2016), but there is no consensus on what method is best. However, all of these strategies are greedy—at each step the program tries to find the optimal next step. Thus, there is always the possibility to detect a local maximum instead of the global maximum, and therefore not the best structure. This can be an issue in large-scale biological network analysis, specifically if one aims to use the community structure to, for example, find similarities between drug targets in a drug-gene interaction network, or to get insights in potential regulatory functions of ncRNAs by analyzing a ncRNA-gene network.

Some of the most widely used strategies for optimizing modularity are discussed below.

4.1. Spectral Optimization (SO)

Spectral optimization methods are algorithms that take advantage of the structure of the various matrices (e.g. the adjacency matrix or the modularity matrix) associated to a network. The most widely used spectral optimization method for bipartite networks is Bipartite Recursively Induced Modules (BRIM) (Barber, 2007). BRIM uses the fact that, if B is the bimodularity matrix of a network, R is a community membership

matrix for the nodes in V_1, and T a community membership matrix for the nodes in V_2, then the formula in Equation (2) can be written as follows:

$$Q_B = \frac{1}{m} \mathrm{Tr}(R^T \tilde{B} T), \tag{6}$$

where Tr is the trace of the matrix. Then, given an initial community structure on V_1, the community assignment in V_2 that maximizes modularity can be calculated. This is done recursively using the new assignment as initial community structure, until the modularity cannot increase further.

BRIM is considerably fast, because uses matrix multiplications, which are optimally implemented in several programming languages. However, it has the drawback that it strongly depends on the initial community structure assignment. In addition, it requires one to know the total number of communities beforehand. In large-scale biological networks, the number of communities is usually unknown (Sah et al., 2014; Gaiteri et al., 2015).

4.2. Projections and Adapted Unipartite Methods

A bipartite network can be projected onto one of its sets of nodes, for example V_1. Its projection is a new unipartite network that has as nodes those in V_1, and weighted edges corresponding to the number of shared neighboring nodes $i, j \in V_1$ have. This projection retains part of the information about the topology of the network and can then be used to find a community structure using unipartite methods. Projections are often applied to large networks, where unipartite methods, such as Louvain (Blondel et al., 2008) or Leiden (Traag et al., 2019) can work very effectively. However, a drawback of projecting a network is that it will lead to a loss in resolution which, as we discuss above, is not ideal when analyzing biological networks. In addition, the relationship between a bipartite network and its projection is not one-to-one. Significantly different bipartite networks can have the same projection and, thus, could result in the same community structure. This could, for example, hamper the identification of differences between networks modeled on disease and control samples.

Some unipartite methods can be adapted to deal with bipartite networks by having a resolution/distance parameter set to two, which forces the method to compare nodes from the same bipartite set. This is a not an optimal approach, as it does not take into account the bipartite structure of the network. In large-scale bipartite biological networks, this structure is important, as we are often interested in understanding how two different types of components, such as transcription factors and their target genes, or diseases and genes, relate to one another. In addition, this approach is not valid for weighted networks, where the distance between the two sets is not uniformly two. Edges in large-scale bipartite biological networks are generally weighted as they are often based on effect sizes or probabilities. For example, in regulatory networks, one often estimates the likelihood of a transcription factor or ncRNA to regulate a target gene. eQTL networks can be built on the strength of SNP-gene associations. While these weighted networks can be transformed into unweighted networks by thresholding them on the edge weights, this approach is not ideal, as subtle changes in edges weights can drive biological differences (Lopes-Ramos et al., 2020). Therefore, methods that can only be applied to unweighted networks are generally not ideal for community structure detection in genomic biological networks.

4.3. Label Propagation (LP)

In label propagation (Liu and Murata, 2009b), each node is initialized in its own community. Then, for each community, the modularity that would be gained if the community were to be merged with another community is computed. Those merges that maximize modularity gain are then applied, and this process is repeated until the modularity cannot increase any further. When this point is reached, a condensation step is applied that generates a new network. In this new network, each node represents a community from the former network. The edges are interactions between the communities, which are weighted, for example, using the sum of weights from all nodes in a community to all nodes in the other. Label propagation can then again be applied to this network to find a new level of community structure. Further condensations can be applied until the modularity gain stabilizes. This is how the unipartite method Louvain works.

For bipartite networks this approach is adapted [for example in LPA (Costa and Hansen, 2014), DIRTLPAwb+ (Beckett, 2020), LP-BRIM (Liu and Murata, 2009a), biLouvain (Pesantez-Cabrera and Kalyanaraman, 2016)] to take the two different types of nodes in the modularity gain function into account.

It should be noted that these methods can have a stochastic component to solve ties in modularity gain. Therefore, it is possible that different runs of the method on the same network result in slightly different community structures. This could be a problem if one wants to compare community structures to, for example, detect phenotype-driven transitions in regulatory networks (Padi and Quackenbush, 2018), as it is difficult to distinguish differences caused by this stochastic component from those that arise due true biological differences in network structure. Also, as mentioned before, this can lead to detecting a local instead of the global maximum, and thereby not detecting the best community structure. Some algorithms, such as *DIRTLPAwb+* run this approach several times and then keep the structure with the highest modularity. However, this comes with additional computational load, and may thus not be ideal for analysis on genome-wide networks.

4.4. Node Similarity (NS)

Node similarity algorithms, such as *ComSim* (Tackx et al., 2018) are different from the methods described above as they are not designed to optimize modularity. They define a similarity function between nodes, for example the number of common neighbors or the Jaccard similarity. They then use this function to find cycles in the network—so-called core communities—that have high similarity. These core communities do not contain all available nodes, as some nodes are left unassigned. To obtain a community structure that includes all nodes, these unassigned nodes are then added to the core community with which they have the highest similarity score.

4.5. Overlapping Community Detection

Overlapping methods for bipartite networks aim to give a covering of the bipartite sets that is not disjoint. This means that some nodes can be present in more than one community. This property makes sense in, for example, regulatory networks, because a transcription factor may regulate different biological functions that could be represented in different communities.

The main strategy for finding overlapping community structures in bipartite networks consist of finding bicliques—sets of nodes that form a complete bipartite graph—and then merging those based on a similarity function (see above). Two methods that implement this strategy for unweighted networks are BiTector (Du et al., 2008) and maxBic (Alzahrani and Horadam, 2019).

4.6. Limitations and Strengths of Published Methods in Their Applications to Genomic Networks

As discussed above, several methods for community detection in bipartite networks exist. In **Table 1**, we list the community detection algorithms described in this mini review, together with their community detection strategy (which we describe above), the modularity scores or similarity measures they maximize (objective function), whether they can be applied to weighted networks, and the programming language that these methods are available in.

Bipartite biological networks all have the same basic properties—two disjoint types of nodes, with interactions only forming between the different node types. Therefore, in principle, any bipartite community detection algorithm can be applied to any type of large-scale bipartite biological network. There is no consensus on what method is best, and to our knowledge no benchmarking study has been performed to evaluate which methods are most appropriate for different types of bipartite genomic networks. However, as we also describe above, certain limitations can hamper community detection in these networks. We describe the most important limitations below.

Some community detection methods can only handle unweighted networks and thus can not be applied to all large-scale bipartite biological networks. Most biological networks can be both modeled in weighted or unweighted form. Gene-disease networks, drug-target networks, or pathway-gene networks have previously mostly been constructed and analyzed in unweighted form Goh et al. (2007), He et al. (2014), and Halu et al. (2019). However, they can also be estimated in weighted form by including, for example, information on predictions or associations in the edge weights (Sumathipala et al., 2019). While regulatory networks and eQTL networks are sometimes unweighted, they are more often based on likelihoods or associations. Weighted networks include more information and allow one to compare the strength, intensity, or capacity of interactions within a network or between different types of networks (Horvath, 2011). Thus, when possible, we recommend to use methods that can be applied to weighted networks.

The high computation load of many community detection methods is also a limitation and will influence the feasibility of applying community detection to genomic networks. This is particularly important in very large genomic networks, such as eQTL networks, which can include hundreds of thousands of SNPs in one of the node sets, and tens of thousands of genes in the other node set. For genome-wide bipartite networks with fewer nodes, such as gene-disease networks or pathway-gene networks, this may be less of a challenge. All methods we reviewed here have worst-case complexity $O(n^3)$, except in special cases where particular properties of the network—for example the presence of nodes in V_2 that are mainly connected to a single node in V_1—can be taken advantage of to reduce complexity to $O(n^2)$. However, this would require a specific implementation of the method for each particular network. The complexity of these methods means that they can be challenging to run on genome-wide biological networks, as we show in the example below.

In addition, as we describe in the section above, detecting communities using methods that rely on maximizing a modularity score may be hampered by the resolution limit. Again, this will be particularly relevant for very large networks, such as those based on eQTLs.

Finally, some community detection algorithms, including biTector and maxBic, the code to run the method is not publicly available. Thus, these methods may be challenging to run as the user would need to implement the code themselves or contact the authors to obtain it.

5. APPLICATION TO A GENE-DRUG INTERACTION NETWORK

In general, most community detection algorithms are tested on small benchmark networks (Lancichinetti et al., 2008) and tests on large-scale bipartite genomic networks are lacking. We therefore wanted to test the performance of community detection methods on a near genome-wide network. As an example, we used a gene-drug interaction network from the The Drug Gene Interaction Database (DGIdb) (Cotto et al., 2018). We selected this network, because it is a well-known example of a large-scale biological network that is known to be modular (Pesantez-Cabrera and Kalyanaraman, 2016). This allows us to showcase the different methods retrieving, as we show below, significant communities.

5.1. Preparation of the Network

We downloaded the *interactions.tsv* file from DGIdb (Cotto et al., 2018) (accessed August 14, 2020). We removed all missing and duplicate data and kept only the confirmed gene-drug interactions. We built an unweighted bipartite network from these data representing the interactions between genes and drugs. Because all methods require the network to be connected, we kept the largest connected component (99% of the network in terms of nodes). This resulted in a network consisting of 22,693 interactions between 2,336 genes and 6,049 drugs.

5.2. Application of the Methods

We applied those community detection methods that had a functioning and available implementation to the gene-drug interaction network. As a means to consistently use the same score, we computed the Murata+ score for all of the methods. For each method, we obtained a partition of the set of genes and

TABLE 1 | Community detection methods with their respective strategies of community detection, the used objective function, whether they allow for weighted networks, and their availability in different programming environments.

Method	Strategy	Objective function	Weighted	Available
BRIM (Barber, 2007)	SO	Bimodularity	Yes	R, Python
LP-BRIM (Liu and Murata, 2009a)	LP + SO	Murata	Yes	R
LPA (Costa and Hansen, 2014)	LP	Bimodularity	Yes	R
DIRTLPAwb+ (Beckett, 2020)	LP	Bimodularity	Yes	R
CONDOR Platig et al., 2016	LP + SO	Bimodularity	Yes	R, Python
ComSim (Tackx et al., 2018)	NS	Common neighbors, Jaccard	Yes	C++
biLouvain (Pesantez-Cabrera and Kalyanaraman, 2016)	LP + SO	Murata+	Yes	C++
biTector (Du et al., 2008)	Overlapping	–	No	Unavailable
maxBic (Alzahrani and Horadam, 2019)	Overlapping	–	No	C++ (not public)

SO, spectral optimization; LP, label propagation; NS, node similarity.

A

Method	Q_{M+}	Runtime	#C*
CONDOR	0.8263	3 sec	46/58
LP-BRIM	0.7058	510 sec	118/671
DIRTLPAwb+	-	>10 days	-
ComSim (common neighbours)	0.4501	2 sec	156/244
biLouvain	0.7462	6262 sec	99/370

B

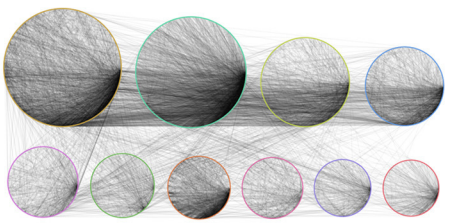

FIGURE 2 | (A) Modularity, runtime of the method with default settings on a high-performance computing server (128 Intel Haswell cores, 1 Tb RAM), and number of communities obtained with running different community detection methods on the gene-drug interaction network. *Number of communities with more than four members/total number of communities in the gene node set. **(B)** Example "shell plot" of the ten largest communities detected in the drug-gene network using CONDOR. Communities are indicated with different colors.

a partition of the set of drugs into communities. We focused on the structure in the gene node set, so that we could explore Gene Ontology enrichment and assess the significance of enriched gene sets in the different communities. Some of the communities revealed by the methods included less than four genes (see **Figure 2A**). We excluded these from the following analysis because they were too small to apply GO term enrichment analyses on.

The obtained modularities are shown in **Figure 2A**, together with the runtime and number of detected communities on the gene node set. We note that ComSim results in a significantly lower modularity score. This does not necessarily mean that the community structure is poorly defined. It is simply a result of the fact that this method does not work to optimize a modularity score. The quality of the community structure might, thus, not be captured by such scores.

An example of the ten largest communities detected with CONDOR is shown in **Figure 2B**. As can be seen, more edges are detected within communities compared to between different communities. However, there are also intra-community edges, indicating that community detection in large-scale networks is a complex problem.

5.3. Results

5.3.1. Information Comparison

Because we lack a ground-truth for this network, we cannot assess the quality of results in terms of discovering a previously known community structure. However we can compare how similar the results are across the different methods. Given two community assignments on the same set of genes, we compared the information they share with the *Normalized Mutual Information* (NMI) score. This score ranges from 0 to 1, with scores closer to 1 indicating higher similarity. We computed pairwise NMIs between each of the methods. We found that the scores were similar, and contained within the $[0.6077, 0.7746]$ range, indicating that the community assignments share a high amount of information.

5.3.2. GO Enrichment

We wanted to evaluate whether the communities we discovered were enriched for specific biological processes. For each method we ran GO enrichment analysis (Klopfenstein et al., 2018) on the selected communities. All methods resulted in communities that were significantly ($p_{\text{fdr}} < 10^{-8}$) enriched for biological pathways. This high level of enrichment confirms that the retrieved communities likely represent true biological information. A t-test concluded that there was no difference between the significance of the results for each method.

5.3.3. Co-cluster Analysis

The final community structure obtained by biLouvain with Murata+ offers a relationship between communities of each of the bipartite sets. Above, we mentioned that this relationship is not necessarily one-to-one, as the co-cluster $D \subset V_2$ of a community $C \subset V_1$ does not necessarily need to have C as

its co-cluster. This allows for higher flexibility when it comes to splitting particular communities in one of the sets without affecting the other. In this particular network, however, we found that the relationship was one-to-one. This might be because the network is already very modular, or the corrections in Murata+ are subtle and do not influence the final community structure strongly enough.

The co-cluster relation between communities of genes and communities of drugs is biologically significant. For example, the three largest co-clusters (based on node size) contained a co-cluster of a gene-community containing GABA genes with a drug-community that contains several benzodiazepines, which enhance the effect of GABA neurotransmitters at $GABA_A$ receptors. There are several other examples of co-clusters between communities of genes of well-known pathways and communities of drugs that are known to act on those pathways (see **Supplementary Table 1**).

6. DISCUSSION

While unipartite community detection has been widely applied to large-scale biological networks, community detection on bipartite networks and, in particular, on genome-wide bipartite networks, has been less studied. However, as many types of biological networks are bipartite, it is important to review community detection approaches that are specifically designed for such networks. Here, we reviewed several community detection strategies, discussed their strengths and weaknesses in the context of their application to genomic bipartite networks, and applied these to a near genome-wide gene-drug interaction network.

Dealing with large-scale networks is a computationally expensive task, and thus not all software packages can deal with the data in a fast manner. Although the communities detected by different methods were highly similar, the modularity scores and, in particular, their runtimes were rather different. Thus, methods that run fast could be prioritized for genomic bipartite networks. For example, as can be seen in **Figure 2A**, CONDOR is relatively fast on such large networks.

We would like to note that the gene-drug interaction network we included in our evaluation is indeed highly modular, and that the advantages and drawbacks of the different community detection methods might be more visible with networks with lower structure. However, there is a lack of large-scale bipartite networks with ground-truth (Peel et al., 2017) and it is very difficult to identify a large biological network that does not suffer from the resolution limit.

The Murata+ score is versatile and the communities detected by the method respect the bipartite structure of the network. However, the only method that implements it is biLouvain, which can be very slow to run on genome-wide networks. We believe that a method that uses a spectral optimizer, such as BRIM, to maximize Murata+ modularity scores would be highly useful in large-scale bipartite biological network analysis and could be a potential direction for future research.

Finally we note that, as most of the algorithms designed for bipartite community detection are focused on optimizing modularity, they may reach the resolution limit. This may render it difficult to detect communities in large-scale genomic networks and is a problem that is currently unsolved and one that warrants further investigation.

AUTHOR CONTRIBUTIONS

GC and MK: conceptualization, investigation, and writing—review and editing. GC: methodology, formal analysis, and writing—original draft. MK: resources, supervision, and funding acquisition.

FUNDING

This work was supported by the Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM).

ACKNOWLEDGMENTS

We would like to thank Annabel Darby for help with language editing and all members from the Kuijjer and Mathelier groups for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.649440/full#supplementary-material

Supplementary Table 1 | Co-clusters in the drug-gene network, detected using biLouvain and Murata+. Rows represent the co-clustered communities (detected on the gene node set in column 1, on the drug node set in column 2), as well as the genes and drugs present in those communities (columns 3 and 4, respectively).

REFERENCES

Alzahrani, T., and Horadam, K. (2019). Finding maximal bicliques in bipartite networks using node similarity. *Appl. Netw. Sci.* 4:21. doi: 10.1007/s41109-019-0123-6

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

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Phys. Rev. E* 76:66102. doi: 10.1103/PhysRevE.76.066102

Beckett, S. J. (2020). Improved community detection in weighted bipartite networks. *R. Soc. Open Sci.* 3:140536. doi: 10.1098/rsos.140536

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Costa, A., and Hansen, P. (2014). A locally optimal hierarchical divisive heuristic for bipartite modularity maximization. *Optimiz. Lett.* 8, 903–917. doi: 10.1007/s11590-013-0621-x

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

Dao, V. L., Bothorel, C., and Lenca, P. (2017). "Community detection methods can discover better structural clusters than ground-truth communities," in *2017 IEEE/ACM International Conference on Advances in Social*

Networks Analysis and Mining (ASONAM) (Sydney, NSW: IEEE), 395–400. doi: 10.1145/3110025.3110053

Diestel, R. (2005). *Graph Theory, 3rd Edn. Graduate Texts in Mathematics*. New York, NY: Springer-Verlag Heidelberg.

Du, N., Wang, B., Wu, B., and Wang, Y. (2008). "Overlapping community detection in bipartite networks," in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1 (Sydney, NSW: IEEE), 176–179. doi: 10.1109/WIIAT.2008.98

Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* 2:38. doi: 10.3389/fcell.2014.00038

Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C. Y., Lopes-Ramos, C. M., et al. (2017). Exploring regulation in tissues with eQTL networks. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7841–E7850. doi: 10.1073/pnas.1707375114

Fortunato, S., and Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A.* 104:36. doi: 10.1073/pnas.0605965104

Gaiteri, C., Chen, M., Szymanski, B., Kuzmin, K., Xie, J., Lee, C., et al. (2015). Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering. *Sci. Rep.* 5:16361. doi: 10.1038/srep16361

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99:7821. doi: 10.1073/pnas.122653799

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Phys. Rev. E* 76:36102. doi: 10.1103/PhysRevE.76.036102

Guo, Y., and Amir, A. (2021). Exploring the effect of network topology, mRNA and protein dynamics on gene regulatory network stability. *Nat. Commun.* 12:130. doi: 10.1038/s41467-021-21415-w

Halu, A., De Domenico, M., Arenas, A., and Sharma, A. (2019). The multiplex network of human diseases. *NPJ Syst. Biol. Appl.* 5, 1–12. doi: 10.1038/s41540-019-0092-5

He, L., Wang, Y., Yang, Y., Huang, L., and Wen, Z. (2014). Identifying the gene signatures from gene-pathway bipartite network guarantees the robust model performance on predicting the cancer prognosis. *Biomed Res. Int.* 2014:424509. doi: 10.1155/2014/424509

Horvath, S. (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. New York, NY: Springer Science & Business Media.

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., et al. (2018). GOATOOLS: a Python library for gene ontology analyses. *Sci. Rep.* 8:10872. doi: 10.1038/s41598-018-28948-z

Koch, L. (2016). A global view of regulatory networks. *Nat. Rev. Genet.* 17, 252–252. doi: 10.1038/nrg.2016.36

Kuijjer, M. L., Fagny, M., Marin, A., Quackenbush, J., and Glass, K. (2020). PUMA: PANDA using microRNA associations. *Bioinformatics* 36, 4765–4773. doi: 10.1093/bioinformatics/btaa571

Lancichinetti, A., and Fortunato, S. (2011). Limits of modularity maximization in community detection. *Phys. Rev. E* 84:066122. doi: 10.1103/PhysRevE.84.066122

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* 78:046110. doi: 10.1103/PhysRevE.78.046110

Liu, X., and Murata, T. (2009a). "Community detection in large-scale bipartite networks," in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Milan), Vol. 1, 50–57. doi: 10.1109/WI-IAT.2009.15

Liu, X., and Murata, T. (2009b). "How does label propagation algorithm work in bipartite networks?" in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Milan), Vol. 3, 5–8. doi: 10.1109/WI-IAT.2009.217

Lopes-Ramos, C. M., Chen, C. Y., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., et al. (2020). Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell Rep.* 31:107795. doi: 10.1016/j.celrep.2020.107795

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., et al. (2020). A reference map of the human binary protein interactome. *Nature* 580, 402–408. doi: 10.1038/s41586-020-2188-x

Murata, T. (2009). "Detecting communities from bipartite networks based on bipartite modularities," in *2009 International Conference on Computational Science and Engineering* (Vancouver, BC), Vol. 4, 50–57. doi: 10.1109/CSE.2009.81

Newman, M. E. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E* 94:052315. doi: 10.1103/PhysRevE.94.052315

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Padi, M., and Quackenbush, J. (2018). Detecting phenotype-driven transitions in regulatory network structure. *NPJ Syst. Biol. Appl.* 4, 1–12. doi: 10.1038/s41540-018-0052-5

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., and Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* 7:giy014. doi: 10.1093/gigascience/giy014

Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Sci. Adv.* 3:e1602548. doi: 10.1126/sciadv.1602548

Pellegrini, M. (2019). "Elsevier Reference Module in Life Sciences," in *Community Detection in Biological Networks*, (Amsterdam: Elsevier).

Pesantez-Cabrera, P., and Kalyanaraman, A. (2016). "Detecting communities in biological bipartite networks," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '16* (New York, NY: Association for Computing Machinery), 98–107. doi: 10.1145/2975167.2975177

Platig, J., Castaldi, P. J., DeMeo, D., and Quackenbush, J. (2016). Bipartite community structure of eQTLs. *PLoS Comput. Biol.* 12:e1005033. doi: 10.1371/journal.pcbi.1005033

Sah, P., Singh, L. O., Clauset, A., and Bansal, S. (2014). Exploring community structure in biological networks with random graphs. *BMC Bioinformatics* 15:220. doi: 10.1186/1471-2105-15-220

Sonawane, A. R., Platig, J., Fagny, M., Chen, C. Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088. doi: 10.1016/j.celrep.2017.10.001

Statello, L., Guo, C. J., Chen, L. L., and Huarte, M. (2020). Gene regulation by long non-coding rnas and its biological functions. *Nat. Rev. Mol. Cell Biol.* 22, 96–118.

Sumathipala, M., Maiorino, E., Weiss, S. T., and Sharma, A. (2019). Network diffusion approach to predict lncRNA disease associations using multi-type biological networks: LION. *Front. Physiol.* 10:888. doi: 10.3389/fphys.2019.00888

Tackx, R., Tarissan, F., and Guillaume, J. L. (2018). "ComSim: a bipartite community detection algorithm using cycle and node's similarity," in *Complex Networks & Their Applications VI*, eds C. Cherifi, H. Cherifi, M. Karsai, and M. Musolesi (Cham: Springer International Publishing), 278–289. doi: 10.1007/978-3-319-72150-7_23

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 9:5233. doi: 10.1038/s41598-019-41695-z

Xu, Y., Chen, L., Li, B., and Liu, W. (2015). Density-based modularity for evaluating community structure in bipartite networks. *Inform. Sci.* 317, 278–294. doi: 10.1016/j.ins.2015.04.049

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L., and Vidal, M. (2007). Drug-target network. *Nat. Biotechnol.* 25, 1119–1127. doi: 10.1038/nbt1338

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

frontiers in Genetics

PERSPECTIVE
published: 23 April 2021
doi: 10.3389/fgene.2021.649942

Gene Targeting in Disease Networks

*Deborah Weighill[1], Marouen Ben Guebila[1], Kimberly Glass[1,2,3], John Platig[2,3], Jen Jen Yeh[4] and John Quackenbush[1,2]**

[1] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA, United States, [2] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, United States, [3] Harvard Medical School, Harvard University, Boston, MA, United States, [4] Departments of Surgery and Pharmacology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Profiling of whole transcriptomes has become a cornerstone of molecular biology and an invaluable tool for the characterization of clinical phenotypes and the identification of disease subtypes. Analyses of these data are becoming ever more sophisticated as we move beyond simple comparisons to consider networks of higher-order interactions and associations. Gene regulatory networks (GRNs) model the regulatory relationships of transcription factors and genes and have allowed the identification of differentially regulated processes in disease systems. In this perspective, we discuss gene targeting scores, which measure changes in inferred regulatory network interactions, and their use in identifying disease-relevant processes. In addition, we present an example analysis for pancreatic ductal adenocarcinoma (PDAC), demonstrating the power of gene targeting scores to identify differential processes between complex phenotypes, processes that would have been missed by only performing differential expression analysis. This example demonstrates that gene targeting scores are an invaluable addition to gene expression analysis in the characterization of diseases and other complex phenotypes.

Keywords: cancer genomics, network medicine, gene targeting, differential targeting, gene regulatory networks

OPEN ACCESS

Edited by:
Pierre De Meyts, Université catholique de Louvain, Belgium

Reviewed by:
James Lim, University of Arizona, United States
Ankush Sharma, University of Oslo, Norway

***Correspondence:**
John Quackenbush
johnq@hsph.harvard.edu

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

Received: *05 January 2021*
Accepted: *15 March 2021*
Published: *23 April 2021*

Citation:
Weighill D, Ben Guebila M, Glass K, Platig J, Yeh JJ and Quackenbush J (2021) Gene Targeting in Disease Networks. Front. Genet. 12:649942. doi: 10.3389/fgene.2021.649942

INTRODUCTION

A core tenet of molecular biology is that phenotypic differences are reflected through patterns of differential expression of key genes involved in relevant biological processes. Since its inception, whole-genome transcript profiling has been an invaluable tool for exploring these associations and has been used in a range of applications, including identification of clinically relevant molecular subtypes in cancer exhibiting different morbidities and implications for treatment together with characteristic genes associated with these phenotypes (Rouzier et al., 2005; Collisson et al., 2011; Moffitt et al., 2015; Bailey et al., 2016; Kwa et al., 2017; Rodriguez-Salas et al., 2017; Rudin et al., 2019; Sjödahl et al., 2019). Studies have also found that complex patterns of association between genes represented as networks can provide additional insight and that network metrics parameterizing these associations can be used to prioritize and identify crucial disease-related genes (Ramadan et al., 2016; Dimitrakopoulos et al., 2018; Horn et al., 2018; Gumpinger et al., 2020). However, there is growing evidence that the processes regulating the expression of phenotype-associated genes can provide a more holistic picture of drivers of disease and other phenotypes. Gene regulatory networks (GRNs) are often represented as directed bipartite graphs that are used to depict inferred relationships between transcription factors (TFs) and their target genes. GRNs can be characterized by calculating "gene targeting scores," a network topology measure that captures

the complex relationships that a gene has with other TFs and genes and represents the extent to which a gene is targeted in a given system. In this perspective, we will present gene targeting scores, discuss their meaning, and show how this network-based measure provides information about disease systems beyond that found using only differential expression analysis in the investigation and characterization of human disease.

GENE REGULATORY NETWORKS: CHARACTERIZATION OF SYSTEMS

Networks are useful tools for representing and analyzing large, complex datasets because they capture information about the *relationships* within a system rather than simply the state of individual components. This is an important distinction, as can be illustrated with a small toy example first described in Glass et al. (2014; **Figure 1**). In this example, we consider the expression of four genes in nine healthy individuals and nine individuals with a disease (**Figures 1A,B**). Comparing expression levels between healthy and diseased individuals, we find that none are differentially expressed (**Figures 1C,D**). However, when looking at the co-expression of these genes in each group of individuals, we see that that the genes are *differentially co-expressed* between groups. For example, in healthy individuals, gene G1 is co-expressed with gene G2 (**Figure 1E**), whereas in diseased individuals, gene G1 is co-expressed with gene G3 (**Figure 1F**). This illustrates that differential expression analysis alone may miss important correlations or regulatory relationships that distinguish biological states such as healthy and disease.

This does not mean that gene expression analysis alone is not useful. Differential expression analyses have contributed to many key advances in our understanding of disease. For example, much of our understanding of the complexities of human cancers is derived from large-scale expression profiling of cancer, such as that carried out by The Cancer Genome Atlas (TCGA)[1], where the expression-based subtypes that have been identified possess distinct clinical characteristics. In pancreatic ductal adenocarcinoma (PDAC), several studies have used expression profiling to identify molecular subtypes (Collisson et al., 2011; Moffitt et al., 2015; Bailey et al., 2016; Rashid et al., 2020; Puleo et al., 2018; Maurer et al., 2019). However, we suggest that a more comprehensive molecular characterization of diseases can be achieved by exploring inferred regulatory network differences and differential gene targeting.

In 2013, Glass et al. (2013) introduced the PANDA (Passing Attributes between Networks for Data Assimilation) framework for GRN construction. This method takes a data integration approach to GRN construction by using message passing to combine multiple data sources. PANDA predicts regulatory relationships between TFs and genes by considering three main sources of information: (1) a TF–gene network "adjacency matrix" representing an initial guess of which TFs regulate which genes based on the presence/absence of TF motif in the promoter regions of genes, (2) a protein–protein interaction network "co-operativity matrix" that recognizes that many TFs exert their influence through regulatory complexes, and (3) a gene co-expression matrix representing gene–gene relationships initially based on correlation in expression patterns across a set of samples. These three different sources of information are iteratively updated using a message-passing algorithm, using the logic that if two genes are co-expressed, they are more likely to be co-regulated by a similar set of TFs (**Figure 2A**), and that if two TFs interact, they are more likely to bind promoter regions as a complex and co-regulate the expression of their target genes (**Figure 2B**). In this process, the TF–gene "edge weights" in the adjacency matrix are updated to reflect the evidence supporting a regulatory interaction; the refinement of edge weights through message passing has been found to improve the prediction accuracy of GRNs, validated through prediction of chromatin immunoprecipitation (ChIP)-seq-derived TF binding.

PANDA has been used to investigate gene regulatory relationships in several disease contexts, including chronic obstructive pulmonary disease (COPD) (Glass et al., 2014), asthma (Qiu et al., 2018), ovarian cancer (Glass et al., 2015), and colorectal cancer (Vargas et al., 2016; Lopes-Ramos et al., 2018). In addition, single-sample versions of PANDA GRNs, derived using a method called LIONESS (Linear Interpolation to Obtain Network Estimates for Single Samples) (Kuijjer et al., 2019b), have been used to study sex-linked differences in colon cancer (Lopes-Ramos et al., 2018) as well sex-related differences in gene regulation (Lopes-Ramos et al., 2020) in tissues from the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013).

GENE TARGETING SCORE: IDENTIFYING INFORMATIVE REGULATORY PROCESSES

The use of GRNs in the analysis of disease relies on analysis of the "gene targeting score," a numerical score representing the extent to which a gene is targeted by TFs in a given biological context. The gene targeting score is calculated by summing the weights of all inbound regulatory edges for a gene (**Figure 2C**). Because of the way in which PANDA estimates edge weights, a gene's targeting score synthesizes multiple lines of evidence—TF motif data, TF–TF interactions, and gene expression correlation. Thus, gene targeting scores are not necessarily correlated with absolute gene expression levels, and consequently, differential targeting is not necessarily correlated with differential gene expression.

Sonawane et al. (2017) used PANDA to construct tissue-specific GRNs for 38 tissues in GTEx and investigated the tissue specificity of TF–gene regulatory relationships. They found many tissue-specific regulatory relationships that would have been missed by using expression information alone. For example, when comparing the tissue-specific regulatory activity of TFs based on gene expression with that deduced using network targeting, they found that TF regulation of tissue-specific function was evident when using gene targeting metrics, but it was largely independent of TF expression level (Sonawane et al., 2017). PANDA analysis also identified unique, tissue-specific

[1]https://www.cancer.gov/tcga

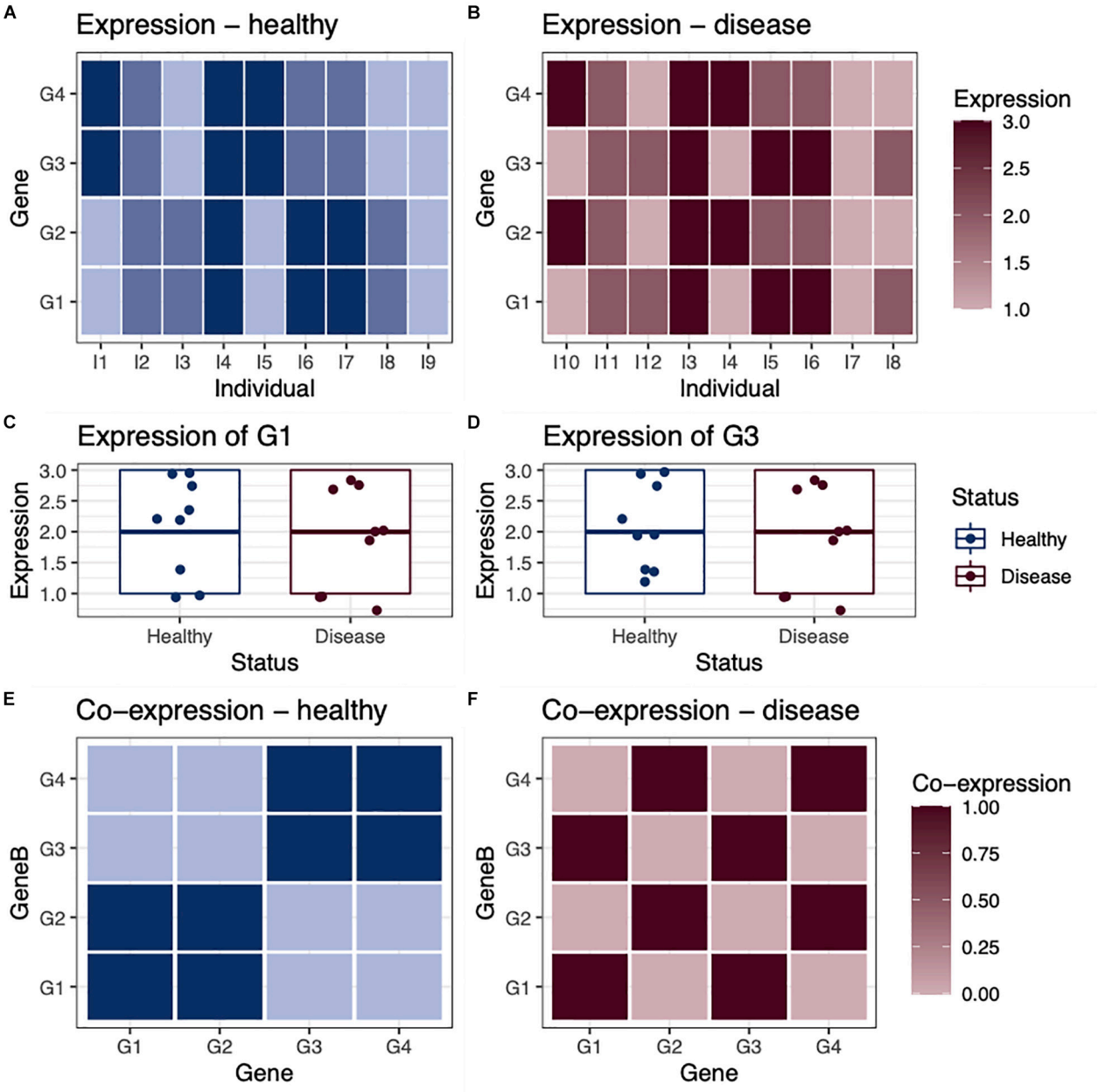

FIGURE 1 | Differential expression vs. differential co-expression. As a toy example, we consider the expression of four genes in **(A)** nine healthy individuals and **(B)** nine individuals with a disease. In this example, none of the genes are differentially expressed, as they have a similar average expression level in both healthy and disease individuals, shown in examples **(C)** gene G1 and **(D)** gene G3. However, when we look at the co-expression between genes within healthy individuals **(E)** and within disease individuals **(F)**, we see that there is obvious differential co-expression between genes in healthy individuals, compared with disease individuals.

targeting patterns in the TF–gene edges and found significant enrichment of tissue-specific regulatory edges targeting tissue-specific expressed genes. This example demonstrates how differences in tissue-specific regulatory relationships between TFs and genes can give rise to the distinct phenotype by altering regulation of key biological processes (Sonawane et al., 2017).

A study of COPD by Glass et al. (2014) found patterns of gene targeting that differed between men and women and offered a possible explanation for higher disease susceptibility of women compared with men. They first compared gene expression between males and females with COPD and found little evidence of differential expression of autosomal genes. Using a resampling approach, they constructed an ensemble of 100 male and 100 female PANDA GRNs. They calculated a targeting score for each gene in each network defined as the sum of all inbound edge weights for the gene. They found several

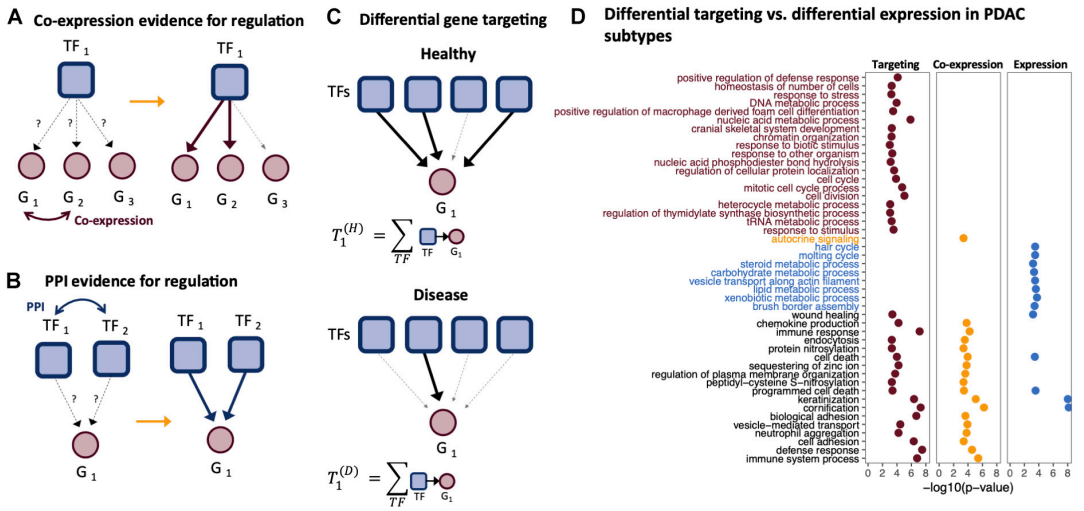

FIGURE 2 | Gene targeting. Gene targeting scores are derived from gene regulatory networks (GRNs) and thus are influenced by the components used to derive the edge weights of a GRN. For example, PANDA GRNs include information regarding the **(A)** co-expression relationships between genes and **(B)** protein–protein interactions between transcription factors (TFs). **(C)** Gene targeting scores are calculated as the sum of the weights across all inbound edges pointing to a gene. **(D)** Gene Ontology (GO) enrichment of ranked differential gene scores comparing the basal-like and classical pancreatic ductal adenocarcinoma (PDAC) subtypes. Genes were ranked by differential targeting (red), differential co-expression (orange), and differential expression (blue).

genes that were not differentially expressed but nevertheless had significantly different targeting scores between the sexes. Pre-ranked gene set enrichment analysis based on targeting scores found many processes associated with mitochondrial function that were more highly targeted in females; these processes had previously been implicated in many aspects of COPD and lung disease (Glass et al., 2014). These results suggest that that differential regulation of processes associated with disease may alter disease development and progression in meaningful ways.

Lopes-Ramos et al. (2020) investigated sex differences in gene expression and gene regulation in 29 human tissues by constructing individual-specific networks for each sample in each tissue. Differential edge weights between males and females were identified, and genes were classified as differentially targeted if at least 5% of their inbound edge weights were significantly different between males and females. This allowed genes to be classified as being male-biased if most (>60%) of the inbound differential edges were higher in males, female-biased if most (>60%) of the inbound differential edges were higher in females, and sex-divergent if the number of inbound differential edges was evenly split between being higher in males and higher in females (Lopes-Ramos et al., 2020). Consistent with previous studies, they found little differential gene expression except in breast tissue, with the median number of differentially expressed genes across tissues equal to 64. However, widespread sex-biased targeting was detected in all tissues, with a median number of differentially targeted genes across tissues equal to 169. Interestingly, the sex hormone receptors ESR1, ESR2, and AR were differentially targeted between male and female individuals in several tissues such as the breast, heart, and blood, despite the fact that those hormone receptors were not differentially expressed.

SPECIFIC EXAMPLE: PANCREATIC DUCTAL ADENOCARCINOMA SUBTYPES

PDAC is a lethal disease involving heterogeneous tumors composed of diverse cell types including tumor epithelial cells and components of the tumor microenvironment such as immune cells and fibroblasts. Molecular subtypes of PDAC have been identified through gene expression analysis (Collisson et al., 2011; Moffitt et al., 2015; Bailey et al., 2016; Puleo et al., 2018; Maurer et al., 2019), and the basal-like and classical subtypes first identified by Moffitt et al. have been associated with both prognosis and treatment response (Moffitt et al., 2015; Aung et al., 2018; Rashid et al., 2020; O'Kane et al., 2020). The basal-like subtype is associated with worse median patient survival and resistance to chemotherapy (Moffitt et al., 2015; O'Kane et al., 2020) and has characteristically high expression of keratins and laminins, both structural proteins also associated with the basal subtypes of breast and bladder cancers (Damrauer et al., 2014; McConkey et al., 2014; Weinstein et al., 2014). The classical subtype shows better response to treatment and better overall survival and is marked by increased expression of GATA binding protein 6 (GATA6), a TF involved in cell differentiation.

To identify factors driving these subtypes, we compared differential gene expression and differential GRN gene targeting scores between basal-like and classical subtypes of 150 PDAC tumors using TCGA (Cancer Genome Atlas Research Network, 2017) transcripts per kilobase million (TPM) expression data processed using Recount (Collado-Torres et al., 2017). We used PANDA and LIONESS to construct sample specific GRNs and chose to limit our analysis to those genes with a high standard

deviation of logTPMs [sd(logTPMs) > 0.4] across samples. For each gene in each individual tumor, a gene targeting score was calculated as the sum of all inbound edge weights surrounding each gene. Separately, we also calculated a sample-specific co-expression network for each tumor (Kuijjer et al., 2019a) and for each gene in each sample, and we calculated a gene co-expression score equal to the sum of each gene's co-expression edges surrounding the gene. We used limma (Ritchie et al., 2015) to compare the expression data, the correlation networks, and GRNs between the basal-like and classical subtypes, allowing us to identify differentially expressed genes, differentially co-expressed genes, and differentially targeted genes, respectively.

The three genes found to be most significantly differentially targeted in GRNs, but not differentially expressed, between basal-like and classical subtypes are folate receptor beta (FOLR2), hedgehog interacting protein (HHIP), and the CD209 antigen C-type lectin domain family 4 member L (CD209). FOLR2 encodes the folate receptor 2 protein and is known to be overexpressed in tumor-associated macrophages (Tie et al., 2020). HHIP codes for the hedgehog interacting protein; the hedgehog signaling pathway regulates cell differentiation and proliferation and is activated in several cancers including PDAC (Yang et al., 2010; Honselmann et al., 2015; Gu et al., 2016). CD209 codes for a C-type lectin domain family 4 protein and is a dendritic cell marker. The roles that these play in PDAC have not yet been explored.

Ranking genes according to three different metrics, differential targeting, differential co-expression, and differential expression, we performed ranked gene set enrichment analysis (Eden et al., 2009; Supek et al., 2011) to identify significantly over-represented biological process Gene Ontology (GO) terms. Both the differential targeting analysis and differential expression analysis identified keratinization, cornification, cell death, and wound healing as differentiating basal-like and classical samples. However, several immune-related processes, epigenetic, and cell cycle processes found by differential targeting analysis were missed using differential expression alone (**Figure 2**). Functional enrichment using co-expression scores to rank genes identified some processes similar to those found using differential targeting but missed several important pathways related to cell cycle and other processes, such as chromatin organization.

The identification of keratinization as enriched in differentially expressed genes is consistent with previous studies that identified genes encoding keratins and laminins as biomarkers for basal-like tumors (Moffitt et al., 2015, O'Kane et al., 2020). The fact that both differential expression and differential targeting identified keratinization and cell adhesion as biological processes distinguishing PDAC subtypes serves as an internal consistency check. Differential targeting alone identified processes related to the immune system and speaks to the importance of the tumor microenvironment, which is known to influence PDAC prognosis and drug response. A high degree of tumor-associated macrophage infiltration has been linked to lower survival (Karamitopoulou, 2019), which is a known hallmark of the basal-like subtype, and it is possible that the differential targeting analysis is detecting cross-talk between the tumor and the tumor microenvironment. The differential targeting of epigenetic functions between subtypes is consistent with reports that the basal-like and classical subtypes have distinct epigenetic landscapes (Lomberk et al., 2018).

This analysis of PDAC subtypes, although abbreviated, demonstrates the power of using GRN inference and gene targeting score analysis to identify regulatory processes that characterize distinct phenotypes—including processes that are distinct from those that are associated with patterns of gene expression. The biologically relevant differences we see in targeting but not expression or co-expression suggest that regulatory control, even if not activated, is important in defining health and disease.

DISCUSSION

There is growing experimental evidence of the importance of complex regulatory processes in distinguishing phenotypes in health and disease. For example, the Wilms tumor-1 (WT1) TF is a master regulator that targets several essential genes in kidney podocyte cells. Ettou et al. (2020) investigated WT1-based gene regulation during podocyte injury and found that WT1 maintained open chromatin in the regions of its target genes but that the expression level of WT1 was not universally associated with the intensity of its binding. They also found that WT1 could cause either an increase or a decrease in the expression of its target genes. The role of complex regulatory processes is further illustrated by the work of Carnesecchi et al. (2020), who investigated how a single TF could regulate different developmental programs in various cell lineages. They showed that the Ubx TF forms different complexes with distinct binding partners in various cell lineages despite the fact that most of the interaction partners showed no differential expression across the lineages.

Taken together, the results reported by Ettou and Carnesecchi illustrate the complexity of regulatory processes and the importance that regulatory targeting plays in defining phenotype, even in instances in which a key regulator does not itself substantially change in its expression levels. Their results also point to the importance of modeling both "direct" and "indirect" regulation of genes by TFs and the complexes they form. Among the methods for GRN inference, PANDA (and by extension, PANDA+LIONESS) is singular in considering interactions between TF proteins in its model. PANDA's integrative approach using TF–TF interactions, predicted TF–gene regulatory relationships, and gene co-expression data, refines the inputs to optimize agreement between them; the resulting networks provide unique insight into regulatory processes that are linked to phenotypes.

The work summarized in this perspective demonstrates the value of the gene targeting score as a metric for assessing the drivers of phenotypic differences. Gene targeting scores not only capture structural characteristics of regulatory networks but also allow for the identification of processes that may be activated in response to appropriate stimuli and in this way help to

define phenotypes and disease subtypes. For example, Lopes-Ramos et al. (2018) performed gene targeting analysis on gene expression data from colon cancer tumor samples and discovered differences between males and females in the regulation of genes involved drug metabolism, suggesting that male and female tumor cells are programmed to respond differently. They found that genes in drug metabolism pathways, particularly those acting through cytochrome P450, had higher targeting scores in female networks than male networks. Furthermore, higher targeting of the drug metabolism pathways was found to correlate with patient survival, indicating a mechanism for sex-divergent response to chemotherapy in colon cancer.

Our application of PANDA and LIONESS in comparing PDAC subtypes demonstrates the value of the GRN-based approach and of using network-based metrics such as gene targeting to characterize properties of biological systems. We constructed sample-specific GRNs for 150 PDAC tumors and used gene targeting to compare the topologies of networks derived from basal-like and classical tumor subtypes. We found that differential targeting analysis identified compelling differences between the two subtypes in the regulation of processes related to cell cycle, immune, and epigenetic functions, none of which were seen in a standard differential expression analysis. Given that PDAC tumors are known to exhibit immune infiltration, and that the subtypes differ in both their epigenetic landscapes and patient survival, our identification of relevant processes illustrates how GRN-based methods can provide important and relevant biological insights into disease-associated processes beyond what is seen using other analytical methods.

PANDA and LIONESS software for GRN analyses and identification of differential targeting are freely available as open-source tools with extensive documentation (netzoo.github.io) and can easily be implemented in most analytical workflows. We hope that this review motivates the broader use and appreciation of gene targeting analysis.

DATA AVAILABILITY STATEMENT

Publicly available data analyzed in this study were obtained from recount2 (https://jhubiostatistics.shinyapps.io/recount/). We provide an interactive Jupyter notebook hosted in netbooks (netbooks.networkmedicine.org) to reproduce the differential targeting analysis and the PDAC gene regulatory networks are available in GRAND database (https://grand.networkmedicine.org/cancers).

AUTHOR CONTRIBUTIONS

DW drafted the manuscript and performed the analysis. All authors discussed, planned, reviewed and edited the manuscript.

FUNDING

DW, MBG, and JQ were supported by a grant from the US National Cancer Institute (NCI), R35CA220523. MBG and JQ were further supported by NCI grant U24CA231846. KG was supported by a grant from the US National Heart, Lung, and Blood Institute (NHLBI), K25HL133599. JP was supported by a grant from the US National Heart, Lung and Blood Institute, K25HL140186. JJY was supported by R01CA199064, CA193650, and U24CA211000.

REFERENCES

Aung, K. L., Fischer, S. E., Denroche, R. E., Jang, G.-H., Dodd, A., Creighton, S., et al. (2018). Genomics-driven precision medicine for advanced pancreatic cancer: early results from the COMPASS trial. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 24, 1344–1354. doi: 10.1158/1078-0432.CCR-17-2994

Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531, 47–52. doi: 10.1038/nature16965

Cancer Genome Atlas Research Network (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 32, 185–203.e13. doi: 10.1016/j.ccell.2017.07.007

Carnesecchi, J., Sigismondo, G., Domsch, K., Baader, C. E. P., Rafiee, M.-R., Krijgsveld, J., et al. (2020). Multi-level and lineage-specific interactomes of the Hox transcription factor Ubx contribute to its functional specificity. *Nat. Commun.* 11:1388. doi: 10.1038/s41467-020-15223-x

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., et al. (2017). Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* 35, 319–321. doi: 10.1038/nbt.3838

Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., et al. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* 17, 500–503. doi: 10.1038/nm.2344

Damrauer, J. S., Hoadley, K. A., Chism, D. D., Fan, C., Tiganelli, C. J., Wobker, S. E., et al. (2014). Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3110–3115. doi: 10.1073/pnas.1318376111

Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., Behr, J., Montazeri, H., Hall, M. N., et al. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 34, 2441–2448. doi: 10.1093/bioinformatics/bty148

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48. doi: 10.1186/1471-2105-10-48

Ettou, S., Jung, Y. L., Miyoshi, T., Jain, D., Hiratsuka, K., Schumacher, V., et al. (2020). Epigenetic transcriptional reprogramming by WT1 mediates a repair response during podocyte injury. *Sci. Adv.* 6:eabb5460. doi: 10.1126/sciadv.abb5460

Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013). Passing Messages between biological networks to refine predicted interactions. *PLoS One* 8, e64832. doi: 10.1371/journal.pone.0064832

Glass, K., Quackenbush, J., Silverman, E. K., Celli, B., Rennard, S. I., Yuan, G.-C., et al. (2014). Sexually-dimorphic targeting of functionally-related genes in COPD. *BMC Syst. Biol.* 8:118. doi: 10.1186/s12918-014-0118-y

Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B., and Yuan, G.-C. (2015). A network model for angiogenesis in ovarian cancer. *BMC Bioinformatics* 16:115. doi: 10.1186/s12859-015-0551-y

Gu, D., Schlotman, K. E., and Xie, J. (2016). Deciphering the role of hedgehog signaling in pancreatic cancer. *J. Biomed. Res.* 30, 353–360. doi: 10.7555/JBR.30.20150107

Gumpinger, A. C., Lage, K., Horn, H., and Borgwardt, K. (2020). Prediction of cancer driver genes through network-based moment propagation of mutation scores. *Bioinformatics* 36, i508–i515. doi: 10.1093/bioinformatics/btaa452

Honselmann, K. C., Pross, M., Wellner, U. F., Deichmann, S., Keck, T., Bausch, D., et al. (2015). Regulation mechanisms of the hedgehog pathway in pancreatic cancer: a review. *JOP* 16, 25–32.

Horn, H., Lawrence, M. S., Chouinard, C. R., Shrestha, Y., Hu, J. X., Worstell, E., et al. (2018). NetSig: network-based discovery from cancer genomes. *Nat. Methods* 15, 61–66. doi: 10.1038/nmeth.4514

Karamitopoulou, E. (2019). Tumour microenvironment of pancreatic cancer: immune landscape is dictated by molecular and histopathological features. *Br. J. Cancer* 121, 5–14. doi: 10.1038/s41416-019-0479-5

Kuijjer, M. L., Hsieh, P.-H., Quackenbush, J., and Glass, K. (2019a). lionessR: single sample network inference in R. *BMC Cancer* 19:1003. doi: 10.1186/s12885-019-6235-7

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. (2019b). Estimating Sample-Specific Regulatory Networks. *iScience* 14, 226–240. doi: 10.1016/j.isci.2019.03.021

Kwa, M., Makris, A., and Esteva, F. J. (2017). Clinical utility of gene-expression signatures in early stage breast cancer. *Nat. Rev. Clin. Oncol.* 14, 595–610. doi: 10.1038/nrclinonc.2017.74

Lomberk, G., Blum, Y., Nicolle, R., Nair, A., Gaonkar, K. S., Marisa, L., et al. (2018). Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nat. Commun.* 9:1978. doi: 10.1038/s41467-018-04383-6

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45:580.

Lopes-Ramos, C. M., Chen, C.-Y., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., et al. (2020). Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell Rep.* 31:107795. doi: 10.1016/j.celrep.2020.107795

Lopes-Ramos, C. M., Kuijjer, M. L., Ogino, S., Fuchs, C. S., DeMeo, D. L., Glass, K., et al. (2018). Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer Res.* 78, 5538–5547. doi: 10.1158/0008-5472.CAN-18-0454

Maurer, C., Holmstrom, S. R., He, J., Laise, P., Su, T., Ahmed, A., et al. (2019). Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. *Gut* 68, 1034–1043. doi: 10.1136/gutjnl-2018-317706

McConkey, D. J., Choi, W., and Dinney, C. P. N. (2014). New insights into subtypes of invasive bladder cancer: considerations of the clinician. *Eur. Urol.* 66, 609–610. doi: 10.1016/j.eururo.2014.05.006

Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., et al. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* 47, 1168–1178. doi: 10.1038/ng.3398

O'Kane, G. M., Grünwald, B. T., Jang, G.-H., Masoomian, M., Picardo, S., Grant, R. C., et al. (2020). GATA6 expression distinguishes classical and basal-like subtypes in advanced pancreatic cancer. *Clin. Cancer Res.* 26, 4901–4910. doi: 10.1158/1078-0432.CCR-19-3724

Puleo, F., Nicolle, R., Blum, Y., Cros, J., Marisa, L., Demetter, P., et al. (2018). Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. *Gastroenterology* 155, 1999–2013.e3. doi: 10.1053/j.gastro.2018.08.033 * 1999-2013.e3,

Qiu, W., Guo, F., Glass, K., Yuan, G. C., Quackenbush, J., Zhou, X., et al. (2018). Differential connectivity of gene regulatory networks distinguishes corticosteroid response in asthma. *J. Allergy Clin. Immunol.* 141, 1250–1258. doi: 10.1016/j.jaci.2017.05.052

Ramadan, E., Alinsaif, S., and Hassan, M. R. (2016). Network topology measures for identifying disease-gene association in breast cancer. *BMC Bioinformatics* 17:274. doi: 10.1186/s12859-016-1095-5

Rashid, N. U., Peng, X. L., Jin, C., Moffitt, R. A., Volmar, K. E., Belt, B. A., et al. (2020). Purity independent subtyping of tumors (PurIST), a clinically robust, single-sample classifier for tumor subtyping in pancreatic cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 26, 82–92. doi: 10.1158/1078-0432.CCR-19-1467

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

Rodriguez-Salas, N., Dominguez, G., Barderas, R., Mendiola, M., García-Albéniz, X., Maurel, J., et al. (2017). Clinical relevance of colorectal cancer molecular subtypes. *Crit. Rev. Oncol. Hematol.* 109, 9–19. doi: 10.1016/j.critrevonc.2016.11.007

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

Rudin, C. M., Poirier, J. T., Byers, L. A., Dive, C., Dowlati, A., George, J., et al. (2019). Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nat. Rev. Cancer* 19, 289–297. doi: 10.1038/s41568-019-0133-9

Sjödahl, G., Jackson, C. L., Bartlett, J. M., Siemens, D. R., and Berman, D. M. (2019). Molecular profiling in muscle-invasive bladder cancer: more than the sum of its parts. *J. Pathol.* 247, 563–573. doi: 10.1002/path.5230

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088. doi: 10.1016/j.celrep.2017.10.001

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

Tie, Y., Zheng, H., He, Z., Yang, J., Shao, B., Liu, L., et al. (2020). Targeting folate receptor β positive tumor-associated macrophages in lung cancer with a folate-modified liposomal complex. *Signal Transduct. Target. Ther.* 5, 1–15. doi: 10.1038/s41392-020-0115-0

Vargas, A. J., Quackenbush, J., and Glass, K. (2016). Diet-induced weight loss leads to a switch in gene regulatory network control in the rectal mucosa. *Genomics* 108, 126–133. doi: 10.1016/j.ygeno.2016.08.001

Weinstein, J. N., Akbani, R., Broom, B. M., Wang, W., Verhaak, R. G. W., McConkey, D., et al. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507, 315–322. doi: 10.1038/nature12965

Yang, L., Xie, G., Fan, Q., and Xie, J. (2010). Activation of the hedgehog-signaling pathway in human cancer and the clinical implications. *Oncogene* 29, 469–481. doi: 10.1038/onc.2009.392

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fgene.2021.630215

Comparing Statistical Tests for Differential Network Analysis of Gene Modules

Jaron Arbet[1], Yaxu Zhuang[1], Elizabeth Litkowski[2], Laura Saba[3†] and Katerina Kechris[1†]*

[1] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [2] Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [3] Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, Aurora CO, United States

OPEN ACCESS

Edited by:
Kimberly Glass,
Brigham and Women's Hospital and Harvard Medical School, United States

Reviewed by:
Guillermo Barturen,
Junta de Andalucía de Genómica e Investigación Oncológica (GENYO), Spain
Ling-Yun Wu,
Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Katerina Kechris
katerina.kechris@cuanschutz.edu

[†]These authors share senior authorship

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

Received: 17 November 2020
Accepted: 19 April 2021
Published: 19 May 2021

Citation:
Arbet J, Zhuang Y, Litkowski E, Saba L and Kechris K (2021) Comparing Statistical Tests for Differential Network Analysis of Gene Modules. Front. Genet. 12:630215. doi: 10.3389/fgene.2021.630215

Genes often work together to perform complex biological processes, and "networks" provide a versatile framework for representing the interactions between multiple genes. Differential network analysis (DiNA) quantifies how this network structure differs between two or more groups/phenotypes (e.g., disease subjects and healthy controls), with the goal of determining whether differences in network structure can help explain differences between phenotypes. In this paper, we focus on gene co-expression networks, although in principle, the methods studied can be used for DiNA for other types of features (e.g., metabolome, epigenome, microbiome, proteome, etc.). Three common applications of DiNA involve (1) testing whether the connections to a single gene differ between groups, (2) testing whether the connection between a *pair* of genes differs between groups, or (3) testing whether the connections within a "module" (a subset of 3 or more genes) differs between groups. This article focuses on the latter, as there is a lack of studies comparing statistical methods for identifying differentially co-expressed modules (DCMs). Through extensive simulations, we compare several previously proposed test statistics and a new p-norm difference test (PND). We demonstrate that the true positive rate of the proposed PND test is competitive with and often higher than the other methods, while controlling the false positive rate. The R package discoMod (differentially co-expressed modules) implements the proposed method and provides a full pipeline for identifying DCMs: clustering tools to derive gene modules, tests to identify DCMs, and methods for visualizing the results.

Keywords: differential network analysis, differentially co-expressed modules, gene co-expression networks, statistical inference, networks

Abbreviations: DiNA, differential network analysis; DCM, differentially co-expressed module; TOM, topological overlap measure; PND, p-norm difference test; DI, dispersion index; MAD, mean absolute difference; paired, paired t-test statistic; wilcoxSRT, Wilcoxon signed rank test statistic; QAP, Quadratic assignment procedure test statistic; GHD, Generalized Hamming distance test statistic; TPR, true positive rate; FPR, false positive rate; CS, compound symmetric correlation structure; AR1, autoregressive order 1 correlation structure; BIC, Bayesian information criterion.

INTRODUCTION

Gene expression studies measure expression levels on thousands of genes, with a goal of identifying individual genes or groups of genes that explain differences between phenotypes of interest (e.g., disease subjects and healthy controls). An extensive literature exists regarding methods for identifying individual genes whose *mean* expression differs between groups (Soneson and Delorenzi, 2013; Huang et al., 2015), often referred to as differentially expressed genes. Pathway analysis (Huang et al., 2009; Emmert-Streib and Glazko, 2011; Ramanan et al., 2012; De Leeuw et al., 2016) aims to identify groups of genes (pathways or gene sets) that are enriched with differentially expressed genes (competitive tests) or whose overall mean structure differs between groups (self-contained tests). However, all of these methods ignore interactions between multiple genes.

In recent years, there is a growing interest in systems or network biology (Barabasi and Oltvai, 2004; Chuang et al., 2010; Barabási et al., 2011) in which one uses a statistical network to model the relationships between multiple genes (or other molecular features). For analyzing networks of gene expression (gene co-expression networks), genes are represented as nodes in the network, with the relationships between genes represented as lines/edges connecting the nodes. The strength of the connections is usually represented by a correlation matrix that measures the pairwise correlations between all genes. An adjacency matrix and the topological overlap measure (TOM) are other common forms of representing the connections between genes (Zhang and Horvath, 2005). See Singh et al. (2018) and van Dam et al. (2018) for review of important terminology and concepts used in gene co-expression network analysis.

In differential network analysis (DiNA), the goal is to determine whether the network structure differs between two or more phenotype groups (see de la Fuente, 2010; Kayano et al., 2014; Singh et al., 2018; Shojaie, 2020 for review). Many of the methods of DiNA of gene co-expression networks can be classified into three categories: (1) Identifying a single node (gene) in the network where the connections at that node differ between phenotype groups. For example (Lichtblau et al., 2017), compare 10 methods for quantifying node specific differences between groups. (2) Identifying *pairs* of genes whose correlation differs between two or more groups (Liu et al., 2010; Dawson et al., 2012; Fukushima, 2013; Ha et al., 2015; McKenzie et al., 2016; Siska et al., 2016), i.e., the focus is on the connection between only two genes at a time. (3) The last category, and the focus of this paper, attempts to identify subsets of co-expressed genes, called modules (also referred to as clusters or communities; Petereit et al., 2016) whose connections differ between phenotypes (Watson, 2006; Choi and Kendziorski, 2009; Gill et al., 2010; Tesson et al., 2010; Langfelder et al., 2011; Rahmatallah et al., 2014; Jardim et al., 2019). Modules are groups of multiple genes that interact in a coordinated manner, e.g., their expression levels are correlated. Two main approaches are used for defining modules: one may obtain *a priori* predefined modules from a database (e.g., KEGG, Kanehisa and Goto, 2000; GO, Ashburner et al., 2000), or one can use clustering methods (Langfelder and Horvath, 2008; Andreopoulos et al., 2009; Tesson et al., 2010; Xu and Wunsch, 2010) to derive data dependent modules. Comparing clustering methods for deriving data-dependent modules is beyond the scope of this paper (see Kakati et al., 2019 for one comparative study). After defining the modules, the final step is to test whether a module's connections differ between phenotype groups, which is known as a "differentially co-expressed module" (DCM). The null hypothesis is that the network structure within the module is equal between the groups being compared. Although several methods have been proposed for testing whether the network structure within a module differs between two groups (Watson, 2006; Choi and Kendziorski, 2009; Gill et al., 2010; Tesson et al., 2010; Langfelder et al., 2011; Rahmatallah et al., 2014; Jardim et al., 2019), there is a lack of simulation studies comparing such methods. Therefore, we attempt to fill this gap by conducting extensive simulations of different network structures to compare existing test statistics for identifying DCMs, as well as a new framework the p-norm difference (PND) test that encompasses previous approaches but also provides more flexibility. Tests in the PND framework demonstrate a true positive rate that is competitive with and often higher than existing methods, while controlling the false positive rate. Lastly, the discoMod R package is made available, which implements a full pipeline for identifying DCMs: clustering tools to derive modules, tests to identify DCMs, and methods to visualize the results.

MATERIALS AND METHODS

Assume one has a list of M number of gene modules, which may have been predefined from a database (Ashburner et al., 2000; Kanehisa and Goto, 2000) or derived using clustering methods (Langfelder and Horvath, 2008; Andreopoulos et al., 2009; Tesson et al., 2010; Xu and Wunsch, 2010). Each module contains three or more genes, and the modules need not be disjoint (e.g., the same gene could appear in more than one module). Although we focus on genes, all the methods discussed can be used for other types of features besides gene expression (e.g., metabolome, epigenome, microbiome, proteome).

Let $X^{(gm)}$ be the gene expression matrix for groups $g = 1, 2$ and modules $m = 1, \ldots, M$, where each gene expression variable may be measured as an integer count (i.e., number of mapped reads) from a sequencing platform or a continuous value from a microarray platform. Next, let $S^{(gm)}$ be a similarity matrix used to represent the network structure of the mth module within the gth group. Note $S^{(gm)}$ is a symmetric $|P_m|^* |P_m|$ matrix where $|P_m|$ represents the number of genes in the mth module, $i = 1, 2, \ldots |P_m|$ is the gene index, and $S_{ij}^{(gm)}$ is a measure of similarity between genes i and j. Several measures of similarity between two genes ($S_{ij}^{(gm)}$) have been used, including: correlation (Pearson, Spearman, or Kendall), partial correlation, or mutual information (Gill et al., 2010; Kumari et al., 2012; Kayano et al., 2014; van Dam et al., 2018). This similarity matrix may be further represented as an adjacency or TOM matrix (Ravasz et al., 2002; Zhang and Horvath, 2005; Langfelder and Horvath, 2008) which will be discussed later.

For the mth module with similarity matrices $S^{(gm)}$ for both groups ($g = 1, 2$), we are interested in testing the following null (H_0) and alternative (H_A) hypotheses:

$$H_0 : S^{(1m)} = S^{(2m)} \; vs. \; H_A : S^{(1m)} \neq S^{(2m)} \quad (1)$$

Test Statistics for Identifying DCMs

We now define several test statistics that will be compared for testing (1). Given that $S^{(gm)}$ is a symmetric $|P_m|^* |P_m|$ matrix, let $V^{(gm)}$ be a vector of the lower triangle of $S^{(gm)}$, thus $V^{(gm)}$ is a vector of length $\lambda_m = \frac{|P_m|(|P_m|-1)}{2}$. Let $k = 1,\ldots, \lambda_m$ be the indexing variable for iterating between the elements of $V^{(gm)}$. Many test statistics can be formulated as functions of the difference (or product) in $V^{(gm)}$ between the two groups. For example, the "Dispersion Index" (DI), used by GSCA (Choi and Kendziorski, 2009) and DiffCoEx (Tesson et al., 2010), for the mth module is defined as:

$$DI\left(V^{(1m)}, V^{(2m)}\right) = \sqrt{\frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m}\left(V_k^{(1m)} - V_k^{(2m)}\right)^2} \quad (2)$$

The mean absolute difference (MAD) (Gill et al., 2010; Ruan et al., 2015), is defined as:

$$MAD\left(V^{(1m)}, V^{(2m)}\right) = \frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m}\left|V_k^{(1m)} - V_k^{(2m)}\right| \quad (3)$$

The DGCA R package (McKenzie et al., 2016) simply uses the mean (or median) of the differences. A potential problem with this approach is that positive and negative differences can cancel out, thus losing power to detect DCMs where some correlations increase while other correlations decrease between conditions. Nevertheless, similar to their approach, we consider the paired t-test statistic (mean of the differences divided by the standard error of the mean difference):

$$pairedT\left(V^{(1m)}, V^{(2m)}\right) = \left[\frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m}\left(V_k^{(1m)} - V_k^{(2m)}\right)\right] * \frac{\sqrt{\lambda_m}}{\sqrt{\frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m}\left(V_k^{(1m)} - V_k^{(2m)}\right)^2}} \quad (4)$$

Similar to the paired t-test statistic, we also consider the Wilcoxon signed rank test statistic, as implemented in the *wilcox.test* base R function (R Core Team, 2018). The Wilcoxon signed rank test statistic ranks the differences of $|V^{(1m)} - V^{(2m)}|$ and then sums the ranks where the sign of $(V^{(1m)} - V^{(2m)})$ is positive.

Three additional statistics are compared that were also considered in (Ruan et al., 2015): the Quadratic Assignment Procedure (QAP), GCOR, and Generalized Hamming Distance (GHD). These statistics are defined as:

$$QAP\left(V^{(1m)}, V^{(2m)}\right) = \frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m} V_k^{(1m)} * V_k^{(2m)} \quad (5)$$

$$GCOR\left(V^{(1m)}, V^{(2m)}\right) = \sum_{k=1}^{\lambda_m}(V_k^{(1m)} - \overline{V}^{(1m)}) * (V_k^{(2m)} - \overline{V}^{(2m)}) \quad (6)$$

$$GHD\left(V^{(1m)}, V^{(2m)}\right) = \frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m}\left[(V_k^{(1m)} - \overline{V}^{(1m)}) - (V_k^{(2m)} - \overline{V}^{(2m)})\right]^2 \quad (7)$$

Where $\overline{V}^{(1m)}$ and $\overline{V}^{(2m)}$ are the means of $V_k^{(1m)}$ and $V_k^{(2m)}$, respectively.

The test statistic from GSNCA (Rahmatallah et al., 2014) is also considered in this manuscript. GSNCA does not fit within the previously described framework of comparing the difference (or product) of the vectors $V^{(1m)}$ and $V^{(2m)}$, thus we refer the reader to the original paper for the formal definition. Nevertheless, GSNCA can still be used to test whether the network structure of a module differs between the two groups. Briefly, GSNCA assigns a weight vector to each group of length $|P_m|$ (one weight per gene) and the test statistic is the sum of the absolute differences of the weight vector between the two groups. The ith gene is given a weight w_i that is proportional to the sum of the correlations between the ith gene with all other genes. Thus, a gene that is highly correlated with many other genes will be given a larger weight, which indicates the gene may have regulatory importance.

We propose a new class of test statistics for identifying DCMs, the p-norm difference test ("PND"), which uses the p-norm (or L^P norm) of the differences between $V^{(1m)}$ and $V^{(2m)}$.

$$PND\left(V^{(1m)}, V^{(2m)}, p\right) = \left(\frac{1}{\lambda_m}\sum_{k=1}^{\lambda_m}\left|V_k^{(1m)} - V_k^{(2m)}\right|^p\right)^{\frac{1}{p}} \quad (8)$$

The motivation of the PND test is, given a "partially differentially co-expressed module" (a module where some of the correlations, but not all, change between groups), then the higher the exponent p, the less weight is given to the null correlations that do not change between groups. Therefore, we expect the PND test with a large value of p (e.g., $p \geq 4$) to be more sensitive for detecting DCMs where only a small proportion of the module correlations change between conditions. In our simulations, we consider four different values for the exponent p: 4, 6, 8, and 20. Note the Dispersion Index is equivalent to the PND test with $p = 2$.

For all previously defined test statistics, the elements of $V^{(gm)}$ are unlikely to be independent since they come from a structured similarity matrix (e.g., a correlation matrix), thus it is challenging to derive the sampling distribution under the null hypothesis without imposing additional assumptions. Therefore, we use a non-parametric permutation method to calculate p-values, which accounts for this complex dependency structure. Specifically, given a test statistic θ_m for the mth module, the permutation p-value is defined as follows:

1. Using the original gene expression matrices for each group, $X^{(1m)}$ and $X^{(2m)}$, and for a particular similarity measure of interest, calculate the similarity matrices $S^{(1m)}$ and $S^{(2m)}$. Then calculate the test statistic θ^m for testing the null hypothesis in (1)

2. For $b = 1, 2, \ldots, B$ (B total number of permutations):

a. Combine the gene expression matrices of both groups, $X^{(1m)}$ and $X^{(2m)}$, and randomly shuffle (permute) the group labels, to create new "permuted" gene expression matrices $X^{(1m)}(b)$ and $X^{(2m)}(b)$.
b. Calculate the new similarity matrices $S^{(1m)}(b)$ and $S^{(2m)}(b)$ based on the permuted gene expression matrices $X^{(1m)}(b)$ and $X^{(2m)}(b)$
c. Calculate the permuted test statistic $\theta^m(b)$ based on $S^{(1m)}(b)$ and $S^{(2m)}(b)$.

3. Calculate the permutation p-value $= \frac{[\sum_{b=1}^{B} I(|\theta^m(b)| \geq |\theta^m|)] + 1}{B+1}$

Lastly, we compare with three tests from the HDtest R package: HD (Chang et al., 2017), CLX (Cai et al., 2013), and Schott (Schott, 2007). These tests are designed to compare a high dimensional covariance matrix between two groups. CLX and Schott use asymptotic approximations to calculate p-values (and are thus much faster than all other methods considered) while HD uses a multiplier bootstrap method.

Similarity Measures for Constructing Test Statistics

For the test statistics we are comparing, one needs to decide what type of similarity measure will be used for $S_{ij}^{(gm)}$ (similarity between any two genes i and j in the mth module of group g), when testing the null hypothesis of (1). As previously mentioned, several similarity measures have been used in practice: correlation (Pearson, Spearman, or Kendall), partial correlation, mutual information (Gill et al., 2010; Kumari et al., 2012; Kayano et al., 2014; van Dam et al., 2018), and adjacency or TOM matrices (Zhang and Horvath, 2005; Langfelder and Horvath, 2008). Gaussian and semi/nonparametric graphical models have also been used to measure the conditional dependence between each pair of genes (i.e., partial correlations) (Friedman et al., 2008; Wang et al., 2016; Zhang et al., 2018; Shojaie, 2020). It is beyond the scope of this paper to compare all of these similarity measures for constructing networks.

This paper will focus on comparing two particular types of unconditional similarity measures: correlation vs. TOM (Ravasz et al., 2002; Langfelder and Horvath, 2008). For correlation, we will use Spearman's correlation (rather than Pearson's correlation), based on recommendations from other studies (Kumari et al., 2012; Siska and Kechris, 2017). When calculating the similarity between two genes, $S_{ij}^{(gm)}$, unconditional correlation only considers the relationship between the two genes i and j, while ignoring any shared relationships these genes might have with other genes. This is true for all of the aforementioned measures of similarity, except for partial correlation and TOM. In contrast to unconditional correlation, TOM captures shared relationships or "neighbors" between the two genes, as defined in Equation (9) (note the signed version of the adjacency measure, a_{ij}, is defined in Langfelder and Horvath, 2008).

$$TOM_{ij} = \frac{a_{ij} + \sum_{u \neq i,j} a_{iu} a_{uj}}{\min(k_i, k_j) + 1 - a_{ij}} \quad (9)$$

$$k_i = \sum_u a_{iu}, \; a_{ij} = \begin{cases} |corr_{ij}|^{\beta} \text{ if unsigned} \\ \left|\frac{1 + corr_{ij}}{2}\right|^{\beta} \text{ if signed} \end{cases}$$

The intuition behind TOM is that if the two genes i and j are connected to a common set of genes, then the similarity between the two genes, $S_{ij}^{(gm)}$, should increase (i.e., the greater the number and strength of the connections that are shared by genes i and j, the larger the TOM value will be for those two genes). To calculate TOM, one must first calculate the correlation matrix, then convert to an adjacency matrix (a_{ij}), and then calculate the TOM matrix. We used the WGCNA R package *adjacency* function with type = "signed", power = 1, and the *TOMsimilarity* function with TOMType = "signed." We used "signed" versions since we want to be able to detect correlations that change from positive to negative between groups, when calculating $\left(V_k^{(1m)} - V_k^{(2m)}\right)$ in Equation (2) (e.g., for unsigned versions, a correlation that changes from 0.5 to −0.5 would result in a $V_k^{(1m)} - V_k^{(2m)} = 0$ which is undesirable when trying to measure differential co-expression). When constructing the test statistics, our motivation for comparing correlation versus TOM, is to assess whether there is any benefit to averaging over the connections with other genes "u," as TOM does through the numerator term $\sum_{u \neq i,j} a_{iu} a_{uj}$. Thus, we keep the exponent $\beta = 1$ for both the correlation and TOM approaches. If we were to set $\beta \neq 1$, than it would be unclear whether any differences in results for correlation versus TOM were due to the exponent, or the neighborhood averaging, and we are interested in the latter. In summary, the motivation for comparing correlation vs. TOM for constructing test statistics is to determine whether TOM is more sensitive to detecting DCMs when the number (and strength) of the connections that are *shared* between genes changes between conditions (e.g., cases vs. controls).

Simulations

Simulations were used to compare the false positive rate (FPR) and true positive rate (TPR) between all of the test statistics under consideration, under several simplified correlation structures. If methods do not perform well under these simple scenarios, then they may not perform well under more complex network structures. The *rmvnorm* function within the mvtnorm R package (Genz et al., 2020) was used to simulate modules. Specifically, when simulating a given module m, an $N^* |P_m|$ gene expression matrix $X^{(gm)}$ (N subjects, $|P_m|$ genes) is simulated for the gth group from a multivariate normal distribution with a zero mean vector, and a $|P_m|^* |P_m|$ correlation matrix $\sum^{(gm)}$ (the variance of each gene equals 1).

Null Simulations

To assess the FPR, a variety of "null" simulations were conducted where modules are simulated such that the correlation matrix is identical between the two groups (i.e., $\sum^{(1m)} = \sum^{(2m)}$). In Equation (10), two different correlation structures are considered for the null simulations: compound symmetric ("CS," i.e., constant pairwise correlation "ρ" between genes), and an "AR1" correlation structure where the correlation between genes "ρ" decays exponentially as genes get further apart.

$$\text{CS correlation}: \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix},$$

$$\text{AR1 correlation}: \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{P-1} \\ \rho & 1 & \rho & \dots & \rho^{P-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{P-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{P-1} & \rho^{P-2} & \rho^{P-3} & \dots & 1 \end{pmatrix} \quad (10)$$

For the CS and AR1 null scenarios, the following parameter values are considered: $\rho = 0.3$ or 0.7, $N = 25$ samples per group, and $P = 10$, 50, or 100 genes within each module. For each setting, 1,000 modules are simulated, and 3,000 permutations are used to calculate p-values.

A final scenario is considered with a "hub gene" network structure, since hub genes are common in many biological applications (Zhang and Horvath, 2005). Here there is a single hub gene, where all other genes have a correlation of $\rho = 0.7$ with the hub gene. To allow for smaller transitive correlation, the correlation between non hub genes is 0.4 in all simulations. A larger sample size (50 or 100 per group) was used for the hub gene simulations, since a larger sample size was needed in the DCM simulations of section "CS Where Correlations Change Direction," in order to have higher TPR to compare between methods. The goal for all of these null simulations is to determine whether each method can control the FPR at level 0.05.

DCM Simulations

The DCM simulation framework is similar to section "Null Simulations," except now the correlation within a module differs between the two groups ($\sum^{(1m)} \neq \sum^{(2m)}$). Specifically, $\sum^{(1m)}$ is fixed as one of the correlation structures from section "Null Simulations" (CS, AR1, or hub), while $\sum^{(2m)}$ changes a randomly selected proportion, γ, of the lower triangle of $\sum^{(1m)}$ (the same changes are then made in the upper triangle to ensure the correlation matrix remains symmetric). We consider $\gamma = 0.1$, 0.4, or 0.7 to represent small, medium and large effects. Five scenarios are considered 1) CS correlation ($\rho = 0.3$ or 0.7) with a proportion of correlations, γ, dropped to zero; 2) AR1 ($\rho = 0.7$) correlation with a proportion of correlations dropped to zero; 3) CS correlation with a proportion of the correlations changed, γ, such that half of the changed correlations increase 50% while the other half decrease by 50%; 4) CS correlation ($\rho = 0.5$) with a proportion of correlations, γ, changed to -0.5; and 5) hub gene structure with a proportion, γ, of the hub gene correlations dropped to zero. For scenario 3, we set $\rho = 0.5$, thus for the subset of changed correlations, half of the correlations increase to 0.75, while half decrease to 0.25. The motivation for scenarios 1–2 is to compare performance between methods given a module with homogeneous (CS) vs. heterogeneous (AR1) correlations that drop to zero. The motivation for scenario 3 and 4 is to compare performance when the changed correlations increase/decrease, or when the correlations change sign. The motivation for scenario 5 is to simulate a module with very sparse changes between groups: i.e., only one row in the correlation matrix changes between groups (the hub gene correlations). Lastly, when changing the population correlation matrices between the two groups, the *make.positive.definite* function from the lqmm R package (Geraci, 2014) is used to ensure that the changed correlation matrix is positive definite, which is necessary in order to simulate the modules from a multivariate normal distribution.

Case Study: Leukemia Microarray Data

All test statistics were compared using data from the leukemia microarray study of Golub et al. (1999). The dataset was downloaded from the multtest R package (Pollard et al., 2005), and contains tumor gene expression measured on 3051 genes from 27 subjects with acute lymphoblastic (ALL) and 11 subjects with acute myeloid leukemia (AML). The data was preprocessed according to Dudoit et al. (2002).

The mclust R package (Scrucca et al., 2016) was used to derive data driven modules within the ALL group, then the corresponding modules were obtained from the AML group and tested for differential co-expression. Then the process was repeated the other way: mclust was used to derive modules in the AML group, then the corresponding modules were obtained from the ALL group and tested for differential co-expression. This approach to module derivation is similar to that taken by CoXpress (Watson, 2006), however, CoXpress uses hierarchical clustering where the researcher must choose the height at which to cut the dendrogram, which determines the number of modules. In contrast, mclust uses the Bayesian Information Criterion (BIC) to determine the number of modules. Specifically, BIC was used to choose between two different diagonal cluster covariance structures (VII or EII), and to estimate the number of modules. The VII and EII covariance structures were chosen since they are the most parsimonious covariance structures included in mclust, which assume a diagonal covariance structure similar to k-means, but with the benefit of being able to use BIC to choose the number of modules. In addition, the assumption of a diagonal covariance structure has been shown to work well in other high dimensional supervised classification settings (Tibshirani et al., 2003; Bickel and Levina, 2004). After deriving the modules, 10,000 permutations were used to calculate p-values and false discovery rate (FDR) adjusted p-values were used to account

for multiple testing (Benjamini and Hochberg, 1995). Example network graphics were generated using Cytoscape version 3.8.2 (Shannon et al., 2003).

RESULTS

A summary of all simulations settings is given in **Supplementary Table 1**.

Null Simulations

Tables 1–3 present the false positive rate (FPR) of each method for the compound symmetric, AR1, and hub gene null simulations, respectively. For each simulation scenario, a one sample proportion test is used to assess whether the FPR of a given test statistic differs from the nominal rate of 0.05. For the one sample proportion tests, a *p*-value cutoff of 0.01 was used due to the large number of statistical tests. Overall, only the HD, CLX, and Schott methods were unable to control the FPR in some scenarios, thus these methods were removed from the DCM simulations of section "DCM Simulations," in order to present a fair comparison of the true positive rates (TPR). For example, the maximum FPR observed was 16.9, 10.7, and 8.7% for the HD, CLX, and Schott tests, respectively. All other tests were able to control the FPR across all scenarios, and fluctuations from the nominal rate of 0.05 across P or ρ values are likely due to random variation.

DCM Simulations

For each simulation setting, a line graph was used to compare the TPR between all methods for small, medium and large correlation effects (corresponding tables are found in **Supplementary Material**). The QAP and GCOR methods were removed from the line graphs to save space, since they consistently have the lowest TPR across all simulations.

CS With Correlations Dropped to Zero

Figure 1 and **Supplementary Table 2** present TPR for the compound symmetric DCM simulations where a proportion, γ, of the correlations are randomly changed to zero between the two groups. PND4 had a TPR within the top three highest TPR 13 times, followed by PND6 and DI (11), PND8 and MAD (9), with all other methods appearing in the top three at most 6 times. The QAP, GCOR and GSNCA methods consistently had the lowest TPRs, while PND20, MAD, pairedT, wilcoxSRT, and GHD methods were often more in the middle, with the GHD

TABLE 1 | Compound symmetric null simulation false positive rates.

ρ	P	PND4	PND6	PND8	PND20	DI	MAD	pairedT	wilcoxSRT	GSNCA	GHD	QAP	GCOR	HD	CLX	Schott
0.3	10	0.045	0.051	0.047	0.048	0.051	0.049	0.058	0.053	0.049	0.050	0.055	0.059	0.077*	0.059	0.062
0.3	50	0.063	0.059	0.048	0.042	0.062	0.063	0.059	0.061	0.043	0.038	0.063	0.043	0.120*	0.073*	0.076*
0.3	100	0.060	0.063	0.062	0.047	0.060	0.058	0.057	0.055	0.049	0.054	0.052	0.050	0.169*	0.099*	0.087*
0.7	10	0.055	0.054	0.048	0.044	0.055	0.057	0.056	0.055	0.039	0.048	0.058	0.050	0.079*	0.026*	0.084*
0.7	50	0.038	0.037	0.040	0.046	0.037	0.039	0.040	0.043	0.052	0.045	0.051	0.054	0.077*	0.013*	0.070*
0.7	100	0.050	0.047	0.048	0.040	0.053	0.051	0.049	0.050	0.052	0.052	0.038	0.054	0.111*	0.019*	0.067

*All settings use N = 25 subjects per group. ρ, compound symmetric correlation parameter; P, number of genes in each module. *False positive rate significantly differs from the nominal rate of 0.05 (one-sample proportion test p-value < 0.01).*

TABLE 2 | AR1 null simulation false positive rates.

ρ	P	PND4	PND6	PND8	PND20	DI	MAD	PairedT	wilcoxSRT	GSNCA	GHD	QAP	GCOR	HD	CLX	Schott
0.3	10	0.048	0.050	0.051	0.051	0.048	0.045	0.049	0.043	0.052	0.046	0.059	0.056	0.085*	0.069*	0.053
0.3	50	0.044	0.048	0.054	0.056	0.048	0.052	0.057	0.052	0.045	0.049	0.054	0.054	0.122*	0.094*	0.051
0.3	100	0.049	0.048	0.050	0.053	0.046	0.045	0.050	0.048	0.053	0.046	0.050	0.050	0.150*	0.107*	0.047
0.7	10	0.046	0.045	0.043	0.042	0.046	0.049	0.049	0.054	0.050	0.053	0.051	0.053	0.072*	0.045	0.063*
0.7	50	0.045	0.051	0.050	0.059	0.044	0.044	0.048	0.052	0.054	0.044	0.054	0.053	0.122*	0.079*	0.057
0.7	100	0.049	0.052	0.053	0.050	0.046	0.045	0.044	0.047	0.047	0.043	0.054	0.051	0.164*	0.105*	0.060

*All settings use N = 25 subjects per group. ρ, AR1 correlation parameter; P, number of genes in each module. *False positive rate significantly differs from the nominal rate of 0.05 (one-sample proportion test p-value < 0.01).*

TABLE 3 | Hub gene null simulation false positive rates.

N	ρ	P	PND4	PND6	PND8	PND20	DI	MAD	pairedT	wilcoxSRT	GSNCA	GHD	QAP	GCOR	HD	CLX	Schott
50	0.7	10	0.038	0.040	0.044	0.043	0.040	0.040	0.049	0.055	0.049	0.051	0.053	0.048	0.054	0.035	0.082*
100	0.7	50	0.064	0.058	0.058	0.050	0.060	0.059	0.057	0.053	0.051	0.039	0.063	0.054	0.056	0.031*	0.084*

*N, number of subjects per group. ρ, correlation between the non-hub genes with the hub gene; P, number of genes in the module. *False positive rate significantly differs from the nominal rate of 0.05 (one-sample proportion test p-value < 0.01).*

test having near zero TPR in **Figure 1F**. For most settings, there was little difference in TPR between PND4-8 and DI (note DI is equivalent to the PND with exponent 2, i.e., "PND2"). One exception was the tenth row of **Supplementary Table 2** ($\rho = 0.7$, $P = 10$, $\gamma = 0.1$), where PND4 had 20% higher TPR than DI, while PND6-20 had $\geq$ 30% higher TPR than DI.

AR1 With Correlations Dropped to Zero

Figure 2 and **Supplementary Table 3** present TPRs for the AR1 DCM simulations where a proportion, γ, of the correlations are randomly changed to zero between the two groups. In contrast to the previous CS simulations, the AR1 simulations consider a more heterogeneous set of population correlation values. In addition, the AR1 simulations have more separation in the TPR when comparing the PND tests with the DI and MAD tests. PND4-20 were in the top 3 highest TPRs 6, 9, 8, and 3 times, respectively, followed by DI which was in the top 3 one time. No other methods had TPR in the top three for any scenarios. PND4-8 were consistently near the top TPR, while PND20, DI, MAD, and GHD had TPR near the middle, with pairedT, wilcoxSRT, GSNCA, QAP, and GCOR consistently having the lowest TPR.

CS Where Half of the Changed Correlations Increase 50%, Half Decrease 50%

Figure 3 and **Supplementary Table 4** present TPRs for compound symmetric simulations where a proportion, γ, of the correlations are randomly changed such that half of the changed correlations increase by 50%, while the other half decrease by 50%. In contrast to previous sections, power was lower for most methods, with GHD as the most powerful test in **Figures 3A,B** and wilcoxSRT was the most powerful in **Figure 3C**. However, GHD became less powerful as the number of genes increased, with most other methods having TPR higher than GHD in **Figure 3C**. The GHD had TPR in the top 3 for 7 settings, followed by PND6 and PND8 (6), and three times for PND4, PND20, and wilcoxSRT.

CS Where Correlations Change Direction

Supplementary Figure 1 and **Supplementary Table 5** present TPRs for compound symmetric ($\rho = 0.5$) simulations where a proportion of correlations, γ, of the correlations are randomly changed to –0.5. PND6 had TPR in the top 3 for all 9 settings, followed by PND8 (8), DI, MAD, wilcoxSRT (6). Similar to section "CS Where Half of the Changed Correlations Increase 50%, Half Decrease 50%," the TPR for the GHD test substantially decreased as the number of genes increased. The PND4-8 tests, DI and MAD usually had the highest TPRs, with the PND tests having higher TPR when only 10% of the correlations were changed between groups.

Hub Gene Setting Where a Proportion of the Hub Gene Correlations Are Dropped to 0

Figure 4 and **Supplementary Table 6** present TPRs for the hub gene correlation structure where a proportion, γ, of the hub gene correlations are dropped to zero. PND6 was in the top three highest TPRs 6 times, followed by PND8 (5), PND20 (4), GHD (3), and PND4 (1). No other methods had TPR in the top three for any scenarios. For most scenarios the PND and GHD tests have substantially higher TPR compared with DI, MAD, and the other remaining tests. Having a higher exponent in the PND tests (6 or higher) resulted in higher TPR compared to PMD4 when only 10% the hub gene correlations changed between the two groups.

Comparing Correlation Versus TOM Similarity Measures

As explained in section "Similarity Measures for Constructing Test Statistics," we were interested in comparing two different similarity measures for constructing the test statistics: correlation versus the TOM. Given that TOM has a higher computational cost compared to correlation, results comparing with TOM are only shown for a subset of tests (PND6, DI, MAD, GHD) and simulation settings, using 100 simulation replicates and 2,000 permutations. **Supplementary Figure 2** displays a line graph comparing the TPR in correlation-based methods (solid lines) to their TOM counterparts (dashed lines). In nearly all simulation settings, the TOM methods had lower TPR than their correlation counterparts. Two exceptions were: **Supplementary Figure 2A** when $\gamma = 0.7$, PND6 had slightly higher TPR when using TOM compared to correlation; and **Supplementary Figure 2E** where the TPR of GHD was higher when using TOM, particularly when $\gamma = 0.1$. Overall, given the increased computational cost of TOM, and the fact that TOM had lower TPR in nearly all simulation settings, the TOM based methods are omitted from the remainder of the paper.

Overall Comparison of Tests Across All Simulation Studies

In summary, we evaluated 51 different simulation scenarios and the median TPR across these scenarios was greater than 0.70 for all PND methods (**Supplementary Table 7**). The DI and MAD methods followed with median TPR of 0.63 and 0.54, respectively. As alternative summaries, we also examined which methods ranked in the top three of all methods based on highest TPR or whether their TPR was within 5% of the highest TPR value (**Supplementary Table 7**). Based on these metrics PND4, PND6 and PND8 were in these top lists 58–80% of the times, followed by PND20, DI, MAD, WilcoxSRT and GHD, which were in these top lists 25–51% of the time.

Case Study: Leukemia Microarray Data

We used the Golub leukemia data set to illustrate the application of DiNA, in addition to the visualization of module results. **Supplementary Table 8** reports the following information for each of 86 derived modules: number of genes, *p*-values and FDR adjusted *p*-values for a subset of the top performing tests from our simulations (PND6, DI, MAD, GHD). Note when deriving the modules in the ALL group, BIC selected 49 modules with the VII covariance structure. The median module size had 64 genes (25th and 75th quantiles: 35 and 79 genes). When deriving the modules in the AML group, BIC selected 37 modules with the EII covariance structure. The median module size had 34 genes (25th and 75th quantiles: 20 and 133 genes).

Figures 5A,B compares the –log10 *p*-values among the PND6, DI, MAD, and GHD tests. The PND6, DI, and MAD

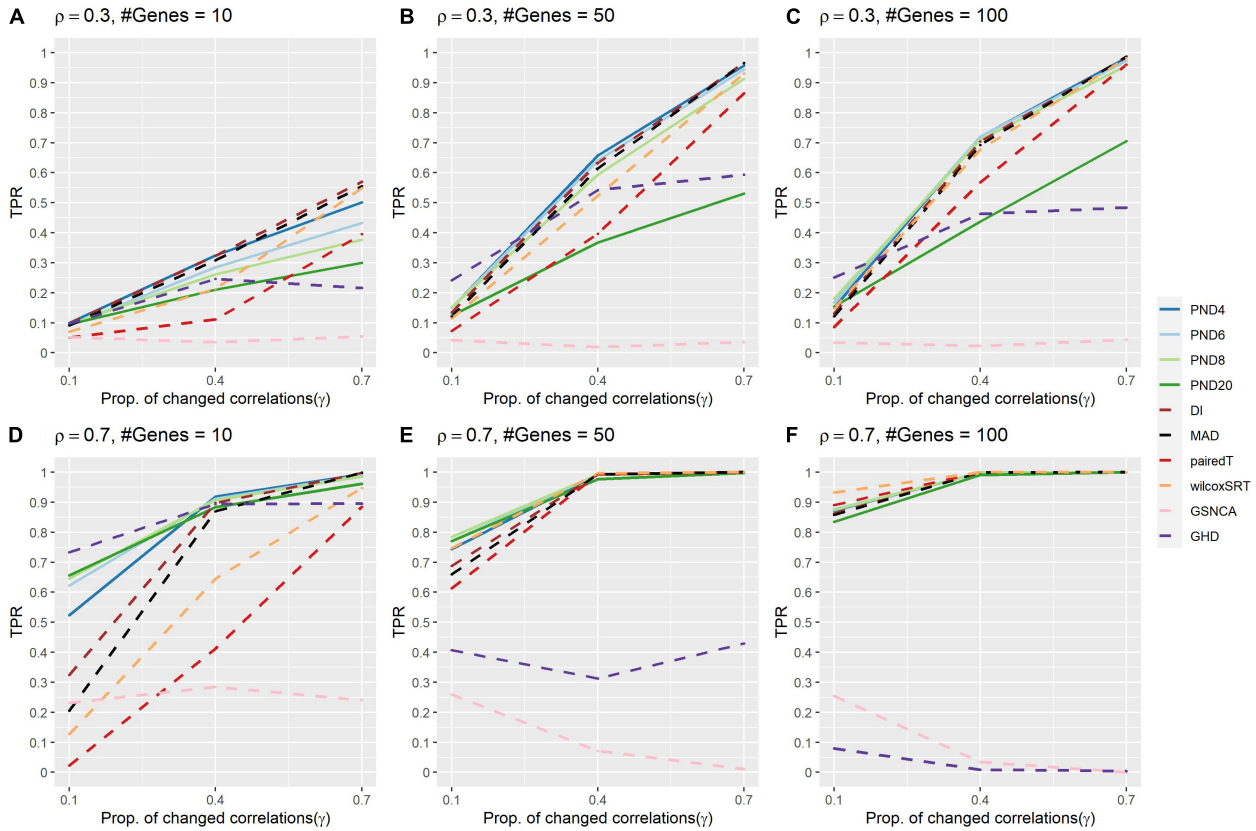

FIGURE 1 | True positive rates for the compound symmetric DCM simulations with a proportion of correlations dropped to zero between the two groups. Solid lines refer to the proposed PND tests, while dashed lines refer to pre-existing methods. **(A–C)** Compound symmetric correlation parameter, ρ= 0.3, and the number of genes in a module ("#Genes") is 10, 50, or 100. **(D–F)** Compound symmetric correlation parameter, ρ= 0.7, and the number of genes in a module ("#Genes") is 10, 50, or 100. N=25 samples per group.

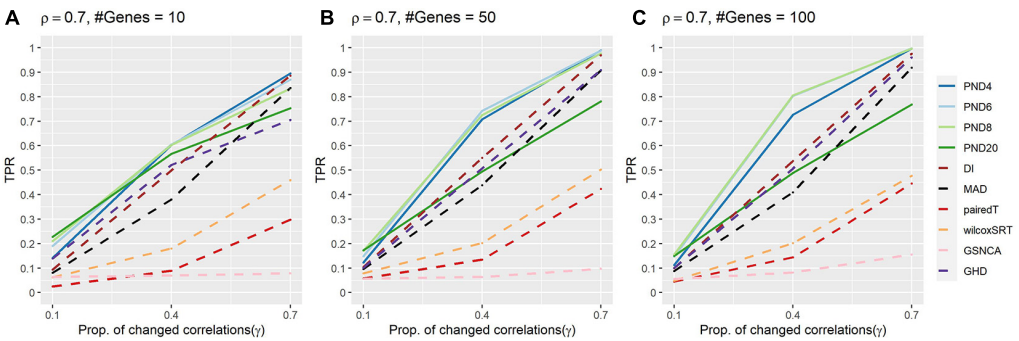

FIGURE 2 | True positive rates for the AR1 DCM simulations with a proportion of correlations dropped to zero between the two groups. Solid lines refer to the proposed PND tests, while dashed lines refer to pre-existing methods. **(A–C)** AR1 correlation parameter, ρ = 0.7, and the number of genes in a module ("#Genes") is 10, 50, or 100. N = 25 samples per group.

tended to produce similar *p*-values, with GHD generally having larger *p*-values, especially for the AML derived modules. **Supplementary Figure 3** presents a Venn diagram for the total number of modules with FDR adjusted *p*-values < 0.01 for each method. The DI and MAD methods had the most overlap with 9 modules that were only identified using these two methods. Of interest is that 2 modules were identified using the PND6 method only and 1 module that was only identified using the MAD method. For the nine modules only identified using the DI and MAD methods, the PND6 FDR ranged from 0.01 to 0.03 with unadjusted *p*-values well within the range of the unadjusted *p*-values for DI and MAD. Likewise, for the MAD only and the

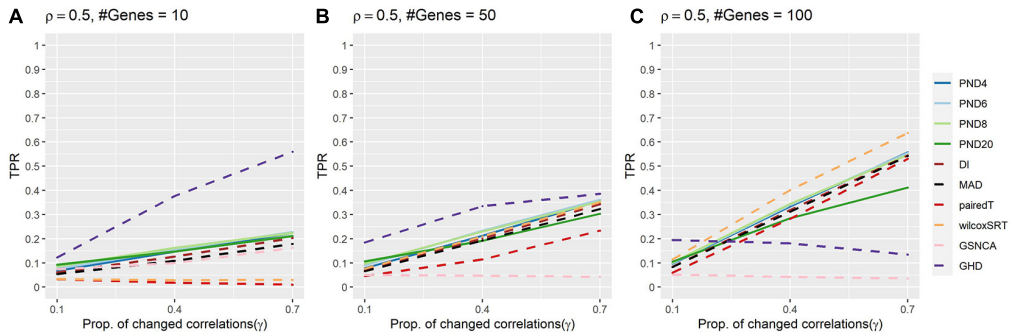

FIGURE 3 | True positive rates for the compound symmetric DCM simulations with a proportion of correlations changed between groups such that half of the changed correlations increased by 50%, while the other half decreased by 50%. Solid lines refer to the proposed PND tests, while dashed lines refer to pre-existing methods. **(A–C)** Compound symmetric correlation parameter, ρ = 0.5, and the number of genes in a module ("#Genes") is 10, 50, or 100. N = 25 samples per group.

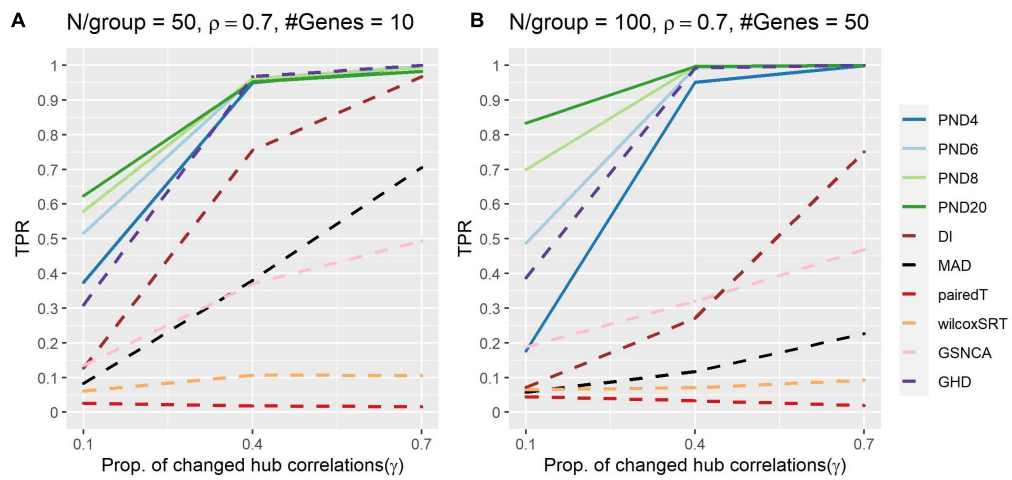

FIGURE 4 | True positive rates shown for the hub gene DCM simulations with a proportion of hub-correlations dropped to zero; true positive rates shown for all methods. Solid lines refer to the proposed PND tests, while dashed lines refer to pre-existing methods. **(A–B)** N=50 or 100 samples per group, correlation between non hub-genes with the hub gene is ρ = 0.7, and the number of genes ("#Genes") is 10 or 50.

PND6 only modules, the FDR ranged from 0.01 to 0.04 for the other methods (not including GHD).

Figure 6 contains two examples of differential co-expression in this data. For ease of visualization, we focused on modules with less than 50 genes that were differentially expressed (FDR < 0.01) in at least three of the four methods explored. This resulted in 8 modules where all modules were differentially co-expressed using PND6, DI, and MAD and none were differentially co-expressed using the GHD method. The module among the 8 with the smallest p-value using the GHD method (ALL_19) and the module with the largest p-value using the GHD method (AML_29) were chosen for visualization. **Figures 6A,B** is a module that was identified in the ALL subjects that was differentially co-expressed in the AML subjects. In the original module (**Figure 6A**) all of the probe sets are positively correlated. Within the AML subjects, many of the correlations increased in intensity (light red to bright red), some correlations were dropped to approximately zero, and a few went from a positive association (red line) to a negative association (blue line). **Figures 6C,D** is a module that was originally identified in the AML subjects and was differentially co-expressed in the ALL subjects. For this module, most of the correlations among the probe sets dropped to values close to zero indicating a co-expression network that was only active in the AML group and not in the ALL group. See **Supplementary Figures 4, 5** for correlation heatmaps as additional visualizations of differential co-expression in ALL_19, AML_29, and several other example modules with FDR adjusted p-values < 0.01.

Because PND6 did marginally better in the simulation studies, we further explored the module identified (FDR < 0.01) only when using the PND6 method whose unadjusted p-values in all other methods were greater than or equal to 0.01, ALL_24. Unlike in the modules depicted in **Figure 6**, the co-expression patterns of only a few genes in ALL_24 changed dramatically rather than all relationship, i.e., edges, changing in a coordinated way (**Supplementary Figures 6A–C**). To quantify this observation, we calculated the median difference in correlations for each gene. A large spread of the median difference between genes within

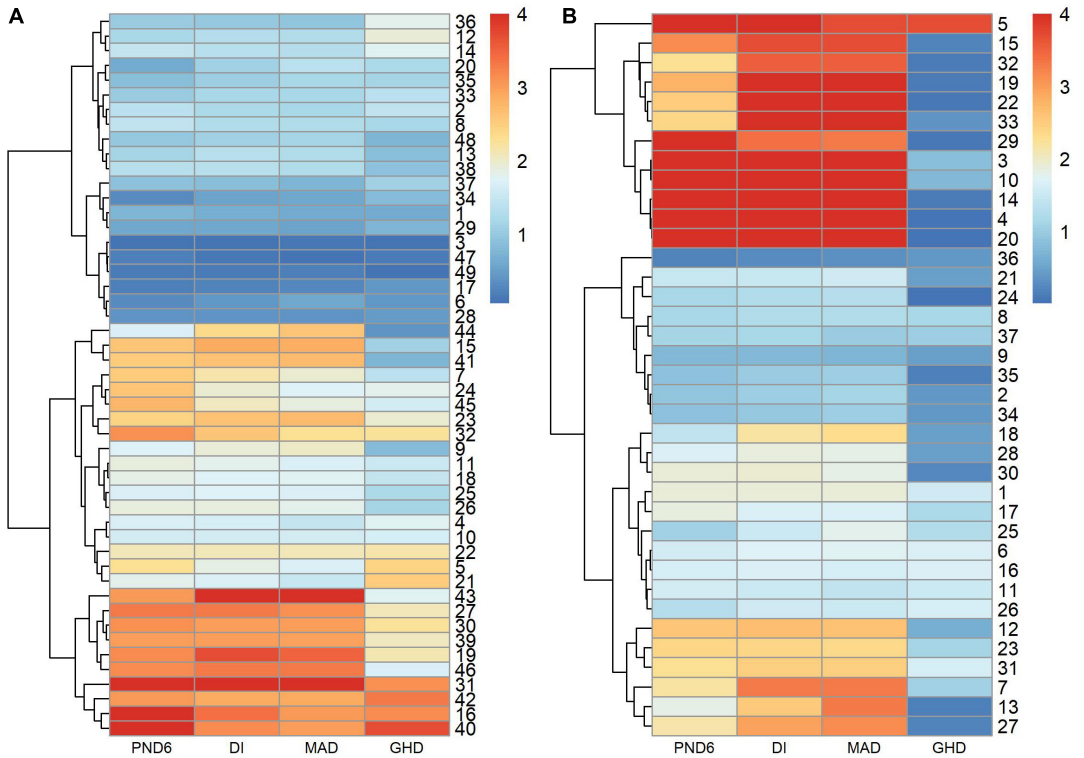

FIGURE 5 | –log10 *p*-values of modules derived in leukemia microarray dataset. Hierarchical clustering was used to sort the modules and tests. **(A)** 49 modules derived in ALL group. **(B)** 37 modules derived in AML group.

a module would indicate connections for only few genes are changing dramatically, but most genes maintain their original connections (similar to simulation 5). When compared to a module that is not differentially co-expressed (ALL_3) and the two differentially co-expressed modules from **Figure 6**, the PND6 exclusive module, ALL_24, has a highly skewed distribution of median correlation differences, i.e., only associations with a few genes are dramatically altered (**Supplementary Figure 6D**). This trend held true among all modules as the ALL_24 had the largest estimated skewness among all modules (skewness = 2.87).

Within ALL_24, *cathepsin G* (*CTSG*) had the largest median difference (median difference = 0.90). Many of its edges changed from strong positive correlations to strong negative correlations among genes. CTSG is a well-established therapeutic target for both AML and ALL cancers (e.g., Jin et al., 2013; Khan et al., 2017). In a functional enrichment through EnrichR (Kuleshov et al., 2016), cellular response to cytokine stimulus (GO:0071345) was significantly enriched (adjusted p-value < 0.01) among the genes within ALL_24. Nine of the 59 genes within the module were associated with this GO term. Although CTSG was not associated directly with this GO term, its role in inflammation can easily be connected to the other genes (e.g., Gao et al., 2018). These results suggest that the role of CTSG in the inflammatory response to leukemia may differ between AML and ALL. Not only does this differentially co-expressed module indicate that this pattern of differential co-expression in present in "real" data, but it also indicates that this pattern can be biologically relevant.

DISCUSSION

Statistical networks provide a convenient framework for representing the interactions between multiple genes (or other molecular features). Differential network analysis (DiNA) quantifies how this network structure differs between two or more groups/phenotypes (e.g., disease subjects and healthy controls), and is a growing field of research (de la Fuente, 2010; Kayano et al., 2014; Singh et al., 2018; Shojaie, 2020). One major application of DiNA is to identify "modules" (subsets of 3 or more genes), where the network connections within a module differ between phenotype groups, known as differentially co-expressed modules (DCMs). Although several statistical tests have been proposed for identifying DCMs (Watson, 2006; Choi and Kendziorski, 2009; Gill et al., 2010; Tesson et al., 2010; Rahmatallah et al., 2014), there is a lack of simulation studies comparing such methods. Thus, the primary motivation of this study was to compare existing methods via simulations, as well as the proposed framework of the p-norm difference test (PND) which encompasses existing methods such as DI and MAD.

In the "Null Simulations" section (where the network structure within the module was identical between groups), all of the permutations based test statistics were able to control the FPR (PND4-20, DI, MAD, pairedT, wilcoxSRT, GSNCA, GHD, QAP, and GCOR). However, the three tests from the HDtest R package (CLX and Schott use asymptotic approximations to calculate

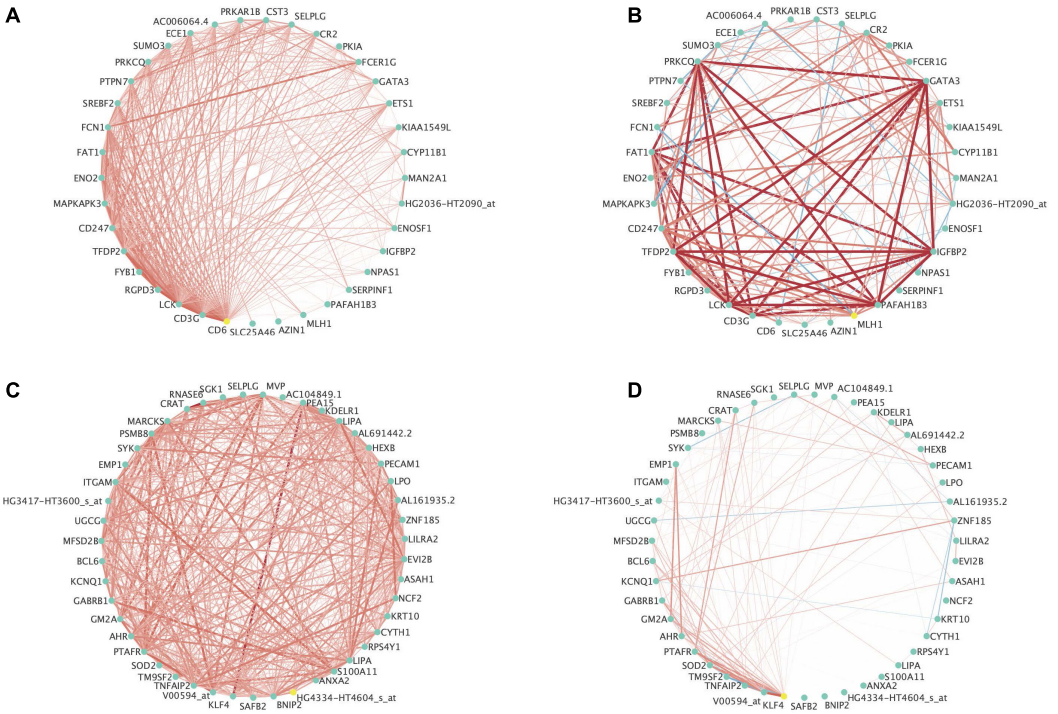

FIGURE 6 | Example differentially expressed networks from the leukemia microarray data. For all networks, circles present individual probe sets. Labels are gene symbols for probe sets with annotation information in the Ensembl database. Otherwise, the original probe set identifier from Affymetrix was used. Red lines connecting circles indicate a positive correlation (correlation coefficient > 0.3) between the two probe sets. Blue lines connecting circles indicates a negative correlation between the two probe sets (correlation coefficient < -0.30). The intensity of the color and thickness of the lines are associated with the magnitude of the correlation between the two probe sets. **(A)** Associations between probe sets among the ALL subjects for Module 19 originally identified among the ALL subjects and **(B)** Associations between probe sets among the AML subjects for Module 19 originally identified among the ALL subjects. This module was significantly differentially co-expressed using the PND6 method (FDR < 0.01), the DI method (FDR < 0.01), and the MAD method (FDR < 0.01). It was borderline significant using the GHD method (FDR $= 0.06$). **(C)** Associations between probe sets within the AML subjects for Module 29 originally identified among the AML subjects and **(D)** Associations between probe sets within the ALL subjects for Module 29 originally identified among the AML subjects. This module was significantly differentially co-expressed using the PND6 method (FDR < 0.01), the DI method (FDR < 0.01), and the MAD method (FDR < 0.01). It was not significant with the GHD method (unadjusted p-value $= 0.84$; FDR $= 0.90$).

p-values, while HD uses a "multiplier bootstrap" method) were often unable to control the FPR at level 0.05, and thus were omitted from the remainder of the manuscript. It is possible these methods may control the FPR given larger sample sizes, however, even with 50 or 100 samples per group (**Table 3**), the CLX and Schott tests did not control the FPR, although the HD test did control the FPR in these settings.

In the DCM simulations, it is worth noting that the TPRs of methods depend on the network structure. In the homogenous correlation structure of section "CS With Correlations Dropped to Zero", the PND4-8, DI and MAD tests had the highest TPRs. In the more heterogeneous correlation structure of section "AR1 With Correlations Dropped to Zero", there was greater separation in TPR when comparing PND4-20 with DI and MAD, with PND4-20 having the highest TPRs. In section "CS Where Half of the Changed Correlations Increase 50%, Half Decrease 50%", the GHD test had the highest TPR in most settings, with the wilcoxSRT and PND tests surpassing the GHD as the number of genes increased. In the hub gene simulations of section "CS Where Correlations Change Direction," the PND4-20 and GHD tests had the highest TPR.

Despite differences based on network structure, on average, test statistics in the PND framework were consistently the best performing (PND4-20, DI, MAD). In many of the scenarios, we found advantages of intermediate values for the power (e.g., PND 6 and 8) Thus for the question of what the exponent value should be for PND, we recommend PND6 as a default choice but recommend users to explore other power values based on their particular data sets.

One of the difficulties of evaluating differential co-expression techniques is to determine if the simulated scenarios are biologically relevant in any or all experimental designs. We have shown the existence of the hub gene framework in the AML/ALL case study. However, we did not observe patterns of differential co-expression in the AML/ALL dataset similar to all simulation scenarios. This observation does not indicate these simulation scenarios are not biologically relevant, they were just not observed under these experimental conditions.

This study is not without limitations, thus we identify five areas for future research. (1) Given that the TPRs of methods depends on the true network structure, it would be interesting to consider methods that combine multiple test statistics, in order to

increase sensitivity across a greater variety of network structures. (2) Further research is needed for comparing other types of similarity measures for constructing the test statistics, such as various types of "conditional" partial correlation measures (Shojaie, 2020), or settings where using the TOM may improve power compared to correlation (3). Although one may use predefined modules from an existing database (e.g., KEGG, Kanehisa and Goto, 2000; GO; Ashburner et al., 2000), further research is needed to compare clustering methods for deriving data dependent modules, and determining the optimal number of modules. In section "Case Study: Leukemia Microarray Data," we used a similar approach to CoXpress (Watson, 2006) but with model based clustering and BIC to select the number of modules. Instead of performing the clustering twice, as in CoXpress, DiffCoEx (Tesson et al., 2010) uses hierarchical clustering only once where the distance matrix is the TOM of the difference between two correlation matrices. WGCNA is another approach for deriving network modules using hierarchical clustering with a dynamic tree-cutting algorithm for choosing the number of modules (also used by DiffCoEx). However, the authors admit that it remains an open research question for how to optimize the tree-cutting parameters to determine the number of modules (Langfelder and Horvath, 2008; Langfelder et al., 2008). (4) This manuscript focused on comparing methods for identifying DCMs between two phenotype groups. Further research is needed for developing methods to identify DCMs for quantitative outcomes, or for categorical outcomes with more than two groups. (5) Lastly, more research is needed for differential network analysis when integrating multiple different types of molecular features (e.g., transcriptome, metabolome, microbiome, proteome). Some existing methods include: (Class et al., 2018; Erola et al., 2019; Shi et al., 2019).

In summary, several test statistics for identifying differentially co-expressed modules (DCMs) were compared via simulations and a leukemia microarray study (Golub et al., 1999). Through extensive simulations, tests in the PND framework had TPR that was competitive with and often higher than the other methods, while controlling the FPR. When comparing two different similarity measures for constructing the test statistics, correlation versus TOM, we found little benefit of using the more computationally expensive TOM. An approach to deriving data dependent modules was demonstrated using the dataset of (Golub et al., 1999), by using Gaussian mixture models with BIC to select the number of modules. However, further research is needed to compare clustering methods for deriving data dependent modules. Nevertheless, after obtaining a list of modules (predefined or data driven), we recommend the user take an intermediate power in the PND framework, such as PND6, for identifying DCMs. All methods considered are implemented in the discoMod R package, available at https://github.com/arbet003/discoMod.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.bioconductor.org/packages/release/bioc/html/multtest.html.

AUTHOR CONTRIBUTIONS

JA and YZ further developed the project direction, wrote R code for simulations, analyzed the leukemia microarray dataset, and developed the discoMod R package. JA wrote the manuscript. KK and LS co-supervised this project. The thesis work of EL influenced the research questions and methods compared. All authors read and approved the final version of the manuscript, made substantial contributions to the conception, design, drafting, and revisions of this project.

FUNDING

This work was supported by the NIH/NCATS Colorado CTSA Grant No. UL1 TR002535, (NIH/NIAAA) NIH R01AA021131, R24 AA013162, and (NIH/NIDA) P30 DA044223 contents are the authors' sole responsibility and do not necessarily represent official NIH views.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.630215/full#supplementary-material

REFERENCES

Andreopoulos, B., An, A., Wang, X., and Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief. Bioinform.* 10, 297–314. doi: 10.1093/bib/bbn058

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

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

Barabasi, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bickel, P. J., and Levina, E. (2004). Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 10, 989–1010.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Am. Stat. Assoc.* 108, 265–277. doi: 10.1080/01621459.2012.758041

Chang, J., Zhou, W., Zhou, W. X., and Wang, L. (2017). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics* 73, 31–41. doi: 10.1111/biom.12552

Choi, Y., and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics* 25, 2780–2786. doi: 10.1093/bioinformatics/btp502

Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* 26, 721–744. doi: 10.1146/annurev-cellbio-100109-104122

Class, C. A., Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2018). iDINGO—integrative differential network analysis in genomics with Shiny application. *Bioinformatics* 34, 1243–1245. doi: 10.1093/bioinformatics/btx750

Dawson, J. A., Ye, S., and Kendziorski, C. (2012). R/EBcoexpress: an empirical Bayesian framework for discovering differential co-expression. *Bioinformatics* 28, 1939–1940. doi: 10.1093/bioinformatics/bts268

de la Fuente, A. (2010). From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. *Trends Genet.* 26, 326–333. doi: 10.1016/j.tig.2010.05.001

De Leeuw, C. A., Neale, B. M., Heskes, T., and Posthuma, D. (2016). The statistical properties of gene-set analysis. *Nat. Rev. Genet.* 17:353. doi: 10.1038/nrg.2016.29

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97, 77–87. doi: 10.1198/016214502753479248

Emmert-Streib, F., and Glazko, G. V. (2011). Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput. Biol.* 7:e1002053. doi: 10.1371/journal.pcbi.1002053

Erola, P., Bonnet, E., and Michoel, T. (2019). Learning differential module networks across multiple experimental conditions. *Methods Mol. Biol.* 1883, 303–321. doi: 10.1007/978-1-4939-8882-2_13

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi: 10.1093/biostatistics/kxm045

Fukushima, A. (2013). DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* 518, 209–214. doi: 10.1016/j.gene.2012.11.028

Gao, S., Zhu, H., Zuo, X., and Luo, H. (2018). Cathepsin G and its role in inflammation and autoimmune diseases. *Arch. Rheumatol.* 33, 498–504. doi: 10.5606/archrheumatol.2018.6595

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., et al. (2020). Package 'mvtnorm'. *J. Comput. Graphic. Stat.* 11, 950–971.

Geraci, M. (2014). Linear quantile mixed models: the lqmm package for Laplace quantile regression. *J. Stat. Softw.* 57, 1–29.

Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinform.* 11:95. doi: 10.1186/1471-2105-11-95

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537. doi: 10.1126/science.286.5439.531

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015). DINGO: differential network analysis in genomics. *Bioinformatics* 31, 3413–3420. doi: 10.1093/bioinformatics/btv406

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, H.-C., Niu, Y., and Qin, L.-X. (2015). Differential expression analysis for RNA-Seq: an overview of statistical methods and computational software: supplementary issue: sequencing platform modeling and analysis. *Cancer Inform.* 14:S21631.

Jardim, V. C., Santos, S. D. S., Fujita, A., and Buckeridge, M. S. (2019). BioNetStat: a tool for biological networks differential analysis. *Front. Genet.* 10:594.

Jin, W., Wu, K., Li, Y. Z., Yang, W. T., Zou, B., Zhang, F., et al. (2013). AML1-ETO targets and suppresses cathepsin G, a serine protease, which is able to degrade AML1-ETO in t(8;21) acute myeloid leukemia. *Oncogene* 32, 1978–1987. doi: 10.1038/onc.2012.204

Kakati, T., Bhattacharyya, D. K., Barah, P., and Kalita, J. K. (2019). Comparison of methods for differential co-expression analysis for disease biomarker prediction. *Comp. Biol. Med.* 113:103380. doi: 10.1016/j.compbiomed.2019.103380

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kayano, M., Shiga, M., and Mamitsuka, H. (2014). Detecting differentially coexpressed genes from labeled expression data: a brief review. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 154–167. doi: 10.1109/tcbb.2013.2297921

Khan, M., Carmona, S., Sukhumalchandra, P., Roszik, J., Philips, A., Perakis, A. A., et al. (2017). Cathepsin G is expressed by acute lymphoblastic leukemia and is a potential immunotherapeutic target. *Front. Immunol.* 8:1975.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97.

Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., et al. (2012). Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One* 7:e50411. doi: 10.1371/journal.pone.0050411

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9:559.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Comput. Biol.* 7:e1001057. doi: 10.1371/journal.pcbi.1001057

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

Lichtblau, Y., Zimmermann, K., Haldemann, B., Lenze, D., Hummel, M., and Leser, U. (2017). Comparative assessment of differential network analysis methods. *Brief. Bioinform.* 18, 837–850.

Liu, B.-H., Yu, H., Tu, K., Li, C., Li, Y.-X., and Li, Y.-Y. (2010). DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data. *Bioinformatics* 26, 2637–2638. doi: 10.1093/bioinformatics/btq471

McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M., and Zhang, B. (2016). DGCA: a comprehensive R package for differential gene correlation analysis. *BMC Syst. Biol.* 10:106.

Petereit, J., Smith, S., Harris, F. C., and Schlauch, K. A. (2016). petal: co-expression network modelling in R. *BMC Syst. Biol.* 10:51.

Pollard, K. S., Dudoit, S., and Van Der Laan, M. J. (2005). "Multiple testing procedures: the multtest package and applications to genomics," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit (New York, NY: Springer), 249–271. doi: 10.1007/0-387-29362-0_15

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics* 30, 360–368. doi: 10.1093/bioinformatics/btt687

Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* 28, 323–332. doi: 10.1016/j.tig.2012.03.004

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

Ruan, D., Young, A., and Montana, G. (2015). Differential analysis of biological networks. *BMC Bioinform.* 16:327.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Stat. Data Anal.* 51, 6535–6542. doi: 10.1016/j.csda.2007.03.004

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8:289. doi: 10.32614/rj-2016-021

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, W. J., Zhuang, Y., Russell, P. H., Hobbs, B. D., Parker, M. M., Castaldi, P. J., et al. (2019). Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics* 35, 4336–4343. doi: 10.1093/bioinformatics/btz226

Shojaie, A. (2020). Differential network analysis: a statistical perspective. *Wiley Interdiscip. Rev. Comput. Stat.* 13:e1508.

Singh, A. J., Ramsey, S. A., Filtz, T. M., and Kioussi, C. (2018). Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.* 75, 1013–1025. doi: 10.1007/s00018-017-2679-6

Siska, C., Bowler, R., and Kechris, K. (2016). The discordant method: a novel approach for differential correlation. *Bioinformatics* 32, 690–696. doi: 10.1093/bioinformatics/btv633

Siska, C., and Kechris, K. (2017). Differential correlation for sequencing data. *BMC Res. Notes* 10:54.

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* 14:91.

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinform.* 11:497. doi: 10.1186/1471-2105-11-497

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* 18, 104–117.

van Dam, S., Vosa, U., Van Der Graaf, A., Franke, L., and De Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* 19, 575–592.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., Macdonald, M. L., et al. (2016). FastGGM: an efficient algorithm for the inference of gaussian graphical model in biological networks. *PLoS Comput. Biol.* 12:e1004755. doi: 10.1371/journal.pcbi.1004755

Watson, M. (2006). CoXpress: differential co-expression in gene expression data. *BMC Bioinform.* 7:509.

Xu, R., and Wunsch, D. C. (2010). Clustering algorithms in biomedical research: a review. *IEEE Rev. Biomed. Eng.* 3, 120–154. doi: 10.1109/rbme.2010.2083647

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17.

Zhang, R., Ren, Z., and Chen, W. (2018). SILGGM: an extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Comput. Biol.* 14:e1006369. doi: 10.1371/journal.pcbi.1006369

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 July 2021
doi: 10.3389/fgene.2021.649764

Filtering of Data-Driven Gene Regulatory Networks Using *Drosophila melanogaster* as a Case Study

Yesid Cuesta-Astroz[1†], *Guilherme Gischkow Rucatti*[2†], *Leandro Murgas*[3,4], *Carol D. SanMartín*[5,6], *Mario Sanhueza*[2,7] *and Alberto J. M. Martin*[3,7*]

[1] Colombian Institute of Tropical Medicine, CES University, Medellin, Colombia, [2] Centro de Biología Integrativa, Facultad de Ciencias, Universidad Mayor, Santiago, Chile, [3] Laboratorio de Biologia de Redes, Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile, [4] Programa de Doctorado en Genómica Integrativa, Vicerrectoría de Investigación, Universidad Mayor, Santiago, Chile, [5] Departamento de Neurología y Neurocirugía, Hospital Clínico Universidad de Chile, Santiago, Chile, [6] Centro de Investigacíon Clínica Avanzada (CICA), Hospital Clínico Universidad de Chile, Santiago, Chile, [7] Escuela de Biotecnología, Facultad de Ciencias, Universidad Mayor, Santiago, Chile

Gene Regulatory Networks (GRNs) allow the study of regulation of gene expression of whole genomes. Among the most relevant advantages of using networks to depict this key process, there is the visual representation of large amounts of information and the application of graph theory to generate new knowledge. Nonetheless, despite the many uses of GRNs, it is still difficult and expensive to assign Transcription Factors (TFs) to the regulation of specific genes. ChIP-Seq allows the determination of TF Binding Sites (TFBSs) over whole genomes, but it is still an expensive technique that can only be applied one TF at a time and requires replicates to reduce its noise. Once TFBSs are determined, the assignment of each TF and its binding sites to the regulation of specific genes is not trivial, and it is often performed by carrying out site-specific experiments that are unfeasible to perform in all possible binding sites. Here, we addressed these relevant issues with a two-step methodology using *Drosophila melanogaster* as a case study. First, our protocol starts by gathering all transcription factor binding sites (TFBSs) determined with ChIP-Seq experiments available at ENCODE and FlyBase. Then each TFBS is used to assign TFs to the regulation of likely target genes based on the TFBS proximity to the transcription start site of all genes. In the final step, to try to select the most likely regulatory TF from those previously assigned to each gene, we employ GENIE3, a random forest-based method, and more than 9,000 RNA-seq experiments from *D. melanogaster*. Following, we employed known TF protein-protein interactions to estimate the feasibility of regulatory events in our filtered networks. Finally, we show how known interactions between co-regulatory TFs of each gene increase after the second step of our approach, and thus, the consistency of the TF-gene assignment. Also, we employed our methodology to create a network centered on the *Drosophila melanogaster* gene *Hr96* to demonstrate the role of this transcription factor on mitochondrial gene regulation.

Keywords: gene regulatory network, transcriptional regulation, transcription factor targets, *Drosophila melanogaster*, *HR96*

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer,
University of Oslo, Norway

Reviewed by:
Tom Michoel,
University of Bergen, Norway
Jaime Castro-Mondragon,
University of Oslo, Norway

***Correspondence:**
Alberto J. M. Martin
alberto.martin@umayor.cl

[†] *These authors share first authorship*

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

Received: *05 January 2021*
Accepted: *30 April 2021*
Published: *28 July 2021*

Citation:
Cuesta-Astroz Y, Gischkow Rucatti G, Murgas L, SanMartín CD, Sanhueza M and Martin AJM (2021) Filtering of Data-Driven Gene Regulatory Networks Using Drosophila melanogaster as a Case Study. Front. Genet. 12:649764. doi: 10.3389/fgene.2021.649764

1. INTRODUCTION

The control of gene expression is one of the key processes that allow living organisms to adapt to their environment. Different regulatory mechanisms determine which gene is expressed and what amount of the product encoded is generated. Among regulatory mechanisms, Transcription Factors (TFs) are deemed to be the most relevant players in the control of transcription, but there are other types of regulation that include ncRNAs and other proteins (Ramírez-Clavijo and Montoya-Ortíz, 2013). TFs bind to specific regions in the DNA to attract or repel RNA polymerase and other components of the transcriptional machinery to modulate the expression of certain genes. When studying the regulation in whole genomes, gene regulation is often represented as a network where nodes represent genes. In this type of network called Gene Regulatory Network (GRN), connections between genes indicate that the product of a gene regulates the expression of another gene, and thus, their direction is important.

Despite the relevance of the processes represented in a GRN, the majority of the different regulators for each gene still remain unknown. For example, in the human GRN there are about 5,400 TF-gene connections of high confidence (Garcia-Alonso et al., 2019), thus, considering there are over 1,600 TFs in this species (Lambert et al., 2018), we still need to verify a large proportion of likely regulators for most of the genes. This lack of knowledge is even worse for other species to a varying degree, including most common model organisms such as *Mus musculus* (Holland et al., 2020), *Caenorhabditis elegans* (Harris et al., 2020), *Drosophila melanogaster* (Thurmond et al., 2019), and even *Escherichia coli* (Santos-Zavaleta et al., 2019). Recent efforts aim to close this gap of knowledge of how genes are regulated. For example, the ENCODE project (Abascal et al., 2020) focuses on the discovery and annotation of cis regulatory elements in human and mouse genomes based on experimental evidence such as TF binding sites. CIS-BP, a database of TF Binding Motifs (TFBMs), employs evolutionary information to infer binding motifs (Weirauch et al., 2014). Another approach to determine TFBMs relies on the detection of motifs from experimentally determined TF Binding Sites (TFBSs) such as those reported by the ENCODE project (Matys et al., 2003; Forrest et al., 2014; Khan et al., 2018; Kulakovskiy et al., 2018). Importantly, even if it is possible to determine where a TFs binds on the DNA by determining occurrences of these motifs (Jayaram et al., 2016), the majority of motifs are not functional (Dror et al., 2015). Even more, the identification of an actual TFBS does not imply knowing which gene or genes are regulated by the binding of the TF to it.

There are several approaches to assign TFs to the regulation of specific genes based on occurrences of TFBMs or experimentally determined TFBMs. Experimental methods to identify TFBSs on DNA are diverse. Non high-throughput methods were initially implemented like DNA footprinting or electrophoretic mobility shift assays (Galas and Schmitz, 1978; Garner and Revzin, 1981; O'Neill and Turner, 1996), these data being a valuable source of several gene regulation databases. According to the genomics advance and DNA sequencing technologies, high-throughput methods were necessary for discovering TFBSs such as Protein binding microarrays, ChIP-chip or ChIP-Seq experiments (Ren et al., 2000; Berger and Bulyk, 2006; Johnson et al., 2007). These methodologies produce large volumes of raw sequence data and different computational strategies need to be implemented for preprocessing and filtering data to find DNA motifs. On the other hand, site-directed mutagenesis (O'Neill et al., 1998) is based on the introduction of modifications in the nucleotide bases that are recognized by the TF residues, restriction enzymes must recognize target sequences with precision to interfere with DNA binding. Nonetheless, once a TFBS is discovered, it still remains to assign its binding to this site to the regulation of a given gene. To do so, one of the techniques is to select targets for a TF if it binds in the respective regulatory region of a gene, e.g., its promoter. Another common way to determine which TFs regulates certain genes is to determine whether their binding motifs or experimentally determined binding sites are near the gene or within a certain distance from the transcription start site (Blatti et al., 2015; Liu et al., 2015; Garcia-Alonso et al., 2019; Qin et al., 2020; Murgas et al., 2021).

There is a fourth approach that aims to assign TFs to genes by identifying regulatory relationships from transcriptional profiles using computational approaches such as GENIE3 (Huynh-Thu et al., 2010) and ARACNE (Margolin et al., 2006). Both tools rely on a relatively large number of transcriptomic experiments, benefiting from the presence of various experimental conditions, and arguable reliability (Marbach et al., 2012; Mochida et al., 2018). While most of these approaches are validated using knowledge driven GRNs such as RegNetwork (Liu et al., 2015), some of the most recent ones employ ChIP-Seq determined TFBSs to estimate their performance (Janky et al., 2014; Desai et al., 2017). Other approaches perform noise reduction in GRNs not only with experimentally determined TFBSs, but also applying GWAS SNPs which are known to alter TF-binding affinities (Chen et al., 2020). Pioneering work in this area related TFBSs to the logfold changes observed in microarray experiments (Bussemaker et al., 2001) or TFs instead of their binding sites once TFBSs were used to assign TF to genes (Gao et al., 2004).

Nowadays, the number of experimentally determined TFBSs keeps steadily growing. This growth is specially relevant for TFBSs determined by high-throughput techniques and made available in general repositories such as GEO (Barrett et al., 2013) and ArrayExpress (Athar et al., 2019) or in specialized portals such as ENCODE (Contrino et al., 2012). Even so, it is still difficult and expensive to prove that any TFBS is involved in the regulation of a gene. To overcome the lack of tools to assign TFs to the regulation of their target genes, we propose a two-step approach to both improve and automate the assignation of TF to the regulation of target genes. The first step of our methodology assigns TF to genes employing a distance threshold between ChIP-Seq derived TFBSs and genes, creating a GRN that over-estimates targets for each TF (Chen et al., 2020). Then, in a second step, this initial GRN is filtered by using a large collection of RNA-Seq data and GENIE3, but instead of using this tool to select regulators

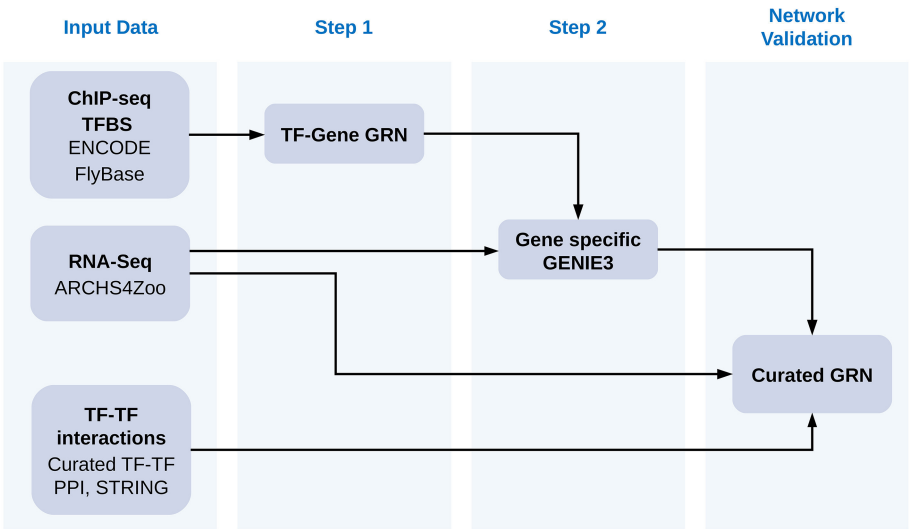

FIGURE 1 | Workflow of our approach. We first gathered a collection of TFBS from ENCODE and FlyBase determined with ChIP-Seq experiments and used them to assign TF to the regulation of specific genes according to their distance to genes. We then used GENIE3 to prune TFs for each gene. We employed as input for GENIE3 all gene counts available for *Drosophila melanogaster* at the ARCHS4ZOO repository for all TFs assigned to the same gene in the first step. We then demonstrated how the results of pruning TF-gene assignments improved the resulting gene regulatory networks by increasing the connectance in the TF-TF interaction networks made of all regulators for the same gene. We employed TF-TF interactions from a curated yeast two hybrids collection, from TF-TF interactions obatined at the STRING database and from TF-TF coexpresion networks calculated from ARCHS4ZOO gene counts. Additionally we also demonstrated that genes sharing more than one TF tend to have expression patterns more correlated after the second step of our approach than by simply using distance cut-offs to assign TF to genes.

from all TFs in the genome for each gene, we use it to select regulators from all TFs assigned to a gene in the first step.

To demonstrate the improved consistency of resulting networks we employed *D. melanogaster* because of its relatively small genome and the availability of experimentally determined TFBS for many TFs. Based on that, TFs that regulate the same gene tend to interact between them (Shokri et al., 2019), forming the so called transcriptional complex (Ogata et al., 2003), we will show how our approach provides an effective method to increase the reliability of TF target assignments. In this way, one expects an increase on the connectance in interaction networks made of all TFs regulating the same gene after using our approach. In addition, as a case example to show the utility of our approach, we studied the role of *D. melanogaster* gene *Hr96* (UniProt Q24143) in the transcriptional control of mitochondrial genes. *Hr96* is a TF orthologous to the human Vitamin D receptor (Fisk and Thummel, 1995). *Hr96* is activated by small lipophilic compounds from dietary signals and metabolic intermediates, acting in the regulation of developmental pathways and cellular metabolism (McKenna and O'Malley, 2002). It is mainly expressed during the mid-embryogenesis stages in the metabolic fat body, excretory organs, and in the central nervous system (Wilk et al., 2013), mostly induced by the ecdysone hormone, the main factor that coordinates molting and metamorphosis (Fisk and Thummel, 1995). *Hr96* plays a role in xenobiotics detection such as the pesticide DDT and phenobarbital, inducing the expression of detoxification and clearance genes (King-Jones et al., 2006). Furthermore, *Hr96* has a key role in lipid metabolism, sensing triacylglycerol levels to facilitate their breakdown, and regulating cholesterol catabolism through modulation of genes involved in its storage, uptake, and trafficking (Horner et al., 2009; Sieber and Thummel, 2009). However, despite these features, little is still known about the role of *Hr96* on the regulation of gene expression associated with mitochondrial function to directly modulate lipid and energy metabolism.

2. MATERIALS AND METHODS

The general workflow of our approach is described graphically in **Figure 1**. Each of the steps described in the figure and how we obtained data is explained in detail below.

2.1. Reference Gene Regulatory Networks

We created reference gene regulatory networks for *D. melanogaster* by combining TFBS information from the ChIP-Seq available at the ENCODE data repository (Contrino et al., 2012) and FlyBase (Thurmond et al., 2019) as were available on July 2019 and March 2020, respectively. In this way, we inferred regulatory relationships based on the distance between the ChIP-Seq determined TFBSs for a total of 350 TFs and the Transcription Start Site (TSS) of each gene in the genome of the fruit fly version 6.32. To determine whether a TF regulates a gene, we chose distance thresholds between TFBSs and the TSS of each gene, so if the TFBS falls within this distance, we assumed it regulates the respective gene. We created three reference networks with different distance thresholds: 1,500, 2,000, and 5,000 nucleotides inspired by other approaches (Dupuy et al., 2004; Blatti et al., 2015) and described in **Table 1**. Further details

TABLE 1 | Description of the networks analyzed in this work.

	Threshold (kb)	Genes	Edges	Avg. Indegree	Avg. Outdegree
Reference networks	1.5	15,576	1,094,130	44.50	3,126.09
	2	15,899	1,190,168	45.43	3,400.48
	5	16,665	1,679,173	47.61	4,797.64
Filtered networks	1.5	11,635	147,203	33.24	420.58
	2	11,968	369,346	34.19	1,055.27
	5	12,994	467,442	37.13	1,335.55

Reference networks were created by assigning TFs to the regulation of specific genes based on a distance threshold between the TFBS and the gene. Filtered networks were created by selecting the TFs for each gene that better predict its expression levels with GENIE3. All networks described in the table contain the same 350 TFs.

on ChIP-Seq data employed and the procedure used are available in Murgas et al. (2021).

2.2. Gene Expression Profiles and Network Inference

To obtain a comprehensive dataset of transcriptomic data, we employed all RNA-Seq experiments of *D. melanogaster* available at ARCHS4ZOO version update 8/2018 (Lachmann et al., 2018) as was available on April 2020 at https://maayanlab.cloud/archs4/archs4zoo.html. This dataset comprises 9,924 RNA-seq samples belonging to 368 series and gene counts were used as available from the data repository without further processing as previously recommended (Aibar et al., 2017). This dataset of gene expression profiles was then employed with GENIE3 (Huynh-Thu et al., 2010) to remove TF-gene regulations from the regulators assigned to each gene in the reference networks. GENIE3 employs a random forest algorithm to select the subset of TF for each gene whose expression better predicts the expression of the gene, assigning them those TFs as regulators of that gene. In our case, we created subsets of expression data with all samples for each gene and for all TFs that were assigned as its regulators using each of the three distance thresholds, and employed GENIE3 to determine which TFs better predicted the expression of the gene, and thus, were actually regulating it. GENIE3 does not use a preset cut-off to select regulators and reports the relevance of each TF sorted by decreasing values. To remove the most unlikely regulators, we implemented a dynamic threshold by which for each gene we removed all TFs with a relevance lower than 10% of that reported for the most relevant TF.

2.3. Improvement of TF-Gene Assignment

We measured connectance in interaction networks made of all TFs that regulate the same gene in networks before and after using GENIE3 and counted for how many genes connectance increased. We define the connectance of a network, or connectivity density, as the fraction of connections present in a network divided by the total number of edges that could take place in the network. The connectance (ρ) lies in the range [0,1], with greater values indicating that nodes are more interconnected between them than with values closer to 0. This way, to estimate the quality of a GRN relies on the fact that TFs controlling the expression of a gene are more likely to interact between them (Shokri et al., 2019).

To validate our approach, we employed several types of TF interaction networks: a curated Protein-Protein Interaction (PPI) network (Shokri et al., 2019); a correlation network calculated with Pearson's correlation coefficient on the same expression data used with GENIE3 with edges defined with different thresholds; and STRING functional networks (Szklarczyk et al., 2019) created querying this database with all 350 TFs on September 2020 and filtering the resulting network at different confidence thresholds for combined score and several evidence types on its own. These networks are described in **Table 2**. Additionally, we also calculated average gene co-expression for all pairs of genes regulated by at least the same two TFs. This is based on the idea that co-regulated genes should have more similar expression patterns than those which are not regulated by the same TFs (Martyanov and Gross, 2010). We calculated average Pearson correlation on the ARCHS4ZOO RNA-Seq data between pairs of genes that share more than one TF in filtered and reference networks. We assumed normality and used a two samples *T*-test to compare if the difference between the average for genes sharing the same number of regulators before and after GENIE3 was significant.

2.4. Hr96 and Its Role in *D. melanogaster* Mitochondrial Function

2.4.1. Selection of Mitochondrial Genes and Functional Characterization

We first assigned all *D. melanogaster* genes as mitochondrial if sub-cellular localization GO terms associated to them available at FlyBase (Thurmond et al., 2019) contained the term "mitochondria." Following, we created GRNs formed by these mitochondrial genes and all TFs in the networks using the regulations present in the global networks.

2.4.2. Network Analysis, Visualization and *Hr96* Centered Subnetworks

All network analyzes were carried out using Cytoscape (Shannon et al., 2003). This platform was also employed to create subnetworks using its graphical interface as follow. Subnetworks centered on *Hr96* were created by selecting its node in each network before and after applying our procedure, and then using Cytoscape to select all nodes connected to *Hr96* by edges arising from it, i.e., regulated by *Hr96*.

3. RESULTS

We first show how our approach improves the consistency of TF-gene assignment created by assigning TFs to genes if a TFBS is near the gene. Following, we demonstrate how using the improved version of the networks leads to edges that are more likely to take place, and which, in fact, allow interpretation and analysis that are precluded in unpruned networks.

TABLE 2 | Description of TF-TF interaction networks employed to verify our approach.

Network	Nodes	Edges
Corr 0.25	349	30,915
Corr 0.45	340	16,584
Corr 0.65	288	6,912
Corr 0.85	137	353
Curated_PPI	271	796
STRING (combined ≥ 0.5)	260	1,065
STRING (combined ≥ 0.8)	150	241
STRING (textmining ≥ 0.4)	265	1,351
STRING (textmining ≥ 0.6)	196	502
STRING (textmining ≥ 0.8)	139	223
STRING (database_annotated ≥ 0.4)	53	69
STRING (experimentally_determined ≥ 0.5)	117	120
STRING (experimentally_determined ≥ 0.7)	65	49
STRING (experimentally_determined ≥ 0.9)	26	15

Correlation networks were created using different thresholds of positive Pearson's correlation coefficient. Curated PPI is the subnetwork of the 350 TFs employed to create our reference GRN verified experimentally in (Shokri et al., 2019). Interaction networks obtained from STRING (Szklarczyk et al., 2019) differ on the criteria employed to define edges: STRING (combined score ≥ 0.5) is the functional interaction network retrieved querying the STRING web with the 350 TF and by default parameters, i.e., combined score ≥ 0.5. All other STRING networks were created by employing different thresholds with single evidence types and thresholds applied.

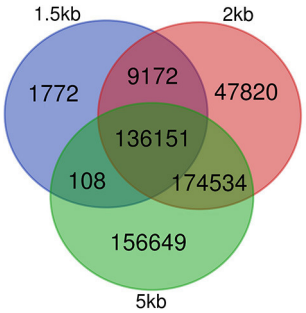

FIGURE 2 | Conservation of edges in GRNs after filtering unlikely edges. Venn diagram showing edges in GENIE3 networks for each of the three distance thresholds employed, 1.5, 2, and 5 kb. Edges were defined by their source and target node IDs.

3.1. Characterization of Networks Before and After Applying Our Approach

Table 1 shows different properties of the networks created using three distance thresholds (1.5, 2, and 5 kb) to assign TFs to the regulation of genes. First, all networks before and after applying our approach contain edges arising from all the 350 different TFs employed in this work. We then looked at the average outdegree and indegree, respectively for TF and non-TF genes in each network. These metrics, averaged connectivity for each node type, serve as indicator of how dense the networks are. While unfiltered networks have average outdegree ranging from 3,126 in the network with the more restrictive distance threshold of 1.5 Kb–4,797 in the 5 kb threshold network, the networks after using our approach have smaller values (420 with 1.5 kb–1,335 with 5 kb), evidencing a significant reduction on the number of genes regulated by the same TFs. Regarding the number of nodes that are connected by at least one edge, there is also a decrease of about 4,000 in the number of genes in the three networks and a reduction in the average indegree.

Regarding the number of TF and nodes, networks made with shorter distance thresholds are included in reference GRNs made with longer distance cut-offs before filtering. For filtered networks, this is not the case. All nodes with at least one connection in the 1.5 kb filtered network are in the network made with the 2 kb threshold, and the same occurs with nodes in the 2 and 5 kb cut-off. Nonetheless, some of the edges in the 1.5 kb network are not present in the 2 kb and the same occurs for edges in the 2 and 5 kb networks (see **Figure 2**). This is caused by the dependence of each edge on the expression patterns of all regulatory nodes for each gene and how GENIE3 combines them.

3.1.1. Connectance Analysis on TF-TF Interaction Networks

Considering the connectance in all TF-TF subnetworks made with all regulators for each gene, there is a clear trend after applying our approach. We observe a greater number of genes with increased connectance in the TF-TF interaction network for all the regulators of each gene, see **Table 3**. Employing the curated PPI network, more genes show an increase in the TF connectance than genes showing a decrease in their TF connectance for all three distance cut-offs. Using the curated PPI the network with the 2 kb distance threshold has the smaller proportion of genes with decreased connectance. Using co-expression networks made at different thresholds of Pearson's correlation, the number of genes with greater connectance is notoriously larger than the number of genes with lower. As the correlation threshold used to define edges increases, the proportion of genes with smaller connectance increases as genes with greater values decrease. With STRING interaction networks and the reference network created with the 1.5 kb threshold, our approach produced TF-TF interaction subnetworks with lower values of connectance for most of the genes. In contrast, with the other two reference networks (2 and 5 kb) we also see the general trend of better connectance after our approach.

3.1.2. Co-expression Analysis of co-regulated Genes

We compared the mean co-expression correlation between all pairs of genes that share at least two TFs in networks before and after filtering them with GENIE3 on the three cut-offs (See excel file provided in **Supplementary Material**). We found a decrease in the number of genes corregulated by the same TFs after filtering the networks, the maximum number of shared TFs between at least five pairs of genes is 25 in the filtered network at 1.5 kb while there are seven pairs of genes sharing 322 TFs before using GENIE3. Greater number of shared TFs between genes are also seen with 2 and 5 kb thresholds, but again there are less shared regulators after filtering the networks. Considering

TABLE 3 | TF interaction connectance comparison between networks before and after using our approach.

		Genie3_1,500			Genie3_2,000			Genie3_5,000			Reference network	
		Better	Worse	Equal	Better	Worse	Equal	Better	Worse	Equal	#edges	#nodes
ARCHS4	Curated PPI	0.373	0.333	0.294	0.346	0.222	0.433	0.381	0.255	0.364	796	271
	corr_0.25	0.620	0.093	0.286	0.511	0.065	0.424	0.546	0.099	0.355	30,915	344
	corr_0.45	0.603	0.111	0.287	0.492	0.084	0.425	0.520	0.125	0.355	16,584	326
	corr_0.65	0.569	0.143	0.287	0.462	0.112	0.425	0.482	0.161	0.356	6,912	256
	corr_0.85	0.492	0.187	0.321	0.412	0.145	0.443	0.430	0.194	0.377	353	95
STRING	combined_0.5	0.352	0.355	0.293	0.306	0.264	0.43	0.341	0.298	0.361	1065	260
	combined_0.8	0.297	0.392	0.311	0.336	0.225	0.439	0.362	0.265	0.373	241	150
	textmining_0.4	0.344	0.366	0.29	0.285	0.287	0.428	0.317	0.323	0.36	1351	265
	textmining_0.6	0.299	0.392	0.309	0.291	0.274	0.435	0.332	0.302	0.366	502	196
	textmining_08	0.191	0.461	0.348	0.275	0.273	0.452	0.315	0.305	0.38	223	139
	experimental_05	0.184	0.468	0.348	0.3	0.241	0.459	0.333	0.278	0.388	120	117
	experimental_07	0.1	0.48	0.419	0.233	0.27	0.498	0.255	0.315	0.43	49	65
	experimental_09	0.047	0.428	0.525	0.148	0.281	0.571	0.165	0.336	0.499	15	26
	database_04	0.228	0.434	0.337	0.334	0.208	0.459	0.358	0.249	0.392	69	53

This table shows the percentage of genes with greater connectance in the interaction network for all its TFs in all interaction networks employed to test how using GENIE3 to filter the networks improved the three GRN based on distance TF assignment (1.5, 2, and 5 kb at maximum between the TFBS and its target gene).

the statistical significance ($p \leq 0.0005$) of the difference between the means, we found that in the 1.5 kb networks, pairs of genes sharing at least 2, at least 3, 4, 5, 6, 7, 8, 9, and up to 10 TFs are significantly more correlated after filtering the networks. At 2 kb cut-off, means of correlated co-expression are greater for pairs of genes sharing from 2 to 18 regulators and from 2 to 20 at 5 kb.

3.2. *Hr96* and Its Role in *D. melanogaster* Mitochondrial Function

Here we report the results of studying the subnetwork centered on *Hr96*. We first looked at the overall changes in this subnetwork before and after filtering it with GENIE3 at the three selected distance thresholds used to assign TFs to genes. We then focus on the analysis of the genes in these subnetworks. The decrease in the number of edges and nodes in the subnetworks centered on *Hr96* is evident in **Table 4**. This reduction in network elements is more notable regarding the number of edges, which show a reduction of more than 90% in all three networks compared to the 58–76% reduction in the number of nodes. Accordingly to what we saw on whole genome GRNs (see **Table 1**), there is also a large decrease in the average outdegree for TFs in the *Hr96* centered subnetworks. As to differences on the three distance thresholds, 2 and 5 kb GRNs behave more similarly between them than when compared with the 1.5 kb GRN. There are six edges exclusively in the 1.5 kb filtered subnetwork of *Hr96* which are absent in the 2 and 5 kb GRNs, and 52 nodes are present only in the 2 kb network and 167 in the 5 kb (see **Figure 3**). However, there is yet a trend of fewer edges in GRNs made with more stringent thresholds that in their majority appear in more relaxed cutoffs.

Based on its reduced number of nodes and edges (see **Supplementary Material**), we selected the subnetwork centered on *Hr96* made with the 1.5 kb threshold to study the function of this TF on the regulation of mitochondrial genes, shown in **Figure 4A** as well as the subnetwork generated in the same way for the 1.5 kb before applying GENIE3 as a filter (top left inset). There are only 14 TFs (all regulated by *Hr96*) that form a densely connected regulatory cascade together with 33 non-TF coding

TABLE 4 | Description of subnetworks centered on *Hr96*.

	Before		After	
Network	Nodes (TFs)	Edges	Nodes (TFs)	Edges
1.5 kb	191 (81)	8,840	47 (14)	135
2 kb	201 (84)	9,859	84 (17)	384
5 kb	253 (109)	17,652	98 (21)	478

Number of nodes and edges in the subnetworks created starting from Hr96 for each of the GRNs created at different distance thresholds before and after applying our approach. The number of nodes depicting TF coding genes is between brackets.

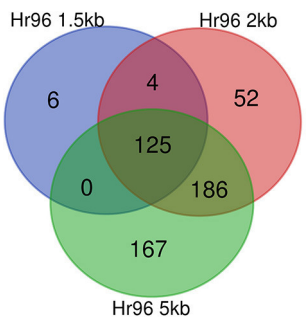

FIGURE 3 | Conservation of edges in *Hr96* centered subnetworks after filtering unlikely edges. Venn diagram showing edges in GENIE3 networks for each of the three distance thresholds employed, 1.5, 2, and 5 kb. Edges were defined by their source and target node IDs.

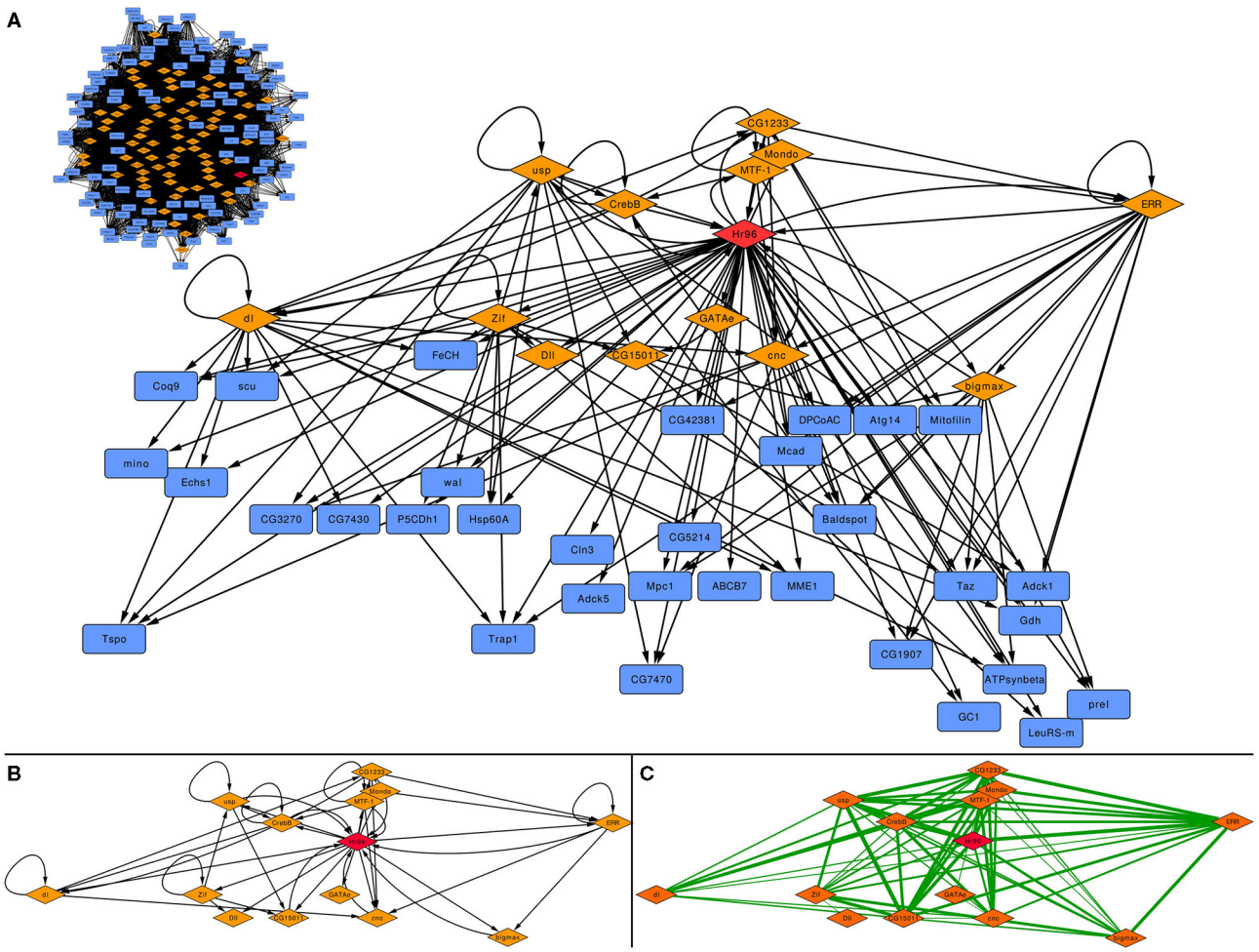

FIGURE 4 | Subnetwork centred on Hr96 depicting its involvement on the regulation of mitochondrial genes. **(A)** depicts the whole subnetwork after applying GENIE3 and before (inset on the top left side of the panel); **(B)** represents the inter-regulation of the 14 TFs present in this subnetwork; and **(C)** is the TF-TF interaction network formed by the same 14 TFs where edges represent Pearson correlation calculated on the 9,924 samples for *D. melanogaster* obtained from the ARCHS4ZOO [thickness is proportional to the value of positive correlation coefficients, all in the range (0.46, 0.88)]. Orange diamonds depict TFs, blue rectangles non-TF genes and the red diamond is *Hr96*.

genes. **Figure 4B** displays how these 14 TFs are interconnected maintaining the same layout as above, while edges between these TFs in **Figure 4C** represent Pearson's correlation calculated using the same expression data previously employed with GENIE3, with their thickness indicating higher coefficients. There are 66 edges in the correlation network, 20 more than in the GRN made with the same TFs, indicating a strong co-expression pattern between these related TFs. The same network generated before applying GENIE3 is formed by 81 TFs and 110 non-TF coding genes (top left of **Figure 4A**). Using 2 kb, the network filtered with GENIE3 centered on *Hr96* contains three more TFs and 34 more non-TF genes, while before GENIE3 it has 84 TFs and 117 genes (See **Supplementary Material**). With the less stringent cut-off of 5 kb, the network filtered with GENIE3 is formed by 21 TFs, the 17 included in the 2 kb network plus another 4, and 77 non-TF. Before using GENIE3 on the 5 kb GRN, the subnetwork has 109 TF and 144 non-TF genes (See **Supplementary material**).

We then studied the function carried out by those 33 genes in the *Hr96* 1.5 kb GRN filtered with GENIE3. Among these, there are several carboxylic acid-related genes, especially involved in its transport and metabolism. This result indeed highlights the *Hr96* regulation of lipid metabolism-related targets in the mitochondria. In the glutamate and fatty acid metabolic and carboxylic catabolic processes, we found that the *Hr96*-mitochondrial network mainly links enzymes such as dehydrogenases, oxidoreductases, and a short-chain enoyl-CoA hydratase (*Echs1*).

4. DISCUSSION

The control of gene transcription is one of the key processes in living organisms. Despite its relevance, we still do not know most of the specific TFs that determine which gene is expressed and which is not. Currently, high throughput techniques such as

ChIP-Seq are routinely employed to annotate TFBSs, but even if this type of knowledge becomes widespread, it still remains to assign TF binding each site to the regulation of target genes. However, even if TF target assignment is carried out routinely in a low-throughput fashion for some TF-gene pairs, whole genome TF target identification remains an expensive and almost impossible task using experimental verification. Here, we propose a two step approach to address this issue: first TFs are assigned to the regulation of certain genes if ChIP-Seq derived binding sites fall within a distance cut-off to the gene. Then, in a second step, for each gene, we remove improbable regulations by using a large collection of RNA-seq data (Lachmann et al., 2018) as input for GENIE3 (Huynh-Thu et al., 2010). Instead of feeding GENIE3 with the expression of all TFs and genes, for each gene we only employed its expression and the expression of all regulators assigned to it in the first step. By doing this, we changed the purpose of GENIE3 from whole genome GRN inference to GRN pruning.

Most eukaryotic genes are regulated by more than one TF that, acting simultaneously, determine whether their target gene expresses or not. TFs, thus, interact forming transcriptional complexes (Ogata et al., 2003) in a cooperative fashion (Hancock et al., 2019) to actively control transcription. Consequently, we assumed that the actual regulatory TFs of each gene would need to interact forming an interconnected TF-TF interaction network. And thus, that the connectance of this TF-TF interaction network would increase if wrongly assigned TFs were removed from the regulation of each gene. We took advantage of a recently released, high confidence, TF-TF PPI network of *D. melanogaster* (Shokri et al., 2019) to test if the connectance between all TFs assigned to each gene increased as expected after using our approach. In addition, to demonstrate the improvement in TF-target assignment deemed to our approach, we also employed several other interaction networks obtained from STRING functional networks (Szklarczyk et al., 2019) and a co-expression network calculated with Pearson's correlation on the same transcriptional dataset employed to remove TF-gene pairs with GENIE3.

We tested if the connectance between TFs regulating the same gene increased with three different distance thresholds of 1.5, 2, and 5 kb for the initial assignation of TFs to genes (**Table 3**). For a 2 kb cut-off, our results indicate a consistent increase of connectance calculated for all regulators that is independent of how the interactions between TFs are defined. This tendency is almost as consistent for 5 kb and can also be seen for 1.5 kb, even if there are few exceptions for these improvement on the connectance. Importantly, these exceptions mainly appear for very stringent definitions of TF-TF interactions, such as a STRING combined score $\geq$0.8, or STRING experimental score $\geq$0.9 for all three cut-offs. Nonetheless, using the high confidence PPI network (Shokri et al., 2019) and all correlated co-expression, a majority of genes had better connectance among their regulators after using GENIE3 than without using it. Even if, biologically, it makes more sense that our approach results in higher connectance between the regulators of each gene, experimentally this can only be tested by comparing our results with a null background. In our case, this would imply the need to randomly remove TF-gene associations for each gene. Nonetheless, it is expected that as TF-TF interaction networks are very sparse, any randomly selected subnetwork is deemed to also be sparse, unless there is biological significance embebed in the approach followed to remove edges.

The observed TF-connectance improvement is more consistent if the TF interaction network has interactions for all regulators. As shown in **Table 2**, the network whose edges are Pearson's correlation $\geq$ 0.25 (*cor*_0.25) contains interactions for 349 out of 350 TFs for which there were ChIP-Seq data available and 62% of the genes show better connectance at 1.5 kb (9.3% worse), 51.1% are better at 2 kb (6.5% worse), and 54.6% at 5 kb (9.9% worse). On the other hand, using Pearson's $\geq$ 0.85 (*cor*_0.85) there are only interactions for 137 TFs and 49.2% of the genes showed improved TF connectance at 1.5kb (18.7% worse), 41.2% are better at 2 kb (14.5%), and 43% at 5 kb (19.4% worse). This previous example indicates that having TF interaction networks with high confidence interactions for all regulators is a key factor to consider when estimating the certainty of the improvement in connectance. It is also very important to take into consideration that a correlation $\geq$ 0.25 is very likely to be significant taking into account it was calculated with 9,924 expression experiments. It should also be considered that the yeast two hybrid experiment, used to determine the PPI curated network, simply does not work for some proteins or it may produce too many false positive or false negative hits (Koegl and Uetz, 2007), and thus, careful curation is indispensable. Similarly, STRING networks are automatically generated and their scores are calculated without any human intervention, making it desirable to carry out manual inspection of each edge and its supporting evidence before using it. Importantly, the results we obtained from the analysis of co-expression between gene pairs that share the same number of TF before and after filtering the networks, also support that our approach does indeed improve the reliability of TF-gene assignment (see excel file in the **Supplementary Material**). These results also showed a notable decrease in pairs of genes that share large numbers of regulators (more than 25 shared TFs), which is caused by the reduction on the number of TF-gene assignments.

We then focused on the subnetwork centered on a specific TF to showcase the utility of the networks generated by our approach. Nuclear hormone receptors (NHR) represent a key hub in the regulation of development, reproduction, and metabolism (Fahrbach et al., 2012). Most NHRs are ligand-regulated TFs activated by lipophilic ligands such as steroid hormones, fatty acids, phospholipids, bile acids, vitamins, and xenobiotics (Huang et al., 2009). Humans present 48 NHR that, despite being widely explored in terms of structure and function, are not fully characterized (Evans and Mangelsdorf, 2014). Approximately half of those remain orphan receptors, a fact that imposes great difficulty to crack down their regulatory network (Weikum et al., 2018). In contrast, the *D. melanogaster* genome carries only 18 nuclear-receptor genes, which represent all six NHR mammalian subfamilies, but importantly showing lower functional redundancy (King-Jones and Thummel, 2005; Palanker et al., 2006). Among *Drosophila* NHRs, *Hr96* (UniProt Q24143) is an interesting case due to its orthology with three vertebrate NHR: Vitamin D Receptor (VDR) (Fisk and Thummel, 1995), Pregnane X

Receptor (PXR), and Constitutive Androstane Receptor (CAR) (Hoffmann and Partridge, 2015).

VDR (UniProt P11473) is widely distributed in mammal tissues (Eyles et al., 2005) and exerts transcriptional control, influenced by vitamin D, in over 3% of the human genome (Ramagopalan et al., 2010; Shirvani et al., 2019). The control that VDR exerts on gene regulation is significantly enriched over the immune functions, cell cycle activity, DNA replication, stress response (Hossein-nezhad et al., 2013) and, also significantly contributes to mitochondrial transcriptional regulating, biogenesis, and metabolism (Lee et al., 2008). Specifically, human skeletal muscle cells treated with the VDR-ligand 1α,25(OH)2D3 showed increased mitochondrial oxygen consumption rate and network mass by down-regulating fission proteins Drp1 and Fis1, and up-regulating the fusion protein OPA1 and the mitochondrial biogenesis modulators MYC, mitogen-activated protein kinase 13 (MAPK13), and endothelial PAS domain-containing protein 1 (EPAS1) (Ryan et al., 2016). In contrast, VDR silencing appears to cause a reduction in cellular respiration, ATP production (Ashcroft et al., 2020) and induces ROS production by up-regulating cytochrome C oxidase subunits proteins (COX2; COX4) and ATP synthase subunits (ATP5B; ATP6), which enhance respiratory membrane potential leading to protons leakage (Ricca et al., 2018). In this way, to test the hypothesis that *Hr96* has the potential to regulate mitochondrial function and improves lipid-based energy production, we used our hybrid protocol to showcase its ability to improve TF factor target assignments.

We analyzed all 33 *Hr96* targeted genes that do not code for TF in the curated 1.5 kb *Hr96* network to further characterize the role of this TF in any specific process. It is important to highlight here that 32 of these genes where also present in the 2 and 5 kb curated subnetworks. In addition, we also disregarded other genes also regulated by the other 13 TFs that are also present in the subnetwork, trying in this way to emphasize the role of this NHR.

The Delta-1-Pyrroline-5-carboxylate dehydrogenase 1 (*P5CDh1*) and Glutamate dehydrogenase (*Gdh*) are enzymes that support energy metabolism by glutamate and α-Ketoglutarate production, to promote the mitochondrial respiration (He and DiMario, 2011; Hohnholt et al., 2018). As well Adck1, which is essential to keep mitochondrial structural organization, energy, and ROS production under control (Yoon et al., 2019). β-oxidation, the catabolic pathway that breaks down fatty acids in the mitochondria, is highly represented in the *Hr96*-network by different genes. Scully (*scu*) and *Mcad* catalyze two different β-oxidation enzymatic steps and are highly conserved (Torroja et al., 1998; Lim et al., 2018). The *wal* gene encodes an electron transfer flavoprotein subunit that works as a specific electron acceptor in the mitochondrial fatty acid β-oxidation of fatty acids (Alves et al., 2012; Chokchaiwong et al., 2019), while ECHS1 is shown to be involved in the second step of mitochondrial β-oxidation (Hirai et al., 2001; Al Mutairi et al., 2017). All these targets operate to maintain the respiratory chain and energy production through carboxylic acid metabolism. To our knowledge, the activity of these enzymes has not been related to *Hr96* until now. In the same line, *Hr96* modulates the Minotaur (*mino*) activity, a conserved glycerol-3-phosphate O-acyltransferase responsible for triglycerides synthesis and lipid droplets biogenesis (Fantin et al., 2019). It has been shown that when this enzyme is down-regulated as observed upon bacterial infection, there is a progressive loss of lipid energy stores (Dionne et al., 2006), meanwhile, its expression is increased in the face of starvation (Fujikawa et al., 2009) possibly promoting a mitochondrial adaptation toward lipid metabolism.

Baldspot (Elovl6) is another fatty acid-related gene regulated by *Hr96*. The Elov16 enzyme extends C16 fatty acids to C18. It has been shown that flies lacking Elovl6 present impaired mitochondrial respiration by promoting a hyper-fragmentation of the mitochondrial network through JNK signaling and mitofusin ubiquitination (Senyilmaz et al., 2015). Regarding anion transport, to properly regulate the mitochondrial β-oxidation, *Hr96* seems to also coordinate the transcription of carboxylic acid transport targets such glutamate carrier (GC1), mitochondrial pyruvate carrier (*mpc1*), and *Cln3*, the Batten disease-associated gene involved in arginine transport and mitochondrial β-oxidation support (Dawson et al., 1996; Chan et al., 2009). Among those, MPC1 has an important role in mitochondrial function since it is found in the inner mitochondrial membrane, and mutant *D. melanogaster* for *mpc1* display impaired pyruvate metabolism, leading to a shortage of intermediates necessary for the tricarboxylic acid cycle, ultimately reducing ATP production (Bricker et al., 2012; Tang, 2019; Rossi et al., 2020). These findings are in line with the most recent research on *Hr96* functionality that points toward its relevance in the regulation of sterol trafficking, housing, and consumption (Sieber and Thummel, 2012). Considering our analyzes, it is possible to postulate that *Hr96* also regulates triacylglycerol metabolism by modulating the transcription of mitochondrial genes to stimulate lipid consumption and mitochondrial respiration to increase ATP production.

Altogether, this analysis highlights the potential effect of Hr96 on key mitochondrial processes such as the catabolism and transport of fatty acids and small molecules.

5. CONCLUSION

We created a two-step approach with the main purpose of helping to assign TF to the regulation of specific genes. We demonstrated that the consistency of TF-gene assignment improves by increasing the number of TFs targeting the same gene that are known to interact between them. In the process of testing our approach, we investigated several distance thresholds to assign TFs to genes. Based on how the number of edges in a GRN varies more by increasing the cut-off distance between the TSS of each gene and the TFBS from 1.5 to 2 kb than by increasing it from 2 to 5 kb, we can say that the best cut-off tested was 2 kb, better than to 1.5 or 5 kb. Our results also indicate that the TF-TF interaction networks are incomplete, and that even if our current results indicate in improvement in TF-gene assignment, more complete interaction networks would help in producing more reliable GRN.

Regarding the example case of *Hr96*, our analysis provides a rational framework for further investigations on *Hr96*-mitochondrial transcriptional regulation and offers an

opportunity to explore a better understanding of *Drosophila* lipid metabolism and signaling pathways for disease mechanisms.

As a final remark, our work proves that the integration of data from different sources is key to produce high quality GRNs, and thus, public data availability must be mandatory for all experimental results.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: *D. melanogaster* Gene counts employed were downloaded from https://maayanlab.cloud/archs4/archs4zoo.html. All GRNs and a Cytoscape session with mitochondrial networks can be found here https://figshare.com/projects/Filtering_of_datadriven_gene_regulatory_networks_using_Drosophila_melanogaster_as_a_case_study/95885. All code employed in this work is now available at https://github.com/networkbiolab/Network-Filtering.git together with a README file explaining all details.

AUTHOR CONTRIBUTIONS

YC-A and GG carried out most of the analysis performed. LM created the GR networks. CS and MS participated in the selection and analysis of *Hr96*. AM had the initial idea, designed the filtering approach, performed *in-silico* experiments and coordinated all people involved in the project. All authors wrote the manuscript.

FUNDING

FONDECYT regular 1181089 from Agencia Nacional de Investigación Científica y Desarrollo (ANID) to AM; ANID Subvención Instalación Academia (PAI77180059) and ANID Fondecyt Iniciación (1120098) to MS; ANID Ph.D. Fellowship 21201856 to LM, and Universidad Mayor Ph.D. scholarship to GG. HPC@CGB-UM: This research was partially supported by the computing infrastructure of the Centro de Genómica y Bioinformática, Universidad Mayor, Chile and from the Chilean National Laboratory of High Performance Computing (ECM-02).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.649764/full#supplementary-material

REFERENCES

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583, 699–710. doi: 10.1038/s41586-020-2493-4

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463

Al Mutairi, F., Shamseldin, H. E., Alfadhel, M., Rodenburg, R. J., and Alkuraya, F. S. (2017). A lethal neonatal phenotype of mitochondrial short-chain enoyl-CoA hydratase-1 deficiency. *Clin. Genet.* 91, 629–633. doi: 10.1111/cge.12891

Alves, E., Henriques, B. J., Rodrigues, J. V., Prudêncio, P., Rocha, H., Vilarinho, L., et al. (2012). Mutations at the flavin binding site of ETF: QO yield a MADD-like severe phenotype in *Drosophila*. *Bioch Biophys Acta*. 1822, 1284–1292. doi: 10.1016/j.bbadis.2012.05.003

Ashcroft, S. P., Bass, J. J., Kazi, A. A., Atherton, P. J., and Philp, A. (2020). The Vitamin D receptor regulates mitochondrial function in C2C12 myoblasts. *Am. J. Physiol.* 318, C536–541. doi: 10.1152/ajpcell.00568.2019

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress update-From bulk to single-cell expression data. *Nucleic Acids Res.* 47, D711–D715. doi: 10.1093/nar/gky964

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets-Update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Berger, M. F., and Bulyk, M. L. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.* 338, 245–260. doi: 10.1385/1-59745-097-9:245

Blatti, C., Kazemian, M., Wolfe, S., Brodsky, M., and Sinha, S. (2015). Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res.* 43, 3998–4012. doi: 10.1093/nar/gkv195

Bricker, D. K., Taylor, E. B., Schell, J. C., Orsak, T., Boutron, A., Chen, Y. C., et al. (2012). A mitochondrial pyruvate carrier required for pyruvate uptake in yeast, *Drosophila*, and humans. *Science* 336, 96–100. doi: 10.1126/science.1218099

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–171. doi: 10.1038/84792

Chan, C. H., Ramirez-Montealegre, D., and Pearce, D. A. (2009). Altered arginine metabolism in the central nervous system (CNS) of the Cln3-/- mouse model of juvenile Batten disease. *Neuropathol. Appl. Neurobiol.* 35, 189–207. doi: 10.1111/j.1365-2990.2008.00984.x

Chen, C. H., Zheng, R., Tokheim, C., Dong, X., Fan, J., Wan, C., et al. (2020). Determinants of transcription factor regulatory range. *Nat. Commun.* 11, 1–15. doi: 10.1038/s41467-020-16106-x

Chokchaiwong, S., Kuo, Y.-T., Hsu, S.-P., Hsu, Y.-C., Lin, S.-H., Zhong, W.-B., et al. (2019). ETF-QO mutants uncoupled fatty acid β-oxidation and mitochondrial bioenergetics leading to lipid pathology. *Cells* 8, 106. doi: 10.3390/cells8020106

Contrino, S., Smith, R. N., Butano, D., Carr, A., Hu, F., Lyne, R., et al. (2012). modMine: Flexible access to modENCODE data. *Nucleic Acids Res.* 40, D1082–D1088. doi: 10.1093/nar/gkr921

Dawson, G., Kilkus, J., Siakotos, A. N., and Singh, I. (1996). Mitochondrial abnormalities in CLN2 and CLN3 forms of Batten disease. *Mol. Chem. Neuropathol.* 29, 227–235. doi: 10.1007/BF02815004

Desai, J. S., Sartor, R. C., Lawas, L. M., Jagadish, S. V., and Doherty, C. J. (2017). Improving gene regulatory network inference by incorporating rates of transcriptional changes. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-17143-1

Dionne, M. S., Pham, L. N., Shirasu-Hiza, M., and Schneider, D. S. (2006). Akt and foxo dysregulation contribute to infection-induced wasting in *Drosophila*. *Curr. Biol.* 16, 1977–1985. doi: 10.1016/j.cub.2006.08.052

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* 25, 1268–1280. doi: 10.1101/gr.184671.114

Dupuy, D., Li, Q. R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., et al. (2004). A first version of the *Caenorhabditis elegans* promoterome. *Genome Res.* 14, 2169–2175. doi: 10.1101/gr.2497604

Evans, R. M., and Mangelsdorf, D. J. (2014). Nuclear receptors, RXR, and the big bang. *Cell* 157, 255–266. doi: 10.1016/j.cell.2014.03.012

Eyles, D. W., Smith, S., Kinobe, R., Hewison, M., and McGrath, J. J. (2005). Distribution of the Vitamin D receptor and 1α-hydroxylase in human brain. *J. Chem. Neuroanat.* 29, 21–30. doi: 10.1016/j.jchemneu.2004.08.006

Fahrbach, S. E., Smagghe, G., and Velarde, R. A. (2012). Insect nuclear receptors. *Ann. Rev. Entomol.* 57, 83–106. doi: 10.1146/annurev-ento-120710-100607

Fantin, M., Garelli, F., Napoli, B., Forgiarini, A., Gumeni, S., De Martin, S., et al. (2019). Flavonoids regulate lipid droplets biogenesis in *Drosophila melanogaster*. *Natural Product Commun*. 14:1934578X1985243. doi: 10.1177/1934578X19852430

Fisk, G. J., and Thummel, C. S. (1995). Isolation, regulation, and DNA-binding properties of three *Drosophila* nuclear hormone receptor superfamily members. *Proc. Natl. Acad. Sci. U.S.A.* 92, 10604–10608. doi: 10.1073/pnas.92.23.10604

Forrest, A. R. R., Kawaji, H., Rehli, M., and Others (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Fujikawa, K., Takahashi, A., Nishimura, A., Itoh, M., Takano-Shimizu, T., and Ozaki, M. (2009). Characteristics of genes up-regulated and down-regulated after 24 h starvation in the head of *Drosophila*. *Gene* 446, 11–17. doi: 10.1016/j.gene.2009.06.017

Galas, D. J., and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* 5, 3157–3170. doi: 10.1093/nar/5.9.3157

Gao, F. G., Foat, B. C., and Bussemaker, H. J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* 5, 31. doi: 10.1186/1471-2105-5-31

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* 29, 1363–1375. doi: 10.1101/gr.240663.118

Garner, M. M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* 9, 3047–3060. doi: 10.1093/nar/9.13.3047

Hancock, S. P., Cascio, D., and Johnson, R. C. (2019). Cooperative DNA binding by proteins through DNA shape complementarity. *Nucleic Acids Res.* 47, 8874–8887. doi: 10.1093/nar/gkz642

Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., Chen, W. J., Cho, J., et al. (2020). WormBase: A modern model organism information resource. *Nucleic Acids Res.* 48, D762–D767. doi: 10.1093/nar/gkz920

He, F., and DiMario, P. J. (2011). *Drosophila* delta-1-pyrroline-5-carboxylate dehydrogenase (P5CDh) is required for proline breakdown and mitochondrial integrity-Establishing a fly model for human type II hyperprolinemia. *Mitochondrion* 11, 397–404. doi: 10.1016/j.mito.2010.12.001

Hirai, K., Aliev, G., Nunomura, A., Fujioka, H., Russell, R. L., Atwood, C. S., et al. (2001). Mitochondrial abnormalities in Alzheimer's disease. *J. Neurosci.* 21, 3017–3023. doi: 10.1523/JNEUROSCI.21-09-03017.2001

Hoffmann, J. M., and Partridge, L. (2015). Nuclear hormone receptors: roles of xenobiotic detoxification and sterol homeostasis in healthy aging. *Crit. Rev. Biochem. Mol. Biol.* 50, 380–392. doi: 10.3109/10409238.2015.1067186

Hohnholt, M. C., Andersen, V. H., Andersen, J. V., Christensen, S. K., Karaca, M., Maechler, P., et al. (2018). Glutamate dehydrogenase is essential to sustain neuronal oxidative energy metabolism during stimulation. *J. Cereb. Blood Flow Metab.* 38, 1754–1768. doi: 10.1177/0271678X17714680

Holland, C. H., Szalai, B., and Saez-Rodriguez, J. (2020). Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochim. Biophys. Acta* 1863:194431. doi: 10.1016/j.bbagrm.2019.194431

Horner, M. A., Pardee, K., Liu, S., King-Jones, K., Lajoie, G., Edwards, A., et al. (2009). The *Drosophila* DHR96 nuclear receptor binds cholesterol and regulates cholesterol homeostasis. *Genes Dev.* 23, 2711–2716. doi: 10.1101/gad.1833609

Hossein-nezhad, A., Spira, A., and Holick, M. F. (2013). Influence of vitamin D status and vitamin D3 supplementation on genome wide expression of white blood cells: a randomized double-blind clinical trial. *PLoS ONE* 8:e58725. doi: 10.1371/journal.pone.0058725

Huang, P., Chandra, V., and Rastinejad, F. (2009). Structural overview of the nuclear receptor superfamily: Insights into physiology and therapeutics. *Annu. Rev. Physiol.* 72, 247–272. doi: 10.1146/annurev-physiol-021909-135917

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 5:e12776. doi: 10.1371/journal.pone.0012776

Janky, R., Verfaillie, A., Imrichová, H., van de Sande, B., Standaert, L., Christiaens, V., et al. (2014). iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput. Biol.* 10:e1003731. doi: 10.1371/journal.pcbi.1003731

Jayaram, N., Usvyat, D., and Martin, C. R. A. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* 17:547. doi: 10.1186/s12859-016-1298-9

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502. doi: 10.1126/science.1141319

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., et al. (2018). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 46, D260–D266. doi: 10.1093/nar/gkx1126

King-Jones, K., Horner, M. A., Lam, G., and Thummel, C. S. (2006). The DHR96 nuclear receptor regulates xenobiotic responses in *Drosophila*. *Cell Metab.* 4, 37–48. doi: 10.1016/j.cmet.2006.06.006

King-Jones, K., and Thummel, C. S. (2005). Nuclear receptors-a perspective from *Drosophila*. *Nat. Rev. Genet.* 6, 311–323. doi: 10.1038/nrg1581

Koegl, M., and Uetz, P. (2007). Improving yeast two-hybrid screening systems. *Brief. Funct. Genomics Proteomics* 6, 302–312. doi: 10.1093/bfgp/elm035

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 46, D252–D259. doi: 10.1093/nar/gkx1106

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., et al. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* 9, 1–10. doi: 10.1038/s41467-018-03751-6

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665. doi: 10.1016/j.cell.2018.01.029

Lee, J., Sharma, S., Kim, J., Ferrante, R. J., and Ryu, H. (2008). Mitochondrial nuclear receptors and transcription factors: Who's minding the cell? *J. Neurosci Res.* 86, 961–971. doi: 10.1002/jnr.21564

Lim, S. C., Tajika, M., Shimura, M., Carey, K. T., Stroud, D. A., Murayama, K., et al. (2018). Loss of the mitochondrial fatty acid β-oxidation protein medium-chain acyl-coenzyme a dehydrogenase disrupts oxidative phosphorylation protein complex stability and function. *Sci. Rep.* 8, 153. doi: 10.1038/s41598-017-18530-4

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015:bav095. doi: 10.1093/database/bav095

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

Martyanov, V., and Gross, R. H. (2010). Identifying functional relationships within sets of co-expressed genes by combining upstream regulatory motif analysis and gene expression information. *BMC Genomics* 11(Suppl. 2):S8. doi: 10.1186/1471-2164-11-S2-S8

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., et al. (2003). TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 374–378. doi: 10.1093/nar/gkg108

McKenna, N. J., and O'Malley, B. W. (2002). Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell* 108, 465–474. doi: 10.1016/S0092-8674(02)00641-4

Mochida, K., Koda, S., Inoue, K., and Nishii, R. (2018). Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. *Front. Plant Sci.* 9:1770. doi: 10.3389/fpls.2018.01770

Murgas, L., Contreras-Riquelme, S., Martínez-Hernandez, J. E., Villaman, C., Santibáñez, R., and Martin, A. J. M. (2021). Automated generation of context-specific gene regulatory networks with a weighted approach in Drosophila melanogaster. *Interface Focus* 11:20200076. doi: 10.1098/rsfs.2020.0076

Ogata, K., Sato, K., and Tahirov, T. (2003). Eukaryotic transcriptional regulatory complexes: cooperativity from near and afar. *Curr Opin Struct Biol.* 13, 40–48. doi: 10.1016/s0959-440x(03)00012-5

O'Neill, L. P., and Turner, B. M. (1996). Immunoprecipitation of chromatin. *Methods Enzymol*. 274:189–197. doi: 10.1016/S0076-6879(96)74017-X

O'Neill, M., Dryden, D. T., and Murray, N. E. (1998). Localization of a protein-DNA interface by random mutagenesls. *EMBO J*. 17, 7118–7127. doi: 10.1093/emboj/17.23.7118

Palanker, L., Necakov, A. S., Sampson, H. M., Ni, R., Hu, C., Thummel, C. S., et al. (2006). Dynamic regulation of *Drosophila* nuclear receptor activity *in vivo*. *Development* 133, 3549–3562. doi: 10.1242/dev.02512

Qin, Q., Fan, J., Zheng, R., Wan, C., Mei, S., Wu, Q., et al. (2020). Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol*. 21, 32. doi: 10.1186/s13059-020-1934-6

Ramagopalan, S. V., Heger, A., Berlanga, A. J., Maugeri, N. J., Lincoln, M. R., Burrell, A., et al. (2010). A ChIP-seq defined genome-wide map of vitamin D receptor binding: Associations with disease and evolution. *Genome Res*. 20, 1352–1360. doi: 10.1101/gr.107920.110

Ramírez-Clavijo, S., and Montoya-Ortíz, G. (2013). "Gene expression and regulation," in *Autoimmunity: From Bench to Bedside, Chapter 1*, eds A. Juan-Manuel, Y. Shoenfeld, A. Rojas-Villarraga, R. A. Levy and R. Cervera (Bogota: El Rosario University Press).

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., et al. (2000). Genome-wide location and function of DNA binding proteins. *Science* 290, 2306–2309. doi: 10.1126/science.290.5500.2306

Ricca, C., Aillon, A., Bergandi, L., Alotto, D., Castagnoli, C., and Silvagno, F. (2018). Vitamin D receptor is necessary for mitochondrial function and cell health. *Int. J. Mol. Sci*. 19, 1672. doi: 10.3390/ijms19061672

Rossi, A., Rigotto, G., Valente, G., Giorgio, V., Basso, E., Filadi, R., et al. (2020). Defective mitochondrial pyruvate flux affects cell bioenergetics in Alzheimer's disease-related models. *Cell Rep*. 30, 2332.e10–2348.e10. doi: 10.1016/j.celrep.2020.01.060

Ryan, Z. C., Craig, T. A., Folmes, C. D., Wang, X., Lanza, I. R., Schaible, N. S., et al. (2016). 1α,25-dihydroxyvitamin D3 regulates mitochondrial oxygen consumption and dynamics in human skeletal muscle cells. *J. Biol. Chem*. 291, 1514–1528. doi: 10.1074/jbc.M115.684399

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., et al. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res*. 47, D212–D220. doi: 10.1093/nar/gky1077

Senyilmaz, D., Virtue, S., Xu, X., Tan, C. Y., Griffin, J. L., Miller, A. K., et al. (2015). Regulation of mitochondrial morphology and function by stearoylation of TFR1. *Nature* 525, 124–128. doi: 10.1038/nature14601

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Shirvani, A., Kalajian, T. A., Song, A., and Holick, M. F. (2019). Disassociation of vitamin D's Calcemic activity and non-calcemic genomic activity and individual responsiveness: a randomized controlled double-blind clinical trial. *Sci. Rep*. 9, 17685. doi: 10.1038/s41598-019-53864-1

Shokri, L., Inukai, S., Hafner, A., Weinand, K., Hens, K., Vedenko, A., et al. (2019). A comprehensive *Drosophila melanogaster* transcription factor interactome. *Cell Rep*. 27, 955.e7–970.e7. doi: 10.1016/j.celrep.2019.03.071

Sieber, M. H., and Thummel, C. S. (2009). The DHR96 nuclear receptor controls triacylglycerol homeostasis in *Drosophila*. *Cell Metab*. 10, 481–490. doi: 10.1016/j.cmet.2009.10.010

Sieber, M. H., and Thummel, C. S. (2012). Coordination of triacylglycerol and cholesterol homeostasis by DHR96 and the *Drosophila* lipa homolog magro. *Cell Metab*. 15, 122–127. doi: 10.1016/j.cmet.2011.11.011

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 47, D607–D613. doi: 10.1093/nar/gky1131

Tang, B. L. (2019). Targeting the mitochondrial pyruvate carrier for neuroprotection. *Brain Sci*. 9, 238. doi: 10.3390/brainsci9090238

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., et al. (2019). FlyBase 2.0: The next generation. *Nucleic Acids Res*. 47, D759–D765. doi: 10.1093/nar/gky1003

Torroja, L., Ortu no-Sahagún, D., Ferrús, A., Hämmerle, B., and Barbas, J. A. (1998). Scully, an essential gene of *Drosophila*, is homologous to mammalian mitochondrial type II L-3-hydroxyacyl-CoA dehydrogenase/amyloid-β peptide- binding protein. *J. Cell Biol*. 141, 1009–1017. doi: 10.1083/jcb.141.4.1009

Weikum, E. R., Liu, X., and Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein Sci*. 27, 1876–1892. doi: 10.1002/pro.3496

Weirauch, M., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Wilk, R., Hu, J., and Krause, H. M. (2013). Spatial profiling of nuclear receptor transcription patterns over the course of *Drosophila* development. *G3* 3, 1177–1189. doi: 10.1534/g3.113.006023

Yoon, W., Hwang, S. H., Lee, S. H., and Chung, J. (2019). *Drosophila* ADCK1 is critical for maintaining mitochondrial structures and functions in the muscle. *PLoS Genet*. 15, e1008184. doi: 10.1371/journal.pgen.1008184

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership