PREDICTION AND EXPLANATION IN BIOMEDICINE USING NETWORK-BASED APPROACHES

EDITED BY: Alessio Martino and Alessandro Giuliani
PUBLISHED IN: Frontiers in Genetics and Frontiers in Artificial Intelligence

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-222-8
DOI 10.3389/978-2-83250-222-8

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

PREDICTION AND EXPLANATION IN BIOMEDICINE USING NETWORK-BASED APPROACHES

Topic Editors:
Alessio Martino, LUISS University, Italy
Alessandro Giuliani, National Institute of Health (ISS), Italy

Citation: Martino, A., Giuliani, A., eds. (2022). Prediction and Explanation in Biomedicine Using Network-Based Approaches. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-222-8

Table of Contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 02 September 2022
DOI 10.3389/fgene.2022.967936

OPEN ACCESS

EDITED AND REVIEWED BY
Simon Charles Heath,
Center for Genomic Regulation, Spain

*CORRESPONDENCE
Alessio Martino,
amartino@luiss.it

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

RECEIVED 13 June 2022
ACCEPTED 02 August 2022
PUBLISHED 02 September 2022

CITATION
Martino A and Giuliani A (2022), Editorial: Prediction and explanation in biomedicine using network-based approaches.
Front. Genet. 13:967936.
doi: 10.3389/fgene.2022.967936

Editorial: Prediction and explanation in biomedicine using network-based approaches

Alessio Martino[1]* and Alessandro Giuliani[2]

[1]Department of Business and Management, LUISS University, Rome, Italy, [2]Department of Environment and Health, Italian National Institute of Health, Rome, Italy

KEYWORDS

network inference, explainable artificial intelligence, pattern recognition, biological network analysis, network modelling

Editorial on the Research Topic
Prediction and explanation in biomedicine using network-based approaches

The complex network paradigm occupies a twilight epistemological status in between data analysis and causal, content based, modelling of complex systems. This status is mirrored by the title of the topic "*Prediction and Explanation in Biomedicine using Network-Based Approaches*" putting together "prediction" (data analysis perspective) and "explanation" (causal modelling perspective). Scientific methodology is aware of the different, albeit related, status of the two perspectives since long time (Shmueli, 2010) and the actual emphasis of machine intelligence community on "explainability" revived the urgency of the issue (Ho et al., 2020).

As aptly stated by Nicosia and others (Nicosia et al., 2014):

"Networks are the fabric of complex systems".

while, at the same time, being a very flexible data analysis tool inheriting from time-honoured multidimensional statistics the focus on correlation matrices (Gorban et al., 2022).

Biological systems are the most evident paradigm of complexity, and this is why it is much more productive to focus on the dynamics of their correlation structure with respect to an in-depth analysis of isolated features. In this Research Topic, this point is made evident by papers exploring correlation structures located at different organization layers: contacts between amino-acid residues of a protein molecule (Uversky and Giuliani), gene expression correlation (Tran et al.) and protein-protein interaction networks (Cesareni et al.; Wang et al.). In particular, Uversky and Giuliani review the most recent results in terms of hierarchical organization of complex biological systems, remarking the benefits of analyzing such systems in a multi-level fashion, hence going beyond the standard causative model where events originate at molecular level and then show up at the 'top' of the hierarchy (e.g., causing a particular disease). Causality is also the key in (Cesareni et al.), where the authors review how causality can help in shaping disease networks, shedding light on using also functional information alongside physical

proximity (i.e., between interacting proteins) for a thoughtful modelling. In Tran et al., the authors question the suitability of MCF-7 cell line for *in vitro* breast cancer research. They use a network-based approach to compare two MCF-7 datasets against a human breast invasive ductal carcinoma dataset taken from The Cancer Genome Atlas (TGCA), showing how they have only minimal similarity in biological processes, hence concluding that using MCF-7 to study breast cancer can hide important gene targets. Finally, TGCA plays an important role also in (Wang et al.), where the authors use a network-based approach to find hub genes related to acute lymphoblastic leukemia.

It is important not to confuse the integration of data analysis and explanatory perspectives with the too-often repeated statement of the substantial irrelevance of the hypothesis-driven approach when in presence of massive amount of data (Mazzocchi, 2015); the situation is exactly the opposite: network paradigm asks for a strict integration between content related and methodological knowledge and the consequent need to overcome research overspecialization. It is not by chance that very interesting new perspectives in statistical mechanics generate from the analysis of biological network systems (Liu et al., 2022): along this line, in this Research Topic, we find papers devoted to theoretical/computational issues (Kuznetsov et al.; Nazarenko et al.) motivated by the solution of relevant biomedical problems. Specifically, Kuznetsov et al. use a variational autoencoder to generate a synthetic 1-cycle ECG which not only looks quite natural, but can also be generated starting from just 25 features automatically learned by the autoencoder. As instead, Nazarenko et al. show an interesting network-based approach based on parenclitic and synolytic networks to describe multidimensional data via a suitable graph that makes the data easier to inspect, visualize and analyze. Tests on synthetic and benchmark data corroborate the competitiveness of using parenclitic and synolytic networks against common machine learning approaches.

In this Research Topic, the application potential to biomedical practice of network-based approaches is explored in (Chen et al.; Jung et al.; Luo et al.; Thomas et al.), that give us the strong impression that network-based approaches are here to stay. In detail, Thomas et al. review how network biology can help in understanding inflammatory bowel disease by discussing different network modelling (notably, protein-protein interaction networks, metabolic networks, gene regulatory and co-expression networks), with some examples also on multi-layered networks. Chen et al. build a predictive model based on network analysis and circular miRNA to address recurrent implantation failure (RIF). Luo et al. also aim at characterizing RIF, but the authors exploit network-based approaches (protein-protein interaction and circRNA–miRNA–mRNA networks) to highlight four hub genes that may be involved in the development of RIF. Finally, Jung et al. briefly review computational models based on machine learning, network modelling and genome-scale metabolic models to characterize drug-resistant cancer cells.

After all, this is not surprising at all, "network biology" is nothing else than "biology as such" given any biological system derives its peculiar behaviour from the interaction of many different element players at different organization layers with no privileged causal layer of explanation (Noble et al., 2019). This is a bare truth (too often overlooked) since the initial definition of "Organism" in classical philosophy (Gotthelf and Lennox, 1987); what is new is the exciting possibility to use these concepts in the day-to-day practice of biomedical sciences by an immediate hands-on approach: we do hope the present Research Topic to transmit this excitement to the reader.

Author contributions

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

Acknowledgments

The guest editors wish to thank all the authors and reviewers for their valuable contributions to this Research Topic and we hope that this collection of articles will be of interest to the medical and genetics community.

Conflict of interest

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

Publisher's note

References

Gorban, A. N., Tyukina, T. A., Pokidysheva, L. I., and Smirnova, E. V. (2022). It is useful to analyze correlation graphs: Reply to comments on "dynamic and thermodynamic models of adaptation". *Phys. Life Rev.* 40, 15–23. doi:10.1016/j.plrev.2021.10.002

Gotthelf, A., and Lennox, J. G. (Editors) (1987). *Philosophical issues in aristotle's biology* (Cambridge: Cambridge University Press). doi:10.1017/CBO9780511552564

Ho, S. Y., Wong, L., and Goh, W. W. B. (2020). Avoid oversimplifications in machine learning: Going beyond the class-prediction accuracy. *Patterns* 1, 100025. doi:10.1016/j.patter.2020.100025

Liu, X., Li, D., Ma, M., Szymanski, B. K., Stanley, H. E., and Gao, J. (2022). Network resilience. *Phys. Rep.Network Resil.* 971, 1–108. doi:10.1016/j.physrep.2022.04.002

Mazzocchi, F. (2015). Could big data be the end of theory in science? *EMBO Rep.* 16, 1250–1255. doi:10.15252/embr.201541001

Nicosia, V., De Domenico, M., and Latora, V. (2014). Characteristic exponents of complex networks. *EPL Europhys. Lett.* 106, 58005. doi:10.1209/0295-5075/106/58005

Noble, R., Tasaki, K., Noble, P. J., and Noble, D. (2019). Biological relativity requires circular causality but not symmetry of causation: So, where, what and when are the boundaries? *Front. Physiol.* 10, 827. doi:10.3389/fphys.2019.00827

Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* 25, 289–310. doi:10.1214/10-STS330

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 February 2021
doi: 10.3389/fgene.2020.627459

Construction of Circular RNA–MicroRNA–Messenger RNA Regulatory Network of Recurrent Implantation Failure to Explore Its Potential Pathogenesis

Jiahuan Luo[1†], *Li Zhu*[2,3†], *Ning Zhou*[1], *Yuanyuan Zhang*[1], *Lirong Zhang*[2,3*] and *Ruopeng Zhang*[2,3*]

[1] Clinical Medical College, Dali University, Dali, China, [2] Department of Reproductive Medicine, The First Affiliated Hospital of Dali University, Dali, China, [3] Institute of Reproductive Medicine, Dali University, Dali, China

OPEN ACCESS

Edited by:
Alessio Martino,
Sapienza University of Rome, Italy

Reviewed by:
Gloria Santoro,
University of Verona, Italy
Yongkang Kim,
University of Colorado Boulder, United States

***Correspondence:**
Ruopeng Zhang
zrp263000@163.com
Lirong Zhang
1398722606@139.com

[†]These authors have contributed equally to this work

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

Received: 09 November 2020
Accepted: 15 December 2020
Published: 16 February 2021

Citation:
Luo J, Zhu L, Zhou N, Zhang Y, Zhang L and Zhang R (2021) Construction of Circular RNA–MicroRNA–Messenger RNA Regulatory Network of Recurrent Implantation Failure to Explore Its Potential Pathogenesis. Front. Genet. 11:627459. doi: 10.3389/fgene.2020.627459

Background: Many studies on circular RNAs (circRNAs) have recently been published. However, the function of circRNAs in recurrent implantation failure (RIF) is unknown and remains to be explored. This study aims to determine the regulatory mechanisms of circRNAs in RIF.

Methods: Microarray data of RIF circRNA (GSE147442), microRNA (miRNA; GSE71332), and messenger RNA (mRNA; GSE103465) were downloaded from the Gene Expression Omnibus (GEO) database to identify differentially expressed circRNA, miRNA, and mRNA. The circRNA–miRNA–mRNA network was constructed by Cytoscape 3.8.0 software, then the protein–protein interaction (PPI) network was constructed by STRING database, and the hub genes were identified by cytoHubba plug-in. The circRNA–miRNA–hub gene regulatory subnetwork was formed to understand the regulatory axis of hub genes in RIF. Finally, the Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the hub genes were performed by clusterProfiler package of Rstudio software, and Reactome Functional Interaction (FI) plug-in was used for reactome analysis to comprehensively analyze the mechanism of hub genes in RIF.

Results: A total of eight upregulated differentially expressed circRNAs (DECs), five downregulated DECs, 56 downregulated differentially expressed miRNAs (DEmiRs), 104 upregulated DEmiRs, 429 upregulated differentially expressed genes (DEGs), and 1,067 downregulated DEGs were identified regarding RIF. The miRNA response elements of 13 DECs were then predicted. Seven overlapping miRNAs were obtained by intersecting the predicted miRNA and DEmiRs. Then, 56 overlapping mRNAs were obtained by intersecting the predicted target mRNAs of seven miRNAs with 1,496 DEGs. The circRNA–miRNA–mRNA network and PPI network were constructed through six circRNAs, seven miRNAs, and 56 mRNAs; and four hub genes (YWHAZ, JAK2, MYH9, and RAP2C) were identified. The circRNA–miRNA–hub gene regulatory subnetwork with nine regulatory axes was formed in RIF. Functional enrichment analysis and reactome

analysis showed that these four hub genes were closely related to the biological functions and pathways of RIF.

Conclusion: The results of this study provide further understanding of the potential pathogenesis from the perspective of circRNA-related competitive endogenous RNA network in RIF.

Keywords: recurrent implantation failure, circRNA, competitive endogenous RNA, GEO, network

INTRODUCTION

Recurrent implantation failure (RIF) refers to infertility in patients younger than 40 years who undergo at least three *in vitro* fertilizations (IVFs) (including fresh embryo transfer and frozen–thawed embryo transfer) or intracytoplasmic sperm injection (ICSI) cycles and implantation of four or more high-quality embryos without embryo implantation or clinical pregnancy (Bashiri et al., 2018). Studies have shown that RIF accounts for about 10% of IVF–embryo transplantation (IVF-ET) (Simur et al., 2009). RIF causes serious mental stress and economic burden to families and even brings a lot of social problems. However, up to now, RIF is still an unsolved problem in assisted reproductive technology. The etiology of RIF has not been elucidated, and there is a lack of effective therapies and no reliable molecular markers to predict the occurrence of RIF. Therefore, elucidating the molecular mechanism of RIF is essential for the development of effective diagnostic and therapeutic targets.

Circular RNAs (circRNAs) are a type of non-coding RNAs that exist in almost all cells of an organism. The 3′ and 5′ ends of circRNA are covalently linked to form a closed circular single-stranded structure, which enables it to resist the hydrolysis of exonucleases and thus has relative stability and conservation (Shao et al., 2017; Shi et al., 2020). In addition, tissue-specific expression and rich diversity of circRNA have made it to be considered the best biomarkers (Chen et al., 2019), some of which have been identified as diagnostic and prognostic biomarkers. Recently, increasing evidence has shown that circRNAs are involved in various cellular processes such as gene expression regulation, cell cycle progression, and chromatin modification (Beermann et al., 2016; Wang Y. et al., 2018; Zang et al., 2020). In summary, the study of circRNAs has become a new hotspot in the field of RNA due to their various functions and specific properties.

Accumulating evidence suggests that circRNAs exert biological processes, including the genesis, translation, and transcriptional regulation of target genes, and extracellular transport, by acting as microRNA (miRNA) sponges, transcriptional activators or inhibitors, and RNA-binding protein (RBP) sponges (Zang et al., 2020). Some circRNAs can even encode polypeptides or proteins to participate in biological regulation (Li et al., 2017; Yang et al., 2017; Han et al., 2018; Xia et al., 2018). Recent studies show that circRNAs exert their functions mainly by adsorbing miRNAs to regulate miRNA expression, thereby regulating the target genes of miRNAs, of which circRNAs are called competing endogenous RNA (ceRNA). In the study of gynecologic tumors, it was found that the expression of circRNA not only promoted cancer but also inhibited cancer. In studies of cervical cancer, hsa_circRNA_101996 was highly expressed in cervical cancer cells. Hsa_circRNA_101996 regulates the proliferation, cell cycle, migration, and invasion of cervical cancer cells mainly through miR-8075 targeting TPX2 (Song et al., 2019), with higher levels of hsa_circRNA_101996 associated with a poor prognosis. Wang H. et al. (2018) found that circRNA-000911 expression was significantly downregulated in breast cancer cells. The high expression of circRNA-000911 could antagonize miR-449a, thereby increasing Notch1 expression to inhibit cell proliferation, migration, and invasion and to promote apoptosis of breast cancer cells. Lu H. et al. (2020) found that CIRS-126 regulated the expression of programmed cell death protein 4 (PDCD4) and inhibited the proliferation of ovarian granulosa cells by acting as a miR-21 sponge in polycystic ovary syndrome. In summary, circRNA–miRNA–messenger RNA (mRNA) regulatory network plays an important role in the occurrence and development of gynecologic diseases, while circRNA has different targets and functions in different tissue cells. Liu et al. (2017) performed microarray sequencing on endometrial biopsies from patients with RIF and found differentially expressed circRNAs (DECs). However, the specific targets and mechanisms of circRNA in RIF have not been reported.

In this study, we explored novel circRNAs and their mechanisms in the endometrium of patients with RIF through bioinformatics analysis. First, RIF-related circRNAs miRNA, and mRNA microarray data were collected from the Gene Expression Omnibus (GEO) database. DECs, differentially expressed miRNAs (DEmiRs), and differentially expressed genes (DEGs) were identified by RStudio software. The circRNA–miRNA–mRNA network was constructed by Cytoscape 3.8.0 software, and then protein–protein interaction (PPI) network was constructed by STRING (Search Tool for the Retrieval of Interacting Genes) (version 11.0) database, and hub genes were identified by cytoHubba plug-in. The circRNA–miRNA–hub gene regulatory subnetwork was formed to understand the regulatory axis of hub genes in RIF. Finally, in order to explore the potential role of hub genes in the development of RIF, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes

Abbreviations: circRNA, circular RNA; miRNA, microRNA; ceRNA, competing endogenous RNA; DECs, differentially expressed circRNAs; DEmiRs, differentially expressed miRNAs; DEGs, differentially expressed genes (mRNAs); IVF-ET, *in vitro* fertilization–embryo transplantation; ICSI, intracytoplasmic sperm injection; FDR, false discovery rate; GEO, Gene Expression Omnibus; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein–protein interaction; RIF, recurrent implantation failure; RBP, RNA-binding protein; PDCD4, programmed cell death protein 4; STRING, Search Tool for the Retrieval of Interacting Gene; MAPK, mitogen-activated protein kinase; cryba2, beta-A2 crystallin; ccdc108, coiled-coil domain-containing protein 108.

(KEGG), and Reactome Functional Interaction (FI) enrichment analyses of hub genes were performed. This flowchart is shown in **Figure 1**.

MATERIALS AND METHODS

Data Extraction

Microarray data of RIF circRNA, miRNA, and mRNA were downloaded from GEO database (http://www.ncbi.nlm.nih.gov/geo/) to identify DECs, DEmiRs, and DEGs. CircRNA expression data were derived from GSE147442 microarray (eight endometrial biopsy tissues from RIF patients and eight endometrial biopsy tissues from healthy controls). And GPL21825 074301 Arraystar Human CircRNA microarray V2 (Agilent Technologies, Inc., Palo Alto, CA) provided annotation information to convert probes into recognizable RNAs. Similarly, GSE71332 microarray and the corresponding GPL18402 Agilent-046064 Unrestricted_Human_miRNA_V19.0_ Microarray (miRNA ID version) were used to extract miRNAs, endometrial biopsy tissues from seven RIF patients and five normal pregnant women. Considering the type of specimen and the availability of data, GSE103465 and the corresponding GPL16043 GeneChip® PrimeView™ Human Gene Expression Array (with External spike-in RNAs) were used for the extraction of mRNA using a total of six samples, including three endometrial biopsies from RIF patients and three from pregnant women. For three microarrays, it can be seen that the difference of general data between the cases and controls is not statistically significant.

Identification of Differentially Expressed Circular RNAs, Differentially Expressed MicroRNAs, and Differentially Expressed Messenger RNAs

Data were extracted and normalized by RStudio software, and then DECs, DEmiRs, and DEGs in the endometrium of RIF patients were obtained by limma package based on the Bioconductor package. The selection criteria for DECs were false discovery rate (FDR) < 0.05, $|\log_2 FC| > 2$, for DEmiRs were FDR < 0.05, $|\log_2 FC| > 0.5$, and for DEGs were FDR < 0.05. $|\log_2 FC| > 1$ was considered to be a statistically significant difference.

Prediction of Circular RNA–MicroRNA Pairs

CircRNAs act as sponges for miRNAs through the miRNA response elements (MREs). The Circular RNA Interactome online tool (https://circinteractome.nia.nih.gov/) was applied to predict target miRNAs of DECs of RIF. Overlapping miRNAs were obtained by intersecting predicted miRNAs and DEmiRs.

Prediction Target Genes of MicroRNAs

The software TargetScan (http://www.targetscan.org/vert_72/), miRDB (http://www.mirdb.org/), and miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/php/search.php) were used to predict the target genes of miRNA, and the intersection part were selected as the predicted mRNAs by Venn diagram in RStudio software. Overlapping mRNAs were obtained by intersecting predicted miRNAs and DEmiRs.

Construction of Circular RNA–MicroRNA–Messenger RNA Network

The above differentially expressed circRNA–miRNA pairs and overlapping mRNAs were used to construct circRNA–miRNA–mRNA network, which were input into the Cytoscape 3.8.0 software program (https://cytoscape.org/) to visualize their circRNA-related ceRNA network.

Construction of Protein–Protein Interaction Network and Identification of Hub Genes

In organisms, although there are multiple genes acting on the same trait, not all expressed genes play an equally important role, and a gene contributes greatly to a certain trait as hub gene. Finding the hub genes acting on RIF would help to understand the molecular mechanisms of this disease. First, a PPI network was built based on DEGs in circRNA–miRNA–mRNA network by STRING (Search Tool for the Retrieval of Interacting Genes) (v11.0) (https://string-db.org/cgi/input.pl) online software and was visualized by Cytoscape 3.8.0 software program. Then the degree, betweenness centrality, and closeness centrality of mRNAs in the PPI network were used to identify RIF-related hub genes by "cytoHubba" plug-in (Chin et al., 2014). We set "hubba nodes" for the top five nodes ranked by degree, closeness, and betweenness. Overlapping, top-ranking genes among the three algorithms were selected as hub genes.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Enrichment Analyses and Reactome Analysis of Hub Genes

GO analysis and KEGG pathway enrichment analyses of hub genes were performed by clusterProfiler package in RStudio software. Reactome pathway analysis was conducted with Reactome FI plug-in to comprehensively analyze the molecular mechanism of hub genes in RIF.

RESULTS

Identification of Differentially Expressed Circular RNAs, Differentially Expressed MicroRNAs, and Differentially Expressed Messenger RNAs

The GSE147442 microarray was extracted and normalized by RStudio software, analyzed by the limma package in RStudio software, and identified 13 DECs, including eight downregulated DECs and five upregulated DECs (**Figures 2A**,**B**, **Supplementary Table 1**). A total of 160 DEmiRs were obtained in the GSE71332 microarray. Of these, 56 were downregulated and 104 upregulated DEmiRs (**Figures 2C,D**, **Supplementary Table 2**). Similarly, we performed the same analysis on GSE103465 microarray and found 1,559 DEGs, including 492 upregulated and 1,067 downregulated DEGs (**Figures 2E,F**, **Supplementary Table 3**).

Prediction of Circular RNA–MicroRNA Pairs

The circRNA–miRNA pairs corresponding to 14 DECs were predicted by Circular RNA Interactome online software.

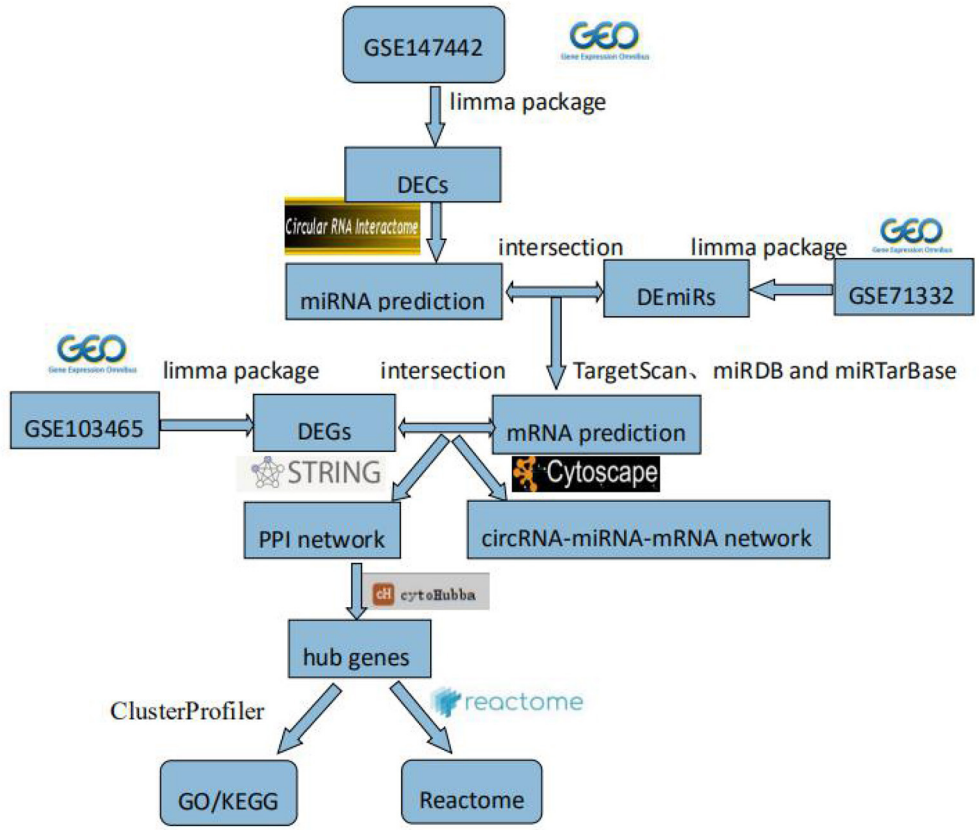


FIGURE 1 | Flowchart. GEO, Gene Expression Omnibus; DECs, differentially expressed circular RNAs; DEmiRs, differentially expressed microRNAs; DEGs, differently expressed genes; PPI, protein–protein interaction; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

The predicted miRNAs and 160 DEmiRs obtained from the microarray were intersected, and finally 11 circRNA–miRNA pairs were identified, including six circRNAs (hsa_circ_0058161, hsa_circ_0033392, hsa_circ_0030162, hsa_circ_0004121, hsa_circ_0034642, and hsa_circ_0034762) and seven miRNAs (hsa-miR-1290, hsa-miR-1305, hsa-miR-375, hsa-miR-370, hsa-miR-887, hsa-miR-1225-5p, and hsa-miR-1825).

Prediction Target Genes of MicroRNAs

The target genes of seven miRNAs were predicted by TargetScan, miRDB, and miRTarBase software; and 562 intersected mRNAs were selected as predicted target genes by Venn diagram in RStudio (**Figure 3A**). The intersection of the predicted 562 miRNA target genes with 1,559 DEGs yielded 56 overlapping mRNAs (**Figure 3B**, **Supplementary Table 4**).

Construction of Circular RNA–MicroRNA–Messenger RNA Network

A circRNA–miRNA–mRNA network was constructed through six DECs, seven DEmiRs, and 56 DEGs and visualized by the Cytoscape 3.8.0 software program (**Figure 4**).

Identification of Four Hub Genes in Protein–Protein Interaction Network by cytoHubba Plug-In

A PPI network based on 56 differentially expressed genes was constructed in circRNA–miRNA–mRNA network to understand the interaction of differentially expressed genes by STRING software, and it was visualized by Cytoscape 3.8.0 software program (**Figure 5A**), which contained 56 nodes and 117 edges. Then the top five genes obtained by the degree, betweenness centrality, and closeness centrality algorithms in cytoHubba plug-in are listed in **Table 1**; and overlapping genes were selected as hub genes YWHAZ, JAK2, MYH9, and RAP2C (**Figure 5B**). Then a circRNA–miRNA–hub gene subnetwork with nine regulatory modules, including hsa_circ_0058161/hsa-miR-1290/YWHAZ regulatory axis, hsa_circ_0058161/hsa-miR-1290/RAP2C regulatory axis, hsa_circ_0030162/hsa-miR-375/JAK2 regulatory axis, hsa_circ_0030162/hsa-miR-375/YWHAZ regulatory axis, hsa_circ_0033392/hsa-miR-375/JAK2 regulatory axis, hsa_circ_0033392/hsa-miR-375/YWHAZ regulatory axis, hsa_circ_0033392/hsa-miR-1305/YWHAZ regulatory axis, hsa_circ_0033392/hsa-miR-1305/MYH9 regulatory axis, and hsa_circ_0033392/hsa-miR-1305/RAP2C regulatory axis, was constructed to

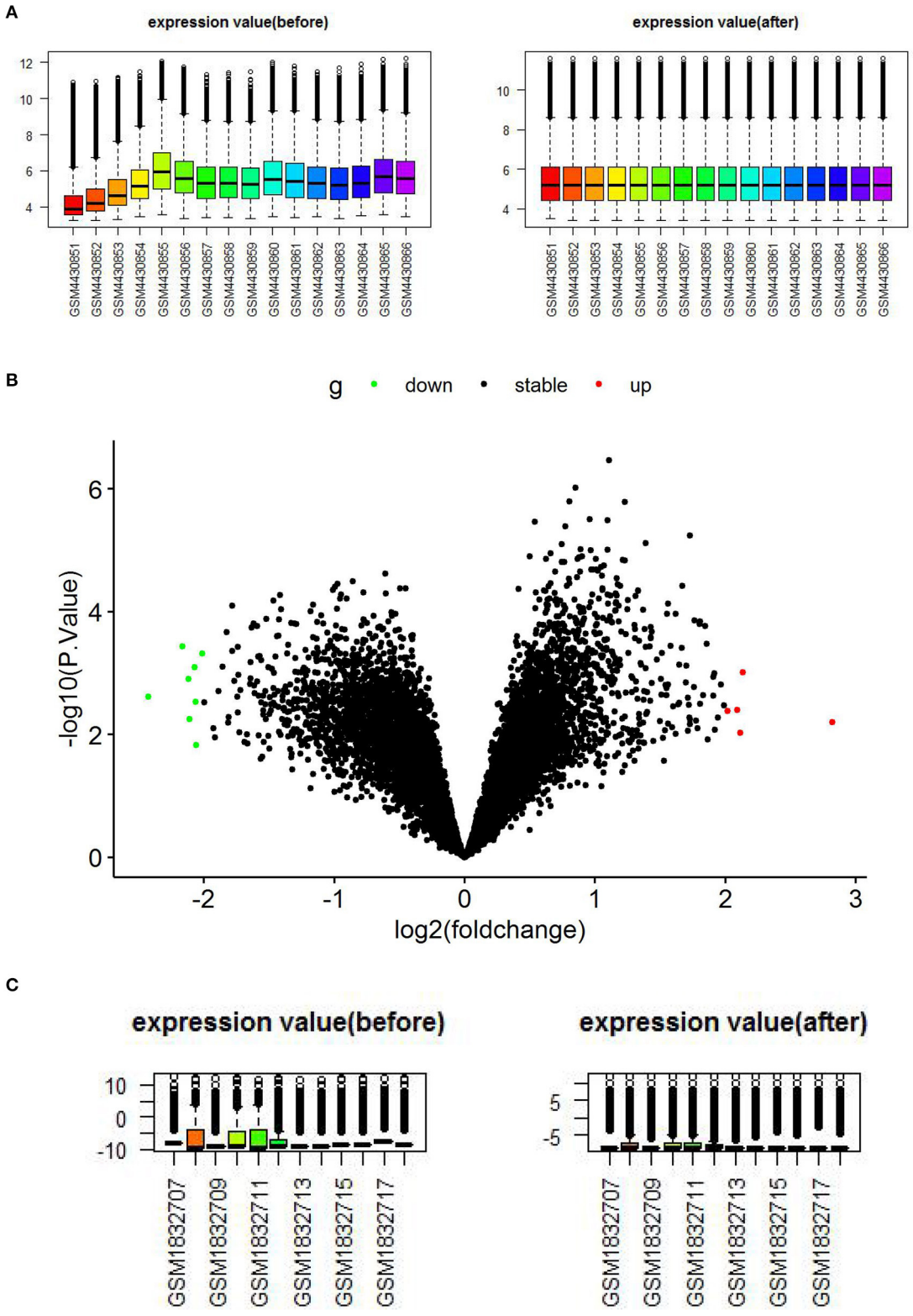


FIGURE 2 | Continued

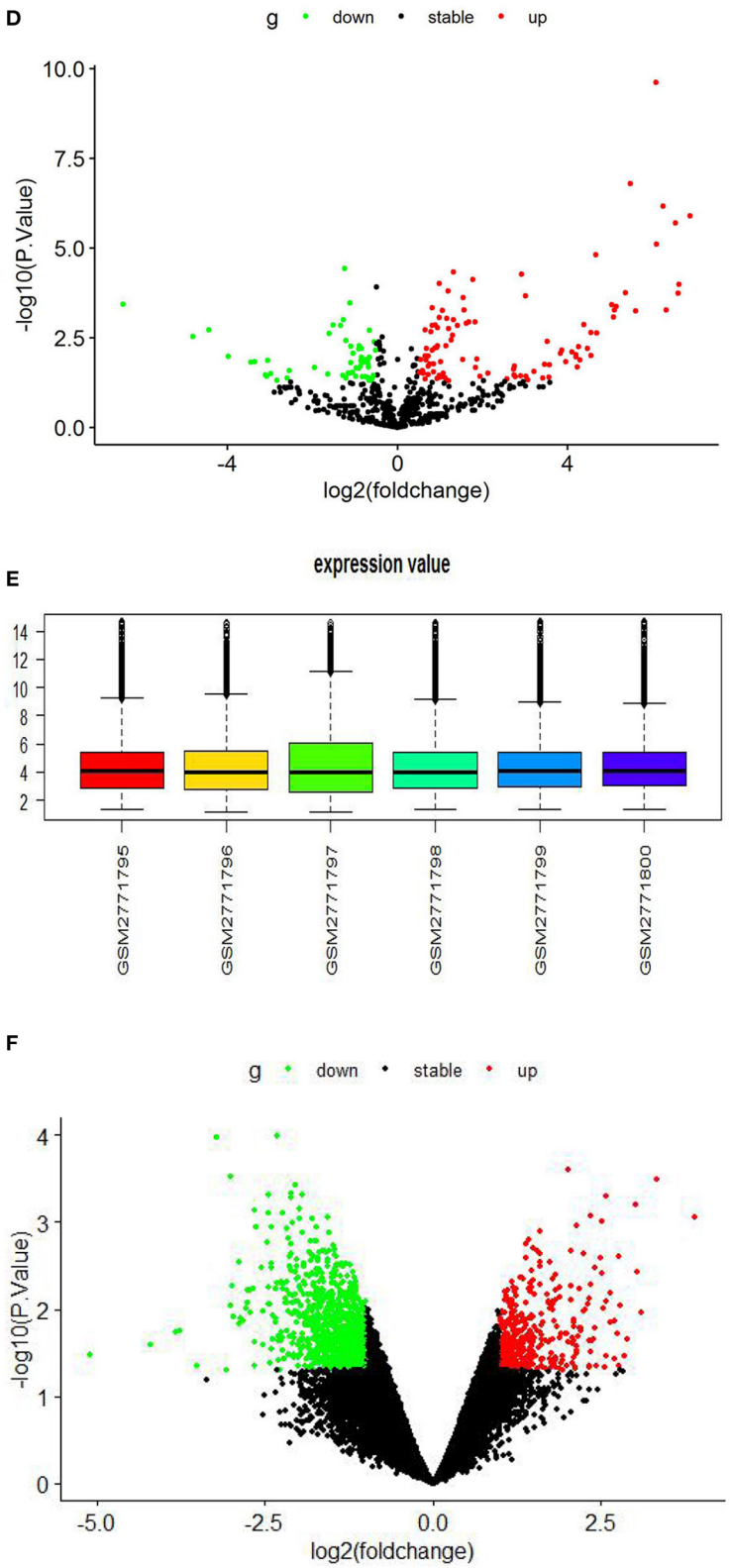


FIGURE 2 | Boxplots and volcano plots for each microarray. **(A)** Boxplot of GSE147442 before and after standardization. **(B)** Volcano plots of DECs based on GSE147442. **(C)** Boxplot of GSE71332 before and after standardization. **(D)** Volcano plots of DEmiRs based on GSE71332. **(E)** Boxplot of GSE103465. **(F)** Volcano plots of DEGs based on GSE103465. DECs, differentially expressed circular RNAs; DEmiRs, differentially expressed microRNAs; DEGs, differently expressed genes.

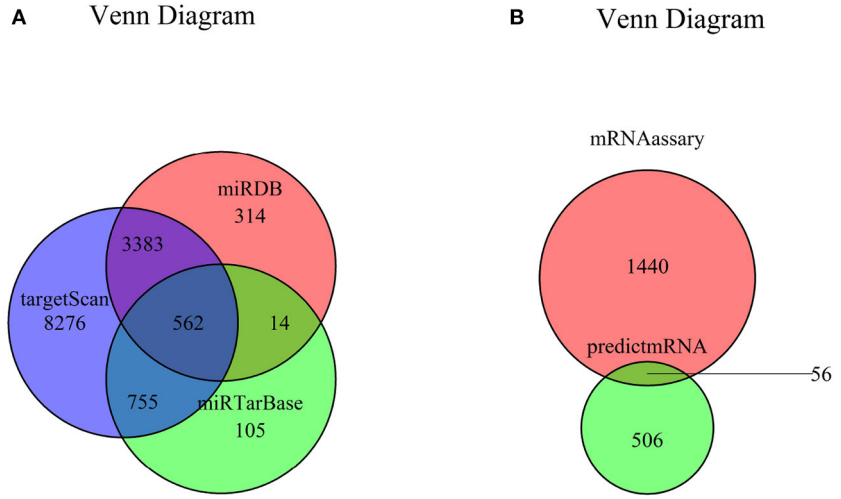


FIGURE 3 | Venn diagram of mRNA. **(A)** Venn diagram of mRNA predicted by TargetScan, miRDB, and miRTarBase. **(B)** Venn diagram of DEGs and predicted mRNAs. mRNA, messenger RNA; DEG, differently expressed genes.

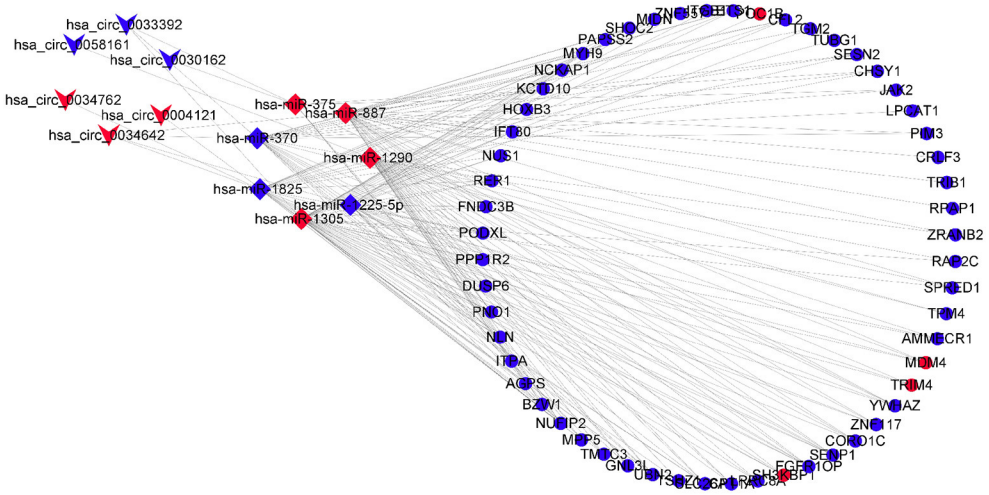


FIGURE 4 | CircRNA–miRNA–mRNA regulatory network, which consists of six DECs, seven DEmiRs, and 56 DEGs. DECs, differentially expressed circular RNAs; DEmiRs, differentially expressed microRNAs; DEGs, differently expressed genes.

depict the relationship between circRNAs, miRNAs, and hub genes (**Figure 5C**).

Gene Ontology Annotation, Kyoto Encyclopedia of Genes and Genomes Pathway, and Reactome Pathway Analyses of Four Hub Genes

Functional annotation of four hub genes was performed by GO analysis. GO terms for biological process (BP), cellular component (CC), and molecular function (MF) are shown in **Figure 6**. The most important GO terms were as follows: "actin filament organization" ($P < 0.01$) in BP, "focal adhesion" ($P < 0.005$) in CC, and "cadherin binding" ($P < 0.024$) in MF. KEGG pathway analysis was also performed to identify the signaling pathways involved in these four hub genes. Two significantly enriched pathways were found ($P < 0.05$) (**Table 2**), in which the "tight junction" was reported to be associated with the progression of RIF (Bellati et al., 2019). In addition, reactome pathway analysis of four hub genes is shown in **Table 3**.

DISCUSSION

CircRNA is a stable non-coding RNA that has long been neglected by transcriptomics due to the lack of a 5′ cap and a 3′ polyadenylated tail. In the past decades, with the development of high-throughput sequencing and bioinformatics analysis, a large number of circRNAs have been unveiled in various tissues

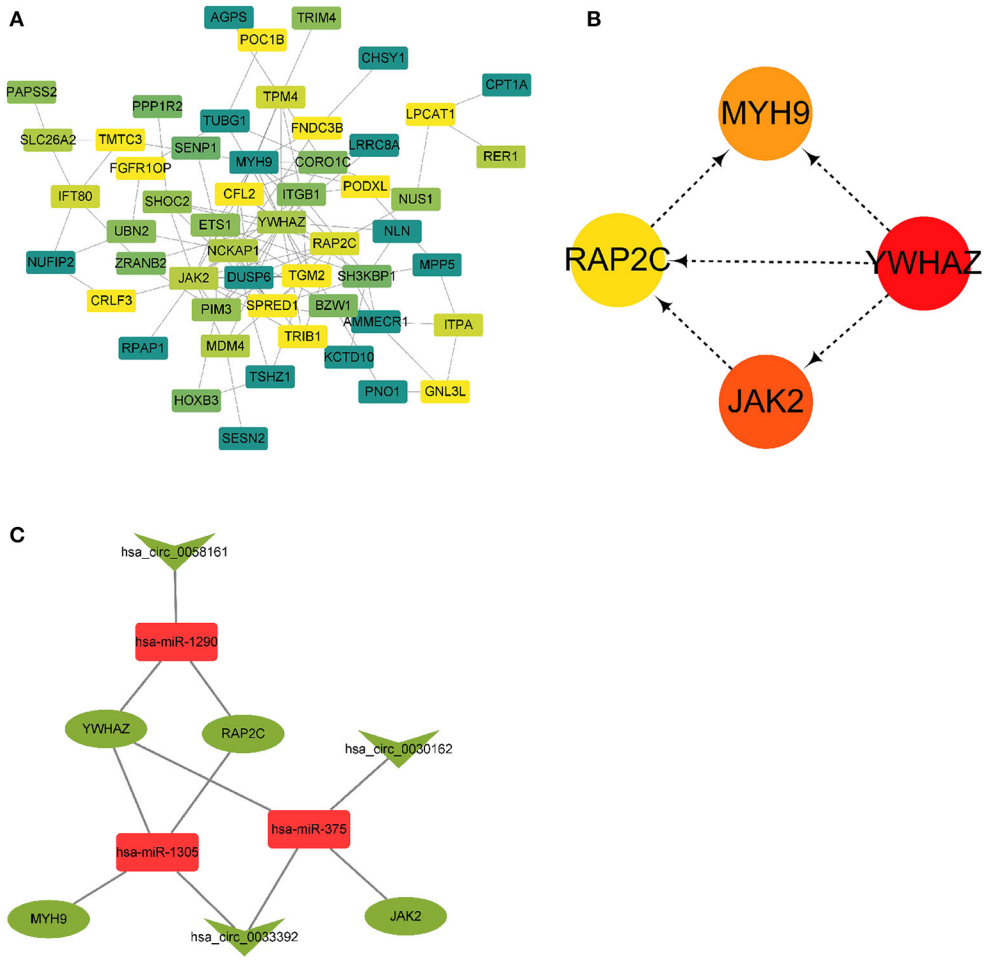


FIGURE 5 | A PPI network and circRNA–miRNA–hub gene regulatory subnetwork. **(A)** A PPI network of the 56 target genes that exert important roles in RIF. This network consists of 53 nodes and 117 edges. **(B)** Four hub genes extracted by cytoHubba plug-in. **(C)** CircRNA–miRNA–hub gene regulatory subnetwork, consisting of three circRNAs, three miRNAs, and four mRNAs. PPI, protein–protein interaction; circRNA, circular RNA; miRNA, microRNA; RIF, recurrent implantation failure.

TABLE 1 | The top five genes obtained by the degree, betweenness centrality, and closeness centrality algorithms in cytoHubba plug-in.

Name	Degree	Name	Betweenness	Name	Closeness
YWHAZ	19	YWHAZ	800.89	YWHAZ	33.42
JAK2	15	MYH9	464.27	JAK2	30.45
MYH9	11	RAP2C	391.39	MYH9	28.50
RAP2C	10	JAK2	373.42	RAP2C	28.08
DUSP6	9	NUS1	296.07	CFL2	26.92

and cells (Chen and Yang, 2015). Accumulating studies have revealed the important role of circRNAs in a variety of human diseases (Hu et al., 2018; Li et al., 2018; Yang et al., 2018). Because circRNAs exhibit specific expression in tissues or developmental stages, the function of circRNAs is still not fully understood (Hu et al., 2018; Li et al., 2018; Yang et al., 2018). Compared with linear RNAs, the higher stability of circRNAs conferred by their circular structure makes these circRNAs potentially valuable as important transcriptional regulators (Meng et al., 2017; Jiang et al., 2018). CircRNAs are commonly used as diagnostic and prognostic biomarkers. However, the exact role of circRNAs in RIF remains largely unknown. To determine whether circRNAs play a role in RIF, we first performed GEO microarray dataset selection and identified 13 DECs.

Current evidence suggests that circRNAs contain multiple MREs that can bind to miRNAs, commonly called "miRNA sponges," which relieve the targeted inhibition of downstream mRNAs by miRNAs (Jin et al., 2019; Lu J. et al., 2020; Qiu et al., 2020), thereby regulating the expression of protein-coding genes. In this study, in order to investigate whether the 13 DECs play a role in RIF as ceRNAs, their related MREs were predicted by Circular RNA Interactome software. The predicted target miRNAs were interacted with DEmiRs from GEO miRNA microarrays, and overlapping miRNAs were taken for further study. Ultimately, 11 circRNA–miRNA pairs were obtained, including six circRNAs (hsa_circ_0058161, hsa_circ_0033392,

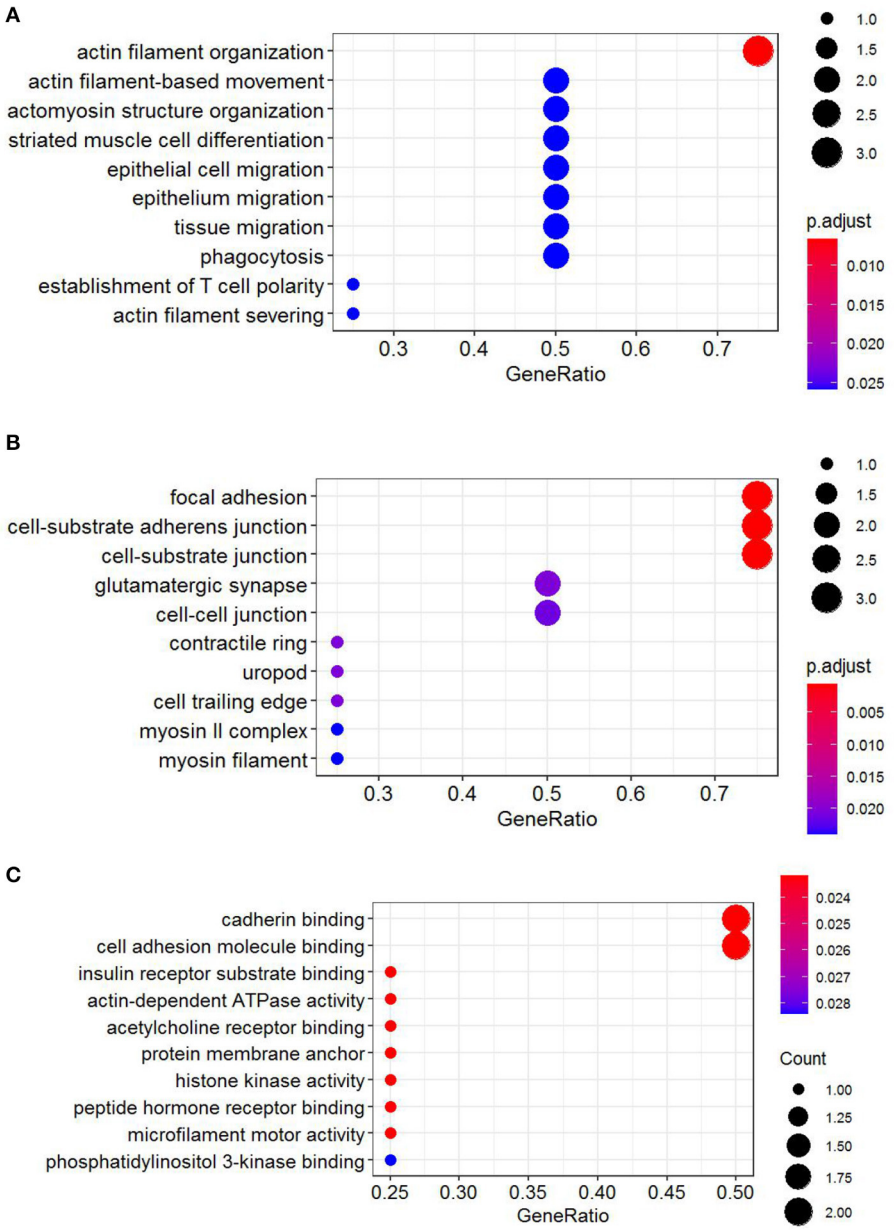


FIGURE 6 | GO functional annotation of four hub genes. **(A)** Biological process (BP). **(B)** Cellular component (CC). **(C)** Molecular function (MF). GO analysis was conducted by R package "clusterProfiler" and visualized by R package "ggplot2." GO, Gene Ontology.

hsa_circ_0030162, hsa_circ_0004121, hsa_circ_0034642, and hsa_circ_0034762) and seven miRNAs (hsa-miR-1290, hsa-miR-1305, hsa-miR-375, hsa-miR-370, hsa-miR-887, hsa-miR-1225-5p, and hsa-miR-1825). After interacting 562 miRNA-related target genes and 1,559 DEGs, 56 overlapping genes were obtained to construct a circRNA-related ceRNA regulatory network. To further identify the key circRNAs involved in the regulatory network, we constructed a PPI network to screen the hub genes. Four hub genes (YWHAZ, JAK2, MYH9, and RAP2C) were identified. Functional annotation and pathway analysis indicated that the four hub genes were involved in multiple cellular functions and signaling pathways in RIF, including "actin filament Organization," "tight junction," and "RHO GTPases activate PKNs."

To investigate the role of circRNA in RIF, a circRNA–miRNA–hub gene regulatory network was constructed based on the circRNA–miRNA–mRNA regulatory network. Hsa_circ_0058161, hsa_circ_0033392, and hsa_circ_0030162 were identified as the key circRNA in this network. GO enrichment analysis showed that the genes in this network were mainly involved in the regulation of actin filament organization, focal adhesion, and cadherin binding. Embryo implantation

TABLE 2 | KEGG pathway analysis of four hub genes.

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
hsa04530	Tight junction	2/4	162/8,063	0.002	0.035	0.025	4,627/57,826	2
hsa05161	Hepatitis B	2/4	162/8,063	0.002	0.035	0.025	3,717/7,534	2

Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was conducted by R package "clusterProfiler".

involves the adhesion of trophoblast cells to the epithelial layer of the endometrium, dependent on cell–cell adhesion molecule interactions (Heneweer et al., 2002). Relevant studies found that the expression of adhesion molecules β-catenin, E-cadherin, and K-cadherin in the endometrium of infertile patients was significantly lower than that of fertile patients, while the expression of β-catenin and E-cadherin was higher at the glandular level than in fertile patients (Koler et al., 2009). However, K-catenin and E-cadherin were lower in glandular levels with recurrent pregnancy loss than fertile patients, suggesting that cadherin is associated with endometrial receptivity and glands (Koler et al., 2009). It is speculated that hub genes affect RIF mainly by acting on the endometrium and related glands at the attachment of embryos through adhesion. KEGG pathway analysis found that hub genes were involved in the development of RIF through the tight junction pathway, which is the part of the interconnection network of adhesion complexes, which generate crosstalk through direct PPIs and interactions affecting their assembly and functional signaling. Karakotchian and Fraser (2007) showed that tight junctions play an important role in the process of embryo implantation, which is consistent with the results of this study. Reactome analysis revealed that MYH9 and YWHAZ could participate in the occurrence of RIF through RHO GTPases activate PKNs. RHO GTPases are important signal transduction molecules involved in a variety of important cell activities, such as actin cytoskeleton remodeling, cell movement, cell adhesion, gene expression, and cell cycle regulation (Bora and Shrivastava, 2017). Heneweer et al. (2002) measured the adhesion of RL95-2 cells of the uterine epithelium to JAR spheres by centrifugal force-based adhesion assay, and they found that the adhesion force depends on RHO GTPases, suggesting that RHO GTPases are most likely to play an important role in the binding of RL95-2 cells to trophoblast in the uterine epithelium. It is speculated that RHO GTPases activate PKNs that mainly affect the adhesion between the endometrial epithelium and gestational trophoblast in this study. These results indirectly suggest that circRNAs in this network may play a key role in the occurrence and development of RIF. This result deserves further study.

YWHAZ, also known as tyrosine 3 monooxygenase/tryptophan 5-monooxygenase activation protein zeta (14-3-3ζ), is a hub gene of many signal transduction pathways and plays a key role in the progression of multiple diseases (Wang et al., 2017; Yang et al., 2019; Gan et al., 2020). More and more studies have shown that YWHAZ is upregulated in breast cancer, ovarian cancer, G2 endometrial adenocarcinoma, prostate cancer, and other types of genitourinary tumors and that it participates in cell growth, cell cycle, apoptosis, migration, and invasion (Jeda et al., 2014; Wang et al., 2017; Yang et al., 2019; Yu et al., 2020). Some studies in placenta and endometrial tissues have considered YWHAZ as a housekeeping gene (Meller et al., 2005; Vestergaard et al., 2011; Sadek et al., 2012a,b; Jeda et al., 2014; Nelissen et al., 2014; Li et al., 2016; Wang et al., 2017; Yang et al., 2019; Yu et al., 2020), and others have found that the expression of YWHAZ is high in the eutopic endometrium of baboons with endometriosis, contributing to the pathophysiology of endometriosis (Joshi et al., 2015). In 12Z cells (immortalized human endometrium), low expression of YWHAZ was also found in ectopic epithelial cell lines, resulting in reducing cell proliferation (Joshi et al., 2015), consistent with the findings of Li et al. (2019), while in our study, YWHAZ expression was found to be downregulated in endometrial tissues of RIF patients, possibly associated with reduced cell proliferation. However, the current research of YWHAZ in RIF is insufficient, so more studies are need for confirmation.

TABLE 3 | Reactome pathway analysis of four hub genes.

Reactome pathway	*P*-value	FDR	HitGenes
RHO GTPases activate PKNs	1.02E−04	4.53E−03	MYH9, YWHAZ
Interleukin-3, Interleukin-5 and GM-CSF signaling	1.13E−04	4.53E−03	JAK2, YWHAZ
Translocation of SLC2A4 (GLUT4) to the plasma membrane	2.03E−04	5.29E−03	MYH9, YWHAZ
Erythropoietin activates STAT5	3.21E−03	0.0155	JAK2
Erythropoietin activates Phospholipase C gamma (PLCG)	3.21E−03	0.0155	JAK2
MAPK1 (ERK2) activation	3.66E−03	0.0155	JAK2
RHO GTPase Effectors	3.92E−03	0.0155	MYH9, YWHAZ
MAPK3 (ERK1) activation	4.12E−03	0.0155	JAK2
Regulation of localization of FOXO transcription factors	4.12E−03	0.0155	YWHAZ
Interleukin-23 signaling	4.12E−03	0.0155	JAK2

Reactome pathway analysis was conducted by Reactome FI plug-in.
FDR, false discovery rate.

JAK2 links to the intracellular domain of many cytokine receptors for signal transduction. When cytokines bind to JAK2 receptors, the phosphorylation of JAK2 leads to the phosphorylation of other intracellular molecules, mainly through the JAK2–STATA3 pathway, which ultimately leads to gene transcription (Roskoski, 2016; Choy, 2019). It plays an important role in cytokine signal transduction and regulation of cell growth and gene expression. JAK2 inhibitors cooperate with

SMO inhibitors to inhibit the growth and metastasis of breast cancer cells (Doheny et al., 2020). Ito et al. (2004) detected the expression of JAK2 in mouse embryos to understand the role of JAK2 in the regulation of early preimplantation development by reverse transcription–polymerase chain reaction analysis and immunocytochemistry and found that JAK2 was mainly localized in single-cell embryos. In the unfertilized oocytes and M-stage single-cell embryos, JAK2 localized on chromosomes. Xu et al. (2017) showed that JAK2-mediated sodium/hydrogen exchange activation regulated acute cell volume changes in the late single-cell stage of mouse preimplantation embryos. Dysregulation of cell volume in early preimplantation embryos may lead to embryonic development arrest. In this study, JAK2 expression reduced, presumably reducing sodium/hydrogen exchange activation leading to dysregulation of cell volume, which affected embryonic development and embryonic adhesion.

The non-myosin heavy chain nine gene (MYH9) is located on chromosome 22q12.3 and encodes a cytoskeletal contractile protein, non-smooth muscle myosin heavy chain IIA (Pecci et al., 2018). Kadam et al. (2006) found that MYH9 protein on gametes interacts with the non-glycosylated N-terminal conserved region of tubal glycoprotein, and one tubal glycoprotein can bind to two gametes, which is associated with capacitated sperm, oocytes, and developing embryos. Lamy et al. (2018) performed proteomic identification in fallopian tube fluid after ovulation and found that MHY9 could regulate sperm function. However, there are few studies on the expression and mechanism of MYH9 in the endometrium of patients with RIF, and more researches are needed to verify it.

RAP2C is a member of the Rap family of small GTP-binding proteins, and a study showed that RAP2C is mainly expressed in the liver, skeletal muscle, prostate, uterus, rectum, stomach, and bladder. The protein is located in the cytoplasm and is involved in regulating cell growth, differentiation, and apoptosis (Guo et al., 2007). RAP2C has been found to be an important molecular switch in the mitogen-activated protein kinase (MAPK) signaling pathway in breast cancer; RAP2C reduces apoptosis and promotes proliferation and migration through the MAPK signaling pathway (Zhu et al., 2020). Zhang et al. (2017) conducted a genome-wide associated study of 43,568 women of European descent and found that variations in the RAP2C locus were associated with duration of pregnancy; and the established roles of these genes in uterine development, maternal nutrition, and vascular control supported their mechanism involvement. Although RAP2C expresses in the uterus, the effect of changes in RAP2C expression on endometrial receptivity and RIF needs further study.

Nine circRNA–miRNA–hub gene regulatory modules, including hsa_circ_0058161/hsa-miR-1290/YWHAZ regulatory axis, hsa_circ_0058161/hsa-miR-1290/RAP2C regulatory axis, hsa_circ_0030162/hsa-miR-375/JAK2 regulatory axis, hsa_circ_0030162/hsa-miR-375/YWHAZ regulatory axis, hsa_circ_0033392/hsa-miR-375/JAK2 regulatory axis, hsa_circ_0033392/hsa-miR-375/YWHAZ regulatory axis, hsa_circ_0033392/hsa-miR-1305/YWHAZ regulatory axis, hsa_circ_0033392/hsa-miR-1305/MYH9 regulatory axis, and hsa_circ_0033392/hsa-miR-1305/RAP2C regulatory axis, were obtained from the final circRNA-related subnetwork. Overall, for four genes, hsa_circ_0033392 and hsa_circ_0030162 had a competitive regulatory relationship. However, so far, there is no research about hsa_circ_0058161, hsa_circ_0033392, and hsa_circ_0030162 on diseases published.

Hsa-miR-1290 overexpression was found in breast cancer (Hamam et al., 2016), glioblastoma (Khalighfard et al., 2020), and fatty liver disease (Tan et al., 2014). Consistent with this study, hsa-miR-1290 is a risk factor for RIF. As the downstream target genes of hsa-miR-1290 in this study, YWHAZ and RAP2C are associated with endometrial cell proliferation. It is speculated that hsa-miR-1290 induces endometrial cell proliferation inhibition and endometrial receptivity impairment leading to RIF by decreasing YWHAZ expression. However, hsa_circ_0058161/hsa-miR-1290/ YWHAZ axis has not been reported in the occurrence and development of RIF. The mechanism of RAP2C in RIF is not clear, so the mechanism of hsa_circ_0058161/hsa-miR-1290/RAP2C axis in RIF cannot be speculated.

Hsa-miR-375 gene is located in the intergenic region between beta-A2 crystallin (cryba2) and coiled-coil domain-containing protein 108 (ccdc108) genes in human chromosome 2q35 region, and the sequence of hsa-miR-375 is highly conserved (Baroukh and van Obberghen, 2009). Further studies have shown that hsa-miR-375 is a multifunctional miRNA involved in islet development, glucose homeostasis, mucosal immunity, pulmonary surfactant secretion, and tumorigenesis (Shao et al., 2014; Yan et al., 2014). In this study, we found that hsa-miR-375 is upregulated in RIF. However, there are few reports on the function of hsa-miR-375 in RIF or its interaction with upstream circRNA. Therefore, more research is necessary.

It has been found in cervical cancer that hsa-miR-1305 regulates the Wnt/β-catenin pathway by binding to Wnt2 to promote cell proliferation, migration, and invasion (Liu et al., 2020). Testing of peripheral blood samples of monozygotic discordant twins for epithelial ovarian carcinoma found that the expression of hsa-miR-1305 was upregulated and that hsa-miR-1305 regulates cell cycle and cell apoptosis (Tuncer et al., 2020). The expression of hsa-miR-1305 is rapidly upregulated after the initiation of pluripotent stem cell differentiation (within 24 h), indicating that it plays a role in early differentiation (Jin et al., 2016). Furthermore, the downregulation of hsa-miR-1035 contributes to the consolidation of the pluripotent phenotype, and its overexpression leads to the initiation of differentiation, thus suggesting that hsa-miR-1305 acts as a regulator to maintain a fine balance between pluripotency and differentiation. Overexpression of hsa-miR-1305 increases cell apoptosis, while its knockdown reduces the number of apoptotic cells (Jin et al., 2016). In this study, miR-1305 was found to be upregulated in RIF. However, there is no report linking miR1305 to RIF or its association with upstream circRNA.

At present, there are a few studies on the mechanism of circRNA in RIF. The novelty of this study is that the circRNA–miRNA–mRNA network was constructed for the first time through the GEO database. However, given that these results are only based on bioinformatics models, further in-depth research is crucial to verify the possible role of these nine axes in RIF.

CONCLUSION

DECs, DEmiRs, and DEGs were identified from publicly available microarray data to construct circRNA-related ceRNA networks. The circRNA–miRNA–hub gene regulatory subnetwork reveals that three important circRNAs and four hub genes may be involved in the development of RIF, provides new insights into the pathogenesis of RIF, and proposes potential therapeutic targets worthy of further study.

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

JL and LZ: conceptualization. NZ and YZ: investigation and validation. LZ and RZ: methodology. JL: software. LZ: supervision. JL and LZ: writing—original draft preparation. LZ and RZ: writing—review and editing. All authors read and approved the final manuscript.

FUNDING

This manuscript was supported by the National Natural Science Foundation of China (81860271), Joint Special Project on Basic Research of Local Undergraduate Universities in Yunnan Province (2017FH001-078), Yunnan Health Training Project of High Level Talents (D-2017020), The Eighth Batch Young and Middle-Aged Academic Target Project Leaders of Dali University (LDYF201702), Reproductive Medicine Innovation Team of Dali University (ZKLX2019320), and Education Department Project of Yunnan Province (2020Y0565).

ACKNOWLEDGMENTS

We would like to thank the academic editor and reviewers for their important contributions that improved the quality of this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.627459/full#supplementary-material

Supplementary Table 1 | 13 DECs were obtained GSE147442 microarray by the limma package in RStudio software. DECs differentially expressed circRNAs.

Supplementary Table 2 | 160 DEmiRs were obtained GSE71332 microarray by the limma package in RStudio software. DEmiRs differentially expressed miRNAs.

Supplementary Table 3 | 1559 DEGs were obtained GSE103465 microarray by the limma package in RStudio software. DEGs differently expressed genes.

Supplementary Table 4 | 56 overlapping mRNAs were obtained by Venn diagram in RStudio software.

REFERENCES

Baroukh, N. N., and van Obberghen, E. (2009). Function of microRNA-375 and microRNA-124a in pancreas and brain. *FEBS J.* 276, 6509–6521. doi: 10.1111/j.1742-4658.2009.07353.x

Bashiri, A., Halper, K. I., and Orvieto, R. (2018). Recurrent implantation failure-update overview on etiology, diagnosis, treatment and future directions. *Reprod. Biol. Endocrinol.* 16:121. doi: 10.1186/s12958-018-0414-2

Beermann, J., Piccoli, M. T., Viereck, J., and Thum, T. (2016). Non-coding RNAs in development and disease: background, mechanisms, therapeutic approaches. *Physiol. Rev.* 96, 1297–1325. doi: 10.1152/physrev.00041.2015

Bellati, F., Costanzi, F., De Marco, M. P., Cippitelli, C., Stoppacciaro, A., De Angelis, C., et al. (2019). Low endometrial beta-catenin and cadherins expression patterns are predictive for primary infertility and recurrent pregnancy loss. *Gynecol. Endocrinol.* 35, 727–731. doi: 10.1080/09513590.2019.1579790

Bora, I., and Shrivastava, N. (2017). ABCs of RhoGTPases indicating potential role as oncotargets. *J. Cancer Res. Ther.* 13, 2–8. doi: 10.4103/0973-1482.204878

Chen, B. J., Huang, S., and Janitz, M. (2019). Changes in circular RNA expression patterns during human foetal brain development. *Genomics* 111, 753–758. doi: 10.1016/j.ygeno.2018.04.015

Chen, L. L., and Yang, L. (2015). Regulation of circRNA biogenesis. *RNA Biol.* 12, 381–388. doi: 10.1080/15476286.2015.1020271

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

Choy, E. H. (2019). Clinical significance of Janus Kinase inhibitor selectivity. *Rheumatology* 58, 953–962. doi: 10.1093/rheumatology/key339

Doheny, D., Sirkisoon, S., Carpenter, R. L., Aguayo, N. R., Regua, A. T., Anguelov, M., et al. (2020). Combined inhibition of JAK2-STAT3 and SMO-GLI1/tGLI1 pathways suppresses breast cancer stem cells, tumor growth, and metastasis. *Oncogene* 39, 6589–6605. doi: 10.1038/s41388-020-01454-1

Gan, Y., Ye, F., and He, X. X. (2020). The role of YWHAZ in cancer: a maze of opportunities and challenges. *Cancer J.* 11, 2252–2264. doi: 10.7150/jca.41316

Guo, Z., Yuan, J., Tang, W., Chen, X., Gu, X., Luo, K., et al. (2007). Cloning and characterization of the human gene RAP2C, a novel member of Ras family, which activates transcriptional activities of SRE. *Mol. Biol. Rep.* 34, 137–144. doi: 10.1007/s11033-006-9023-9

Hamam, R., Ali, A. M., Alsaleh, K. A., Kassem, M., Alfayez, M., Aldahmash, A., et al. (2016). microRNA expression profiling on individual breast cancer patients identifies novel panel of circulating microRNA for early detection. *Sci. Rep.* 6:25997. doi: 10.1038/srep25997

Han, B., Chao, J., and Yao, H. (2018). Circular RNA and its mechanisms in disease: from the bench to the clinic. *Pharmacol. Ther.* 187, 31–44. doi: 10.1016/j.pharmthera.2018.01.010

Heneweer, C., Kruse, L. H., Kindhauser, F., Schmidt, M., Jakobs, K. H., Denker, H. W., et al. (2002). Adhesiveness of human uterine epithelial RL95-2 cells to trophoblast: rho protein regulation. *Mol. Hum. Reprod.* 8, 1014–1022. doi: 10.1093/molehr/8.11.1014

Hu, W., Bi, Z. Y., Chen, Z. L., Liu, C., Li, L. L., Zhang, F., et al. (2018). *Emerging landscape of circular RNAs in lung cancer. Cancer Lett.* 427, 18–27. doi: 10.1016/j.canlet.2018.04.006

Ito, M., Nakasato, M., Suzuki, T., Sakai, S., Nagata, M., and Aoki, F. (2004). Localization of janus kinase 2 to the nuclei of mature oocytes and early cleavage stage mouse embryos. *Biol. Reprod.* 71, 89–96. doi: 10.1095/biolreprod.103.023226

Jeda, A., Witek, A., Janikowska, G., Cwynar, G., Janikowski, T., Cialon, M., et al. (2014). [Expression profile of genes associated with the histaminergic system estimated by oligonucleotide microarray analysis HG-U133A in women with endometrial adenocarcinoma]. *Ginekol. Pol.* 85, 172–179. doi: 10.17772/gp/1709

Jiang, X. M., Li, Z. L., Li, J. L., Xu, Y., Leng, K. M., Cui, Y. F., et al. (2018). A novel prognostic biomarker for cholangiocarcinoma: circRNA Cdr1as. *Eur. Rev. Med. Pharmacol. Sci.* 22, 365–371. doi: 10.26355/eurrev_201801_14182

Jin, C., Shi, L., Li, Z., Liu, W., Zhao, B., Qiu, Y., et al. (2019). Circ_0039569 promotes renal cell carcinoma growth and metastasis by regulating miR-34a-5p/CCL22. *Am. J. Transl. Res.* 11, 4935–4945.

Jin, S., Collin, J., Zhu, L., Montaner, D., Armstrong, L., Neganova, I., et al. (2016). A novel role for miR-1305 in regulation of pluripotency-differentiation balance, cell cycle, and apoptosis in human pluripotent stem cells. *Stem Cells.*. 34, 2306–2317. doi: 10.1002/stem.2444

Joshi, N. R., Su, R. W., Chandramouli, G. V., Khoo, S. K., Jeong, J. W., Young, S. L., et al. (2015). Altered expression of microRNA-451 in eutopic endometrium of baboons (*Papio anubis*) with endometriosis. *Hum. Reprod.* 30, 2881–2891. doi: 10.1093/humrep/dev229

Kadam, K. M., D'Souza, S. J., Bandivdekar, A. H., and Natraj, U. (2006). Identification and characterization of oviductal glycoprotein-binding protein partner on gametes: epitopic similarity to non-muscle myosin IIA. MYH 9. *Mol. Hum. Reprod.* 12, 275–282. doi: 10.1093/molehr/gal028

Karakotchian, M., and Fraser, S. I. (2007). An ultrastructural study of microvascular inter-endothelial tight junctions in normal endometrium. *Micron* 38, 632–636. doi: 10.1016/j.micron.2006.09.010

Khalighfard, S., Kalhori, M. R., Haddad, P., Khori, V., and Alizadeh, M. A. (2020). Enhancement of resistance to chemo-radiation by hsa-miR-1290 expression in glioblastoma cells. *Eur. J. Pharmacol.* 880:173144. doi: 10.1016/j.ejphar.2020.173144

Koler, M., Achache, H., Tsafrir, A., Smith, Y., Revel, A., and Reich, R. (2009). Disrupted gene pattern in patients with repeated *in vitro* fertilization (IVF) failure. *Hum. Reprod.* 24, 2541–2548. doi: 10.1093/humrep/dep193

Lamy, J., Nogues, P., Combes-Soia, L., Tsikis, G., Labas, V., Mermillod, P., et al. (2018). Identification by proteomics of oviductal sperm-interacting proteins. *Reproduction* 155, 457–466. doi: 10.1530/REP-17-0712

Li, M., Zhou, Y., and Taylor, S. H. (2019). miR-451a inhibition reduces established endometriosis lesions in Mice. *Reprod. Sci.* 26, 1506–1511. doi: 10.1177/1933719119862050

Li, P., Yang, X., Yuan, W., Yang, C., Zhang, X., Han, J., et al. (2018). CircRNA-Cdr1as exerts anti-oncogenic functions in bladder cancer by sponging MicroRNA-135a. *Cell. Physiol. Biochem.* 46, 1606–1616. doi: 10.1159/000489208

Li, Y., Lu, H., Ji, Y., Wu, S., and Yang, Y. (2016). Identification of genes for normalization of real-time RT-PCR data in placental tissues from intrahepatic cholestasis of pregnancy. *Placenta* 48, 133–135. doi: 10.1016/j.placenta.2016.10.017

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., et al. (2017). Corrigendum: exon-intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* 24:194. doi: 10.1038/nsmb0217-194a

Liu, L., Li, L., Ma, X., Yue, F., Wang, Y., Wang, L., et al. (2017). Altered circular RNA expression in patients with repeated implantation failure. *Cell. Physiol. Biochem.* 44, 303–313. doi: 10.1159/000484887

Liu, W., Zhuang, R., Feng, S., Bai, X., Jia, Z., Kapora, E., et al. (2020). Long non-coding RNA ASB16-AS1 enhances cell proliferation, migration and invasion via functioning as a ceRNA through miR-1305/Wnt/beta-catenin axis in cervical cancer. *Biomed. Pharmacother.* 125:109965. doi: 10.1016/j.biopha.2020.109965

Lu, H. C., Yao, J. Q., Yang, X., Han, J., Wang, J. Z., Xu, K., et al. (2020). Identification of a potentially functional circRNA-miRNA-mRNA regulatory network for investigating pathogenesis and providing possible biomarkers of bladder cancer. *Cancer Cell Int.* 20, 31. doi: 10.1186/s12935-020-1108-3

Lu, J., Xue, Y., Wang, Y., Ding, Y., Zou, Q., Pan, M., et al. (2020). CiRS-126 inhibits proliferation of ovarian granulosa cells through targeting the miR-21-PDCD4-ROS axis in a polycystic ovarian syndrome model. *Cell Tissue Res.* 381, 189–201. doi: 10.1007/s00441-020-03187-9

Meller, M., Vadachkoria, S., Luthy, D. A., and Williams, A. M. (2005). Evaluation of housekeeping genes in placental comparative expression studies. *Placenta* 26, 601–607. doi: 10.1016/j.placenta.2004.09.009

Meng, S., Zhou, H., Feng, Z., Xu, Z., Tang, Y., Li, P., et al. (2017). CircRNA: functions and properties of a novel potential biomarker for cancer. *Mol. Cancer* 16:94. doi: 10.1186/s12943-017-0663-2

Nelissen, E. C., Dumoulin, J. C., Busato, F., Ponger, L., Eijssen, L. M., Evers, J. L., et al. (2014). Altered gene expression in human placentas after IVF/ICSI. *Hum. Reprod.* 29, 2821–2831. doi: 10.1093/humrep/deu241

Pecci, A., Ma, X., Savoia, A., and Adelstein, S. R. (2018). MYH9: Structure, functions and role of non-muscle myosin IIA in human disease. *Gene* 664, 152–167. doi: 10.1016/j.gene.2018.04.048

Qiu, Y., Pu, C., Li, Y., and Qi, B. (2020). Construction of a circRNA-miRNA-mRNA network based on competitive endogenous RNA reveals the function of circRNAs in osteosarcoma. *Cancer Cell Int.* 20:48. doi: 10.1186/s12935-020-1134-1

Roskoski, R. J. (2016). Janus kinase (JAK) inhibitors in the treatment of inflammatory and neoplastic diseases. *Pharmacol. Res.* 111, 784–803. doi: 10.1016/j.phrs.2016.07.038

Sadek, K. H., Cagampang, F. R., Bruce, K. D., Macklon, N., and Cheong, Y. (2012a). Variation in stability of housekeeping genes in healthy and adhesion-related mesothelium. *Fertil. Steril.* 98, 1023–1027. doi: 10.1016/j.fertnstert.2012.06.033

Sadek, K. H., Cagampang, F. R., Bruce, K. D., Shreeve, N., Macklon, N., and Cheong, Y. (2012b). Variation in stability of housekeeping genes in endometrium of healthy and polycystic ovarian syndrome women. *Hum. Reprod.* 27, 251–256. doi: 10.1093/humrep/der363

Shao, Y., Geng, Y., Gu, W., Huang, J., Ning, Z., and Pei, H. (2014). Prognostic significance of microRNA-375 downregulation in solid tumors: a meta-analysis. *Dis. Markers* 2014:626185. doi: 10.1155/2014/626185

Shao, Y., Li, J., Lu, R., Li, T., Yang, Y., Xiao, B., et al. (2017). Global circular RNA expression profile of human gastric cancer and its clinical significance. *Cancer Med.* 6, 1173–1180. doi: 10.1002/cam4.1055

Shi, Y., Jia, X., and Xu, J. (2020). The new function of circRNA: translation. *Clin. Transl. Oncol.* 22, 2162–2169. doi: 10.1007/s12094-020-02371-1

Simur, A., Ozdemir, S., Acar, H., Colakoglu, M. C., Gorkemli, H., Balci, O., et al. (2009). Repeated *in vitro* fertilization failure and its relation with thrombophilia. *Gynecol. Obstet. Invest.* 67, 109–112. doi: 10.1159/000165776

Song, T., Xu, A., Zhang, Z., Gao, F., Zhao, L., Chen, X., et al. (2019). CircRNA hsa_circRNA_101996 increases cervical cancer proliferation and invasion through activating TPX2 expression by restraining miR-8075. *J. Cell. Physiol.* 234, 14296–14305. doi: 10.1002/jcp.28128

Tan, Y., Ge, G., Pan, T., Wen, D., and Gan, J. (2014). A pilot study of serum microRNAs panel as potential biomarkers for diagnosis of nonalcoholic fatty liver disease. *PLoS ONE* 9:e105192. doi: 10.1371/journal.pone.0105192

Tuncer, S. B., Erdogan, O. S., Erciyas, S. K., Saral, M. A., Celik, B., Odemis, D. A., et al. (2020). miRNA expression profile changes in the peripheral blood of monozygotic discordant twins for epithelial ovarian carcinoma: potential new biomarkers for early diagnosis and prognosis of ovarian carcinoma. *J. Ovarian Res.* 13:99. doi: 10.1186/s13048-020-00706-8

Vestergaard, A. L., Knudsen, U. B., Munk, T., Rosbach, H., and Martensen, M. P. (2011). Transcriptional expression of type-I interferon response genes and stability of housekeeping genes in the human endometrium and endometriosis. *Mol. Hum. Reprod.* 17, 243–254. doi: 10.1093/molehr/gaq100

Wang, H., Xiao, Y., Wu, L., and Ma, D. (2018). Comprehensive circular RNA profiling reveals the regulatory role of the circRNA-000911/miR-449a pathway in breast carcinogenesis. *Int. J. Oncol.* 52, 743–754. doi: 10.3892/ijo.2018.4265

Wang, W., Zhang, L., Wang, Y., Ding, Y., Chen, T., Wang, Y., et al. (2017). Involvement of miR-451 in resistance to paclitaxel by regulating YWHAZ in breast cancer. *Cell Death Dis.* 8:e3071. doi: 10.1038/cddis.2017.460

Wang, Y., Lu, T., Wang, Q., Liu, J., and Jiao, W. (2018). Circular RNAs: crucial regulators in the human body (review). *Oncol. Rep.* 40, 3119–3135. doi: 10.3892/or.2018.6733

Xia, P., Wang, S., Ye, B., Du, Y., Li, C., Xiong, Z., et al. (2018). A circular RNA protects dormant hematopoietic stem cells from DNA sensor cGAS-mediated exhaustion. *Immunity* 48, 688-701.e7. doi: 10.1016/j.immuni.2018.03.016

Xu, B., Zhou, C., Meredith, M., and Baltz, M. J. (2017). Acute cell volume regulation by Janus kinase 2-mediated sodium/hydrogen exchange activation develops at the late one-cell stage in mouse preimplantation embryos. *Biol. Reprod.* 96, 542–550. doi: 10.1095/biolreprod.116.143974

Yan, J. W., Lin, J. S., and He, X. X. (2014). The emerging role of miR-375 in cancer. *Int. Cancer J.* 135, 1011–1018. doi: 10.1002/ijc.28563

Yang, B., Sun, L., and Liang, L. (2019). MiRNA-802 suppresses proliferation and migration of epithelial ovarian cancer cells by targeting YWHAZ. *J. Ovarian Res.* 12:100. doi: 10.1186/s13048-019-0576-3

Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., et al. (2017). Extensive translation of circular RNAs driven by N(6)-methyladenosine. *Cell Res*. 27, 626–641. doi: 10.1038/cr.2017.31

Yang, Y., Gao, X., Zhang, M., Yan, S., Sun, C., Xiao, F., et al. (2018). Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. *J. Natl. Cancer Inst*. 110, 304–315. doi: 10.1093/jnci/djx166

Yu, C. C., Chen, L. C., Lin, W. H., Lin, V. C., Huang, C. Y., Lu, T. L., et al. (2020). Genetic association analysis of cell cycle regulators reveals ywhaz has prognostic significance in prostate cancer. *Cancer Genomics Proteomics* 17, 209–216. doi: 10.21873/cgp.20181

Zang, J., Lu, D., and Xu, A. (2020). The interaction of circRNAs and RNA binding proteins: an important part of circRNA maintenance and function. *J. Neurosci. Res*. 98, 87–97. doi: 10.1002/jnr.24356

Zhang, G., Feenstra, B., Bacelis, J., Liu, X., Muglia, L. M., Juodakis, J., et al. (2017). Genetic associations with gestational duration and spontaneous preterm birth. *N. Engl. J. Med*. 377, 1156–1167. doi: 10.1056/NEJMoa1612665

Zhu, X., Qiu, J., Zhang, T., Yang, Y., Guo, S., Li, T., et al. (2020). MicroRNA-188-5p promotes apoptosis and inhibits cell proliferation of breast cancer cells via the MAPK signaling pathway by targeting Rap2c. *J. Cell. Physiol*. 235, 2389–2402. doi: 10.1002/jcp.29144

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: 01 April 2021
doi: 10.3389/fgene.2021.638191

Interpretable Feature Generation in ECG Using a Variational Autoencoder

V. V. Kuznetsov[1], *V. A. Moskalenko*[1,2], *D. V. Gribanov*[3] *and Nikolai Yu. Zolotykh*[1,2]*

[1] *Institute of Information Technologies, Mathematics, and Mechanics, Lobachevsky State University of Nizhni Novgorod, Nizhni Novgorod, Russia,* [2] *Mathematics of Future Technologies Center, Lobachevsky State University of Nizhni Novgorod, Nizhni Novgorod, Russia,* [3] *Laboratory of Algorithms and Technologies for Networks Analysis, National Research University Higher School of Economics, Nizhni Novgorod, Russia*

We propose a method for generating an electrocardiogram (ECG) signal for one cardiac cycle using a variational autoencoder. Our goal was to encode the original ECG signal using as few features as possible. Using this method we extracted a vector of new 25 features, which in many cases can be interpreted. The generated ECG has quite natural appearance. The low value of the Maximum Mean Discrepancy metric, 3.83×10^{-3}, indicates good quality of ECG generation too. The extracted new features will help to improve the quality of automatic diagnostics of cardiovascular diseases. Generating new synthetic ECGs will allow us to solve the issue of the lack of labeled ECG for using them in supervised learning.

Keywords: feature extraction, variational autoencoder, ECG, electrocardiography, deep learning, explainable AI

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Rajesh Kumar Tripathy,
Birla Institute of Technology and Science, India
Jijun Tang,
University of South Carolina, United States

***Correspondence:**
Nikolai Yu. Zolotykh
nikolai.zolotykh@itmm.unn.ru

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

Received: *05 December 2020*
Accepted: *01 March 2021*
Published: *01 April 2021*

Citation:
Kuznetsov VV, Moskalenko VA, Gribanov DV and Zolotykh NY (2021) Interpretable Feature Generation in ECG Using a Variational Autoencoder. Front. Genet. 12:638191. doi: 10.3389/fgene.2021.638191

1. INTRODUCTION

All the experience gained by the machine learning community shows that the quality of the decision rule largely depends on what features of samples are used. The better the feature description, the more accurately the problem can be solved. The features are used to require their interpretability, since it means the adequacy of the features to the real-world problem.

The traditional way to build a good feature description was to use an expert knowledge. Specialists in a particular subject area offer various methods for constructing the feature descriptions, which are then tested in solving practical problems. Another approach for constructing a good feature description is automatic feature extraction (also called dimensionality reduction).

There is a lot of methods for automatic feature extraction, such as principal component analysis, independent component analysis, principal graphs and manifolds, kernel methods, autoencoders, embeddings, etc. Among the most powerful and perspective approaches, we mention principal graphs and manifolds (Gorban et al., 2008; Albergante et al., 2020) and methods using deep learning (LeCun et al., 2015; Goodfellow et al., 2016).

Variational autoencoders (VAE) are neural networks which allow you to encode the source information and later, on the basis of the encoded information, to obtain a specific object, and further to generate similar objects but from a random set of coded characteristics (Kingma and Welling, 2013; Rezende et al., 2014; Doersch, 2016). Here we examine this method for the problem of automatic electrocardiogram (ECG) generation.

The electrocardiogram is a record of the electrical activity of the heart, obtained with the help of electrodes placed on the human body. Electrocardiography is one of the most important methods in cardiology. Schematic representation of the main part of ECG is shown in **Figure 1**. One cardiac

cycle (the performance of the heart from the beginning of one heartbeat to the beginning of the next) contains P, T, U waves and QRS complex, consisting of Q, R, and S peaks. The size, shape, location of these parts give great diagnostic information about the work of the heart and about the presence/absence of certain diseases.

Recently, machine learning (especially deep learning) methods have been widely used for automatic ECG analysis; see the recent review by Hong et al. (2020). The application tasks include ECG segmentation, disease detection, sleep staging, biometric human identification, denoising, and the others (Hong et al., 2020). A variety of classical and new methods are used. Among them there are discriminant analysis, decision trees, support vector machine, fully-connected and convolutional neural networks, recurrent neural networks, generative adversarial networks, autoencoders, etc. (Schläpfer and Wellens, 2017; Hong et al., 2020).

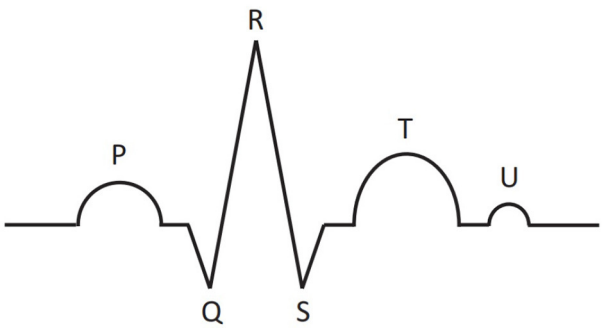


FIGURE 1 | Schematic representation of main parts of the ECG signal for one cardiac cycle: P, T, U waves and QRS complex, consisting of Q, R, and S peaks.

The most interesting and fruitful directions in applying deep learning methods to ECG analysis are generating synthetic ECGs and automatic extracting new interpretable features. Delaney et al. (2019), Golany and Radinsky (2019), and Zhu et al. (2019) study the problem of ECG generation. The authors of those papers used different variants of generative adversarial networks (GANs) (Goodfellow et al., 2014). The best results concerning the ECG generation were obtained by Delaney et al. (2019). The authors report on the Maximum Mean Discrepancy (MMD) metric equals to 1.05×10^{-3}.

Our approach in generating ECG is based on VAE. We propose a neural network architectures for an encoder and a decoder for generating synthetic ECGs and extracting new features. The generated synthetic ECGs look quite natural. MMD equals to 3.83×10^{-3}, which is worse than the value obtained by Delaney et al. (2019) using GAN, but we note that the comparison of these two metric values is not absolutely correct, since the values were obtained on different training sets and for solving similar, but different problems. Qualitatively, the results obtained by the VAE differ from the GAN, but our model is lighter and simpler, and the difference is not colossal. On the other hand, we use VAE, not a regular autoencoder, because VAE will generate signals from a random dataset, which will expand the training sample due to artificially generated ECGs.

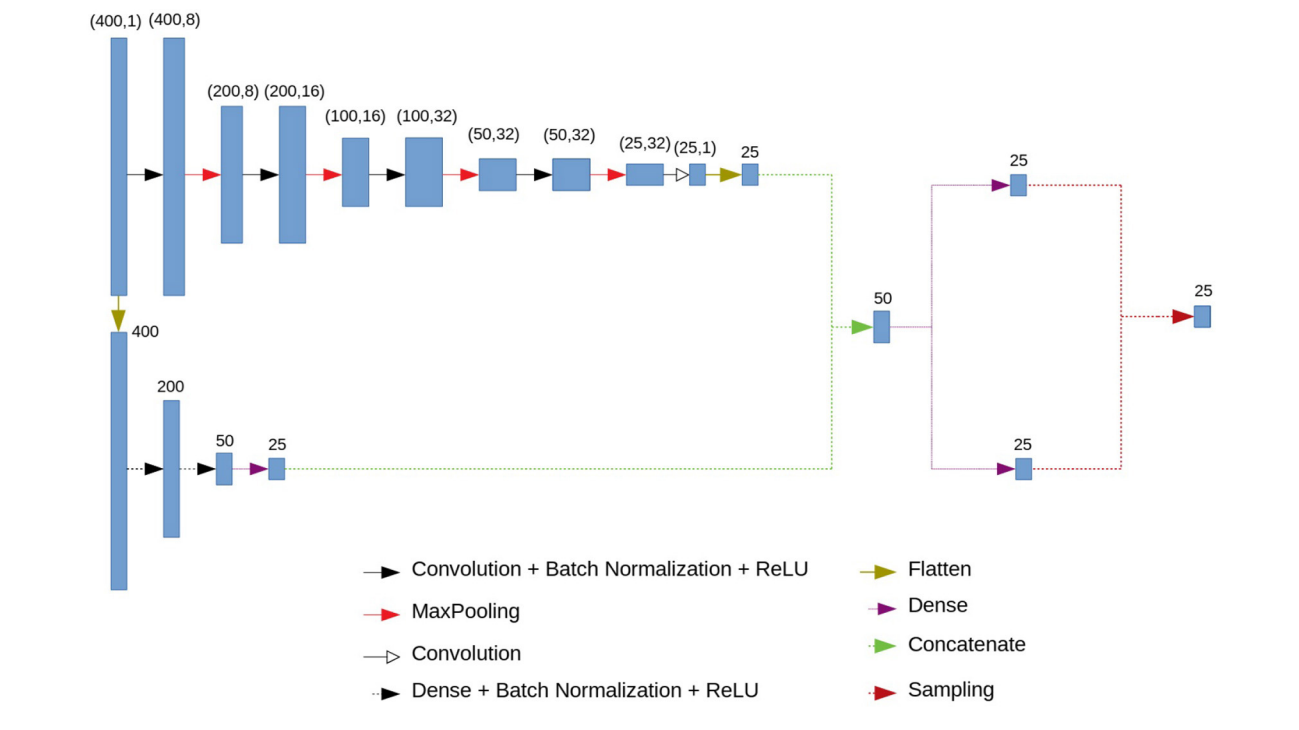


FIGURE 2 | Encoder architecture.

The main advantage of our work is the proposal of the method for extracting new features. The goal is to encode data on the signal with the smallest possible number of features. Our experiments show that these features are quite interpretable. This fact allows us to hope that using these features will help to improve the quality of automatic diagnostics of cardiovascular diseases. Generating new synthetic ECGs will allow us to fix the issue of the lack of labeled ECG for using them in supervised learning.

We note that the RR interval is an extremely important parameter of the ECG. Nevertheless, the aim of the study was to generate one cardiac cycle. On the other hand, our approach allows one to generate an ECG and extract features for one cardiac cycle of any duration. Our model is not as large as for the whole signal, and it is convenient to use it in various subtasks related to ECG diagnostics.

Besides VAE, other autoencoders are also used for ECG analysis. In particular, Gyawali et al. (2019) uses f-SAE to capture relevant features from the 12-lead ECG for the downstream task of VT localization. The subject of the work is very different from ours. In our work, we want to use specifically VAE, which can be used for many tasks related to ECG analysis, including for solving our problem.

2. ALGORITHM

2.1. Pre-processing

Our original ECG is a 10-s 12-lead signal with a frequency of 500 Hz. Using the segmentation algorithms described by Moskalenko et al. (2019), we determine beginnings and endings of all P and T waves and all the picks R. Then, we do the step forward and backward from the R pick at an equal distance. Thus, we obtain the set of cardiac cycles, each of which of vectors length is 400 (800 ms).

2.2. Neural Network Architecture: Encoder

A variational autoencoder (Kingma and Welling, 2013; Doersch, 2016) consists of an encoder and a decoder. We propose the following architecture for them. The encoder consists of a convolutional and a fully connected blocks. The architecture of the encoder is presented in **Figure 2**. The input vector of length 400 is fed to the input of the encoder. The next step is branching into a fully connected and convolutional chains. This branching occurs immediately in order to simultaneously highlight small local features and features based on the entire signal. Otherwise, using only fully connected blocks, we would get smooth ideal signals, and using only convolutional ones—signals close to a simple set of numbers.

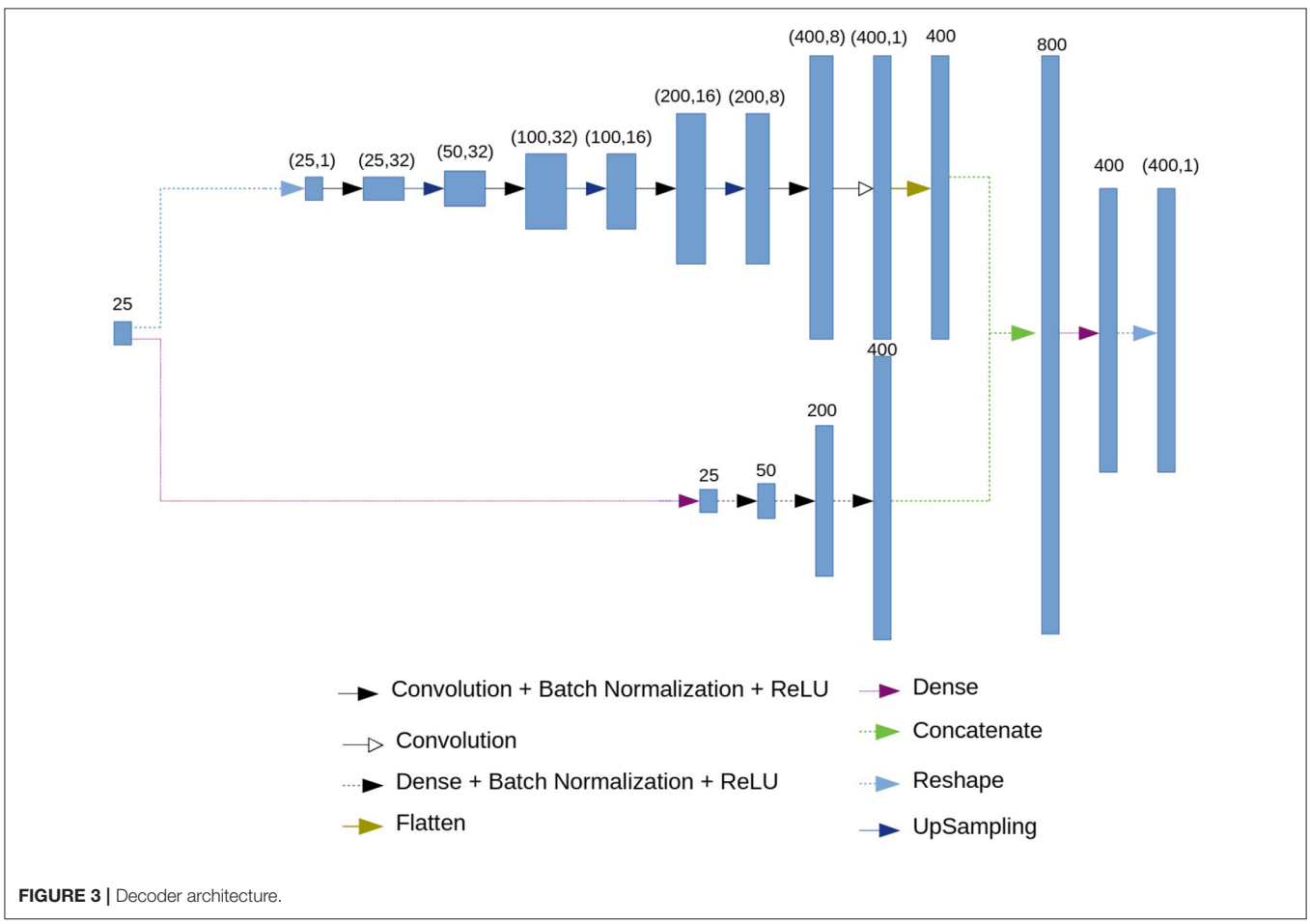


FIGURE 3 | Decoder architecture.

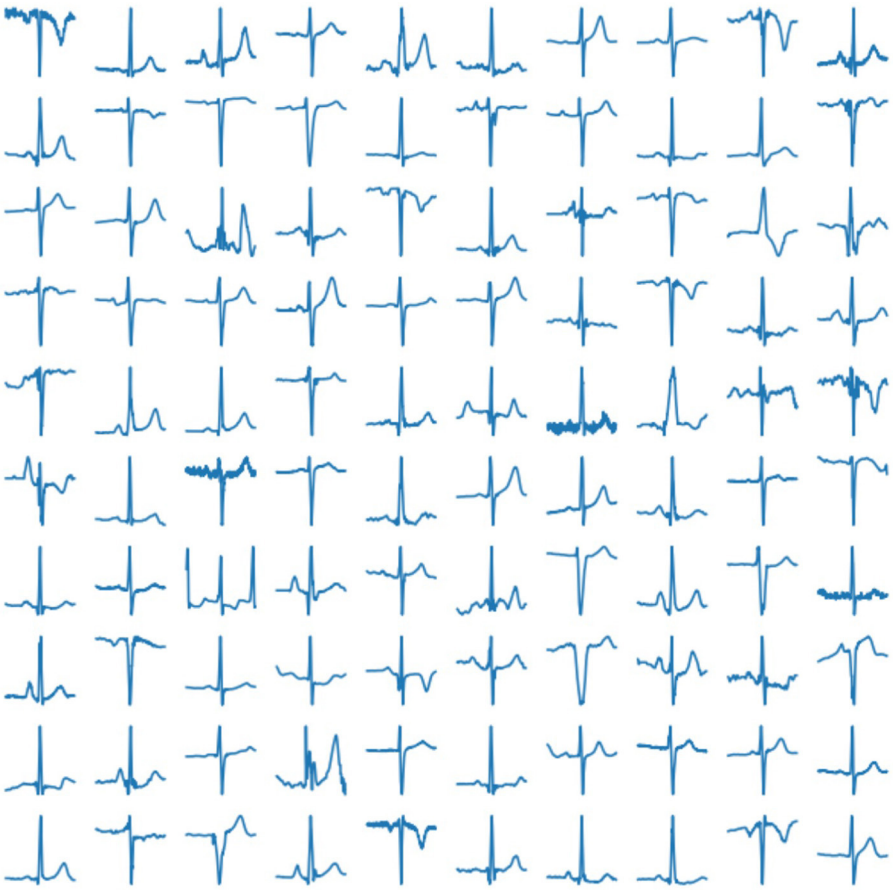


FIGURE 4 | Examples of real cardiac cycles obtained from ECG signals and used in the training of VAE.

The convolutional chain (at the top of the circuit in **Figure 2**) consists of four series-connected blocks, each of which consists of a convolution layer, a batch normalization layer, a ReLU activation function and a MaxPooling layer. In addition, we have another convolution layer. At the output of this block we get 25 neurons.

The fully connected chain of the encoder (at the bottom of the circuit in **Figure 2**) consists of three fully connected (dense) layers, interconnected by a batch normalization and ReLU activation functions. At the output of the last fully connected layer we have 25 neurons.

The outputs of the convolutional and fully connected chains are concatenated, which gives us a vector of length 50. Using two fully connected layers we get two 25-dimensional vectors which are interpreted as a vector of means and a vector of logarithms of variances for 25 normal distributions (or for one 25-dimensional normal distribution with a diagonal covariance matrix). The output of the encoder is a vector of length 25 in which each component is sampled from those normal distributions with specified means and variance.

We will interpret this 25-dimensional vector as a vector of new features sufficient to describe and restore with small error the one cardiac cycle. Note that with fewer features, the results were noticeably worse (the MMD metric was significantly higher). On the other hand, this number of features was enough to restore the signal with sufficient quality.

As the loss function, the Kullback–Leibler distance

$$D_{KL}(P \parallel Q) = \int_X p \log \frac{p}{q} d\mu \tag{1}$$

is used. Due to this fact those 25 new features are of normal distribution. In (1) μ is any measure on X for which there exists a function absolutely continuous with respect to μ: $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$, P is the initial distribution, Q is the new distribution we have obtained.

2.3. Neural Network Architecture: Decoder

The architecture of the decoder is presented in **Figure 3**. As an input, the decoder accepts the 25-dimensional vector of features. Then, similarly to the encoder, branching into convolutional and fully connected chains occurs.

The fully connected chain (at the bottom of the circuit in **Figure 3**) consists of four blocks, each of which contains a fully

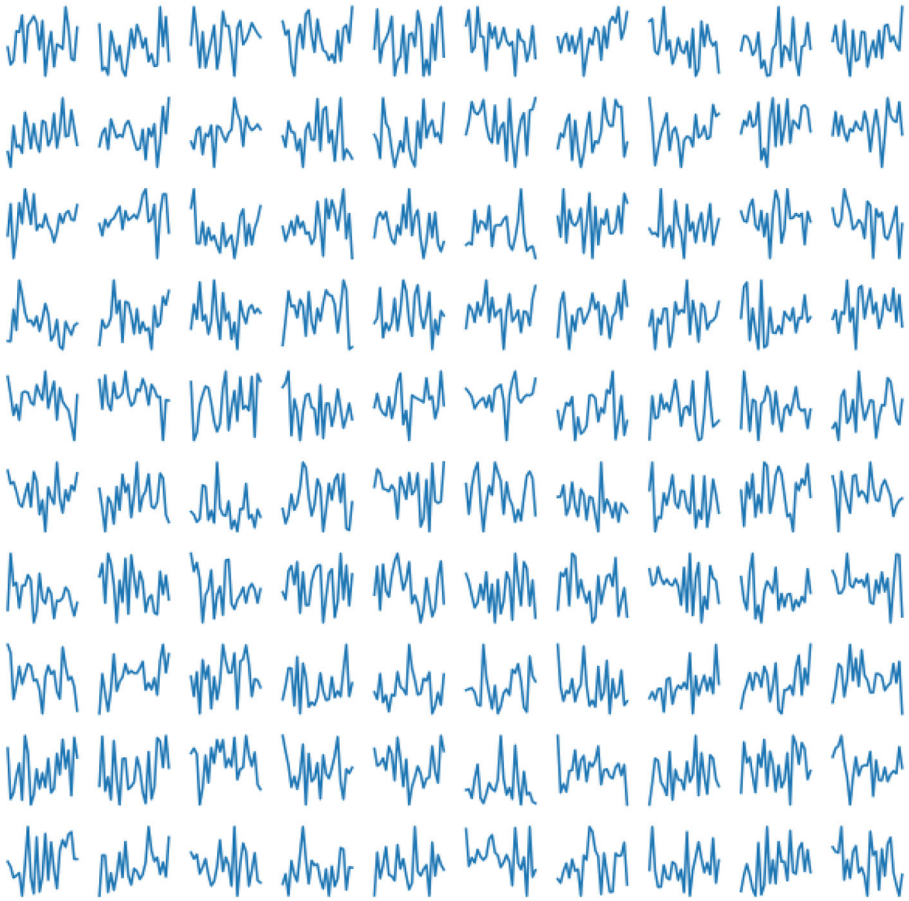


FIGURE 5 | Examples of generated normal distribution features for obtaining a cardio cycle based on them.

connected (dense) layer, batch normalization layer and the ReLU activation function.

The convolutional chain (at the top of the circuit in **Figure 3**) performs a deconvolution. It consists of four blocks which include a convolutional layer, a batch normalization layer, and ReLU activation function, followed by an upsampling layer.

As a result of the convolutional and the fully connected chains, we get 400 neurons from each. Then, we concatenate two results, obtaining 800 neurons. Using a dense layer we get 400 neurons which represent the restored ECG.

As a loss function for the output of the decoder, we use the mean squared error.

The models for the encoder and the decoder can be downloaded from https://github.com/VlaKuz/ecg_cycle_VAE.

3. EXPERIMENTAL RESULTS

In our experiment, we use 2,033 10-s ECG signals of frequency 500 Hz (Kalyakulina et al., 2019, 2020a,b). We process them according to the principles as described above (see section 2.1) and train our network on the obtained 252,636 cardiac cycles. Examples of those real human cardiac cycles derived from ECG signals are presented in **Figure 4**.

To train the model we used 720 epochs of Adaptive Moment Estimation (Adam) algorithm proposed by Kingma and Ba (2014) and implemented in TensorFlow Framework (Abadi et al., 2016). No data augmentation was not performed.

The trained network produce 25 features describing the cardiac cycle. The examples are shown in **Figure 5**.

After having trained the network we may test the decoder by supplying random (generated according to the standard normal distribution) numbers to its input. The examples of the produced results are given in **Figure 6**. These synthetic generated ECG looks quite natural.

To evaluate our results we calculated the Maximum Mean Discrepancy (MMD) metric (Delaney et al., 2019) on the set of 3,000 generated ECG. The value of MMD is equal to 3.83×10^{-3}. Keep it in mind that the best value of MMD obtained by Delaney et al. (2019) by GAN is 1.05×10^{-3}. The value obtained by us is slightly less than the value from Delaney et al. (2019). However, it shouldn't be argued that this metric is a reference. There are no illustrations in Delaney et al. (2019) confirming the correctness of the result. We note that the comparison of these two metric values is not absolutely correct, since these values were obtained on different training sets and for solving similar, but different problems. Unfortunately, the papers (Golany and

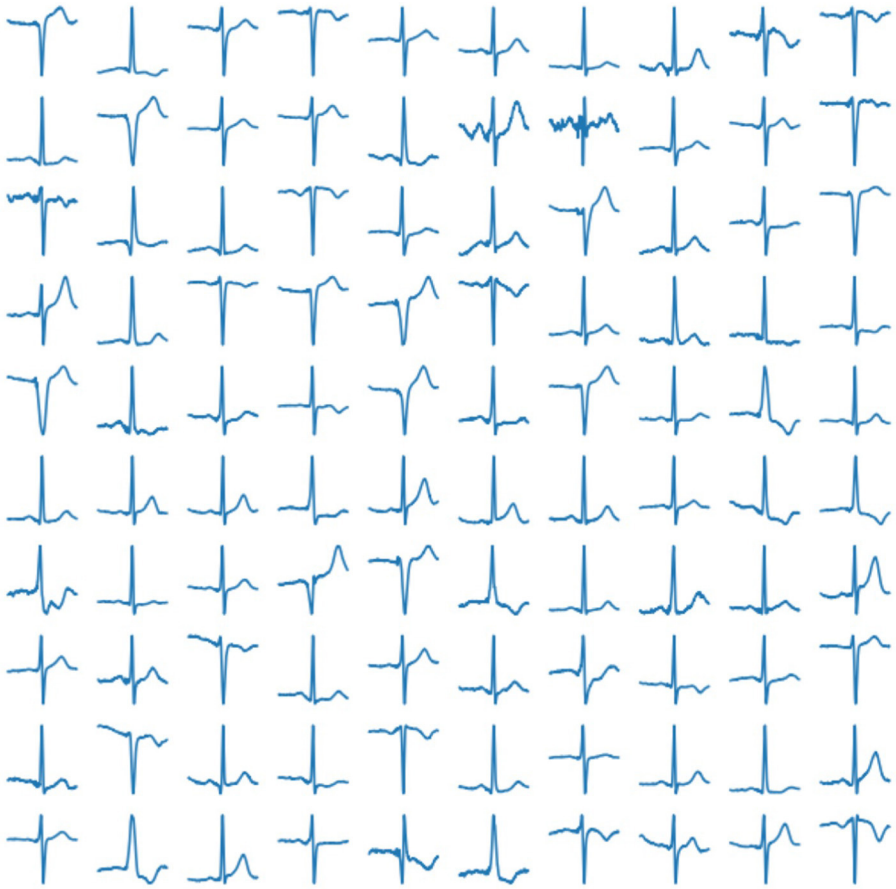


FIGURE 6 | Examples of generated heart cycles based on 25 features.

Radinsky, 2019; Zhu et al., 2019) don't contain (applicable to our problem) values of similar metrics.

Interesting results were obtained when generating ECG with a varying feature. Some generated ECG signals are presented in **Figure 7**. Twenty-four features were fixed for each test when the remaining feature was changing. It was possible to find a parameter responsible, for example, for the height of the wave T, the depression of the ST wave, etc. Thus, in some cases, the extracted features may be interpreted, which also confirms the high quality of the constructed feature description. So, from the figures it can be seen that when fixing the 6th sign of changes in the behavior of the QRS complex. When the 14th feature changes, the amplitude of the P wave changes, and when the 24th feature changes, the behavior of the T wave changes. Other signs have a similar effect. In all cases, it can be seen that with an increase in the value of the feature, the peak rises up, and with a decrease, it goes down.

The variational autoencoder models for each lead were also trained. Examples of the results of trained models in the **Figure 8**. The figure shows the leads I, II, III.

4. CONCLUSIONS AND FURTHER RESEARCH

In this paper, we proposed a neural network (variational autoencoder) architecture that is used to generate an ECG corresponding to a single cardiac cycle. Our method generates synthetic ECGs using rather small number (25) of features, with completely natural appearance, which can be used to augment the training sets in supervised learning problems involving ECG. Our method allowed us to extract new features that accurately characterize the ECG. Experiments show that the extracted features are usually amenable to good interpretation.

Our approach has both advantages and disadvantages.

The advantages include relative simplicity, lightness and small size of the system, which makes it very mobile and convenient; the information content of the extracted features by the encoder; the ability to obtain signals from a random distribution of a relatively small number of features; the ability to generate individual

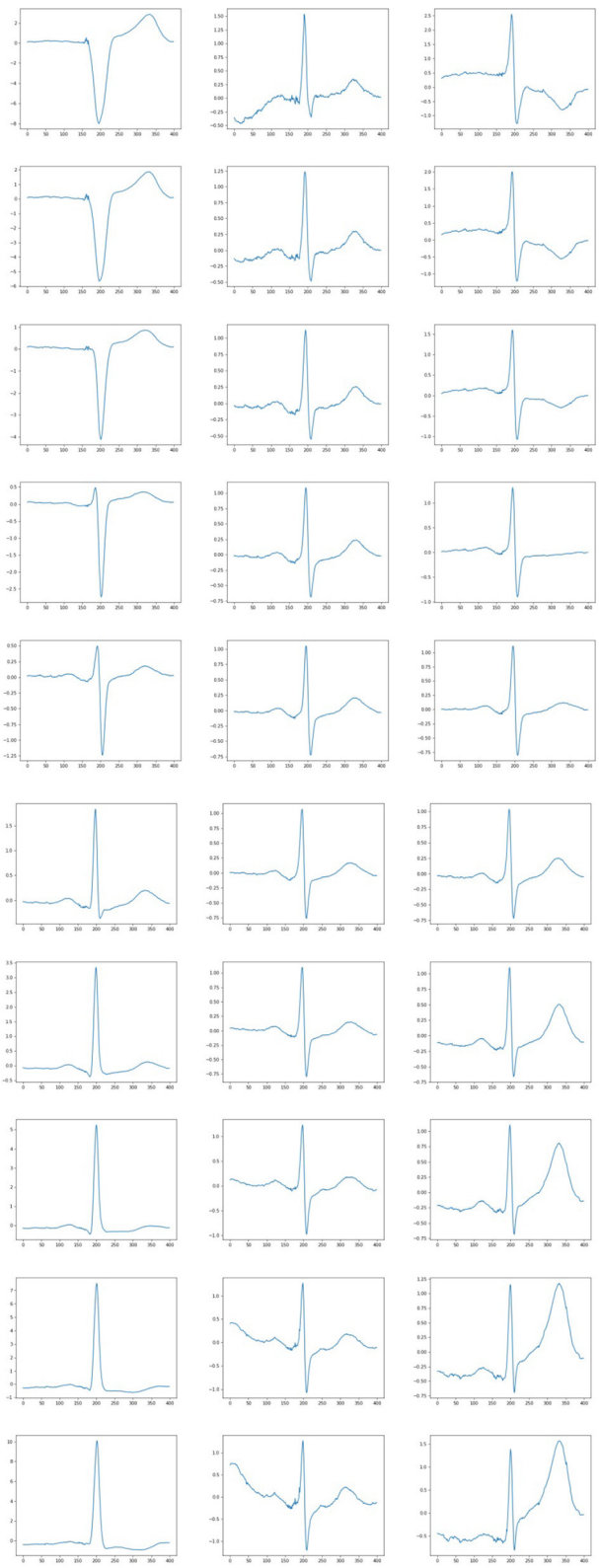


FIGURE 7 | Examples of ECG generated when a parameter is varying. Each column correspond to the set of fixed 24 features and varying other feature (6, 14, and 24 feature, respectively).

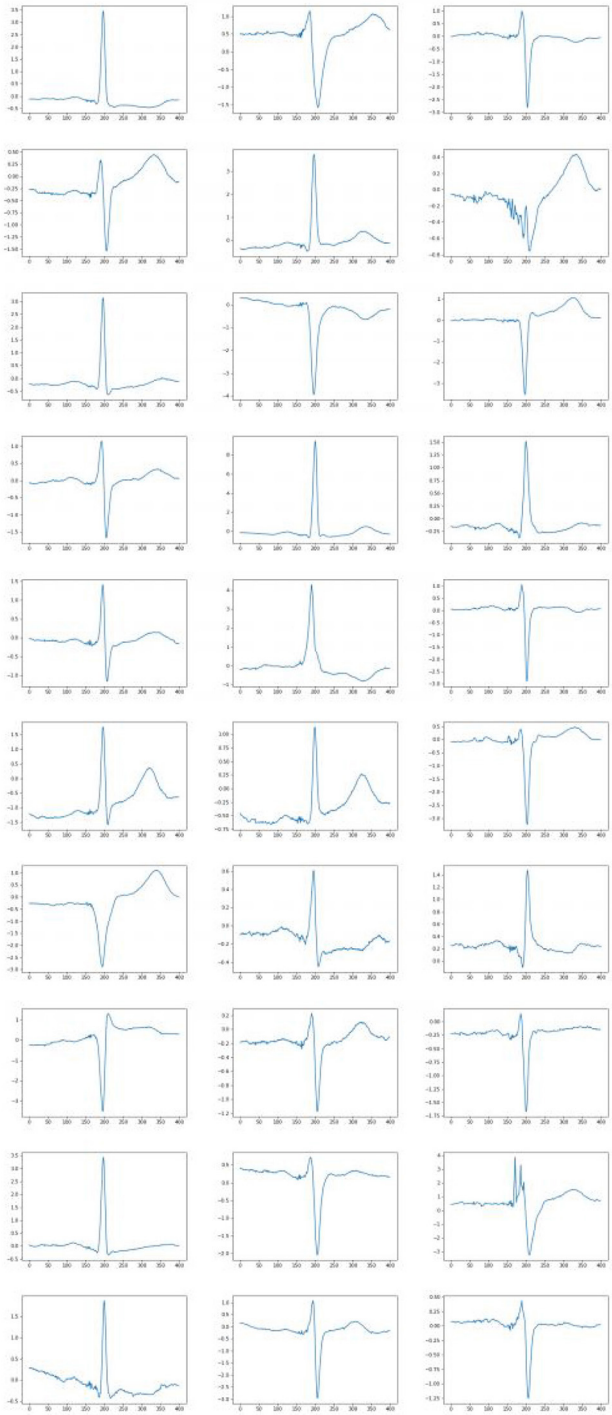


FIGURE 8 | Examples of ECGs generated by a VAE that has been trained in only one lead (I, II, III).

signals from a random distribution, as well as generating pathological signals.

The main of the disadvantages is inability to generate a whole ECG signal.

We plan to use our approach to generate the entire ECG, not just one cardiac cycle and, separately, for normal and pathological ECGs cases. We will also use the extracted features to improve the quality of automatic diagnosis of cardiovascular diseases.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://physionet.org/content/ludb/1.0.1/.

AUTHOR CONTRIBUTIONS

NZ conceived and supervised the study. VK and VM developed the method, performed the experiments and analysis. VK, NZ, and DG wrote the paper. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (agreement number 075-15-2020-808).

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Albergante, L., Mirkes, E., Bac, J., Chen, H., Martin, A., Faure, L., et al. (2020). Robust and scalable learning of complex intrinsic dataset geometry via elpigraph. *Entropy* 22:296. doi: 10.3390/e22030296

Delaney, A. M., Brophy, E., and Ward, T. E. (2019). Synthesis of realistic ECG using generative adversarial networks. *arXiv* 1909.09150.

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv* 1606.05908.

Golany, T., and Radinsky, K. (2019). "Pgans: personalized generative adversarial networks for ECG synthesis to improve patient-specific deep ECG classification," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), Vol. 33, 557–564. doi: 10.1609/aaai.v33i01.3301557

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*, Vol. 1. Cambridge, MA: MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial networks," in *Advances in Neural Information Processing Systems. arXiv [Preprint]* arXiv:1406.2661.

Gorban, A. N., Kégl, B., Wunsch, D. C., and Zinovyev, A. Y. (eds.). (2008). Principal manifolds for Data visualization and dimension reduction," in *Lecture Notes in Computational Science and Engineering* (Berlin: Springer), Vol. 58. 96–130.

Gyawali, P., Li, Z., Knight, C., Ghimire, S., Horacek, B. M., Sapp, J., et al. (2019). "Improving disentangled representation learning with the beta bernoulli process," in *2019 IEEE International Conference on Data Mining (ICDM)* (Beijing: IEEE), 1078–1083. doi: 10.1109/ICDM.2019.00127

Hong, S., Zhou, Y., Shang, J., Xiao, C., and Sun, J. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. *Comput. Biol. Med.* 122:103801. doi: 10.1016/j.compbiomed.2020.103801

Kalyakulina, A., Yusipov, I. I., Moskalenko, V. A., Nikolskiy, A. V., Kozlov, A. A., Kosonogov, K. A., et al. (2020a). *Lobachevsky University Electrocardiography Database (version 1.0.0)*. Cambridge, MA: PhysioNet.

Kalyakulina, A. I., Yusipov, I. I., Moskalenko, V. A., Nikolskiy, A. V., Kosonogov, K. A., Osipov, G. V., et al. (2020b). Ludb: a new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access* 8, 186181–186190. doi: 10.1109/ACCESS.2020.3029211

Kalyakulina, A. I., Yusipov, I. I., Moskalenko, V. A., Nikolskiy, A. V., Kozlov, A. A., Zolotykh, N. Y., et al. (2019). Finding morphology points of electrocardiographic-signal waves using wavelet analysis. *Radiophys. Quant. Electron.* 61, 689–703. doi: 10.1007/s11141-019-09929-2

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv* 1412.6980.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv* 1312.6114.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Moskalenko, V., Zolotykh, N., and Osipov, G. (2019). "Deep learning for ECG segmentation," in *International Conference on Neuroinformatics* (Cham: Springer), 246–254. doi: 10.1007/978-3-030-30425-6_29

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv* 1401.4082.

Schläpfer, J., and Wellens, H. J. (2017). Computer-interpreted electrocardiograms: benefits and limitations. *J. Am. Coll. Cardiol.* 70, 1183–1192. doi: 10.1016/j.jacc.2017.07.723

Zhu, F., Ye, F., Fu, Y., Liu, Q., and Shen, B. (2019). Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. *Sci. Rep.* 9:6734. doi: 10.1038/s41598-019-42516-z

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 Artificial Intelligence

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/frai.2021.674370

Similarities and Differences in Gene Expression Networks Between the Breast Cancer Cell Line Michigan Cancer Foundation-7 and Invasive Human Breast Cancer Tissues

Vy Tran[1], *Robert Kim*[1], *Mikhail Maertens*[1], *Thomas Hartung*[1,2,3] *and Alexandra Maertens*[1]*

[1]*Department of Environmental Health and Engineering, Center for Alternatives to Animal Testing, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States,* [2]*Department of Biology, Center for Alternatives to Animal Testing–Europe, University of Konstanz, Konstanz, Germany,* [3]*Department of Environmental Health and Engineering, Doerenkamp-Zbinden Professor and Chair for Evidence-Based Toxicology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States*

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Carolyn M. Klinge,
University of Louisville, United States
Sabine Matou-Nasri,
King Abdullah International Medical Research Center (KAIMRC), Saudi Arabia

***Correspondence:**
Alexandra Maertens
amaerte1@jhu.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *01 March 2021*
Accepted: *23 April 2021*
Published: *13 May 2021*

Citation:
Tran V, Kim R, Maertens M, Hartung T and Maertens A (2021) Similarities and Differences in Gene Expression Networks Between the Breast Cancer Cell Line Michigan Cancer Foundation-7 and Invasive Human Breast Cancer Tissues. Front. Artif. Intell. 4:674370. doi: 10.3389/frai.2021.674370

Failure to adequately characterize cell lines, and understand the differences between *in vitro* and *in vivo* biology, can have serious consequences on the translatability of *in vitro* scientific studies to human clinical trials. This project focuses on the Michigan Cancer Foundation-7 (MCF-7) cells, a human breast adenocarcinoma cell line that is commonly used for *in vitro* cancer research, with over 42,000 publications in PubMed. In this study, we explore the key similarities and differences in gene expression networks of MCF-7 cell lines compared to human breast cancer tissues. We used two MCF-7 data sets, one data set collected by ARCHS4 including 1032 samples and one data set from Gene Expression Omnibus GSE50705 with 88 estradiol-treated MCF-7 samples. The human breast invasive ductal carcinoma (BRCA) data set came from The Cancer Genome Atlas, including 1212 breast tissue samples. Weighted Gene Correlation Network Analysis (WGCNA) and functional annotations of the data showed that MCF-7 cells and human breast tissues have only minimal similarity in biological processes, although some fundamental functions, such as cell cycle, are conserved. Scaled connectivity—a network topology metric—also showed drastic differences in the behavior of genes between MCF-7 and BRCA data sets. Finally, we used canSAR to compute ligand-based druggability scores of genes in the data sets, and our results suggested that using MCF-7 to study breast cancer may lead to missing important gene targets. Our comparison of the networks of MCF-7 and human breast cancer highlights the nuances of using MCF-7 to study human breast cancer and can contribute to better experimental design and result interpretation of study involving this cell line.

Keywords: WGCNA, TCGA, cell line relevance, network analysis, human breast cancer tissues, MCF-7

INTRODUCTION

Cell lines have been extensively used as models for human biology and have contributed to many insights: from the development of vaccines and toxicology screening, to the study of disease mechanisms and treatments. Despite these achievements, there have been growing concerns about the quality of cell lines (Hartung 2007), ranging from cell-line misidentification, unreproducible studies, to failed clinical trials (Schweppe et al., 2008; Gillet et al., 2013; Hartung, 2013). In 2012, Amgen researchers attempted to replicate 53 landmark cancer papers and found that 47 studies were not reproducible (Begley and Ellis, 2012); the result is in keeping with a broader estimate that most research studies are likely to be not reproducible (Ioannidis, 2005). This leads to wasteful use of financial resources and labor, with an estimation of 28 billion dollars a year spent on irreproducible research (Freedman et al., 2015). While various reasons contribute to the irreproducibility of research, including study power, technical and biological variability, cell line reproducibility has been considered as one of the major factors contributing to the failure to reproduce preclinical studies. For instance, cell line misidentification has been a long standing problem in cell culture, with controversies for HeLa cells dating back to the 1970s (Nelson-Rees et al., 1974). In addition, the usefulness of cell lines as models for human biology has been questioned. Not all cancer cell lines have the same value as models to study cancer in humans (Gillet et al., 2013).

Michigan Cancer Foundation-7 cells (MCF-7) have been used widely in labs as a model for human breast cancer for over 40 years. It is estrogen receptor (ER)-postive, progesterone receptor (PR)-positive, poorly aggresive, and non-invasive, with low metastatic capacity (Comsa et al., 2015). Since its creation in 1973, MCF-7 has resulted in the highest number of scientific papers compared to other breast cancer cell lines (Sweeney et al., 2012), with over 42,000 publications on PubMed related to this cell line. MCF-7 has played an important role in studying estrogen receptor (ER) in tumor growth, characterization of cancer drug candidates, and endocrine disruption screening (Comsa et al., 2015). Since cancer cell lines greatly contribute to our understanding of cancer molecular mechanisms, investigating their relevance of cancer cell lines to human cancer is critical. Noticeably, even MCF-7 cells from a single cell bank batch can exhibit heterogeneity: previous work in our lab at the Center for Alternatives to Animal Testing showed that MCF-7 cells coming from the same ATCC lot still displayed marked differences in cellular and phenotypic characteristics, such as proliferation, and expression of estrogen-related genes that escaped routine cell line authentication techniques (Kleensang et al., 2016). A more recent study on MCF-7 also shows variations in expression of reference genes among sub-clones of this cell line (Jain et al., 2020).

In this study, we used large-scale data analysis to examine the similarities and differences between MCF-7—a cell line belonging to the luminal A molecular subtype (Dai et al., 2017)—and invasive breast cancer tissues including four subtypes—luminal A, luminal B, HER2-enriched, and basal-like (Cancer Genome Atlas Network, 2012). To our knowledge, this is one of only a few studies that use network analysis to compare an immortalized cancer cell line to human cancer tissues. The bioinformatics pipeline established in this study was made available and can potentially be applied to similar analysis between cell lines and their corresponding tissues in humans.

MATERIALS AND METHODS

Data

MCF-7 ARCHS4 data set. Gene expression level RNA-seq data of the human adenocarcinoma cell line MCF-7 was obtained from the ARCHS4 database (All RNA-seq and ChIP-seq Sample and Signature Search). For detailed description of data processing workflow, readers are invited to read the ARCHS4 article (Lachmann et al., 2018). Briefly, raw RNA-seq data was collected from Gene Expression Omnibus (GEO) by the authors of ARCHS4, aligned to the reference genome, mapped to the gene level, and uploaded to the ARCHS4 database. The MCF-7 data set contained 1032 samples from 107 GEO series. Gene expression data was downloaded as an expression matrix using the R script provided by ARCHS4 and was $\log_2$-transformed. Since the data set came from multiple experimental series, data sets were checked for batch effects using *Combat* (Johnson et al., 2007) before downstream analysis.

MCF-7 GSE50705 data set. RNA microarray data were downloaded from Gene Expression Omnibus (GEO) (Shioda et al., 2013). In the original study, MCF-7 cells were treated with various concentrations of natural and xenobiotic estrogens. We extracted samples treated for 48 h with the steroid hormone 17β-estradiol (n = 88), converted probes to gene symbols, and removed probes that were matched to multiple gene names.

BRCA data set. Pre-processed, RSEM-normalized Level 3 RNA-seq data of breast invasive ductal carcinoma tissues from The Cancer Genome Atlas was downloaded from FireBrowse. The data set included 1,212 human tissue samples.

For all three MCF-7 and BRCA data sets, samples were checked for outliers using hierarchical clustering, as well as missing values using the *goodSamplesGenes* function in the Weighted Correlation Network Analysis (WGCNA) package. No obvious outliers and missing values were found. Before constructing the co-expression networks for the MCF-7 and BRCA data sets, genes were filtered for the top 10,000 mostly highly variant genes using median absolute deviation (MAD) to exclude the large fraction of genes that are expressed at low level, as two genes with low variance would result in high correlation that would not be biologically meaningful. The resulting gene expression matrices were then analyzed with the WGCNA approach, a popular network analysis algorithm. The analysis workflow for this study can be viewed in **Figure 1**.

Weighted Gene Co-expression Network Analysis

WGCNA is a systems biology approach that describes the correlation among genes (Langfelder and Horvath, 2008). It

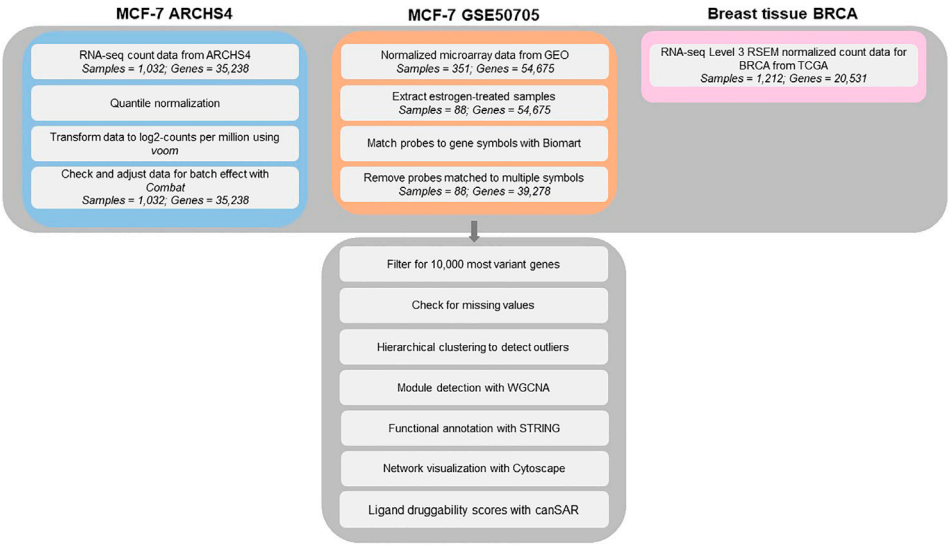


FIGURE 1 | Analysis workflow for the MCF-7 ARCHS4, MCF-7 GSE50705 and BRCA data sets.

uses network language to describe the pairwise correlation between genes in a data set, based on the assumption that genes with similar expression levels tend to belong to similar pathways. Rather than using a hard threshold for the co-expression similarity s_{ij}, which does not reflect the continuous property of gene expression levels and may lead to loss of information, WGCNA uses a soft threshold approach. It raises the co-expression similarity s_{ij} to a power β ($\beta \geq 1$) to obtain the adjacency matrix a_{ij}, allowing the adjacency to having continuous values between 0 and 1:

$$a_{ij} = s_{ij}^{\beta}$$

We set $\beta = 5$ for ARCHS4, $\beta = 7$ for GSE50705, and $\beta = 6$ for BRCA based on the scale-free topology criterion (**Supplementary Figure S1**). After network construction, modules in each data set were detected using hierarchical clustering implemented through the function *blockwiseModules*, with the parameter *minModuleSize* set to 100, 80, and 100 for ARCHS4, GSE50705 and BRCA respectively.

Functional Annotation

Modules detected by WGCNA can have true biological meaning or can be results of noises in the data, such as sample contamination, technical artifacts, or experimental design. Therefore, we performed functional enrichment of biological processes for genes in each module to identify modules with biological meaning. We used the package *STRINGdb* which provides an R interface to the STRING protein-protein interactions database. The annotation was adjusted for *Homo sapiens* background. Enrichment *p*-values were calculated based on over-representation analysis using hypergeometric tests and were adjusted for multiple hypothesis testing with Benjamini-Hochberg procedure (Szklarczyk et al., 2019).

Data Visualization

The modules of interest were visualized with the network visualization software Cytoscape version 3.7.0 (Shannon et al., 2003). For **Figures 2** and **3**, networks were plotted with Group Attributes Layout in Cytoscape to highlight gene module membership. For **Figure 4**, the network was plotted with Prefuse Force Directed Layout. Node color was correlated with the number of gene PubMed publications, obtained by querying the Entrez IDs to obtain a raw count of PMIDs on the PubMed database, as described in our previous publication (Maertens et al., 2020).

Scaled Connectivity

The scaled connectivity for each gene in the MCF-7 and BRCA networks were calculated from the adjacency matrices using the function *fundamentalNetworkConcepts* from the WGCNA package. Scaled connectivity is calculated as K = *Connectivity*/max*(Connectivity)*. Full tables of scaled connectivity of 10,000 most variant genes in GSE50705 and BRCA are available in **Supplementary Table 1A,B**.

Ligand-Based Druggability

The ligand druggability scores for the 10,000 most variant genes in the MCF-7 and BRCA data sets were queried using the Protein Annotation Tool from the canSAR knowledgebase. The canSAR database integrates genomic information, structural biology, and properties of compounds to estimate likely "druggability" of chemicals (Tym et al., 2016; Coker et al., 2019). Ligand-based druggability is calculated by looking at the small molecule compounds that have been tested against the protein or its homologues. The ligand-based druggability score for each protein was calculated based on ligand efficiency, med-chem friendliness, and molecular weight of these compounds. Top 30 genes with highest positive ligand druggability scores were selected for each data set for **Figure 5**. Full tables of ligand

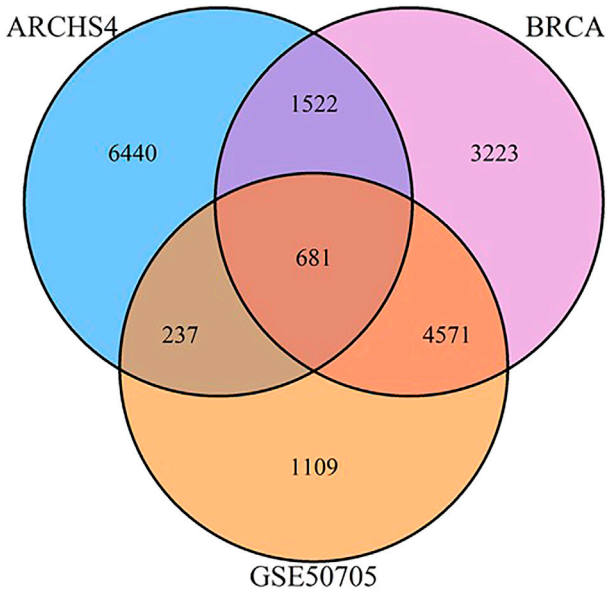


FIGURE 2 | Venn diagram of shared and unique genes between MCF-7 ARCHS4, GSE50705, and BRCA data sets.

druggability for the 10,000 genes in GSE50705 and BRCA are available in **Supplementary Table S2A,B**.

RESULTS

Minimal Overlapping Genes Between Michigan Cancer Foundation-7 and Human Breast Tissues

We selected three data sets based on human breast cancer tissues: 1) the TCGA data set of invasive breast cancer biopsies (henceforth BRCA), which has the advantage of reflecting human *in vivo* samples, although biopsies by their nature include a mix of different tissues 2) the ARCHS4 collection of MCF-7 samples, which is an attempt to massively mine publicly available RNA-seq experiments, and consists of 1032 samples combined from GEO, and 3) a smaller study of MCF-7 cells exposed to estrogen in a dose response curve. As the data sets involve a range of different technologies, preprocessing strategies, and in the case of ARCHS4, potentially many different biological conditions, we began with the basic initial step of reducing the gene expression set to the top 10,000 most variant genes, to eliminate genes that were minimally or inconsistently expressed and would therefore confound the use of a correlation-based approach. Surprisingly, even this initial step indicated minimal conservation of gene expression signatures - only 681 genes were conserved amongst the three datasets, and of the top 10,0000 genes from the ARCHS4 data set, fully 6,440 were unique to that data set (**Figure 6**).

In the case of the genes found in all three data sets, annotation analysis revealed that they were enriched for genes annotated to mitotic cell cycle (adjusted-p value = 1.08E-21), regulation of cell migration (adjusted p = 1.89E-18), and response to endogenous stimulus (adjusted-p value = 2.33E-18) (**Supplementary Table S3**), suggesting that of the highly expressed genes, the common genes are likely annotated to fundamental cell processes.

In order to understand how and why the data sets diverged even at this fundamental level, we explored the genes that were unique to each data set. For the BRCA data set, we suspected that one cause of the difference was likely the fact that cancer biopsies always reflect a mixture of cell-types and typically have a significant component of immune infiltration. Our data support this to a limited extent: genes unique to BRCA were enriched for immune-related GO annotations, such as regulation of immune response (adjusted-p value = 0.002023) and regulation of innate immune response (adjusted-p value = 0.007635)

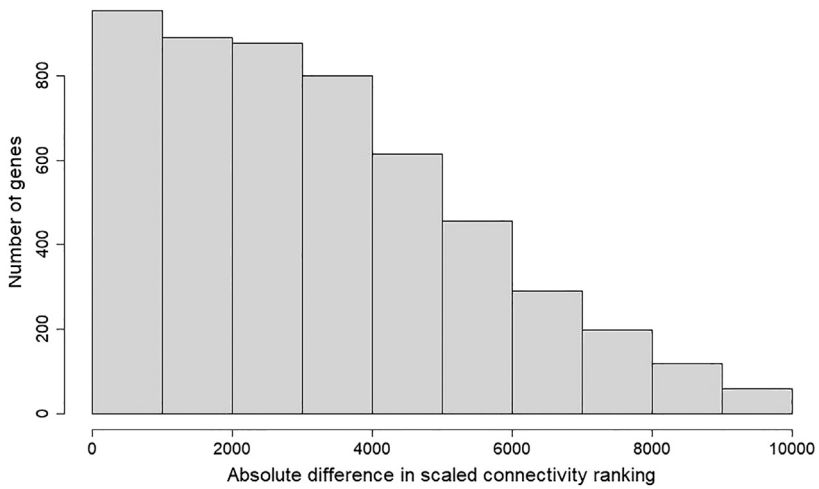


FIGURE 3 | Histogram of absolute difference in scaled connectivity ranking between GSE50705 and BRCA. The scaled connectivity scores obtained with the *fundamentalNetworkConcepts* function from WGCNA was ranked, and the absolute difference between GSE50705 and BRCA were calculated.

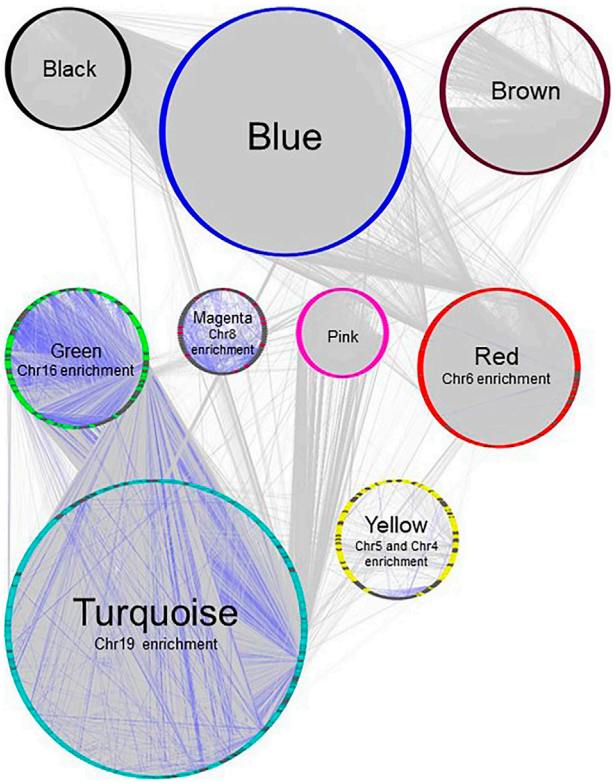


FIGURE 4 | Gene network of BRCA data set, plotted with Group Attributes Layout in Cytoscape. Module membership was determined using WGCNA. All modules except the grey module containing unassigned genes were exported to Cytoscape using the *exportNetworkToCytoscape* function, with a threshold of 0.05. Node color = module color. Genes associated with chromosome enrichment with an adjust-*p* value ≪ 0.05 from Enrichr are highlighted in dark grey. Purple edges indicate interactions between genes enriched for chromosomes.

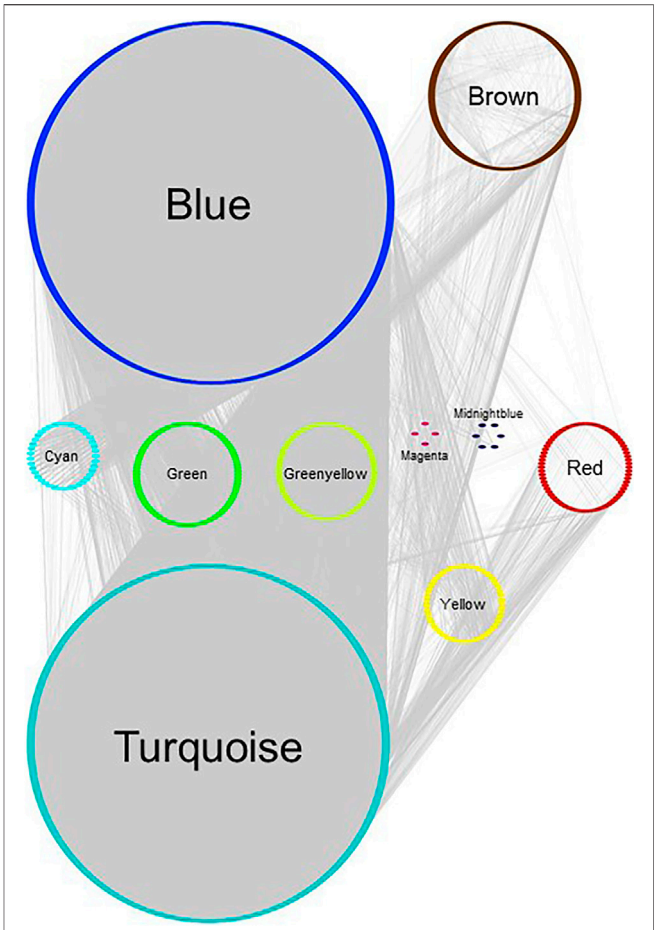


FIGURE 5 | Gene network of MCF-7 GSE50705 estrogen-treated data set, plotted with Group Attributes Layout in Cytoscape. Module membership was determined using WGCNA. All modules except the grey module were exported to Cytoscape using the *exportNetworkToCytoscape* function, with a threshold of 0.2. Node color = module color. There was no gene significantly enriched for chromosomes in Enrichr.

(**Supplementary Table S3**). In addition, within genes mapped to cell types via the Human Gene Atlas, there was a modest level of enrichment for immune-cell related genes (**Supplementary Table S3**).

More striking, however, was a marked presence of ribosomal subunit genes unique to the BRCA data set, annotated via STRING as ribosome biogenesis (adjusted *p*-value = 1.81E-08) (**Supplementary Table S3**), and Bioplanet as Cytoplasmic ribosomal proteins (adjusted-*p* value = 1.61E-10). While these ribosomal subunit genes are ubiquitously expressed in most breast cancer cell lines as well as most tissues (Ebright et al., 2020), they were in neither the MCF7-derived GSE50705 data set nor the ARCHS4 data set, likely owing to some extent to the chip design for the GSE50705, and the high noise level in ARCSH4. Within the BRCA data set, several ribosomal proteins showed a high level of patient-to-patient variation (**Supplementary Figure S2A**). Of the top most variant ribosomal proteins within the BRCA data set, only RPS3 was also in the GSE50705 data, and a much narrower dynamic range (**Supplementary Figure S2B**). Strong ribosomal signatures in a subset of circulating tumor cells have been associated with poor clinical outcomes in breast cancer patients (Ebright et al., 2020), and it seems likely that MCF-7 cells may not capture the effects of the variation in ribosomal protein expression patterns.

The genes unique to the MCF-7 GSE50705 estrogen dose-response curve were enriched for genes related to non-coding RNA processing (adjusted-*p* value = 6.64E-09) and mitochondrial respiratory chain complex IV biogenesis (adjusted-*p* value = 1.77E-07). Meanwhile, the large set of genes unique to the ARCHS4 dataset are most significantly enriched for cell-cell signaling (adjusted-*p* value = 2.23E-59), synaptic transmission (adjusted-*p* value = 8.52E-56), and ion transport (adjusted-*p* value = 9.11E-41) (**Supplementary Table S3**).

As breast cancer cell lines, it is surprising that genes unique to ARCHS4 MCF-7 cells are enriched for generation of neurons (adjusted *p*-value 5.98E-21) — a process unique to neuronal cells. In addition, there were some genes annotated to the meiotic chromosome segregation in this data set and even a few Y chromosome genes (**Supplementary Table S3**). This can be

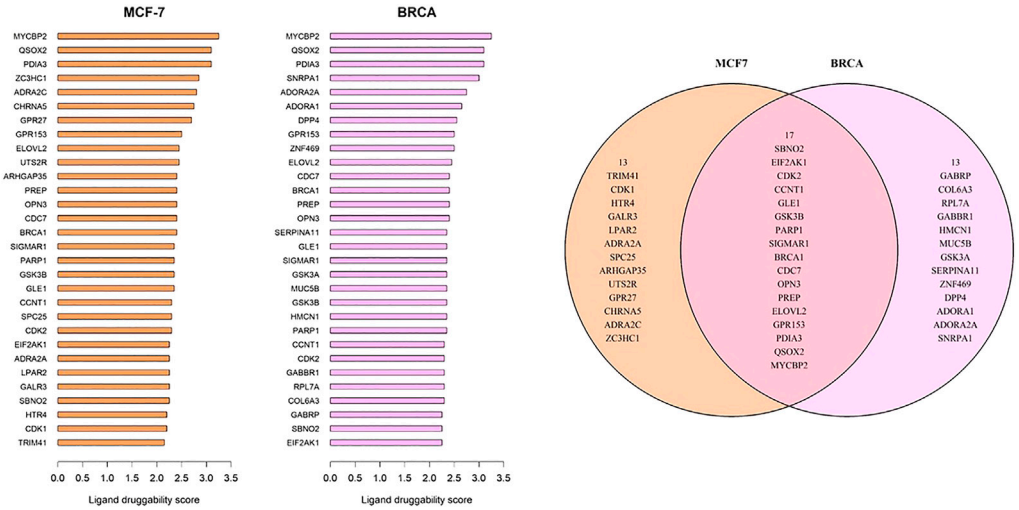


FIGURE 6 | Genes with top ligand druggability scores for **(A)** GSE50705 and **(B)** BRCA **(C)** Among the top 30 genes, 17 genes are overlapping.

caused by artifacts of annotations data or possibly contamination of other cell lines during experimental design of GEO studies or inclusion of non-MCF7 samples during data mining for ARCHS4. Overall, we observed a higher degree of similarity between GSE50705 and BRCA (5252 overlapping genes) than between ARCHS4 and BRCA (2203 overlapping genes).

Network Signatures Indicate Substantial Differences Between Data sets

In order to investigate similarities and differences between the data sets at a more intricate level, we used WGCNA for the 10,000 most variant genes in each data set to assign the genes to functional modules and see if, broadly speaking, interactions amongst genes were conserved. WGCNA uses correlations amongst gene expressions and groups genes in an unsupervised way to determine potential interactions. In keeping with our previous studies (Maertens et al., 2018; Maertens et al., 2020), the modules produced by WGCNA were input into STRING for biological annotations to verify whether WGCNA had produced modules of genes that were known to interact and were enriched for annotations.

For both the BRCA and the GSE50705 data sets, the modules were enriched for known protein interactions in STRING as well as highly significant adjusted *p*-values for GO Biological Processes, indicating that for most of the genes, WGCNA indeed clustered genes with similar biological functions and on the same pathways. However, the modules of the ARCHS4 dataset were unsatisfactory: most genes could not be classified into modules and ended up in the grey module for unassigned genes (**Supplementary Figure S3A**). These modules were also small and lacked distinguishing enrichments in STRING (**Supplementary Table S4A**). Due to the lack of meaningful biological signals in ARCHS4, we decided to focus subsequent analyses on the GSE50705 and BRCA data sets.

Annotations by modules in the GSE50705 and BRCA data sets indicated that the BRCA data set had a red module (**Supplementary Table S4B**) that was enriched for genes annotated as immune-related or cytokine-response biological processes. As expected, the immune component was not present in the GSE50705 data (**Supplementary Table S4C**). Our finding is supported by another study showing consistent upregulation of immune processes in primary tumors, possibly a result of immune infiltration in tumor tissues that is not present in cell lines (Yu et al., 2019).

Interestingly, even modules annotated for the same pathways in the two data sets varied substantially in their gene constituents: the largest module annotated for cell cycle processes in the BRCA data set (brown module, 1013 genes) only has 354 genes overlapping with its counterpart in the MCF-7 GS550705 data set (turquoize module, 2244 genes).

To determine whether the MCF-7 and BRCA data sets differ regarding network topology, we calculated the scaled connectivity, a metric for gene significance in a network, which asks if the gene is acting as a hub, or highly connected gene. We observed substantial difference between the two data sets: the average mean absolute difference between scaled connectivity was 3,223 and a small cluster of genes had a difference in scaled connectivity greater than 9,000, in each case ranking significantly higher in the MCF-7 dataset than the TCGA (**Figure 5**). The genes with the highest scaled connectivity in MCF-7 *vs.* TCGA showed relatively rare over- or under-expression within the BRCA data set, in each case with transcriptional perturbations less than 10 percent (**Table 1**, **Supplementary Figure S4**). The relatively high connectivity in MCF-7 may reflect a combination of the lineage of MCF-7 (perhaps a cancer type that over-expressed one or more of these genes), or the result of cellular instability and the evolution of MCF-7 over time.

TABLE 1 | Top 10 genes with highest difference in scaled connectivity ranking between GSE50705 and BRCA.

GeneSymbol	Scaled connectivity in MCF7	Rank in MCF7	Scaled connectivity in BRCA	Rank in BRCA	Absolute ranking difference
SUSD2	0.787126168	68	6.48E-05	9995	9927
CLU	0.878769987	16	0.000544507	9933	9917
BCAR3	0.78900237	67	0.001091298	9861	9794
TMPRSS3	0.767847573	82	0.000989588	9875	9793
OLFM1	0.70620407	181	0.000370951	9960	9779
SLC24A3	0.748824023	108	0.000982308	9877	9769
DEGS1	0.72618254	148	0.00099857	9874	9726
NPY1R	0.752667373	103	0.001839083	9775	9672
PLK2	0.722418194	156	0.001623592	9802	9646
CYP2J2	0.674845098	238	0.000986705	9876	9638

Breast Invasive Ductal Carcinoma Data Modules Contain Substantially More Cis-Regulated Genes Compared to GSE50705

In our previous study using WGCNA to analyze genes for functional assignment (Maertens et al., 2020), we reported that some modules were significantly enriched for genes on a single chromosome and that these modules tended to have no statistically significant enrichments for protein-protein interactions (PPI) or functional enrichments. Others have similarly reported significant clustering of genes based on intra-chromosomal distance and this feature has been demonstrated to be specific for different phenotypes (Garcia-Cortes et al., 2020). We found that the BRCA data had several modules that, although enriched for both PPI interactions and functional enrichments, had a statistically significant enrichment for genes on a single chromosome (**Figure 2**), likely reflecting increased co-expression from neighboring genes that results from a disruption in the regulatory elements that control gene transcription. Interestingly, this is not true of the MCF-7 sample - no module was significantly enriched for any one chromosome - despite the fact that it originates from a cancer cell line (**Figure 3**). However, it should be noted that the difference may simply be due to the greater dynamic range of RNA-sequencing compared to microarray.

Potential Drug Targets Missed by Using Michigan Cancer Foundation-7

As MCF-7 is often used for drug discovery research, we wanted to explore whether any potentially druggable candidates would be missed. To identify potential drug targets, we used the CanSAR database, which is a protein structure-based model that predicts ligand-based druggability scores based on the predicted cavities of over 144,000 proteins (Coker et al., 2019). While there was some overlap between the top-ranked druggable genes, there were several genes that would likely have been missed if MCF-7 were exclusively used for protein targets. However, it should be kept in mind that this approach is merely looking for potential candidates based on protein accessibility, not cancer biology, and there was no available information about classes of drugs. Therefore, our results merely suggest that it might be useful to look outside the MCF-7 model when screening for cancer drug targets.

As an example, of the genes that ranked highly for ligand-based druggability, CCNT1 has the 23rd highest score in the BRCA data set and was ranked in 277th in scaled connectivity in this data set yet ranked 5,824th in the GSE50705 data set. Altered CCNT1 expression (defined as z-score > +/−1.5) is not significantly associated with any mutations or copy number variations. Interestingly, altered expression is significantly associated with race, being more common in African Americans compared to Whites or Asian Americans (**Supplementary Figure S5**) - while the reason for this may range from SNP polymorphisms to different environmental exposures amongst populations, it does indicate one intrinsic short-coming of MCF-7 cells: they were isolated from a White female (Comsa et al., 2015), and therefore a model of breast cancer based on this tissue type alone will miss much of the molecular diversity of breast cancer in a population with mixed genetic backgrounds, diverse environmental exposures and clinical histories (Makki, 2015; Koual et al., 2020).

In the BRCA data set, CCNT1 was located in the yellow module enriched for several biological processes (intracellular transport - adjusted p value 4.63E-12, protein modification by small protein conjugation or removal - adjusted p value 4.63E-12, and protein transport - adjusted p value 2.11E-09) as well as Chromosome 4 (adjusted p value 0.0009227) and Chromosome 5 (adjusted p value 7.435E-20) enrichments (**Supplementary Table S5**). In order to predict the most likely trans-activated co-expressed genes, we looked at the top 50 genes correlated with CCNT1 expression by Spearman rank correlation (coefficient of correlation > 0.789; q-value < p = 6.27e-235). While these genes were enriched for known protein-protein interactions via STRING (p-value 0.000268), many of the genes were unconnected, and the only significant GO annotations were each based on a maximum of three genes - likely because many of these genes have minimal literature (**Figure 7**), and 31 of the genes were "unclassified" in GO SLIM Biological Process and therefore invisible in any annotation-based approach.

Using the FANTOM EdgeExressDB (Lizio et al., 2015; Lizio et al., 2019) to explore potential connections suggested that the

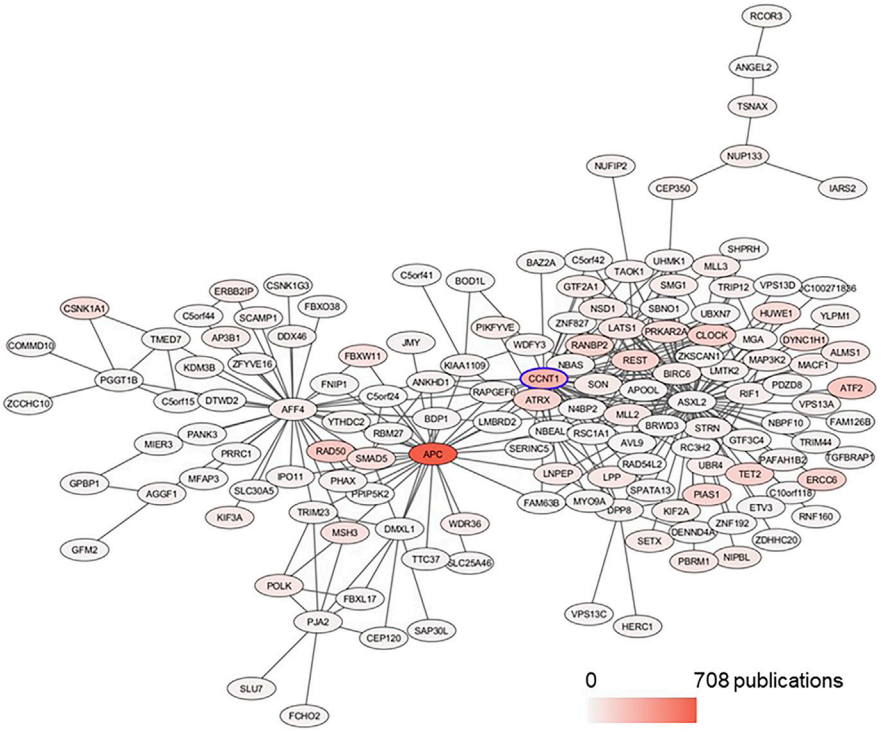


FIGURE 7 | Subnetwork of CCNT1 in the BRCA data set plotted with Prefuse Force Directed Layout in Cytoscape. From the module yellow in BRCA, the first neighbors of CCNT1 were selected. The color of the node corresponds to the number of PubMed publications associated with the genes.

coordinated co-expression of many of the genes was possible due to the transcription factor CLOCK, which had a very high correlation with CCNT1 - indeed, the two genes are significantly correlated in every cancer within TCGA, both before and after adjusting for tumor purity. So, while the mechanism of the correlated co-expression is unknown, it is a fairly robust finding within cancer tissues. The Spearman correlation between CCNT1 and CLOCK in the GSE50705 dataset was 0.130 with a p-value of 0.2285, and the top 50 candidates based on Spearman rank were shown in **Supplementary Table S6**.

While CCNT1 is not associated with survival in breast cancer as a whole, low expression is associated with longer survival in luminal A cancer (HR 1.98, $p = 0.0155$); similarly, low CLOCK expression was associated with increased survival in all breast cancer subtypes as well as the luminal A subtype. Like CCNT1, altered CLOCK expression is more common in African Americans.

The ultimate molecular function of CCNT1 and its interactions with CLOCK and the other predicted genes remains elusive, and any potential role in breast cancer is largely unremarked, as only 3 papers within PubMed mention CCNT1 and breast cancer, and in fact only 39 papers mention CCNT1 and cancer. The significance of CLOCK in cancer is better understood as it is thought to be a molecular link between disrupted circadian rhythm and cancer (Trujillo and Muotri, 2018), including breast cancer (Cadenas et al., 2014; Xiao et al., 2014).

DISCUSSION

Cell lines are often used as models for cancer research, but recent studies have drawn attention to the ways in which cell-lines can introduce artifacts. MCF-7 is not the only cell line that expresses heterogeneity. Other commonly used breast cancer cell lines such as T47D, BT474, and SKBR3 have also been shown to develop chromosomal alterations through cluster analysis (Rondon-Lagos et al., 2014). A recent study on the reproducibility of a perturbational assay in anti-cancer drugs using the human mammary epithelial cell line MCF10A shows variability of findings among five research centers, although it should be noted that the observed variability can be due to multiple biological, experimental and computational factors (Niepel et al., 2019).

One interesting result of this study is the lack of concordance between the ARCHS4 and TCGA data sets. Above and beyond the obvious reasons for differences between *in vivo* cancer tissue and a larger collection of *in vitro* studies with varying experimental conditions, there are likely differences introduced from the data analysis pipelines. Nonetheless, it remains surprising that even at the basic level of sorting by variant genes, there was very little in common with other MCF7-based studies. Correlation based approaches are often used on large data sets to find commonly expressed genes - and this function is built into ARCHS4 - but in this application, the size of the data set did not appear to compensate for the increased noise when it came to teasing out possible interactions.

Our examination of the smaller data set based on MCF-7 cells treated with estradiol and breast cancer tissues shows that although there are some conserved genes between the two networks, the majority of genes were non-overlapping. This issue has been raised in other studies: for instance, the Wellcome Sanger Institute used the CRISPR-cas9 screens on 324 human cell lines to priority gene targets for 30 cancer types, and found 628 priority targets (Behan et al., 2019); however, the vast difference between these cell lines and *in vivo* patient data meant that at least some of the targets predicted by CRISPR-cas9 screens were irrelevant to human cancer biology (Lyu et al., 2020). In our study, the substantial differences in co-expression networks between the MCF-7 cell line and human breast cancer tissues could have many explanations - cell line evolution of the MCF7 cells after multiple generations, as well as the increase in cis-regulated expression in *in vivo* cancer, and technical differences between the transcriptomic approaches. Nonetheless, the heterogeneity among BCRA patients likely contributes to a great deal to the difference. Our study did, like many studies of *in vitro* tumors, underscore that tumors contain multiple cell-types, as evidenced by the presence of a module of immune-related genes.

Similarly, our finding of marked correlations amongst genes based on chromosome distance within the TCGA data set, but not in the GSE data set, suggests that correlation based approaches used on cancer-derived tissues requires caution, as cis-activation (or uncontrolled transcription) will cause markedly strong associations not driven by a common transcription factor binding sites or pathways (Garcia-Cortes et al., 2020), and this complicates any interpretation of the scaled connectivity score or assuming any correlation reflects a specific interaction. The observation of cis-activated genes in the TCGA data set and not in the GSE data set could be due to the difference in dynamic range between RNA-seq and microarray. However, these correlations could also be *biologically* meaningful, potentially caused by pervasive copy number variations within the 19p13 chromosome region in breast cancer tissues, and the increased transcription almost certainly has biological consequences.

There are several limitations in our study: although WGCNA is a powerful bioinformatics method, it can result in false positives and spurious correlations. We addressed this issue by examining the gene modules resulting from the WGCNA algorithm for biologically meaningful annotations. Another limitation is the shortcoming of annotation databases, as mentioned in our previous publication (Maertens et al., 2020). Finally, the difference between MCF-7 and BRCA we observed could also be accounted for by the difference in mRNA sequencing technologies. The MCF-7 GSE50705 data set was measured with the microarray platform Affymetrix Human Genome U133 Plus 2.0 Array, and the BRCA TCGA data set was measured with the RNA seq technology Illumina HiSeq 2000 RNA Sequencing Version 2. RNA-seq is more sensitive than microarray to low-abundance transcripts (Wang et al., 2014), and the concordance between RNA-seq and microarray technologies can vary from low to high (Zhao et al., 2014; Trost et al., 2015).

Nevertheless, our study indicates that both models have limitations: MCF-7 cells lack genetic diversity and are known to have a significant lack of reproducibility; at the same time *in vivo* tumors will have greater cellular heterogeneity and artifacts intrinsic to cancers, such as greater cis-regulation. This is perhaps a useful reminder of the truism that all models are wrong, but some models are useful - and that the models are more useful when we know in what ways they are likely to mislead us.

DATA AVAILABILITY STATEMENT

The ARCHS4 MCF-7 data set was obtained from the ARCHS4 database. The GSE50705 MCF-7 data set was obtained from the GEO database. The BRCA data set was obtained from the FireBrowse database. R codes for our data analysis are available on GitHub. ARCHS4 database: https://maayanlab.cloud/archs4/data.html GEO database: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50705 FireBrowse database: http://firebrowse.org/?cohort=BRCA&download_dialog=true GitHub: https://github.com/vy-p-tran/Similarities-and-differences-in-gene-expression-networks-between-MCF7-and-BRCA.

AUTHOR CONTRIBUTIONS

VT: methodology, data analysis, writing -original draft, review and editing. RK: support data analysis. MM: support data analysis. TH: review and editing. AM: conceptualization, supervision, review and editing.

FUNDING

VT was supported by NIEHS training grant (T32 ES007141).

ACKNOWLEDGMENTS

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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.674370/full#supplementary-material

REFERENCES

Begley, C. G., and Ellis, L. M. (2012). Drug Development: Raise Standards for Preclinical Cancer Research. *Nature* 483 (7391), 531–533. doi:10.1038/483531a

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., et al. (2019). Prioritization of Cancer Therapeutic Targets Using CRISPR-Cas9 Screens. *Nature* 568 (7753), 511–516. doi:10.1038/s41586-019-1103-9

Cadenas, C., van de Sandt, L., Edlund, K., Lohr, M., Hellwig, B., Marchan, R., et al. (2014). Loss of Circadian Clock Gene Expression Is Associated with Tumor Progression in Breast Cancer. *Cell Cycle* 13 (20), 3282–3291. doi:10.4161/15384101.2014.954454

Cancer Genome Atlas Network (2012). Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* 490 (7418), 61–70. doi:10.1038/nature11412

Coker, E. A., Mitsopoulos, C., Tym, J. E., Komianou, A., Kannas, C., Di Micco, P., et al. (2019). canSAR: Update to the Cancer Translational Research and Drug Discovery Knowledgebase. *Nucleic Acids Res.* 47 (D1), D917–D922. doi:10.1093/nar/gky1129

Comsa, S., Cimpean, A. M., and Raica, M. (2015). The Story of MCF-7 Breast Cancer Cell Line: 40 years of Experience in Research. *Anticancer Res.* 35 (6), 3147–3154.

Dai, X., Cheng, H., Bai, Z., and Li, J. (2017). Breast Cancer Cell Line Classification and its Relevance with Breast Tumor Subtyping. *J. Cancer* 8 (16), 3131–3141. doi:10.7150/jca.18457

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 (6485), 1468–1473. doi:10.1126/science.aay0939

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *Plos Biol.* 13 (6), e1002165. doi:10.1371/journal.pbio.1002165

Garcia-Cortes, D., de Anda-Jauregui, 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

Gillet, J.-P., Varma, S., and Gottesman, M. M. (2013). The Clinical Relevance of Cancer Cell Lines. *J. Natl. Cancer Inst.* 105 (7), 452–458. doi:10.1093/jnci/djt007

Hartung, T. (2007). Food for Thought ... On Cell Culture. *ALTEX* 24 (3), 143–147. doi:10.14573/altex.2007.3.143

Hartung, T. (2013). Look Back in Anger - what Clinical Studies Tell Us about Preclinical Work. *ALTEX* 30 (3), 275–291. doi:10.14573/altex.2013.3.275

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *Plos Med.* 2 (8), e124. doi:10.1371/journal.pmed.0020124

Jain, N., Nitisa, D., Pirsko, V., and Cakstina, I. (2020). Selecting Suitable Reference Genes for qPCR Normalization: A Comprehensive Analysis in MCF-7 Breast Cancer Cell Line. *BMC Mol. Cel. Biol.* 21 (1), 68. doi:10.1186/s12860-020-00313-x

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* 8 (1), 118–127. doi:10.1093/biostatistics/kxj037

Kleensang, A., Vantangoli, M. M., Odwin-DaCosta, S., Andersen, M. E., Boekelheide, K., Bouhifd, M., et al. (2016). Genetic Variability in a Frozen Batch of MCF-7 Cells Invisible in Routine Authentication Affecting Cell Function. *Sci. Rep.* 6, 28994. doi:10.1038/srep28994

Koual, M., Tomkiewicz, C., Cano-Sancho, G., Antignac, J. P., Bats, A. S., and Coumoul, X. (2020). Environmental Chemicals, Breast Cancer Progression and Drug Resistance. *Environ. Health* 19 (1), 117. doi:10.1186/s12940-020-00670-2

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), 1366. doi:10.1038/s41467-018-03751-6

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

Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C. C., et al. (2019). Update of the FANTOM Web Resource: Expansion to Provide Additional Transcriptome Atlases. *Nucleic Acids Res.* 47 (D1), D752–D758. doi:10.1093/nar/gky1099

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., et al. (2015). Gateways to the FANTOM5 Promoter Level Mammalian Expression Atlas. *Genome Biol.* 16, 22. doi:10.1186/s13059-014-0560-6

Lyu, J., Li, J. J., Su, J., Peng, F., Chen, Y. E., Ge, X., et al. (2020). DORGE: Discovery of Oncogenes and tumoR Suppressor Genes Using Genetic and Epigenetic Features. *Sci. Adv.* 6 (46). doi:10.1126/sciadv.aba6784

Maertens, A., Tran, V., Kleensang, A., and Hartung, T. (2018). Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated with Bisphenol A Dose-Response. *Front. Genet.* 9, 508. doi:10.3389/fgene.2018.00508

Maertens, A., Tran, V. P., Maertens, M., Kleensang, A., Luechtefeld, T. H., Hartung, T., et al. (2020). Functionally Enigmatic Genes in Cancer: Using TCGA Data to Map the Limitations of Annotations. *Sci. Rep.* 10 (1), 4106. doi:10.1038/s41598-020-60456-x

Makki, J. (2015). Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clin. Med. Insights Pathol.* 8, 23–31. doi:10.4137/cpath.s31563

Nelson-Rees, W. A., Flandermeyer, R. R., and Hawthorne, P. K. (1974). Banded Marker Chromosomes as Indicators of Intraspecies Cellular Contamination. *Science* 184 (4141), 1093–1096. doi:10.1126/science.184.4141.1093

Niepel, M., Hafner, M., Mills, C. E., Subramanian, K., Williams, E. H., Chung, M., et al. (2019). A Multi-Center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. *Cel. Syst.* 9 (1), 35–48. doi:10.1016/j.cels.2019.06.005

Rondón-Lagos, M., Verdun Di Cantogno, L., Marchiò, C., Rangel, N., Payan-Gomez, C., Gugliotta, P., et al. (2014). Differences and Homologies of Chromosomal Alterations within and between Breast Cancer Cell Lines: A Clustering Analysis. *Mol. Cytogenet.* 7 (1), 8. doi:10.1186/1755-8166-7-8

Schweppe, R. E., Klopper, J. P., Korch, C., Pugazhenthi, U., Benezra, M., Knauf, J. A., et al. (2008). Deoxyribonucleic Acid Profiling Analysis of 40 Human Thyroid Cancer Cell Lines Reveals Cross-Contamination Resulting in Cell Line Redundancy and Misidentification. *J. Clin. Endocrinol. Metab.* 93 (11), 4331–4341. doi:10.1210/jc.2008-1102

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 (11), 2498–2504. doi:10.1101/gr.1239303

Shioda, T., Rosenthal, N. F., Coser, K. R., Suto, M., Phatak, M., Medvedovic, M., et al. (2013). Expressomal Approach for Comprehensive Analysis and Visualization of Ligand Sensitivities of Xenoestrogen Responsive Genes. *Proc. Natl. Acad. Sci.* 110 (41), 16508–16513. doi:10.1073/pnas.1315929110

Sweeney, E. E., McDaniel, R. E., Maximov, P. Y., Fan, P., and Jordan, V. C. (2012). Models and Mechanisms of Acquired Antihormone Resistance in Breast Cancer: Significant Clinical Progress Despite Limitations. *Horm. Mol. Biol. Clin. Investig.* 9 (2), 143–163. doi:10.1515/hmbci-2011-0004

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 Data sets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Trost, B., Moir, C. A., Gillespie, Z. E., Kusalik, A., Mitchell, J. A., and Eskiw, C. H. (2015). Concordance between RNA-Sequencing Data and DNA Microarray Data in Transcriptome Analysis of Proliferative and Quiescent Fibroblasts. *R. Soc. Open Sci.* 2 (9), 150402. doi:10.1098/rsos.150402

Trujillo, C. A., and Muotri, A. R. (2018). Brain Organoids and the Study of Neurodevelopment. *Trends Mol. Med.* 24 (12), 982–990. doi:10.1016/j.molmed.2018.09.005

Tym, J. E., Mitsopoulos, C., Coker, E. A., Razaz, P., Schierz, A. C., Antolin, A. A., et al. (2016). canSAR: an Updated Cancer Research and Drug Discovery Knowledgebase. *Nucleic Acids Res.* 44 (D1), D938–D943. doi:10.1093/nar/gkv1030

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., et al. (2014). The Concordance between RNA-Seq and Microarray Data Depends on Chemical Treatment and Transcript Abundance. *Nat. Biotechnol.* 32 (9), 926–932. doi:10.1038/nbt.3001

Xiao, L., Chang, A. K., Zang, M. X., Bi, H., Li, S., Wang, M., et al. (2014). Induction of the CLOCK Gene by E2-ERalpha Signaling Promotes the Proliferation of Breast Cancer Cells. *PLoS One* 9 (5), e95878. doi:10.1371/journal.pone.0095878

Yu, K., Chen, B., Aran, D., Charalel, J., Yau, C., Wolf, D. M., et al. (2019). Comprehensive Transcriptomic Analysis of Cell Lines as Models of Primary Tumors across 22 Tumor Types. *Nat. Commun.* 10 (1), 3574. doi:10.1038/s41467-019-11415-2

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One* 9 (1), e78644. doi:10.1371/journal.pone.0078644

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

frontiers in Genetics

REVIEW
published: 11 June 2021
doi: 10.3389/fgene.2021.694468

Assembling Disease Networks From Causal Interaction Resources

Gianni Cesareni[1], *Francesca Sacco*[1] *and Livia Perfetto*[2]*

[1] Department of Biology, University of Rome Tor Vergata, Rome, Italy, [2] Department of Biology, Fondazione Human Technopole, Milan, Italy

The development of high-throughput high-content technologies and the increased ease in their application in clinical settings has raised the expectation of an important impact of these technologies on diagnosis and personalized therapy. Patient genomic and expression profiles yield lists of genes that are mutated or whose expression is modulated in specific disease conditions. The challenge remains of extracting from these lists functional information that may help to shed light on the mechanisms that are perturbed in the disease, thus setting a rational framework that may help clinical decisions. Network approaches are playing an increasing role in the organization and interpretation of patients' data. Biological networks are generated by connecting genes or gene products according to experimental evidence that demonstrates their interactions. Till recently most approaches have relied on networks based on physical interactions between proteins. Such networks miss an important piece of information as they lack details on the functional consequences of the interactions. Over the past few years, a number of resources have started collecting causal information of the type protein A activates/inactivates protein B, in a structured format. This information may be represented as signed directed graphs where physiological and pathological signaling can be conveniently inspected. In this review we will (i) present and compare these resources and discuss the different scope in comparison with pathway resources; (ii) compare resources that explicitly capture causality in terms of data content and proteome coverage (iii) review how causal-graphs can be used to extract disease-specific Boolean networks.

Keywords: network medicine, logic modeling, causality resources, prior knowledge network, causal interactions

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Adriaan-Alexander Ludl,
University of Bergen, Norway
Ylva Ivarsson,
Uppsala University, Sweden

***Correspondence:**
Livia Perfetto
livia.perfetto@fht.org

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

Received: *13 April 2021*
Accepted: *19 May 2021*
Published: *11 June 2021*

Citation:
Cesareni G, Sacco F and Perfetto L (2021) Assembling Disease Networks From Causal Interaction Resources. Front. Genet. 12:694468. doi: 10.3389/fgene.2021.694468

INTRODUCTION

The term precision or personalized medicine reflects the motivation of using high content molecular information for disease diagnosis and for the design of effective personalized therapies (Ginsburg and Phillips, 2018). Advances in experimental methods, such as deep sequencing and high content proteomics (Nilsson et al., 2010; Goldman and Domschke, 2014), have enabled the comprehensive assessment of a patient's molecular profile in a time- and cost-effective manner. Patients' genomic and expression profiles are becoming increasingly more important diagnostic readouts and are likely to become soon compatible with clinical practice in most public hospitals. Whether patients can benefit from this promising treatment strategy on a large scale still remains uncertain (Zhang et al., 2020).

One main limitation of this genomic-approach is the lack of an effective strategy to extract clinically relevant information from these dense and noisy datasets. Network representation of biological complexity and graph theory are playing an increasingly important role in dealing with the intricacy of human physiology and pathology. Network-based approaches are used, in the context of a relatively new discipline dubbed "network medicine," to address the interplay of the molecular mechanisms underlying complex diseases (Barabási et al., 2011). The main idea behind is that a network, where interactions between its components are represented in the form of a graph, provides a powerful mathematical framework for analysis and visualization of experimental results.

According to this vision, gene products govern cell physiology by interacting in a large interconnected network whose equilibrium is responsible for the dynamic homeostasis of "healthy" cells. The network properties are believed to be rather robust and resilient to perturbations of many of the nodes and some of their connecting edges. A few nodes of the network, however, are quite sensitive and their knock out or hyperactivation may cause large changes of the network properties leading to disease (Brinkman et al., 2006). Alternatively, and more frequently, a combination of alterations of the activities of nodes that, on their own, have little effect may synergize to alter the properties of sensitive regions of the network thereby leading to a pathological condition (Barabási et al., 2011). It is anticipated that the overlay of a patient genomic profile onto such a comprehensive cell network, or part of it, will provide a framework to help in patient diagnosis and therapy choice.

Cell networks are assembled from experimental evidence of physical or functional links between biological entities. This information is often difficult to retrieve and organize as it is dispersed in millions of scientific reports. In addition, experimental results are mainly reported in natural language that is not easily processed by computers. Thus, network approaches mostly rely on the work of database curators that, assisted by natural language processing tools, identify relevant reports in literature repositories and annotate the interaction evidence in a structured machine-readable format. Over the past few decades different players have engaged in the task of capturing evidence of protein interactions. Protein interaction is a generic term including different types of physical and functional relationships between proteins as identified by diverse experimental approaches (Zhou et al., 2016). Databases that aim at capturing this information have distinct focus and adopt models that best adapt to their scope. As a consequence, comparing and merging the data from the diverse databases is made difficult by the heterogeneity of the interaction types and the models to represent them.

A recent review by Touré et al. (2020) has discussed the different types of protein interaction resources focusing on a comparison of the adopted data structures and the data exchange and conversion procedures. Here we go over the models that have been adopted to represent experimental evidence of protein relationships mediating physiological and pathological processes. More specifically, we confront physical and causal interactions by briefly describing their characteristics and the resources that aim at capturing and organizing the two different interaction types.

We focus on resources that annotate causal interactions modeled as "activity-flow" (AF) networks (**Figure 1**) by considering and comparing their coverage and merits in different use cases. We will also present tools and strategies that make use of networks assembled from prior knowledge (PKNs) to produce executable logic models replicating phenotypes of clinical relevance. Finally, we discuss whether the evidence on causal relationships that is presently reported in the scientific literature is adequate to assemble a cell network of sufficiently high coverage and accuracy to be of clinical relevance.

RESOURCES CAPTURING SIGNALING INTERACTIONS

Physical and Causal Interactions

Proteins interact in the cell forming a complex ordered functional mesh. Some of these interactions are necessary for maintaining cell organization whereas others support the cell response to internal and external stimuli, and are often transient (Acuner Ozbabacan et al., 2011). A variety of approaches suitable for high throughput analysis have been used to reveal the physical contacts between proteins without informing on the dynamic of signal propagation (Xing et al., 2016). More than 400 K "physical interactions" between human proteins have been reported in the literature by using these methods and for 85% of the proteins in the human proteome we know at least one physical partner in public databases (Orchard et al., 2014; Oughtred et al., 2021). Physical interactions are symmetrical by nature and, having no directionality, are represented as "undirected" graphs (**Figure 1A**). Transient signaling interactions, on the other hand, are often short lived and as such may not be revealed by the methods developed for physical interactions. They are often causal as one of the partners, the regulator, causes a functionally relevant modification of the target protein. These latter types of interactions may be modeled in two ways that are often referred to as "process descriptions" (PD) and "activity-flow" (AF) (**Figure 1**) (Le Novère, 2015; Türei et al., 2016; Touré et al., 2020).

Let us consider, as an example, the experimental observation that the phosphatase *PTPRJ* (DEP1) binds to *MAPK1* (ERK2) and inactivates it by removing a phosphate (Sacco et al., 2009). As shown in **Figure 1B**, an "undirected PPI" model represents this statement as a link between *PTPRJ* and *MAPK1* that has no direction. The "activity-flow" (AF) model, on the other hand, renders this information as a binary interaction where the two proteins are connected by an edge that has direction from *PTPRJ* to *MAPK1* and a sign which is graphically symbolized with a specific edge-form or color. This representation captures the evidence that *PTPRJ* is the regulator and *MAPK1* the target and that this interaction has the consequence of inactivating *MAPK1*. AF models offer the advantage of being represented as a set of binary interactions in a signed directed graph which is more informative than an undirected graph. Finally, the PD model captures additional mechanistic details. In this representation the

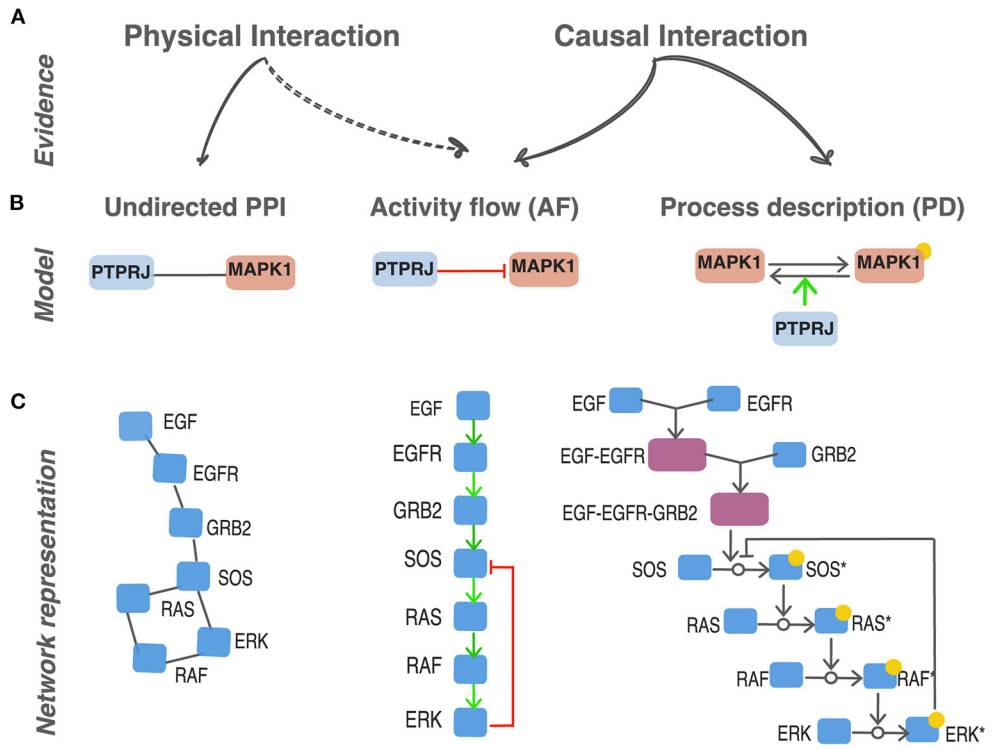


FIGURE 1 | Different representations of protein interactions. **(A)** Experimental methods can either provide evidence that support a physical contact between two proteins to form a complex (physical interaction) or a modulation of the activity of a target protein caused by the activity of a parent protein (causal interaction). **(B)** Different graphical representation of the same biological statement: PTPRJ dephosphorylates and inhibits MAPK1 (Sacco et al., 2009). Three distinct models to represent protein relationships supported by different experimental evidence: undirected PPI, activity-flow and process description. **(C)** The EGFR signaling pathway represented as an undirected protein-protein interaction network (PPI), as activity-flow network (AF) and a process-description network (PD). *indicates the modified form of a given protein node.

target entity, *MAPK1* in our example, is split into two nodes representing the phosphorylated and unphosphorylated forms of the protein. The two forms are connected by a directed edge symbolizing the transition from one form to the other. The activity of the regulatory protein *PTPRJ* is represented as an edge promoting the removal of the phosphate from *MAPK1*. A limitation of this latter model is that the impact of the phosphorylation on the activation status of *MAPK1* cannot be directly derived; it is only implicit as it can only be inferred from the reconstruction of the downstream chains of reactions.

The different representations serve different purposes and answer different questions. For instance, analysis of highly connected regions of an undirected protein interaction network may reveal the formation of macromolecular complexes (Wang et al., 2009; Havugimana et al., 2012). Similarly, the function of a protein that is trapped in a subnetwork formed by proteins that are annotated to a specific biological process may provide hints on its function (Oliver, 2000). On the other hand, process description and activity-flow networks are appropriate to sketch the information flow from a receptor sensing a stimulus to activation of a transcription factor driving phenotype modulation.

Another major difference between physical and causal interaction datasets is proteome coverage as the latter have significantly lower coverage. This is partly due to incomplete curation of reported experimental evidence and partly to the lack of appropriate high throughput experimental approaches to reveal causal interactions on a large scale. In addition, many resources annotating PPI have, in recent years, joined their efforts forming a consortium (Orchard et al., 2014; Porras et al., 2020) for distributing curation investment and using common standards and curation rules, whereas causal resources have not reached such an agreement yet. For these reasons, many of the network approaches presently rely on networks based on physical protein interactions (PPI) (Zhang and Itan, 2019).

Approaches based on networks assembled by using information on causal relationships, however, are gaining momentum as they provide information that can be relatively easily converted into Boolean or ordinary differential equation models thus enabling users to compute the behavior of a system in different conditions (Le Novère, 2015).

Although there is no strict separation between the experimental evidence that can be captured by the different models, it is crucial to understand the data structure adopted by each resource as analyses built on information extracted from distinct databases may lead to different biological conclusions (Mubeen et al., 2019).

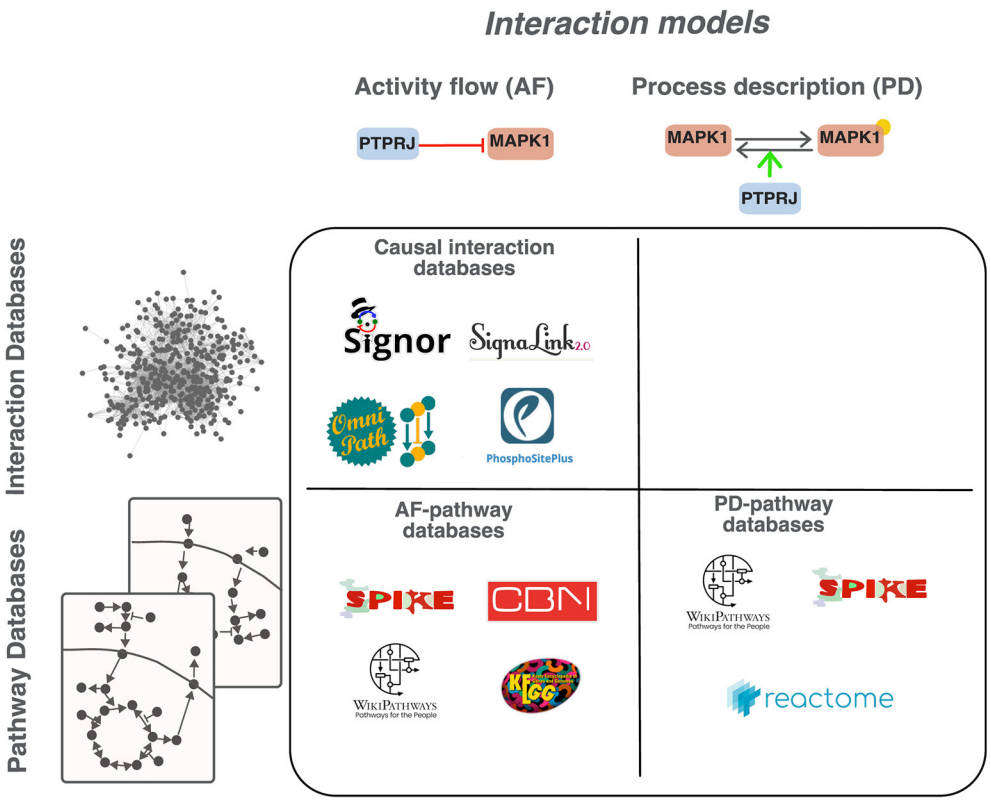


FIGURE 2 | Classification of causality resources. Resources can be grouped according to the model adopted to represent causality in AF or PD (see **Figure 1**) and according to the organization of the information in "interaction databases" or "pathway databases." In interaction databases relationships are annotated separately and not necessarily in the context of higher-level organizational structures, such as pathways. In "pathway databases" interactions are exclusively shown in the context of the pathway they participate in.

Pathway Databases and Interaction Databases

Cell physiology is governed by a large connected network of physical and causal interactions. Nevertheless, biologists sometimes prefer to consider the cell model as an ensemble of unconnected pathways that, in a first approximation, function in isolation and do not crosstalk. However, this approximation neglects the effects of the cell network as a whole that may significantly affect the behavior of the pathway subnetworks. Although networks are useful abstractions, their functional integration into a cell model remains an important challenge. Capturing the experimental information for the assembly of protein interaction networks from primary literature data is an intimidating task. To assist scientists, over the past 20 years, a number of resources have set out to annotate an excerpt of the experimental facts related to protein interactions in structured formats in public repositories. However, different databases have been developed to serve different purposes, they adopt different curation policies and describe the same biological fact at different levels of abstraction and granularity.

We here focus on resources that capture causality, hereafter referred to as "causality resources" (**Figure 2**). Considering the chosen representation model, databases can be grouped into two broad classes (**Figure 2**, **Supplementary Table 1**): "interaction databases," where relationships are integrated in a global network and "pathway databases," where interactions are curated and displayed in the context of the pathway they participate in.

The three most popular pathway resources are KEGG, Reactome and WikiPathways (Kanehisa and Goto, 2000; Slenter et al., 2018; Jassal et al., 2020). Among the pathway databases those adopting AF as interaction model are KEGG, SPIKE (Paz et al., 2011) and CBN (Boué et al., 2015) (**Supplementary Table 1**). The signaling information annotated in these databases is reviewed by domain experts and covers more than 50% of the human proteome. Aside from their descriptive value in the representation of cell physiology, they have proven useful in the analysis and interpretation of -omics data when coupled with algorithmic approaches such as gene set enrichment analysis GSEA and signaling pathway impact analysis (SPIA) (Subramanian et al., 2005; Tarca et al., 2009;

Sprent, 2011). However, they do not provide an integrated picture of cell functioning as interactions are accessible only in the context of pathways and miss to offer a holistic view. Another class of resources, including Cell Collective (Helikar et al., 2012), Biomodels (Malik-Sheriff et al., 2020), the GINsim repository (Naldi et al., 2009) and the PyBoolNet repository (Klarner et al., 2017) collect assembled logical models. Briefly, these resources store models curated and tested by different groups for specific projects. Users can download the models and adapt them to different purposes. However, the models do not necessarily follow a common annotation standard. As a consequence, the integration into larger models is often not straightforward.

A third class of resources, such as SIGNOR (Licata et al., 2020), SignaLink (Csabai et al., 2018), OmniPath (Türei et al., 2016; Ceccarelli et al., 2020) or PhosphoSitePlus (Hornbeck et al., 2019) annotate interactions without necessarily listing them as members of a pathway. We will refer to these with the generic term "causal interaction databases" (**Figure 2**). This organization of the interaction data, which is not pathway centric, allows users to assemble an integrated cell network where all pathways are connected, thereby allowing to monitor pathway crosstalk.

ACTIVITY-FLOW RESOURCES COMPARISON

With our contribution we intend to show how AF interactions from the different resources can be used to build logic networks to support modeling studies. To this end, we compare four major AF resources, KEGG, PhosphoSitePlus, SignaLink and SIGNOR.

These resources were selected as they are open-source, established, and popular as evinced from citation counts. In addition, they exclusively adopt "activity-flow" as a representation model (**Supplementary Table 1**). These databases are, however, highly heterogeneous in scope, and do not follow a common standard for the annotation and the export of the data (Dräger and Palsson, 2014). To address this issue the proteomic standard initiative for molecular interaction (PSI-MI) (Orchard, 2014) and the Gene Regulation Ensemble Effort for the Knowledge Commons (GREEKC) (https://www.greekc.org/) communities have recently developed CausalTAB, a common standard for exchange of causal information (Perfetto et al., 2019). However, of the four databases considered here, only SIGNOR presently offers to download its curated dataset in this format. As a consequence, the organization of the datasets for the comparison reported here turned out to be a substantial effort (Supplementary File 1 in **Supplementary Material**). To facilitate the task of integrating the information that can be downloaded from the different resources, OmniPath has embarked on a project aimed at merging the causal information from a large number of primary resources. This resource was also included in our analysis.

We designed this comparison to help non-computational scientists to incorporate computational modeling into their experimental practice. We point out that the comparison is limited to the portion of AF interactions that satisfy specific criteria and that some datasets (e.g., KEGG) might represent a subset of the total number of interactions that are annotated in the database.

The four primary resources considered here have a different focus and include different entity types as nodes in the network. For instance, KEGG and SIGNOR also annotate complexes. In addition, SIGNOR considers a wider range of entities including "phenotypes," "stimuli" and "chemicals." SignaLink and SIGNOR also curate indirect interactions. To harmonize the data in order to attain a fair comparison, we filtered the datasets to retain only direct causal interactions between human protein pairs (Supplementary File 1 in **Supplementary Material**). In addition, we only considered those relationships that are annotated with a literature reference. In this first comparison two entries are considered coincident if they involve the same protein pair with matching directionality, irrespective of the effect (activation/inhibition) of the interaction.

In **Figure 3A** we show in an UpSet plot (Lex et al., 2014) the number of causal relationships that are annotated only in each of the databases or are common to all dataset combinations. We first notice that the four primary databases are largely complementary as more than 70% of the information is captured by only one database while fewer than 4% of the interactions (510) are annotated in three or four resources. SIGNOR with its 9,845 entries is the primary database with the highest number of entries. Still 5,003 entries of the remaining three primary resources are not in SIGNOR (**Figure 3A**). This complementarity of the datasets has motivated the OmniPath team to integrate all the causal information in a single dataset.

In **Figure 3B** we have reported the results of the comparison as Venn diagrams. Each resource is represented as a circle of different color whose size is proportional to data content. The circles overlap for an area that is proportional to the number of interactions that are present in both databases. The largest overlap between primary databases is observed in the comparison between SIGNOR and PhosphoSitePlus as both resources have put investment in the coverage of phosphorylation reactions. As PhosphoSitePlus does not curate other types of causal relationships its overlap with the other resources is negligible.

OmniPath, which integrates information from more than 100 different primary databases, is by far the most inclusive resource. However, although OmniPath claims full integration of interaction data, only 39% of the KEGG dataset is included in OmniPath (**Figure 3B**). This is because the standard OmniPath dataset only takes into consideration referenced protein relationships, whereas a large fraction of KEGG interactions is not linked to the manuscripts providing the supporting experimental evidence. Other inconsistencies are the consequence of an infrequent synchronization of the OmniPath dataset with the release of the primary resources. Of note over 20% of the interactions in SIGNOR are not present in OmniPath (**Figure 3B**).

By adding to the OmniPath dataset the missing data from the four primary resources it is possible to assemble a network of causal interactions linking nearly 5,800 proteins (28% of the proteome) connected by 27,040 edges (**Supplementary Table 2**). Eighty four percent of these are only curated in one or two resources, while the remaining 16% in three or more.

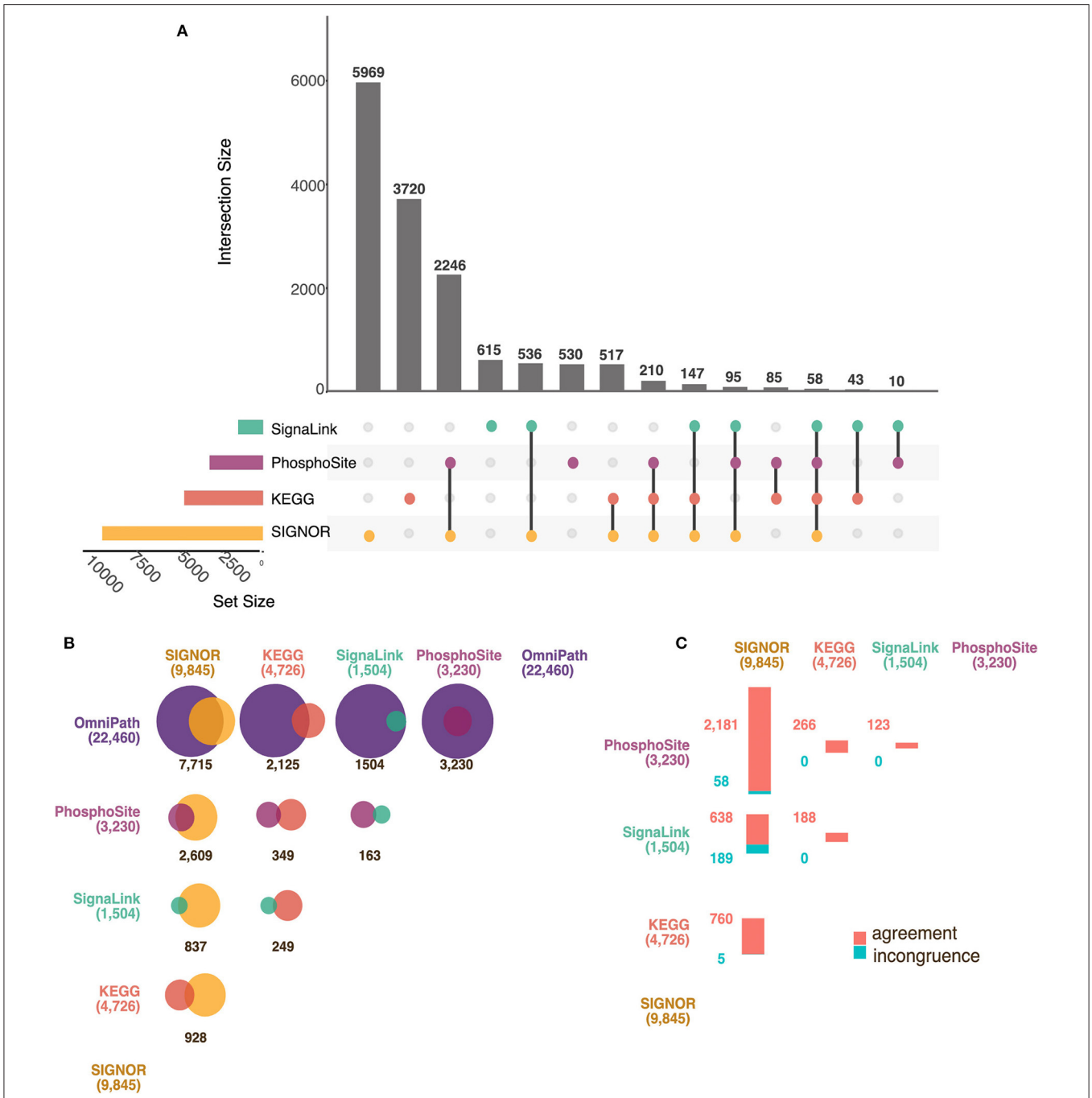


FIGURE 3 | Comparison of AF Databases. **(A)** UpSet Plot showing the overlaps between four primary AF resources: SIGNOR (in yellow), KEGG (in red), PhosphoSitePlus (in purple) and SignaLink (in green). The vertical bars show the number of intersecting protein pairs (regulator-target) between resources, identified as connected colored circles below the histogram. The length of the horizontal bars is proportional to the dataset size of each resource. As an example, PhosphoSitePlus, SIGNOR and KEGG share 210 interactions. **(B)** Proportional Venn diagrams showing the overlap between the datasets of the four primary AF resources and OmniPath: SIGNOR (in yellow), KEGG (in red), PhosphoSitePlus (in purple), SignaLink (in green) and OmniPath (dark purple). Individual set sizes are in parenthesis. **(C)** Matrix of bar plots showing the number of interactions between pairs of proteins whose effect, up- down-regulation is annotated in an opposite way in each pair of primary resources. Agreement and disagreement are shown in red and blue, respectively.

Consistency of Data Curation in the Different Resources

The conclusion of the analysis in the previous section is that, in order to increase coverage, users should consider collating datasets from different resources. However, in large curation efforts, in some instances, the same experimental evidence can lead different curators to different interpretations. In addition, experimental reports addressing the same biological

question reach sometimes contrasting conclusions. Thus, it is not surprising to observe that a causal relation between protein A and protein B is annotated as activating by one database and inactivating by another. However, this represents a problem in the assembly of AF networks from an integrated dataset. To investigate how serious this issue was, we next assessed the fraction of causal relationships that are inconsistently annotated.

Ninety five percent of the edges that are curated by more than one database are consistently associated with either up- or down- regulation (**Supplementary Table 2**). About 3,200 interactions are annotated with the same consensus effect in at least three resources, thereby accounting for a high-confidence subset of causal interactions. Conversely, 5% of the pairs are associated with both up- and down- regulation in different databases. Besides trivial curation errors, some discrepancies might reflect differences in the annotation policies of the different primary resources. Alternatively, it could be the consequence of conflicting literature reports or complex effects of an interaction leading to clashing consequences on the target protein function. For instance, *GSK3*-mediated *MAF* phosphorylation leads both to transcriptional activation and to degradation of the target (Rocques et al., 2007).

To quantify this lack of consistency, we compared the datasets from the four primary repositories. For this analysis, we first filtered out from each dataset those pairs that in each database are annotated with both a positive and a negative effect as in the *GSK3-MAF* example mentioned earlier (internal "inconsistencies"). As shown in **Figure 3C**, the percentage of incongruent pairs between DBs is relatively small, SignaLink and SIGNOR are the two repositories showing the highest number (and percentage) of contradicting interaction annotation. This subset of conflicting pairs has already been discussed (Perfetto et al., 2016) and can be explained by the differences in annotation granularity adopted by the two resources. For instance, SIGNOR annotates the mechanisms (such as ubiquitination, phosphorylation, etc.) involved in the interaction and, when provided, also the modified residues, while SignaLink only provides information about the causal effect.

DISEASE NETWORKS

Assembly of Large Disease Networks

We next asked whether the combined causal information captured by the different primary resources is sufficiently complete to be used to assemble informative disease networks linking most of the genes that are found mutated in patients. We used the expert curated information of the Cancer Gene Census (Sondka et al., 2018) that annotates 389 cancer types with lists of genes observed to be significantly mutated in cancers. Similarly, we used the information collected by the DisGeNET resource (Piñero et al., 2020) to download lists of gene-disease associations (GDAs) for 4,713 polygenic diseases (DisGeNET score > 0.5). These lists have different sizes ranging from one up to 83 genes in the case of "Malignant neoplasm of breast." We filtered the lists by selecting diseases with at least two genes annotated, 163 and 823 diseases in Cancer Gene Census and DisGeNET, respectively (**Figure 4**). These disease-gene lists were used to query the AF resources for interactions linking the disease genes. We also included in the network the proteins that by forming a bridge between the query proteins, allow to connect them. The rationale for inclusion of "bridge proteins" is further discussed in the next paragraph.

The results of the approach are shown in **Figure 4** as violin plots illustrating the distribution in the number of edges in the networks assembled by this automatic procedure. As proteome coverage is far from being complete, not all disease gene lists could be connected to form a network in the different resources. Above each violin we have indicated the number of diseases for which it was possible to assemble a network by interrogating each of the resources together with the average network size (average number of edges). As a larger coverage corresponds to a higher number of connections, retrieving interactions from SIGNOR allowed the assembly of a higher number of disease networks (446 and 107 from the DisGeNET and Cancer Gene Census lists, respectively) in comparison with the other primary resources. Similarly, SIGNOR-derived networks tend to be larger, in terms of number of connections. OmniPath that integrates all the primary databases allows an even higher coverage (both in terms of number of diseases and in average network size). However, as already noted, by integrating the data of the four primary resources and OmniPath an even higher number of disease genes could be assembled into connected networks.

It is finally to note that among the resources compared here, only SIGNOR and OmniPath have implemented a web tool to extract connections between a list of input proteins and to return the results either in graph or table format. To apply a similar procedure to the dataset offered by the other databases dataset-manipulation and/or parsing is necessary.

The Gray Platelet Syndrome

As an example of the results that one obtains by the procedure detailed in the previous section we will describe in more detail the networks retrieved in the case of the Gray Platelet syndrome (GPS). GPS is a rare recessive autoimmune disorder characterized by a variety of symptoms including the absence of platelet alpha-granules, bleeding disorders and bone marrow fibrosis (Gunay-Aygun et al., 2010). *NBEAL2* is the most frequently mutated gene in patients affected by this condition. However, due to the rarity of GPS, the molecular mechanisms underlying the disease are still poorly understood (Gunay-Aygun et al., 2011). We first assembled a list of 36 GPS associated genes and used this list to interrogate the different primary datasets and OmniPath (**Figure 5A**). As shown in **Figure 5**, in network assembly we also included "bridge proteins," nodes that link two "disease proteins." One advantage of using "bridge proteins" is that they allow for the expansion of the search space and for the retrieval of a graph connecting most of the disease proteins. By applying the aforementioned method, we succeeded in retrieving networks with a relevant (>2) number of interactions only by interrogating SIGNOR and OmniPath (**Figures 5B,C**).

By combining all the interactions, we obtain the most detailed graph incorporating 41 nodes and 91 edges and connecting 19 of the 36 input proteins (**Figure 5D**). As phenotypes are entities in the SIGNOR dataset, the integrated network also includes

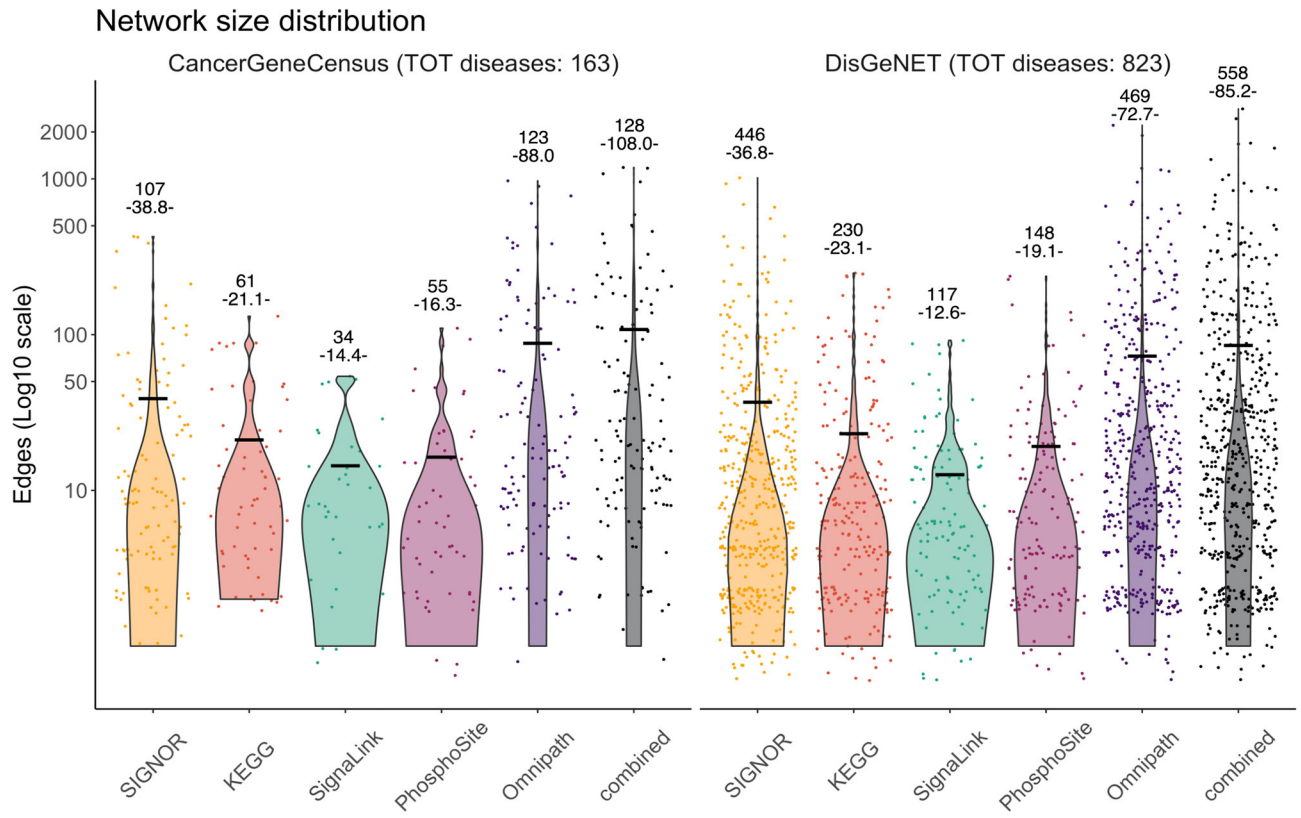


FIGURE 4 | Violin plots illustrating the size distribution of the disease-networks that can be assembled by linking disease genes via causal interactions annotated in AF Databases. Disease-networks were assembled by using gene-disease associations (GDAs) downloaded from the Cancer Gene Census (left panel) (Sondka et al., 2018) and from DisGeNET selecting GDAs with score > 0.5 (right panel) (Piñero et al., 2020). The disease-networks also include proteins that directly connect disease gene products (bridge proteins) (Lo Surdo et al., 2018). Only diseases with at least two GDAs were considered in this analysis. Each dot represents a disease network and its size (y-axis) is defined as the number of edges that can be extracted from the five AF resources: SIGNOR (in yellow), KEGG (in red), PhosphoSitePlus (in purple), SignaLink (in green) and OmniPath (dark purple); and from a network derived by taking into considerations all the relationships annotated in at least one resource, combined (black). On top of each violin the total number of disease-networks that can be assembled by using the annotated causal relationships from each corresponding resource is displayed. In brackets we show the average size of the network, also indicated by a horizontal black bar.

the "Platelet alpha granule formation" phenotype. Including phenotype entities improves the readability of the graph and strengthens the biological significance of the derived network.

GPS is a rare and poorly characterized disease. The advantage of this approach is that it compensates the lack of information annotated in the literature about pathways and perturbed molecular events.

To compare the results obtained in the case of GPS to that of a highly characterized disease, we applied a similar strategy to "Malignant neoplasm of breast." This tumor type has the highest number of GDAs (83) in the DisGeNET resource. Not surprisingly, the retrieved networks are larger than the ones obtained for GPS, including 2,562, 533, 115, and 563 edges for SIGNOR, KEGG, SignaLink and PhosphoSitePlus respectively; 8,430 for OmniPath; and 11,513 for the five resources combined together. Such networks are extremely complex and difficult to interpret and might require stricter search parameters or filtering options that provide contextualization of the network (see next paragraphs).

These observations support the notion that there is no unique strategy to extract a diseases-PKN from AF repositories and the choice of a search method should be guided by quality and amount of information available for that specific pathology.

LOGIC MODELS FROM PRIOR KNOWLEDGE NETWORKS

AF networks provide mechanistic details on the information flow in a biological system in physiological and pathological conditions thereby allowing one to explore the functional consequences of modulating the activity of any specific node. However, they are of little practical value if one wants to identify the equilibrium states of a system in varying contextual conditions. Different approaches have been developed to obtain predictive models, including differential equation-based models, rule-based, Bayesian network inference and logic-based models. Despite their simplicity, logic-based models

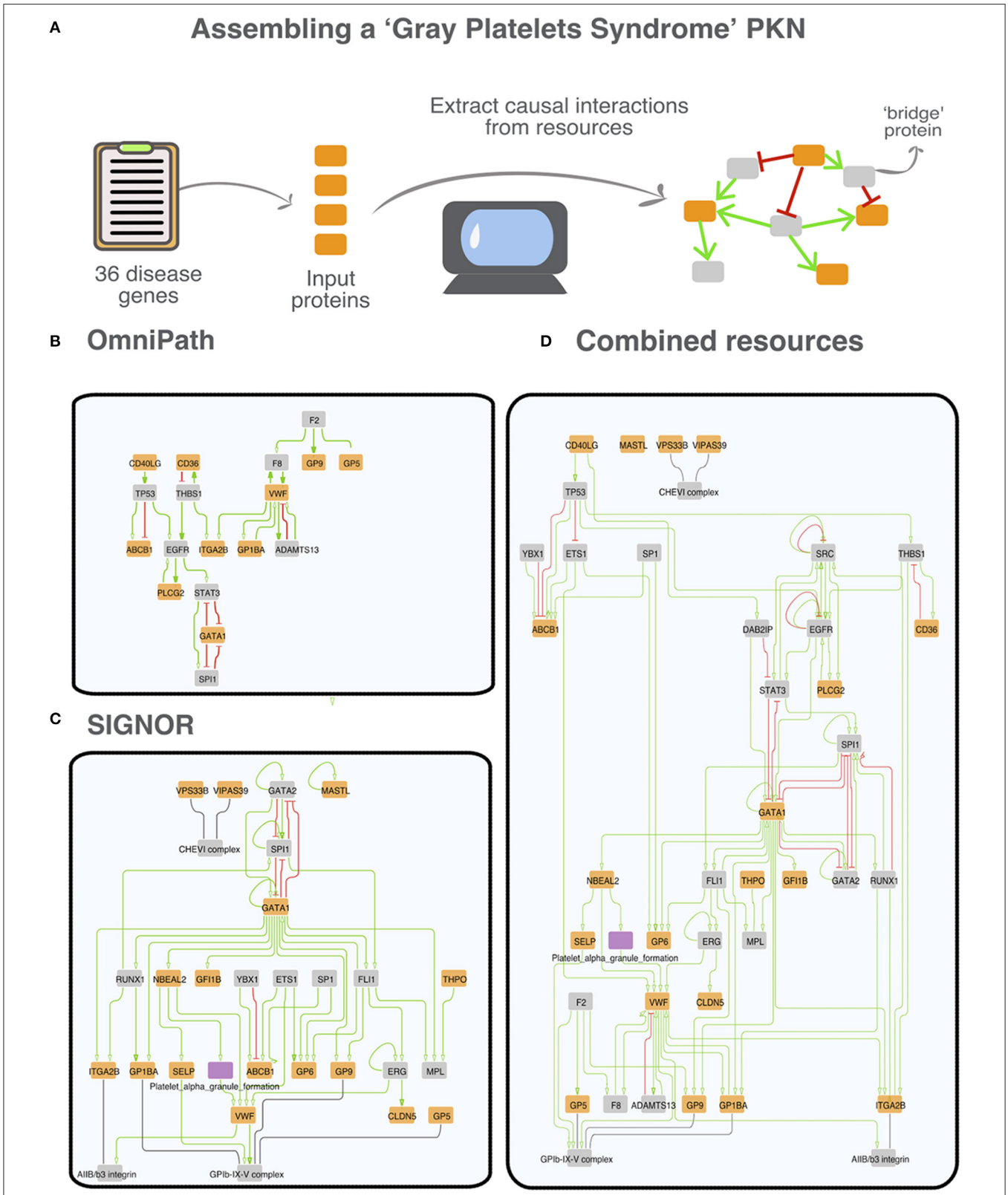


FIGURE 5 | A prior knowledge network (PKN) associated with the "Gray Platelet syndrome." **(A)** Strategy to derive the networks from the causal data in each resource. Thirty six gene-disease associations for the Gray Platelet syndrome were downloaded from MalaCards (Rappaport et al., 2017). Disease genes are used as

(Continued)

FIGURE 5 | seeds (orange nodes) to assemble the networks by searching causal resources for connecting relationships. To implement this strategy, we searched data from primary resources, from OmniPath; and from a virtual resource integrating all the datasets. Up - or down-regulations are illustrated in the graphs as green arrows and red t-shaped edges, respectively. We also included bridge proteins (gray nodes). Bridge proteins are proteins that connect at least two seed proteins (Lo Surdo et al., 2018). We were not able to obtain a significant network (>2 interactions) from KEGG, PhosphoSitePlus and SignaLink. **(B)** Network extracted from OmniPath: 18 nodes and 27 edges. **(C)** Network extracted from SIGNOR: 29 nodes and 53 edges. The purple node corresponds to the phenotype "platelet alpha granule formation," annotated in SIGNOR (Licata et al., 2020). **(D)** Network that can be derived by combining the datasets annotated by the five combined resources: 41 nodes and 96 edges.

(Boolean) have gained attention as, differently from modeling approaches based on ordinary differential equations, they can be applied to relatively large biological networks (Morris et al., 2010; Wang et al., 2012). Boolean models provide a simple yet powerful qualitative approach to describe how a system responds to contextual changes. The problem of assembling and contextualizing predictive Boolean models from prior knowledge and/or experimental data has been discussed and is further reviewed in section Conclusions and Perspectives (Vinayagam et al., 2011; Wang et al., 2012; Lages et al., 2018; Aghamiri et al., 2020; Dugourd et al., 2021).

The information embodied in activity-flow networks can be relatively easily converted into Boolean rules, where biological entities are modeled as Boolean variables whose activities are characterized by a simple On/Off behavior and where multiple incoming regulatory signals are integrated by logic gates. This qualitative approach approximates the response of a system and permits to address simple–albeit relevant- questions related to the phenotype that are favored in specific initial conditions or to the impact of a loss or gain of function mutations on any clinically pertinent phenotype.

Selvaggio and colleagues defined a logic model of the epithelial-to-mesenchymal transition that enabled the identification of new potential paths connecting microenvironmental signals to cancer cell plasticity (Selvaggio et al., 2020). Logic-based models have also been used to understand the molecular mechanisms underlying complex diseases. As an example, the group of Saez-Rodriguez has recently developed an approach combining *ex-vivo* high-throughput screenings of colon cancer biopsies with logic-based models. Their approach enabled them to generate patient-specific predictive models of apoptosis that can be used to rationally design personalized therapies (Eduati et al., 2020). Logic-based models have also been applied to explore whether and how the genomic context affects the behavior of a patient specific system. Béal et al. integrated mutation data, copy number alterations, and expression data into a breast-cancer logical model for clinical stratification of patients (Béal et al., 2018). Palma et al. built a Boolean model of acute myeloid leukemia whose predictions, once combined with patients' genomic profiles, correlate with clinical parameters, including patient life expectancy (Palma et al., 2021). Complex physiological processes such as hematopoiesis or macrophage differentiation can also be described by logic-based models of the different cell populations along the differentiation process (Collombet et al., 2017; Palma et al., 2018). Interestingly, logic-based models have also been used to discover novel anti-cancer drug combinations that efficiently kill cancer cell lines (Flobak et al., 2015).

CONCLUSIONS AND PERSPECTIVES

Resources that organize in a structured computer-readable format causal information between gene/proteins assist in the assembly of networks linking disease genes by logical connections. These in turn can be converted into logic models to predict phenotype modulation in different genomic contexts and under drug treatment.

Here we have focused on network strategies that make use of prior knowledge derived from low throughput experiments as annotated in public databases. These methods are somewhat biased as they depend on curators' decisions. It should be mentioned that alternative approaches based on reverse engineering allow researchers to draw networks in an unbiased manner by using genome wide gene expression data to infer relationships between genes (Pe'er and Hacohen, 2011). By these strategies, if two genes are co-expressed they are inferred to be functionally correlated and are linked in a gene regulatory network. Reverse engineering approaches, however, relying mostly on genome-wide expression studies, provide information on gene regulatory networks but say little about signaling networks where protein modification and modulation of stability play an important role that cannot be inferred from transcriptomics.

Although strategies based on prior knowledge have already shown some success, as reviewed here, we would like to conclude this contribution by discussing the current limits of these approaches and by identifying the areas where investment should be directed in the near future.

Incomplete Coverage

At the time of our survey only ∼28% of the proteome is integrated into a global cell network by the information captured in AF repositories. This represents a severe limitation as for many disease-genes we do not have any clue about the functional consequences of modulating their activities. This can, to some extent, be addressed by increasing the curation effort and perhaps by establishing a collaborative consortium of resources similar to the IMEx consortium in the PPI domain (Porras et al., 2020). However, we also have to accept that for many proteins we have hardly any experimental evidence about their functions, let alone their causal connections with the activity of other proteins in the cell network.

Editing Automatically Generated Models

The networks that are derived by the strategy that we have delineated here are highly connected and complex and as such sometimes difficult to understand and model. Some interactions

that are not supported by thorough evidence and repeatability or are implausible can be removed after a detailed review of the model connections by a domain expert. However, the development of automatic pruning methods is also desirable. For instance, not all the causal edges are equally supported by experimental evidence. The SIGNOR resource assigns to each causal relationship a score that reflects its experimental support. This can be used to filter the models and delete the connections with little experimental support. However, causal relationships are likely to depend on biological context. Thus, the scoring system should be made context/tissue specific. The increasing availability of tissue specific proteomic and (single cell) transcriptomic data (Fagerberg et al., 2014; Uhlén et al., 2015; Fernandez et al., 2019) should make this possible in a reasonably near future. Computational optimization methods such as CellNetOpt (Terfve et al., 2012), PRUNET (Rodriguez et al., 2015) or MetaReg (Ulitsky et al., 2008) can be used to identify the causal connections that are important to adapt models to context by monitoring their ability to reproduce the response of different cell systems to perturbations.

Logic Gates

As briefly discussed in this review, an AF network can be easily converted into simple Boolean models. This conversion process is set back by the observation that proteins in an AF network often receive multiple inputs from upstream proteins and these inputs govern the activity of a node as a function of the activity of the upstream nodes at each cycle of a simulation. To establish the logic functions determining node activity one needs information on how to combine these inputs. For instance, if both the kinase and the phosphatase modulating the phosphorylation state of a substrate site are active, will the substrate be phosphorylated or not? This information cannot be extracted easily from the limited available experimental evidence and approximate approaches are often used. For instance, an inhibitor win approach was often used with some success (Dorier et al., 2016; Palma et al., 2021). Alternatively, once a PKN model has been assembled the connections and the logic gates can be optimized from the ability of different models to reproduce results of perturbation experiments (Terfve et al., 2012). Developments of reasonably high-throughput experimental methods to address this limitation are highly needed.

These considerations underscore the present limits of the approach that we have discussed. Nevertheless, some initial successes in modeling clinically relevant phenotypes, as we have detailed in this review, and the delineation of a strategy to address the current limits provide confidence that cell/disease specific logic models should soon contribute to diagnosis and therapeutic decisions in clinical practice.

AUTHOR CONTRIBUTIONS

GC and LP: conceptualization and supervision. LP: formal analysis and visualization. GC, FS, and LP: writing—original draft preparation and review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Italian association for cancer research (AIRC) with a grant to GC (IG 2017 n. 20322) and by the Italian association for cancer research (AIRC) with a grant to FS (Start-Up Grant n. 21815). FS was supported by MIUR, Rita Levi Montalcini grant.

ACKNOWLEDGMENTS

We thank Denes Turei for his help in accessing OmniPath data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.694468/full#supplementary-material

REFERENCES

Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A., and Keskin, O. (2011). Transient protein-protein interactions. *Protein Eng. Des. Sel. PEDS* 24, 635–648. doi: 10.1093/protein/gzr025

Aghamiri, S. S., Singh, V., Naldi, A., Helikar, T., Soliman, S., and Niarakis, A. (2020). Automated inference of Boolean models from molecular interaction maps using CaSQ. *Bioinforma. Oxf. Engl.* 36, 4473–4482. doi: 10.1093/bioinformatics/btaa484

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

Béal, J., Montagud, A., Traynard, P., Barillot, E., and Calzone, L. (2018). Personalization of logical models with multi-omics data allows clinical stratification of patients. *Front. Physiol.* 9:1965. doi: 10.3389/fphys.2018.01965

Boué, S., Talikka, M., Westra, J. W., Hayes, W., Di Fabio, A., Park, J., et al. (2015). Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database J. Biol. Databases Curation* 2015:bav030. doi: 10.1093/database/bav030

Brinkman, R. R., Dubé, M.-P., Rouleau, G. A., Orr, A. C., and Samuels, M. E. (2006). Human monogenic disorders - a source of novel drug targets. *Nat. Rev. Genet.* 7, 249–260. doi: 10.1038/nrg1828

Ceccarelli, F., Turei, D., Gabor, A., and Saez-Rodriguez, J. (2020). Bringing data from curated pathway resources to cytoscape with omnipath. *Bioinforma. Oxf. Engl.* 36, 2632–2633. doi: 10.1093/bioinformatics/btz968

Collombet, S., van Oevelen, C., Sardina Ortega, J. L., Abou-Jaoudé, W., Di Stefano, B., Thomas-Chollier, M., et al. (2017). Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proc. Natl. Acad. Sci. U.S.A.* 114, 5792–5799. doi: 10.1073/pnas.1610622114

Csabai, L., Ölbei, M., Budd, A., Korcsmáros, T., and Fazekas, D. (2018). SignaLink: multilayered regulatory networks. *Methods Mol. Biol. Clifton NJ* 1819, 53–73. doi: 10.1007/978-1-4939-8618-7_3

Dorier, J., Crespo, I., Niknejad, A., Liechti, R., Ebeling, M., and Xenarios, I. (2016). Boolean regulatory network reconstruction using literature based knowledge with a genetic algorithm optimization method. *BMC Bioinformatics* 17:410. doi: 10.1186/s12859-016-1287-z

Dräger, A., and Palsson, B. Ø. (2014). Improving collaboration by standardization efforts in systems biology. *Front. Bioeng. Biotechnol.* 2:61. doi: 10.3389/fbioe.2014.00061

Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K. B., et al. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol. Syst. Biol.* 17:e9730. doi: 10.15252/msb.20209730

Eduati, F., Jaaks, P., Wappler, J., Cramer, T., Merten, C. A., Garnett, M. J., et al. (2020). Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. *Mol. Syst. Biol.* 16:e8664. doi: 10.15252/msb.209690

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 MCP* 13, 397–406. doi: 10.1074/mcp.M113.035600

Fernandez, D. M., Rahman, A. H., Fernandez, N. F., Chudnovskiy, A., Amir, E.-A. D., Amadori, L., et al. (2019). Single-cell immune landscape of human atherosclerotic plaques. *Nat. Med.* 25, 1576–1588. doi: 10.1038/s41591-019-0590-4

Flobak, Å., Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., et al. (2015). Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS Comput. Biol.* 11:e1004426. doi: 10.1371/journal.pcbi.1004426

Ginsburg, G. S., and Phillips, K. A. (2018). Precision medicine: from science to value. *Health Aff. Proj. Hope* 37, 694–701. doi: 10.1377/hlthaff.2017.1624

Goldman, D., and Domschke, K. (2014). Making sense of deep sequencing. *Int. J. Neuropsychopharmacol.* 17, 1717–1725. doi: 10.1017/S1461145714000789

Gunay-Aygun, M., Falik-Zaccai, T. C., Vilboux, T., Zivony-Elboum, Y., Gumruk, F., Cetin, M., et al. (2011). NBEAL2 is mutated in gray platelet syndrome and is required for biogenesis of platelet α-granules. *Nat. Genet.* 43, 732–734. doi: 10.1038/ng.883

Gunay-Aygun, M., Zivony-Elboum, Y., Gumruk, F., Geiger, D., Cetin, M., Khayat, M., et al. (2010). Gray platelet syndrome: natural history of a large patient cohort and locus assignment to chromosome 3p. *Blood* 116, 4990–5001. doi: 10.1182/blood-2010-05-286534

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., et al. (2012). A census of human soluble protein complexes. *Cell* 150, 1068–1081. doi: 10.1016/j.cell.2012.08.011

Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., et al. (2012). The cell collective: toward an open and collaborative approach to systems biology. *BMC Syst. Biol.* 6:96. doi: 10.1186/1752-0509-6-96

Hornbeck, P. V., Kornhauser, J. M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., et al. (2019). 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.* 47, D433–D441. doi: 10.1093/nar/gky1159

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi: 10.1093/nar/gkz1031

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

Klarner, H., Streck, A., and Siebert, H. (2017). PyBoolNet: a python package for the generation, analysis and visualization of boolean networks. *Bioinforma. Oxf. Engl.* 33, 770–772. doi: 10.1093/bioinformatics/btw682

Lages, J., Shepelyansky, D. L., and Zinovyev, A. (2018). Inferring hidden causal relations between pathway members using reduced google matrix of directed biological networks. *PLoS ONE* 13:e0190812. doi: 10.1371/journal.pone.0190812

Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* 16, 146–158. doi: 10.1038/nrg3885

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Licata, L., Lo Surdo, P., Iannuccelli, M., Palma, A., Micarelli, E., Perfetto, L., et al. (2020). SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update. *Nucleic Acids Res.* 48, D504–D510. doi: 10.1093/nar/gkz949

Lo Surdo, P., Calderone, A., Iannuccelli, M., Licata, L., Peluso, D., Castagnoli, L., et al. (2018). DISNOR: a disease network open resource. *Nucleic Acids Res.* 46, D527–D534. doi: 10.1093/nar/gkx876

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., et al. (2020). BioModels-15 years of sharing computational models in life science. *Nucleic Acids Res.* 48, D407–D415. doi: 10.1093/nar/gkz1055

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., and Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry* 49, 3216–3224. doi: 10.1021/bi902202q

Mubeen, S., Hoyt, C. T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H., and Domingo-Fernández, D. (2019). The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front. Genet.* 10:1203. doi: 10.3389/fgene.2019.01203

Naldi, A., Berenguier, D., Fauré, A., Lopez, F., Thieffry, D., and Chaouiya, C. (2009). Logical modelling of regulatory networks with GINsim 2.3. *Biosystems* 97, 134–139. doi: 10.1016/j.biosystems.2009.04.008

Nilsson, T., Mann, M., Aebersold, R., Yates, J. R., Bairoch, A., and Bergeron, J. J. M. (2010). Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* 7, 681–685. doi: 10.1038/nmeth0910-681

Oliver, S. (2000). Guilt-by-association goes global. *Nature* 403, 601–603. doi: 10.1038/35001165

Orchard, S. (2014). Data standardization and sharing-the work of the HUPO-PSI. *Biochim. Biophys. Acta* 1844, 82–87. doi: 10.1016/j.bbapap.2013.03.011

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project–intAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–363. doi: 10.1093/nar/gkt1115

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.* 30, 187–200. doi: 10.1002/pro.3978

Palma, A., Iannuccelli, M., Rozzo, I., Licata, L., Perfetto, L., Massacci, G., et al. (2021). Integrating patient-specific information into logic models of complex diseases: application to acute myeloid leukemia. *J. Pers. Med.* 11:117. doi: 10.3390/jpm11020117

Palma, A., Jarrah, A. S., Tieri, P., Cesareni, G., and Castiglione, F. (2018). Gene regulatory network modeling of macrophage differentiation corroborates the continuum hypothesis of polarization states. *Front. Physiol.* 9:1659. doi: 10.3389/fphys.2018.01659

Paz, A., Brownstein, Z., Ber, Y., Bialik, S., David, E., Sagir, D., et al. (2011). SPIKE: a database of highly curated human signaling pathways. *Nucleic Acids Res.* 39, D793–799. doi: 10.1093/nar/gkq1167

Pe'er, D., and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell* 144, 864–873. doi: 10.1016/j.cell.2011.03.001

Perfetto, L., Acencio, M. L., Bradley, G., Cesareni, G., Del Toro, N., Fazekas, D., et al. (2019). CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination. *Bioinform. Oxf. Engl.* 35, 3779–3785. doi: 10.1093/bioinformatics/btz132

Perfetto, L., Briganti, L., Calderone, A., Cerquone Perpetuini, A., Iannuccelli, M., Langone, F., et al. (2016). SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* 44, D548–554. doi: 10.1093/nar/gkv1048

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855. doi: 10.1093/nar/gkz1021

Porras, P., Barrera, E., Bridge, A., Del-Toro, N., Cesareni, G., Duesbury, M., et al. (2020). Towards a unified open access dataset of molecular interactions. *Nat. Commun.* 11:6144. doi: 10.1038/s41467-020-19942-z

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 45, D877–D887. doi: 10.1093/nar/gkw1012

Rocques, N., Abou Zeid, N., Sii-Felice, K., Lecoin, L., Felder-Schmittbuhl, M.-P., Eychène, A., et al. (2007). GSK-3-mediated phosphorylation enhances maf-transforming activity. *Mol. Cell* 28, 584–597. doi: 10.1016/j.molcel.2007.11.009

Rodriguez, A., Crespo, I., Androsova, G., and del Sol, A. (2015). Discrete Logic Modelling Optimization to contextualize prior knowledge networks using PRUNET. *PLoS ONE* 10:e0127216. doi: 10.1371/journal.pone.0127216

Sacco, F., Tinti, M., Palma, A., Ferrari, E., Nardozza, A. P., van Huijsduijnen, R. H., et al. (2009). Tumor suppressor density-enhanced phosphatase-1 (DEP-1) inhibits the RAS pathway by direct dephosphorylation of ERK1/2 kinases. *J. Biol. Chem.* 284, 22048–22058. doi: 10.1074/jbc.M109.002758

Selvaggio, G., Canato, S., Pawar, A., Monteiro, P. T., Guerreiro, P. S., Brás, M. M., et al. (2020). Hybrid epithelial-mesenchymal phenotypes

are controlled by microenvironmental factors. *Cancer Res.* 80, 2407–2420. doi: 10.1158/0008-5472.CAN-19-3147

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46, D661–D667. doi: 10.1093/nar/gkx1064

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696–705. doi: 10.1038/s41568-018-0060-1

Sprent, P. (2011). "Fisher exact test," in *International Encyclopedia of Statistical Science,* ed M. Lovric (Berlin, Heidelberg: Springer), 524–525. doi: 10.1007/978-3-642-04898-2_253

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

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., et al. (2009). A novel signaling pathway impact analysis. *Bioinforma. Oxf. Engl.* 25, 75–82. doi: 10.1093/bioinformatics/btn577

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M. K., et al. (2012). CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst. Biol.* 6:133. doi: 10.1186/1752-0509-6-133

Touré, V., Flobak, Å., Niarakis, A., Vercruysse, S., and Kuiper, M. (2020). The status of causality in biological databases: data resources and data retrieval possibilities to support logical modeling. *Brief. Bioinform.* doi: 10.1093/bib/bbaa390. [Epub ahead of print].

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967. doi: 10.1038/nmeth.4077

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

Ulitsky, I., Gat-Viks, I., and Shamir, R. (2008). MetaReg: a platform for modeling, analysis and visualization of biological systems using large-scale experimental data. *Genome Biol.* 9:R1. doi: 10.1186/gb-2008-9-1-r1

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* 4:rs8. doi: 10.1126/scisignal.2001699

Wang, H., Kakaradov, B., Collins, S. R., Karotki, L., Fiedler, D., Shales, M., et al. (2009). A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. *Mol. Cell. Proteomics MCP* 8, 1361–1381. doi: 10.1074/mcp.M800490-MCP200

Wang, R.-S., Saadatpour, A., and Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol.* 9:055001. doi: 10.1088/1478-3975/9/5/055001

Xing, S., Wallmeroth, N., Berendzen, K. W., and Grefen, C. (2016). Techniques for the analysis of protein-protein interactions *in vivo*. *Plant Physiol.* 171, 727–758. doi: 10.1104/pp.16.00470

Zhang, P., and Itan, Y. (2019). Biological network approaches and applications in rare disease studies. *Genes* 10:797. doi: 10.3390/genes10100797

Zhang, Q., Fu, Q., Bai, X., and Liang, T. (2020). Molecular profiling-based precision medicine in cancer: a review of current evidence and challenges. *Front. Oncol.* 10:532403. doi: 10.3389/fonc.2020.532403

Zhou, M., Li, Q., and Wang, R. (2016). Current experimental methods for characterizing protein-protein interactions. *ChemMedChem* 11, 738–756. doi: 10.1002/cmdc.201500495

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

frontiers in Genetics

REVIEW
published: 21 June 2021
doi: 10.3389/fgene.2021.706260

Networks of Networks: An Essay on Multi-Level Biological Organization

Vladimir N. Uversky[1] and Alessandro Giuliani[2]**

[1] Department of Molecular Medicine, Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, United States, [2] Department of Environment and Health, Istituto Superiore di Sanità, Rome, Italy

The multi-level organization of nature is self-evident: proteins do interact among them to give rise to an organized metabolism, while in the same time each protein (a single node of such interaction network) is itself a network of interacting amino-acid residues allowing coordinated motion of the macromolecule and systemic effect as allosteric behavior. Similar pictures can be drawn for structure and function of cells, organs, tissues, and ecological systems. The majority of biologists are used to think that causally relevant events originate from the lower level (the molecular one) in the form of perturbations, that "climb up" the hierarchy reaching the ultimate layer of macroscopic behavior (e.g., causing a specific disease). Such causative model, stemming from the usual genotype-phenotype distinction, is not the only one. As a matter of fact, one can observe top-down, bottom-up, as well as middle-out perturbation/control trajectories. The recent complex network studies allow to go further the pure qualitative observation of the existence of both non-linear and non-bottom-up processes and to uncover the deep nature of multi-level organization. Here, taking as paradigm protein structural and interaction networks, we review some of the most relevant results dealing with between networks communication shedding light on the basic principles of complex system control and dynamics and offering a more realistic frame of causation in biology.

Keywords: network, interaction network, protein-protein interactions, protein structure, protein function, intrinsically disordered proteins

OPEN ACCESS

Edited by:
Jun Chen,
Mayo Clinic, United States

Reviewed by:
Guang Hu,
Soochow University, China
George F. R. Ellis,
University of Cape Town, South Africa

****Correspondence:***
Vladimir N. Uversky
vuversky@usf.edu
orcid.org/0000-0002-4037-5857
Alessandro Giuliani
alessandro.giuliani@iss.it
orcid.org/0000-0002-4640-804X

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

Received: *07 May 2021*
Accepted: *31 May 2021*
Published: *21 June 2021*

Citation:
Uversky VN and Giuliani A (2021) Networks of Networks: An Essay on Multi-Level Biological Organization. Front. Genet. 12:706260. doi: 10.3389/fgene.2021.706260

INTRODUCTION

The network formalism is probably the most natural way to represent biological systems. Even if in the last decades the analysis of complex networks became a very widespread paradigm to face problems going from macromolecular structures (Di Paola et al., 2013) to genetic regulation circuits (Lopez-Kleine et al., 2013), neuroscience (Petersen and Sporns, 2015), and ecological systems (Bascompte, 2010), this is not a new idea. In 1948 Warren Weaver (1948), one of the fathers of mathematical information theory, sketched a very intriguing synthetic tripartite description of science into problems of "organized simplicity," "disorganized complexity," and "organized complexity" with biology located in the last class.

The first class (simplicity) refers to the case of very few elements interacting among them with largely invariant relations. Class 1 problems allow for an extreme abstraction (e.g., a planet can be thought as a dimensionless 'material point"). The possibility to take into consideration only very few basic (and object independent) features, such as mass and distance, is at the basis of the extreme precision and generality of classical mechanics.

Problems of Disorganized Complexity (class 2) allow for an analogous generalization power by means of a very different style of reasoning. Here, the predictive power stems from the abandoning of the goal to reach the elemental scale shifting to a population level statistical knowledge corresponding to gross averages (like pressure, volume, and temperature are) on a transfinite number of atomic elements. Thermodynamics is the brightest example of this style of reasoning. Both the approaches must fulfill very stringent constraints. Class 1 approach asks for few involved elements interacting in a stable way, class 2 style needs a very large number of identical particles with only negligible (or very stable and invariant) interactions among them. Biological systems, only in a very few cases do satisfy these constraints, so we step into Weaver's third class (Organized Complexity). Organized Complexity arises whenever many (even if not so many as in class 2) non-identical elements interact with each other by means of links endowed with time-varying correlation strength. The interaction of "non-identical elements" with "varying correlation strengths" corresponds to a network of links (correlations) with variable strength, connecting different nodes that in turn are "non-identical" being themselves networks with variable wiring structure.

Weaver (1948) commented that while science was at home (relying on the usual repertoire of laws and boundary conditions deciding for their application) in both Class 1 and Class 2 phenomena, the overwhelming importance of contextual information with respect to lawful invariant behavior, of Class 3 systems, makes the situation much more uncomfortable. After more than 70 years from Weaver's article, we made some steps ahead in Organized Complexity studies and the present work deals with some of these advancements. The article is organized as follows: in the first part (biodynamic interfaces), we will discuss the basic principles of the interaction between complex systems, with an emphasis on the need of an intermediate layer shared by the two interacting systems with a partially independent nature with respect to the two interactors. In the second part (the middle way), we will introduce the concept of mesoscopic or "middle-out" organization demonstrating why the "network representation" allows for a natural, hypothesis-free formalization of the meso-scale. The third part will be devoted to the transit of information across a network system and the consequent discrimination from noise of the relevant (signal) perturbations able to "climb-up" or "stepping-down" the multilevel organization. In the fourth part, we will put at work the above considerations analyzing protein-protein interaction (PPI) networks in consideration of the wiring structure of participating proteins. The essay will end with some general conclusions and future possible research trends.

BIODYNAMIC INTERFACES

There is no interaction without information exchange, and there is no information exchange without an efficient communication channel. This channel is exactly what we call "interface." If Mary calls Peter by means of her smartphone, the establishing of a contact strictly depends on the existence of an electromagnetic field endowed with a band of frequencies devoted to cell phone communication. Peter smartphone corresponds to a very specific frequency modulation of the field that is elicited by the digits Mary composes on her phone and sends on the specific band of frequencies. Consequently, Peter's smartphone rings and the communication begins. We do not enter into the actual content of communication (that only pertains to Mary and Peter), instead we focus on two crucial points of the process:

1. The existence of a medium (the field) that cannot be considered as a discrete entity with a specific location in both space and time but as a "global feature" covering the space and assuming different values in different locations. The interactors (here the Mary and Peter phones) are causally connected in both directions only because they share the same field. From basic physics we know that a point charge embedded into an electromagnetic field both senses (i.e., is influenced by the field) and modifies (i.e., influences) the field. This is exactly what happens in human-environment interaction, in which environment influences physiology (e.g., toxic effects and sensory information. . .) and is in turn influenced by humans. Both human beings and environment are complex systems and for their interaction they need a shared interface (Arora et al., 2020).
2. The interface (field) oscillates with a specific frequency, this implies it has both a "spatial" and a "temporal" structure, it is a dynamic interface. The frequency of oscillation is not independent from the spatial features of the interface, more in general, any network system (even a field can be imagined as a grid with some focal points, the "cells" in the case of mobile phones) has characteristic oscillation modes originating from its wiring structure. We will go back on this point when dealing with protein structures "resonating" with specific modes that are the carriers of across levels information. The specificity of the interaction (Mary's phone call elicits a response only in the Peter apparatus) depends on the resonance phenomenon: an oscillator with a characteristic ω frequency only "recognizes" (e.g., by amplifying its potency) an incoming stimulus with the same (or very similar) frequency.

Both these issues are at work in multi-level organization and, more in general, in biological regulation by networks of networks.

THE MIDDLE WAY

The majority of biological explanations and models are made of statements like this: "*gene A provokes the phenotype E by the activation of pathway A-B-C-D-E,*" with B, C, and D being relevant biological players, such as proteins or metabolites, whose concentration (expression) is increased (decreased) or structure is changed (e.g., via posttranslational modifications) by the action of the preceding player. This kind of "pathway" (IF:THEN for informatics) models take for granted the existence of a single "explanatory layer" located at the most microscopic level

(gene) that, thanks to a sort of domino effect, ends up into a phenotypic consequence.

This view is in sharp contrast with what we know about complex structured systems, where a multi-layer causality is at work. One of the most clear falsifications of the obliged "bottom-up" character of biological causation, comes from a 1945 article (Fankhauser, 1945) by the German (but United States based) embryologist Gerhard Fankhauser. He considered cell size in polyploid triton larvae that have a doubled chromosome number with respect to their diploid counterpart. The polyploid individuals have a doubled cell size with respect to the diploid ones, notwithstanding that, they have exactly the same dimension of organs and ducts (Fankhauser, 1945). This comes from the fact that the polyploid organism uses half the number of cells, though each cell was itself double in size, to build up its organs. This is crucial for life—the optimization of the caliber of a biological structure (the duct) is finely tuned to fit with the flow of biological fluids (a top-down constraint) and cannot be established by either its constituent cells or the genome. While this is an intuitive tenet (after all, we do not decide the size of our house based solely on the size of the bricks!), it was considered as a largely unexpected finding by Albert Einstein (a colleague of Fankhauser at Princeton) that admitted he was expecting the double size cells should give rise to double size ducts and that the Fankhauser observation pointed to still largely unknown principles (Fankhauser, 1972). The brilliant Fankhauser experiment was largely overlooked and obscured by the successes of molecular biology in the years to come, but it is a clear example of a top-down causative model, in which a "high-level" constraint "slaved" the microscopic cellular/genomic level.

It is important to stress that the "bottom-up only" obsession is not shared by all the biological fields of investigation. Ecologists recognized since many years that the most microscopic level of organization is not necessarily the place where "the most relevant facts do happen." On the contrary, the most fruitful scale of investigation is where "non-trivial determinism is maximal" (Pascual and Levin, 1999). That is to say, the scale more rich in meaningful correlations between features pertinent to micro and macro- scale or, to use an ecological term: the mesoscopic realm (Cheng et al., 2014).

Non-trivial determinism can be defined in terms of prediction error as (Pascual and Levin, 1999):

$$\text{Prediction } r^2 \;=\; 1 - E^2/S^2$$

In the above formula, E is the mean prediction error and S the standard deviation. In the case of a simple linear regression, in which a dependent variable Y must be predicted by an independent variable X, the non-trivial determinism is nothing else than the usual squared Pearson correlation between the two X and Y variables. The formula can be extended to any other situation, in which we wish to predict a system feature Y, both X and Y do not need to represent single variables but any suitable set of information at any definition scale.

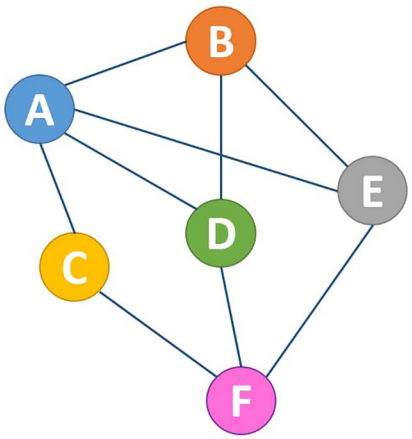
The "non-trivial" attribute of determinism stands for the need of "explaining the variance" of the system at hand (the statistic r^2 corresponds to the proportion of variance explained by a model) and not its "average" (or most stable/frequent) pattern: the aim is to account for the actual behavior of the system in both space and time and not to describe a "frozen" ideal configuration.

The individuation (i.e., description of the manner in which a thing is identified as distinguished from other things) of "mesoscopic principles" largely independent from the material constitution of the studied system and only dependent on their relational structure was the theme of an important work written by 1998 Nobel prize in Physics Robert B. Laughlin and colleagues appeared in year 2000 entitled "The Middle Way" that aptly recognized in the discovery of universal mesoscopic principles the next frontier of science (Laughlin et al., 2000). As pointed out by Nicosia et al. (2014): "*Networks are the fabric of complex systems.*" This is why different investigation fields from protein science (Di Paola et al., 2013) to neuroscience (Sporns, 2018) make use of network formalization. The basic idea here is that shared organization rules (i.e., similar wiring patterns) give rise to similar phenomenology, independently of the nature of the constituting elements. In other words, complex network invariants promise to be the place, where to look for universal mesoscopic principles, the viewpoint that maximizes "non-trivial determinism" (Pascual and Levin, 1999).

The Dutch electrical engineer Bernard Tellegen (Mikulecky, 2001) developed a sort of conservation principle of both potential and flux across a network analogous to Kirchoff's laws. The flux does not need to be an electrical current, and the same holds for the potential, a system represented by a set of nodes linked by edges with a given topology has similar emerging properties independently of the physical nature of nodes and edges. As aptly stressed in Mikulecky (2001), the theorem opens the way to a sort of "network thermodynamics," whose principles are strictly dependent on the wiring architecture, while largely independent of the constitutive laws governing the single elements.

Complex network invariants (Strogatz, 2001) catch the essence of multi-level organization for the simple fact that their estimation merges different level of definition of the system at hand. Mathematically speaking, a network corresponds to a graph, whose entire information is caught by its adjacency matrix (see **Figure 1**): a binary matrix having as rows and columns the nodes and at each *i*, *j* position a unit value if the *i* and *j* nodes have a direct link between them and 0 otherwise.

Graph invariants are relative to local (single nodes), global (entire network), and mesoscopic (clusters of nodes and optimal paths) levels. The "degree" (how many links are attached to a given node) is a local descriptor, the "average shortest path" (characteristic length) is the average length of minimal paths connecting all the node pairs, and can be considered as a mesoscopic feature, while the general connectivity of the network (density of links) is a global property (Csermely et al., 2013; Giuliani et al., 2014). All these descriptors (and many others) are strictly intermingled across different organization layers. In fact, characteristic length inherits from the "bottom" the information of the single node degree (higher degree nodes have a higher probability to enter into shortest paths), while betweenness (the number of shortest paths passing by a node, thus a strictly speaking a microscopic feature of the network) inherits from the "top" the existence of clusters (modules) of nodes so that a



	A	B	C	D	E	F
A	-	1	1	1	1	0
B	1	-	0	1	1	0
C	1	0	-	0	0	1
D	1	1	0	-	0	1
E	1	1	0	0	-	1
F	0	0	1	1	1	-

FIGURE 1 | Mathematically, every network **(right)** can be expressed in the form of an adjacency matrix **(left)**. In this case, a network with undirected, unweighted edges is shown, which is represented by a symmetric adjacency matrix containing only the values 0 and 1 to indicate the absence and presence of connections, respectively.

node in between two different clusters A and B is traversed by all the shortest paths linking the A,B node pairs so scoring a high betweenness (**Figure 2**).

In other terms, describing a system by network formalism implies a multi-level structural representation without the need of "imposing" a particular bottom-up or top-down causative pattern.

INFORMATION FLUXES ACROSS NETWORKS

Biological systems are complex systems that both adapt to their environment and interact with other systems. Provided we are able to find a meaningful formalization in terms of interacting parts, each complex system can be intended as a network. Therefore, it is crucial to understand the peculiarities of information transfer across networks, in order to understand the basic principles of biological organization.

Probably the most straightforward paradigm of information transfer through a network in proteins is the allosteric effect. Allostery is a neologism coming from Greek language, which has to do with the ability of proteins to transmit a signal from one site of molecule to another in response to environmental stimuli. This ability is related to the transmission of information across the protein molecule from a sensor (allosteric) site to the effector (binding or active) site (Hilser et al., 2012). The molecule, hence, perceives ligand binding (or any other micro-environmental perturbation) at distance from the active site, and adapts its configuration accordingly. For example, hemoglobin molecule senses at the allosteric site the partial pressure of oxygen ($p[O_2]$): when $p[O_2]$ is high, the affinity of hemoglobin for oxygen increases and the protein binds oxygen molecules at active site. On the contrary, when $p[O_2]$ is low, affinity decreases and bound oxygen is released to the cells. This process is crucial for life: in lungs, there is a very high oxygen pressure and the red blood cells containing hemoglobin must catch oxygen molecules that in turn must be released in peripheral tissues (low $p[O_2]$) so to make oxidative metabolism possible. How the protein molecule can discriminate such a relevant signal from the continuous motions coming from thermal noise and transmit the information at distance so to reach the active site?

To answer this question is useful to consider a protein molecule as a network (**Figure 3**). In the left panel, the 3D structure of a small protein (recoverin) follows the usual "ribbon" style: the polypeptide chain is represented in terms of contiguous segments of "secondary structure" namely α-helices, irregular structure, and β-sheets (Di Paola et al., 2013). In this particular representation, the parallel segments give an idea of the flexibility of the different tracts of the molecule related to the thermal motion. The right panel represents the same protein in terms of the adjacency matrix of the corresponding network (PCN = Protein Contact Network) whose nodes are the constituent amino acids, while the darkened pixels mark the unit values of adjacency matrix (**Figure 1**) pointing to an effective pairwise contact between amino acid residues. The amino acid residues are ordered along the protein sequence and the "trivial" contacts between amino acids adjacent along the chain are eliminated. This implies the scored contacts (links of the PCN) correspond to non-covalent intermolecular bonds putting different parts of the molecule into close contact by the action of folding process. This intra-protein interactability is illustrated by **Figure 4**, where a protein molecule is represented as a bracelet having amino acid residues as pearls and active contacts as dashed red lines.

In PCNs, the shortest paths passing by the network edges mediate concerted motions and energy transmission upon stimulation of allosteric site (Di Paola and Giuliani, 2015; Gadiyaram et al., 2021). The topological metrics of shortest paths (minimum number of links separating two residues) is thus the actual metrics for signaling (Gadiyaram et al., 2021). The discrimination between relevant signals to be transmitted

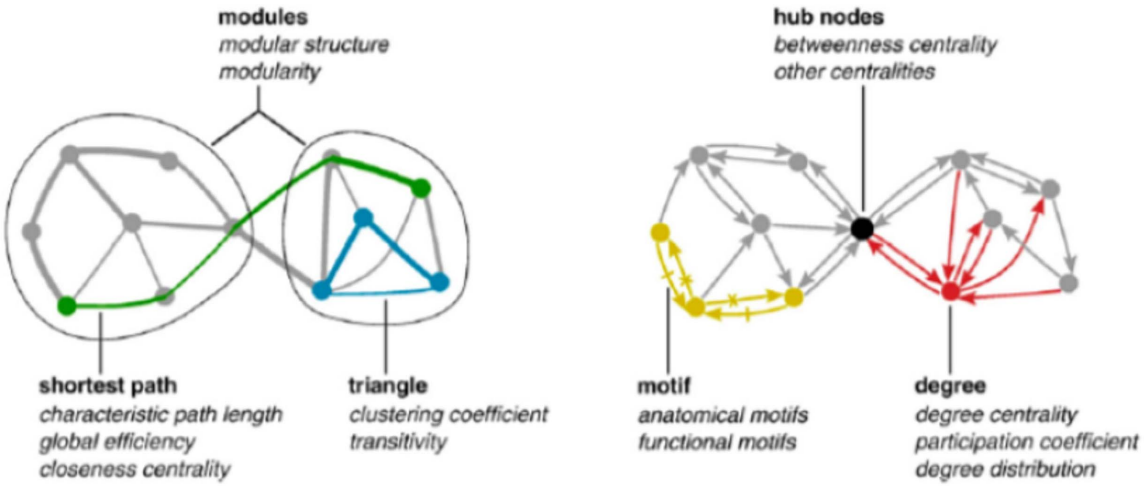


FIGURE 2 | Major features (graph invariants) used to describe networks. It is worth noting the mutual dependence of network descriptors across different scales going from single nodes to the entire network.

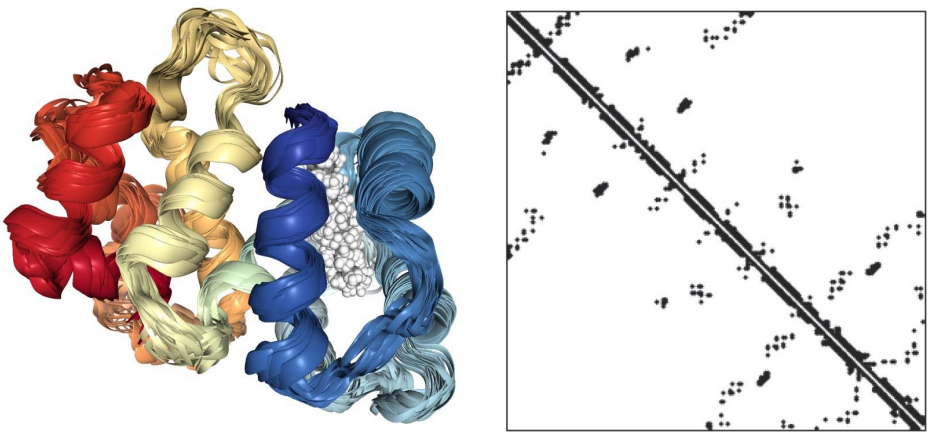


FIGURE 3 | NMR solution structure of myristoylated recoverin in the calcium-free state (PDB ID: 1IKU, left) and correspondent adjacency matrix.

at distance without loss of information and non-informative perturbations to be dissipated without relevant changes in the 3D structure, relies upon two very important mesoscopic network descriptors: "Guimera and Amaral' z and P indexes (Guimera and Nunes Amaral, 2005). The index z quantifies the number of contacts a given node (amino acid residue in this case) has with other nodes of its own cluster (local contacts), while P scales with the number of edges linking the node to amino acid residues pertaining to different clusters.

A perturbation affecting specifically a "high-P" node travels a long distance across the network passing by subsequent "high-P" nodes and arriving at the destination, thereby supporting allosteric effects. On the contrary, generic (noisy) thermal motion rapidly dissipates distributing across non-directional cycles through intra-module motions.

High-P nodes create a "fast lane" for relevant information neatly separated by noise. This is exactly the role of biodynamic interfaces: some proteins (multimeric proteins) are made by distinct chains held together by intermolecular contacts. This is the case of hemoglobin that is made by four distinct polypeptide chains: the allosteric effect ends up into a different re-arrangement of the relative positions of the four chains that go back and forth between two different patterns (R and T for Relaxed and Tense) with high and low affinity for oxygen. The interface between these four chains is made of high-P amino acid residues that allow concerted motions among the chains. **Figure 5** gives a pictorial description of the situation by showing the adjacency matrix of hemoglobin (**Figure 5**).

Here, the adjacency matrix of hemoglobin is described by a color code. As usual in such presentation, the axes of the figure report the order of the residues along the chains (each chain contains 150 residues). The dark blue corresponds to the lack of contacts, and the different colors correspond to the four chains. It is evident, the presence of "displaced contacts" in the form of residues that, while pertaining to a given chain (module of the network) have the majority of their contacts with amino

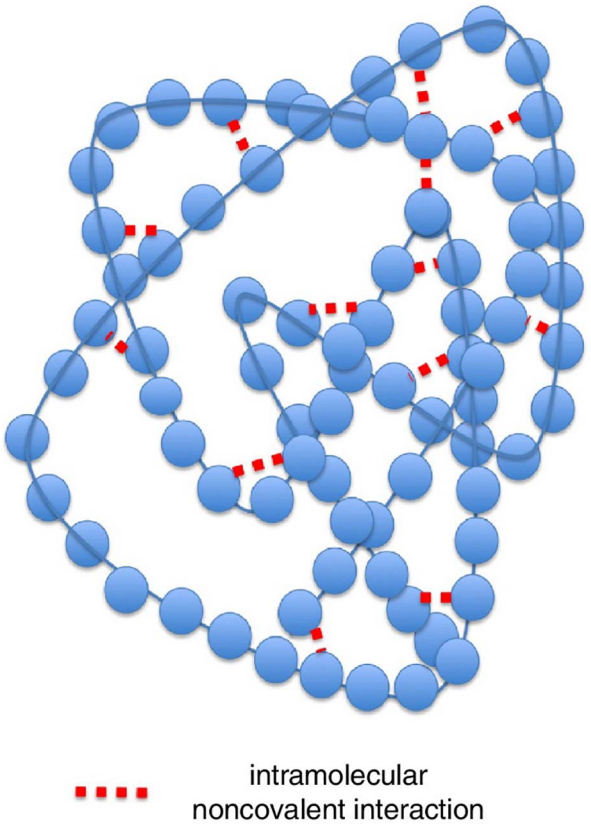


FIGURE 4 | Representation of the intra-protein interactability for a model protein. Here a protein molecule is shown as a bracelet having amino acid residues as pearls and active contacts as dashed red lines. Modified from Di Paola and Giuliani (2015).

acids pertaining to different chains. These "displaced contacts" are the long "whiskers" contacting zones different from their own cluster [e.g., the pale blue line pertaining to the first chain (1-150)] that is in contact with the orange (second chain) module). These whiskers correspond to the high-*P* nodes that generate "something in between" the interacting systems with a "shared ontology" across the interacting systems (polypeptide chains). Very similar models allow for the synchronization of interacting networks thereby passing from single stimulus effects to sustained periodic oscillations.

NETWORK OF NETWORKS: FROM SINGLE PROTEINS TO PROTEIN-PROTEIN INTERACTIONS

Any protein can be considered as a specific network of residue-residue interactions. Importantly, such network consideration works for both monomeric proteins (e.g., aforementioned recoverin and hemoglobin monomer) and oligomeric proteins [e.g., hemoglobin heterotetramer $(\alpha\beta)_2$]. These and many other similar examples can be used as illustrations of information flow within ordered proteins and ordered protein complexes. In fact, in such cases, protein (protein complex) is characterized by a unique, relatively stable crystal-like 3D structure whose Ramachandran angles vary only slightly around their equilibrium positions with occasional cooperative conformational switches and with almost constant and very specific residue-residue interactions that are relatively fixed in time and space. The stability of such a uniquely folded structure of an ordered protein is defined by the tight packing of its interior achieved by multiple specific residue-residue interactions (Pace et al., 2014). There is very little free space in the protein interior (Richards, 1963; Klapper, 1971; Lee and Richards, 1971), which is closer to a solid than to a liquid (Klapper, 1971), since it is twice as tightly packed as water and possesses a packing density, which exceeds that of closely packed spheres (Pace et al., 2014). This tight packing is achieved during protein folding by burying about 85% of the non-polar side groups, 65% of the polar side chains, and 70% of the peptide groups (Lesser and Rose, 1990), and due to the formation of 1.1 hydrogen bonds per residue (Stickle et al., 1992). This stable structural organization, supported by the numerous crystal structures of proteins solved by X-ray diffraction, resulted in a very common use of terms "unique 3D structure" and "rigid 3D structure" for the description of the structural properties of ordered proteins. Furthermore, the relative rigidity of structures of globular proteins was further supported by their high conformational stability and cooperative folding-unfolding behavior, where, for example, denaturant-induced unfolding was described as a reversible and highly cooperative "all-or-none"-type transition between native and denatured states (Tanford, 1968), and where the temperature-induced melting was shown to be accompanied by the cooperative heat absorption related to the sharp change in the state of a protein on heating (Privalov, 1979, 1982).

However, it is recognized now that considering a protein molecule as a static entity with "rigid 3D structure" and a unchanging PCN is an oversimplification, as proteins are rather dynamic biological systems that have some degree of flexibility, as a matter of fact we observe changes in PCN of apo- and holo-forms and in response to allosteric effectors (Di Paola et al., 2013; Di Paola and Giuliani, 2015). In fact, the importance of conformational flexibility and the need of dynamics for the successful functionality of globular proteins (even enzymes) was emphasized in many studies over the past 65 years or so (e.g., Koshland, 1958; Villa et al., 2000; Agarwal et al., 2002, 2004; Eisenmesser et al., 2002, 2005; Rajagopalan and Benkovic, 2002; Sutcliffe and Scrutton, 2002; Tousignant and Pelletier, 2004; Agarwal, 2005; Yang and Bahar, 2005; Olsson et al., 2006; Frauenfelder et al., 2009). The internal dynamics of enzymes (i.e., movement of their parts including individual amino acid residues, a group of amino acids, or even an entire domain that occurs in a wide range of time-scales, from femto-seconds to seconds) has been suggested to be linked to their mechanism of catalysis (Eisenmesser et al., 2002, 2005; Agarwal, 2005). Furthermore, the existence of conformational sub-states (which were detected based on the atomic displacements involved in the inter-conversion of different local configurations of the same overall protein structure) in globular proteins potentially

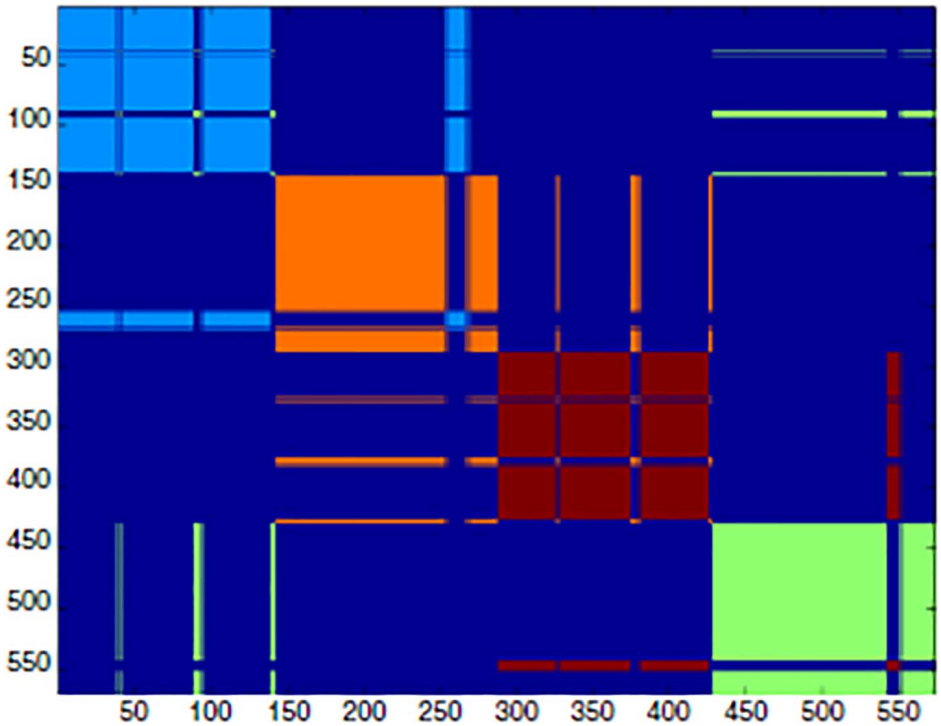


FIGURE 5 | Spectral clustering of hemoglobin. The adjacency matrix is shown as a clustering color map that reports the cluster partition along the sequence. The spectral clustering technique decomposes the space through the adjacency matrix eigenvalues, so that the partition relies on the topological role of residues in the interaction network, rather than on their spatial positioning [modified from Di Paola and Giuliani (2015)].

related to their functional conformational changes and allosteric behavior has been established (Austin et al., 1975; Artymiuk et al., 1979; Frauenfelder et al., 1979; Beece et al., 1980; Frauenfelder and Petsko, 1980; Parak et al., 1981; Hartmann et al., 1982). It was also pointed out that although the entire protein molecule is rather flexible, the flexibility is not homogeneously distributed within a molecule, and some structural parts of ordered proteins are more rigid than others (Ma et al., 1999). Such more rigid parts or structural units (which could be structural domains, sub-domains or any other sub-structure) are typically more compactly packed, have a stronger hydrophobic effect and have a larger stabilizing electrostatic contribution (Ma et al., 1999).

A protein with a set of stable structural units can form a range of conformational isomers, structural peculiarities of which (and corresponding PCNs) would depend on the extent of the overall structural flexibility and the locations of the more flexible joints, whereas, in a protein with unstable structural units, the thermal motions of the backbone could generate an entirely flexible molecule (Ma et al., 1999). Notable, in PCN formalism, the residues devoted to structural stability (high z, low P) are the less flexible, while the opposite holds for high P residues. Obviously, the presence of such structural flexibility changes the PCN perception and transforms its representation from a static mesh into a network with spatio-temporal dynamics, where residue-residue contacts are not fixed in time and space, but change over time. This, in turn, complicates information transmission, which cannot be considered as a passage through a rigid bridge or tunnel anymore, but represents an attempt to cross the river by a suspension bridge in a very windy day.

Furthermore, complications and complexity are not stopped there, as in their functional states, many proteins can be disordered to different degree. In fact, recent years provided solid evidence of the existence of the entirely different class of biologically active proteins, which do not have unique structures as a whole or in some parts. These are intrinsically disordered proteins (IDPs) and hybrid proteins containing ordered and intrinsically disordered protein regions (IDPRs), the existence of which has changed protein science. Such proteins are commonly found in proteomes of all the oganisms in all kingdoms of life and all viral proteomes analyzed so far (Dunker et al., 2000; Ward et al., 2004; Tompa et al., 2006; Krasowski et al., 2008; Shimizu and Toh, 2009; Tokuriki et al., 2009; Pentony and Jones, 2010; Tompa and Kalmar, 2010; Uversky, 2010; Xue et al., 2010, 2012, 2014; Schad et al., 2011; Hegyi and Tompa, 2012; Korneta and Bujnicki, 2012; Midic and Obradovic, 2012; Pancsa and Tompa, 2012; Di Domenico et al., 2013; Kahali and Ghosh, 2013; Peng et al., 2015; Kumar et al., 2021). They have crucial roles in various biological processes and their penetrance increases with the increase in the organism complexity (Dunker et al., 2000; Ward et al., 2004; Oldfield et al., 2005; Uversky, 2010; Xue et al., 2012). As a result, the putative fraction of sequences with predicted long IDPRs (30 residues or longer) increases in the order: Bacteria $\sim$ Archaea $<<$ Eukaryota (Dunker et al., 2000; Ward et al., 2004; Xue et al., 2010; Na et al., 2013; Peng et al., 2015), and

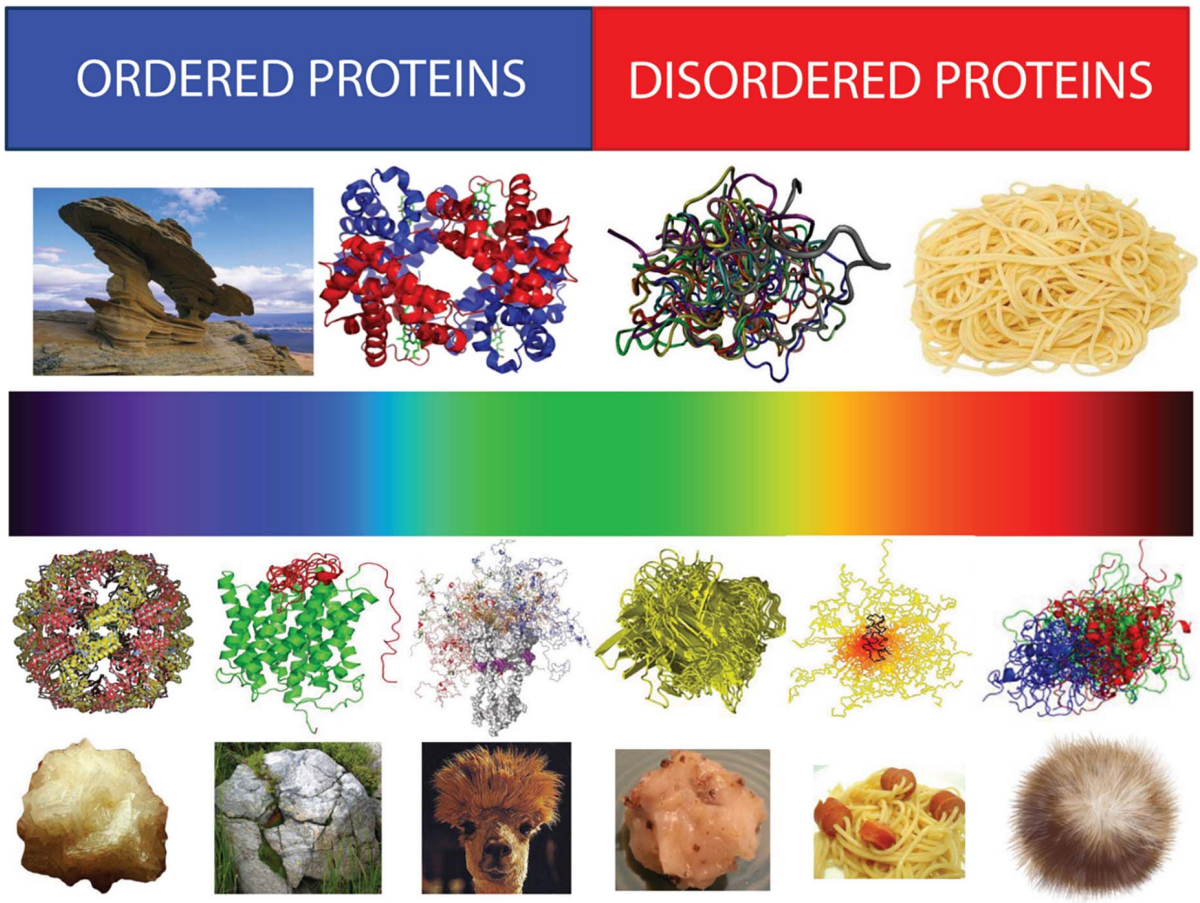


FIGURE 6 | Structural spectroscopy of proteins representing structural heterogeneity of IDPs/IDPRs. Top half: Bi-colored view of functional proteins which are considered to be either ordered (folded, blue) or completely structure-less (disordered, red). Ordered proteins are taken as rigid rocks, whereas IDPs are considered as completely structure-less entities, kind of cooked noodles. Bottom half: A continuous emission spectrum representing the fact that functional proteins can extend from fully ordered to completely structure-less proteins, with everything in between. Intrinsic disorder can have multiple faces, can affect different levels of protein structural organization, and whole proteins, or various protein regions can be disordered to a different degree. Some illustrative examples includes ordered proteins that are completely devoid of disordered regions (rock-like type), ordered proteins with limited number of disordered regions (grass-on-the rock type), ordered proteins with significant amount of disordered regions (llama/camel hair type), molten globule-like collapsed IDPs (greasy ball type), pre-molten globule-like extended IDPs (spaghetti-and-sausage type), and unstructured extended IDPs (hairball type). Adapted from Uversky (2013a).

this increase in the penetrance of protein disorder is linked to the increased roles of structure-less proteins and protein regions in cellular signaling, regulation, and recognition (Wright and Dyson, 1999; Dunker and Obradovic, 2001; Dunker et al., 2001, 2002a,b; Dyson and Wright, 2002, 2005; Tompa, 2002).

One of the characteristic features of IDPs/IDPRs is their exceptionally complex and heterogeneous spatio-temporal structural organization, where different parts of a molecule are dynamically ordered (or disordered) to a different degree (**Figure 6**). In fact, within the highly dynamic conformational ensembles of IDPs/IDPRs one can find foldons (independent foldable units of a protein), inducible foldons (disordered regions that can fold at least in part due to the interaction with binding partners), inducible morphing foldons (disordered regions that can differently fold at interaction with different binding partners), non-foldons (non-foldable protein regions), semi-foldons (regions that are always in a semi-folded form), and unfoldons (ordered regions that have to undergo an order-to-disorder transition to become functional) (Uversky, 2013a,b,c, 2015, 2016a,b, 2019a,b; Jakob et al., 2014; Deforte and Uversky, 2016), whose distribution is constantly changing over time (Uversky, 2013c, 2016c).

This behavior of an IDP/IDPR as a highly frustrated system without single folded state, is reflected in its free energy landscape, which is relatively flat, lacks a deep energy minimum seen in the landscape of an ordered protein, and represents instead a "hilly plateau," with multiple local minima corresponding to a multitude of conformations and multiple hills that correspond to the forbidden conformations (Uversky et al., 2008; Turoverov et al., 2010; Fisher and Stultz, 2011). Such energy landscape is extremely sensitive to different environmental changes that can modify landscape in a number of very different ways, making some energy minima deeper and some energy barriers higher. This explains the conformational plasticity of an

IDP/IDPR, its extreme sensitivity to changes in the environment, its ability to specifically interact with many partners of different nature, and to fold differently as a result of these interactions (Uversky, 2013c).

Obviously, intrinsic disorder plays a crucial role in the organization of the intra-protein networks. In fact, the aforementioned exceptionally complex and heterogeneous spatio-temporal structural organization of a protein molecule with all its foldons, inducible foldons, inducible morphing foldons, non-foldons, semi-foldons, and unfoldons can be presented in the form of an intra-protein network, where residues are involved in transient or more stable conformational interactions. This network is highly dynamic and extremely sensitive to the environment and interaction with partners. Therefore, the aforementioned sensitivity of IDPs to the subtle changes in their environment and capability to fold, often differently, at interaction with binding partners or differently respond to different post-translational modifications (PTMs) or other stimuli, can be considered as a kind of condition-driven rewiring of their intra-molecular networks, where new paths (new connections) can emerge in a condition-specific manner. It is worth noting the strong resemblance of IDPs with the features of biodynamic interfaces we sketched above: this is fully consistent with their role of taking care of physiologically relevant interactions.

Therefore, this complex structural organization of IDPs/IDPRs defines their exceptional multi-functionality and serves as a foundation for "protein structure-function continuum" model, where protein exists as a dynamic conformational ensemble comprised of interchanging foldons, inducible foldons, inducible morphing foldons, non-foldons, semi-foldons, and unfoldons and containing multiple proteoforms (conformational/basic, inducible/modified, and functioning) characterized by a broad spectrum of structural features and possessing various functional potentials (Uversky, 2016a, 2019a,b) (see **Figure 7**).

From the viewpoint of information flow, multi-functionality of such highly dynamic conformational ensembles can be understood if they are depicted as inter-converting ensembles of multi-component systems (networks), whose configurations show extreme sensitivity to the environment. Constituents of these networks are the aforementioned foldons, inducible foldons, inducible morphing foldons, non-foldons, semi-foldons, and unfoldons, which exist transiently ("now you see me, now you don't") and define dynamic nature of the network by forming transient contacts with other constituents in the environment-dependent manner. All this places IDPs/IDPRs in the category of the "edge of chaos" systems that operate in the region of maximal complexity (i.e., in a region between order and complete randomness or chaos), where even small changes in the environment might generate large and diversified changes in protein structure and function (Uversky, 2013c, 2019a), defining the ability of a system to differently channel information and to behave as moving staircases in the Hogwarts Castle.

Therefore, a protein molecule represents a complex system that exists as a dynamic, multilevel network of networks. In fact, one can represent a protein molecule as nesting doll

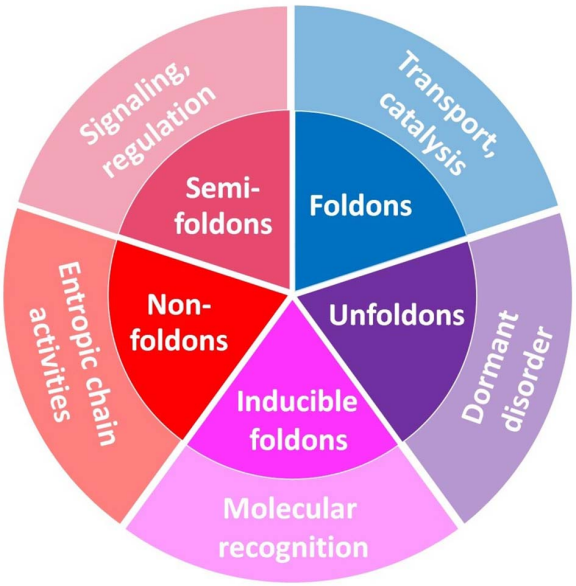


FIGURE 7 | Schematic representation of the mosaic nature of the protein structure–function space. It should be noted that "Dormant disorder" is different from the other "outer-ring" functional grouping because the corresponding segment does not describe a particular functional group but rather represents the means by which the functionality is achieved. Adopted from Uversky (2015).

(Matryoshka) of the networks of increasing size. Here, at the lowest level, different segments of polypeptide chain form secondary structure elements that represent local networks of hydrogen bonds and residue-residue interactions. The next level of the network is formed by interactions between the elements of secondary structure, which are local networks themselves. This generates foldons, inducible foldons, inducible morphing foldons, non-foldons, semi-foldons, and unfoldons. Next, interactions between these second-tier networks generate higher level networks, proteins domains. Finally, a functional monomeric protein represents seemingly highest level network that includes inter-domain interactions and interactions between domains and second-tier networks. However, formation of an oligomeric protein (and engagement in the temporary protein-protein interactions) would require a new level of inter-subunit interactions, where the inter-protein interaction network might include interactions between the networks of various lower levels.

Despite being a complex system with a complex fate, a single protein is not life *per se*, while protein-protein interactions (PPIs) and their networks are the core of biological regulation. Biological PPI networks belong to the category of the "scale-free" or "small-world" networks, which are neither completely regular (i.e., networks, where each node has exactly the same number of links) or completely random (Erdös and Rényi, 1960). An example of the random networks is given by the highway system, in which despite the random placement of links most nodes have approximately same number of links (Erdös and Rényi, 1960; Barabasi and Bonabeau, 2003). Because

the nodes follow a Poisson distribution with a bell shape, such a system almost do not have nodes that have significantly more or fewer links than the average (Barabasi and Bonabeau, 2003). Topology of the PPI networks (as well the airline routes, the author-collaboration network, the metabolic network, gene network, the protein domain network, social networks, and the World Wide Web) is different, as they have hubs, with many connections, and ends, that aren't connected to anything but a hub (Watts and Strogatz, 1998; Goh et al., 2002). Scale-free networks combine the local clustering of connections characteristic of regular networks with occasional long-range connections between clusters, as can be expected to occur in the random networks. As a result, as a whole, such network has a power-law distribution of the number of links connecting to a node, with some popular nodes possessing a very large number of connections to other nodes, and with the most nodes having just a few (Barabasi and Bonabeau, 2003). Such popular nodes, known as hubs, might have hundreds, thousands or even millions of links depending on the type of network being described. It has been emphasized that from this perspective, the network appears to have no scale (Barabasi and Bonabeau, 2003), and in such scale-free networks, the distance between nodes also follows a power-law distribution (Barabasi and Albert, 1999). This defines the "small world" nature of these networks, as the average distance between two vertices in scale-free network is very small relative to a highly ordered network (e.g., regular lattice), but clustering coefficient is large. As a result, although most nodes are not neighbors of one another, they can be reached from every other node by a small number of steps, since the neighbors of any given node are likely to be neighbors of each other (Watts and Strogatz, 1998) (e.g., in a social network, the small world phenomenon is reflected by a short chain of acquaintances needed to link strangers).

Due to their scale-free nature, PPI networks contains several hubs, which are multitasking proteins that have multiple links. Binding promiscuity of hubs is mostly determined by the intrinsic disorder phenomenon (Dunker et al., 2005; Dosztanyi et al., 2006; Haynes et al., 2006; Oldfield et al., 2008; Hu et al., 2017). In fact, some protein hubs are disordered as a whole, others are hybrid proteins containing both ordered and disordered regions, and very few hubs can be highly structured proteins. Many (but not all) interactions of hybrid hubs are mapped to their IDPRs (Dunker et al., 2005; Uversky and Dunker, 2010), whereas the binding regions of the partners of ordered hubs are intrinsically disordered (Bustos and Iglesias, 2006; Radivojac et al., 2006). These observations clearly indicate that hub proteins commonly use disordered regions (either their own or of their binding partners) to bind to multiple partners (Uversky et al., 2005; Dosztanyi et al., 2006; Ekman et al., 2006; Haynes et al., 2006; Patil and Nakamura, 2006; Singh et al., 2006). The presence of inducible foldons within the conformational ensembles of hubs allow them to (at least partially) fold at interaction with binding partners, whereas the presence of inducible morphing foldons defines the capability of hubs to fold differently at interaction with different partners. All this creates the means for binding promiscuity of hub proteins that relies on intrinsic disorder and related binding-induced disorder-to-order transitions enabling one protein to interact with multiple partners (one-to-many signaling) or to enable multiple partners to bind to one protein (many-to-one signaling) (Dunker et al., 1998).

With respect to the temporal structure of the PPI networks and the roles of intrinsic disorder in maintaining network topology, some proteins have multiple simultaneous interactions ("party hubs"), while others have multiple sequential interactions ("date hubs") (Han et al., 2004). From a functional perspective, date hubs may connect biological modules to each other (Hartwell et al., 1999), whereas party hubs may form scaffolds that enable the assembly of functional modules (Han et al., 2004). As far as information flow is concerned, PPI network represents a clear example of the "network of the networks of the networks" concept, as it is formed by the interacting Matryoshkas, each being a network of networks itself. Due to the presence of high-*P* and low-*P* nodes and high sensitivity to environment, the topology of a protein PCN (at least PCNs of IDPs/IDPRs) is likely to be described as dynamic inter-converting scale-free networks with the characteristics of the edge of chaos systems, where information can be channeled to different nodes depending on the peculiarities of the protein environment or due to the introduction of post-translational modifications (PTMs). This, in turn, makes PPI network a higher level dynamic non-linear system of the inter-converting scale-free networks possessing the edge of chaos features. This also defines the ability of PPI networks to show the peculiar chaos signature named "butterfly effect," where a small change in the state of one component of one Matryoshka (e.g., conformational changes induced by binding of a ligand or PTM of a region in one of proteins) can result in large differences in later states (i.e., leading to initiation of different cellular responses). This feature can be considered as the structural counterpart of a "bottom-up" causative chain where a seemingly minor perturbation (e.g., point mutation, ligand binding at a specific receptor, or PTM) gives rise to macroscopic effects. Here it is in action a "permissive" and not an "instructive" (as often implicitly assumed) causative model: the incoming stimulus does not embed the "instructions" for the subsequent process, it only impinges over a "permissive" context (the particular network structure) allowing for the subsequent signal amplification.

CONCLUSION

Protein molecules are the most elementary complex systems, lying in the borderline between simple and complex systems physics (Frauenfelder and Wolynes, 1994), they present the basic features of "Weaver organized complexity" (Weaver, 1948): multiple stable states, wiring structure changing in time, adaptation to changing environmental conditions. All these features are acquired by means of biodynamic interfaces (Arora et al., 2020) that, in the case of protein molecules, can be traced down to "high-*P*" residues (and consequently by IDP/IDPR elements). Such features are amplified at the next organization level (PPI), where the same basic principles hold but at an higher level of complexity and, consequently allowing for a much wider repertoire of possible configurations. A coarse grain estimation

of the possible "allowed interfaces" between the 25,000 yeast proteins (a very low number with respect to the more than 100,000 protein species of human cells) gives the astronomical number of 10^{7200} (Tompa and Rose, 2011). Notwithstanding that, we observe a relatively low number of "allowed configurations" out of the transfinite number of possible ones: e.g., the actual estimates of "different cell kinds" each with a specific asset of protein-protein interaction pattern tells us of only 411 different human cell types (Vickaryous and Hall, 2006). This dramatic collapse of the number of discrete phenotypes starting from a huge variety of "solutions" at the "bottom of the scale" asks for very strict thermodynamic-like constraints granting for multiple "phenotypically equivalent" solutions at the molecular scale.

The "two-way" interactions between PCN and PPI uncovers some empirical organization principles of the multi-layer networks-of-networks organization of life, here we suggest two of these seminal principles:

1. The "between domains" communication is mainly the duty of "flexible elements." This creates a partition between structure preserving "conservative" nodes and "creative flexible elements" at each organization layer (Csermely, 2008). This separation is at the basis of biological evolution: it is not by chance that the "structure preserving" amino acid residues are the most conserved, while allosteric signaling are much more prone to mutations along evolutionary scale. Leander et al. (2020) by use of deep mutational scanning, elucidated the molecular basis and underlying functional landscape of allostery. The authors showed that allosteric signaling exhibits a high degree of functional plasticity and redundancy through myriad of mutational pathways. Residues critical for allosteric signaling are poorly conserved, while those required for structural integrity are highly conserved. This result seems at first sight paradoxical: evolution seems to preserve fold over function. But this conundrum is only apparent, if we think that allostery (and, more in general, communication among different layers/domain) has a distributed nature. The presence of multiple equivalent solutions to the thermodynamic conditions of cooperativity (i.e., the collapse to very few phenotypic forms at higher levels) guarantees a much higher resilience (multiple equivalent solutions) with respect to a fine-tuned much more deterministic solution. In the same way, the multiplicity of "quasi-equivalent" communication channels allows for a much more rapid adaptation to a continuously changing environment.
2. Any system made by interacting parts and constrained in a finite size environment oscillates. The frequency of such oscillations roughly (inversely) scales with the size of the system at hand. The entrainment of two oscillators with similar frequencies is the basis of resonance phenomenon provoking a huge amplification of the combined output signal. Resonance phenomena are present at every level of biological organization (Lerner et al., 2018; van der Groen et al., 2018) and are at the basis of the new (and very promising) research avenue of "allosteric drugs" that are able to "generalize" an incoming pharmacological stimulus thanks to resonance phenomena similar to what happens in musical instruments (Zhang et al., 2018; Ni et al., 2019). In a recent work, we found that an ensemble of interacting proteins made of many IDP/IDPR elements was able to greatly enhance the global phenotypic plasticity of yeast cells (Camponeschi et al., 2021). This is an example of a microscopic level stimulus made evident at the macroscopic phenotypic scale thanks to resonance phenomena with external oscillatory stimuli.

All in all, we can affirm the exploration of networks-of-networks can promote a new integrative view of biology at both theoretical and applicative levels. In our opinion, the "bottom-line" of such hierarchy constituted by the analysis of protein and protein complexes, is a perfect playground for generating organization principles universally valid for different organization scales. The "Middle Way" (Laughlin et al., 2000) attitude shifts the "shared foundation of different sciences" from the recognition that "*all the entities are made of the same fundamental particles*" (orienting the various "theories of everything" flourished in the last century) to the statement "*all the entities can be considered as networks of interacting parts*" (Giuliani, 2021). This shift implies the "Universality" of mesoscopic organization principles and the consequent presence of the same wiring rules and emerging properties at different organization layers. This is why, protein science, with its unique mixture of plenty of good quality data and the natural link to a baseline of established chemico-physical properties coming from the adjacent "organized simplicity realm" (Frauenfelder and Wolynes, 1994), is a privileged vantage point for initiating a new avenue of biological research.

AUTHOR CONTRIBUTIONS

VU and AG: conceptualization, literature search and analysis, writing – original draft preparation, and writing – reviewing and editing. Both authors contributed to the article and approved the submitted version.

REFERENCES

Agarwal, P. K. (2005). Role of protein dynamics in reaction rate enhancement by enzymes. *J Am Chem Soc* 127, 15248–15256. doi: 10.1021/ja055251s

Agarwal, P. K., Billeter, S. R., Rajagopalan, P. T., Benkovic, S. J., and Hammes-Schiffer, S. (2002). Network of coupled promoting motions in enzyme catalysis. *Proc Natl Acad Sci U S A* 99, 2794–2799. doi: 10.1073/pnas.052005999

Agarwal, P. K., Geist, A., and Gorin, A. (2004). Protein dynamics and enzymatic catalysis: investigating the peptidyl-prolyl cis-trans isomerization activity of cyclophilin A. *Biochemistry* 43, 10605–10618. doi: 10.1021/bi0495228

Arora, M., Giuliani, A., and Curtin, P. (2020). Biodynamic Interfaces Are Essential for Human-Environment Interactions. *Bioessays* 42, e2000017.

Artymiuk, P. J., Blake, C. C., Grace, D. E., Oatley, S. J., Phillips, D. C., and Sternberg, M. J. (1979). Crystallographic studies of the dynamic properties of lysozyme. *Nature* 280, 563–568. doi: 10.1038/280563a0

Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I. C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry* 14, 5355–5373. doi: 10.1021/bi00695a021

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

Barabasi, A. L., and Bonabeau, E. (2003). Scale-free networks. *Sci Am* 288, 60–69. doi: 10.1093/acprof:oso/9780199211517.003.0004

Bascompte, J. (2010). Ecology. Structure and dynamics of ecological networks. *Science* 329, 765–766. doi: 10.1126/science.1194255

Beece, D., Eisenstein, L., Frauenfelder, H., Good, D., Marden, M. C., Reinisch, L., et al. (1980). Solvent viscosity and protein dynamics. *Biochemistry* 19, 5147–5157.

Bustos, D. M., and Iglesias, A. A. (2006). Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins. *Proteins* 63, 35–42. doi: 10.1002/prot.20888

Camponeschi, I., Damasco, A., Uversky, V. N., Giuliani, A., and Bianchi, M. M. (2021). Phenotypic suppression caused by resonance with light-dark cycles indicates the presence of a 24-hours oscillator in yeast and suggests a new role of intrinsically disordered protein regions as internal mediators. *J Biomol Struct Dyn* 39, 2490–2501. doi: 10.1080/07391102.2020.1749133

Cheng, H., Yao, N., Huang, Z. G., Park, J., Do, Y., and Lai, Y. C. (2014). Mesoscopic interactions and species coexistence in evolutionary game dynamics of cyclic competitions. *Sci Rep* 4, 7486.

Csermely, P. (2008). Creative elements: network-based predictions of active centres in proteins and cellular and social networks. *Trends Biochem Sci* 33, 569–576. doi: 10.1016/j.tibs.2008.09.006

Csermely, P., Korcsmaros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* 138, 333–408. doi: 10.1016/j.pharmthera.2013.01.016

Deforte, S., and Uversky, V. N. (2016). Order, Disorder, and Everything in Between. *Molecules* 21, 1090. doi: 10.3390/molecules21081090

Di Domenico, T., Walsh, I., and Tosatto, S.C. (2013). Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. *BMC Bioinformatics* 14(Suppl. 7), S3.

Di Paola, L., and Giuliani, A. (2015). Protein contact network topology: a natural language for allostery. *Curr Opin Struct Biol* 31, 43–48. doi: 10.1016/j.sbi.2015.03.001

Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., and Giuliani, A. (2013). Protein contact networks: an emerging paradigm in chemistry. *Chem Rev* 113, 1598–1613. doi: 10.1021/cr3002356

Dosztanyi, Z., Chen, J., Dunker, A. K., Simon, I., and Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 5, 2985–2995. doi: 10.1021/pr060171o

Dunker, A. K., and Obradovic, Z. (2001). The protein trinity–linking function and disorder. *Nat Biotechnol* 19, 805–806. doi: 10.1038/nbt0901-805

Dunker, A. K., Brown, C. J., and Obradovic, Z. (2002b). Identification and functions of usefully disordered proteins. *Adv Protein Chem* 62, 25–49. doi: 10.1016/s0065-3233(02)62004-2

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002a). Intrinsic disorder and protein function. *Biochemistry* 41, 6573–6582.

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005). Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS Journal* 272, 5129–5148. doi: 10.1111/j.1742-4658.2005.04948.x

Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., et al. (1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* 473–484.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. *J Mol Graph Model* 19, 26–59.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 11, 161–171.

Dyson, H. J., and Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 12, 54–60. doi: 10.1016/s0959-440x(02)00289-0

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6, 197–208. doi: 10.1038/nrm1589

Eisenmesser, E. Z., Bosco, D. A., Akke, M., and Kern, D. (2002). Enzyme dynamics during catalysis. *Science* 295, 1520–1523. doi: 10.1126/science.1066176

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., et al. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438, 117–121. doi: 10.1038/nature04105

Ekman, D., Light, S., Bjorklund, A. K., and Elofsson, A. (2006). What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol* 7, R45.

Erdös, P., and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* 5, 17–61.

Fankhauser, G. (1945). Maintenance of normal structure in heteroploid salamander larvae, through compensation of changes in cell size by adjustment of cell number and cell shape. *J Exp Zool* 100, 445–455. doi: 10.1002/jez.1401000310

Fankhauser, G. (1972). Memories of great embryologists. Reminiscences of F. Baltzer, H. Spemann, F. R. Lillie, R. G. Harrison, and E. G. Conklin. *Am Sci* 60, 46–55.

Fisher, C. K., and Stultz, C. M. (2011). Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol* 21, 426–431.

Frauenfelder, H., and Petsko, G. A. (1980). Structural dynamics of liganded myoglobin. *Biophys J* 32, 465–483. doi: 10.1016/s0006-3495(80)84984-8

Frauenfelder, H., and Wolynes, P. G. (1994). Biomolecules: where the physics of complexity and simplicity meet. *Physics Today* 47, 58. doi: 10.1063/1.881414

Frauenfelder, H., Chen, G., Berendzen, J., Fenimore, P. W., Jansson, H., Mcmahon, B. H., et al. (2009). A unified model of protein dynamics. *Proc Natl Acad Sci U S A* 106, 5129–5134.

Frauenfelder, H., Petsko, G. A., and Tsernoglou, D. (1979). Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* 280, 558–563. doi: 10.1038/280558a0

Gadiyaram, V., Dighe, A., Ghosh, S., and Vishveshwara, S. (2021). Network Re-Wiring During Allostery and Protein-Protein Interactions: A Graph Spectral Approach. *Methods Mol Biol* 2253, 89–112. doi: 10.1007/978-1-0716-1154-8_7

Giuliani, A. (2021). The statistical mechanics of life: Comment on "Dynamic and thermodynamic models of adaptation" by A.N. Gorban et al. *Phys Life Rev* 37, 100–102.

Giuliani, A., Filippi, S., and Bertolaso, M. (2014). Why network approach can promote a new way of thinking in biology. *Front Genet* 5:83. doi: 10.3389/fgene.2014.00083

Goh, K. I., Oh, E., Jeong, H., Kahng, B., and Kim, D. (2002). Classification of scale-free networks. *Proc Natl Acad Sci U S A* 99, 12583–12588.

Guimera, R., and Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature* 433, 895–900. doi: 10.1038/nature03288

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi: 10.1038/nature02555

Hartmann, H., Parak, F., Steigemann, W., Petsko, G. A., Ponzi, D. R., and Frauenfelder, H. (1982). Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. *Proc Natl Acad Sci U S A* 79, 4967–4971. doi: 10.1073/pnas.79.16.4967

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

Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., et al. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* 2:e100. doi: 10.1371/journal.pcbi.0020100

Hegyi, H., and Tompa, P. (2012). Increased structural disorder of proteins encoded on human sex chromosomes. *Mol Biosyst* 8, 229–236. doi: 10.1039/c1mb05285c

Hilser, V. J., Wrabl, J. O., and Motlagh, H. N. (2012). Structural and energetic basis of allostery. *Annu Rev Biophys* 41, 585–609. doi: 10.1146/annurev-biophys-050511-102319

Hu, G., Wu, Z., Uversky, V. N., and Kurgan, L. (2017). Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. *Int J Mol Sci* 18, 2761. doi: 10.3390/ijms18122761

Jakob, U., Kriwacki, R., and Uversky, V. N. (2014). Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem Rev* 114, 6779–6805. doi: 10.1021/cr400459c

Kahali, B., and Ghosh, T. C. (2013). Disorderness in *Escherichia coli* proteome: perception of folding fidelity and protein-protein interactions. *J Biomol Struct Dyn* 31, 472–476. doi: 10.1080/07391102.2012.706071

Klapper, M. H. (1971). On the nature of the protein interior. *Biochim Biophys Acta* 229, 557–566. doi: 10.1016/0005-2795(71)90271-6

Korneta, I., and Bujnicki, J. M. (2012). Intrinsic disorder in the human spliceosomal proteome. *PLoS Comput Biol* 8:e1002641. doi: 10.1371/journal.pcbi.1002641

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A* 44, 98–104. doi: 10.1073/pnas.44.2.98

Krasowski, M. D., Reschly, E. J., and Ekins, S. (2008). Intrinsic disorder in nuclear hormone receptors. *J Proteome Res* 7, 4359–4372. doi: 10.1021/pr8003024

Kumar, N., Kaushik, R., Tennakoon, C., Uversky, V. N., Longhi, S., Zhang, K. Y. J., et al. (2021). Comprehensive Intrinsic Disorder Analysis of 6108 Viral Proteomes: From the Extent of Intrinsic Disorder Penetrance to Functional Annotation of Disordered Viral Proteins. *J Proteome Res*. 20, 2704–2713. doi: 10.1021/acs.jproteome.1c00011

Laughlin, R. B., Pines, D., Schmalian, J., Stojkovic, B. P., and Wolynes, P. (2000). The middle way. *Proc Natl Acad Sci U S A* 97, 32–37.

Leander, M., Yuan, Y., Meger, A., Cui, Q., and Raman, S. (2020). Functional plasticity and evolutionary adaptation of allosteric regulation. *Proc Natl Acad Sci U S A* 117, 25445–25454. doi: 10.1073/pnas.2002613117

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55, 379–400. doi: 10.1016/0022-2836(71)90324-x

Lerner, E., Cordes, T., Ingargiola, A., Alhadid, Y., Chung, S., Michalet, X., et al. (2018). Toward dynamic structural biology: Two decades of single-molecule Forster resonance energy transfer. *Science* 359, eaan1133. doi: 10.1126/science.aan1133

Lesser, G. J., and Rose, G. D. (1990). Hydrophobicity of amino acid subgroups in proteins. *Proteins* 8, 6–13. doi: 10.1002/prot.340080104

Lopez-Kleine, L., Leal, L., and Lopez, C. (2013). Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. *Brief Funct Genomics* 12, 457–467. doi: 10.1093/bfgp/elt003

Ma, B., Kumar, S., Tsai, C. J., and Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng* 12, 713–720. doi: 10.1093/protein/12.9.713

Midic, U., and Obradovic, Z. (2012). Intrinsic disorder in putative protein sequences. *Proteome Sci* 10(Suppl. 1), S19.

Mikulecky, D. C. (2001). Network thermodynamics and complexity: a transition to relational systems theory. *Comput Chem* 25, 369–391. doi: 10.1016/s0097-8485(01)00072-9

Na, I., Redmon, D., Kopa, M., Qin, Y., Xue, B., and Uversky, V. N. (2013). Ordered disorder of the astrocytic dystrophin-associated protein complex in the norm and pathology. *PLoS One* 8:e73476. doi: 10.1371/journal.pone.0073476

Ni, D., Lu, S., and Zhang, J. (2019). Emerging roles of allosteric modulators in the regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug discovery. *Med Res Rev* 39, 2314–2342. doi: 10.1002/med.21585

Nicosia, V., Domenico, M. D., and Latora, V. (2014). Characteristic exponents of complex networks. *Europhys Lett* 106, 58005. doi: 10.1209/0295-5075/106/58005

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 44, 1989–2000. doi: 10.1021/bi047993o

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., and Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9(Suppl 1):S1. doi: 10.1186/1471-2164-9-S1-S1

Olsson, M. H., Parson, W. W., and Warshel, A. (2006). Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis. *Chem Rev* 106, 1737–1756. doi: 10.1021/cr040427e

Pace, N. C., Scholtz, J. M., and Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS Lett* 588, 2177–2184. doi: 10.1016/j.febslet.2014.05.006

Pancsa, R., and Tompa, P. (2012). Structural disorder in eukaryotes. *PLoS One* 7:e34687. doi: 10.1371/journal.pone.0034687

Parak, F., Frolov, E. N., Mossbauer, R. L., and Goldanskii, V. I. (1981). Dynamics of metmyoglobin crystals investigated by nuclear gamma resonance absorption. *J Mol Biol* 145, 825–833. doi: 10.1016/0022-2836(81)90317-x

Pascual, M., and Levin, S. A. (1999). FROM INDIVIDUALS TO POPULATION DENSITIES: SEARCHING FOR THE INTERMEDIATE SCALE OF NONTRIVIAL DETERMINISM. *Ecology* 80, 2225–2236. doi: 10.1890/0012-9658(1999)080[2225:fitpds]2.0.co;2

Patil, A., and Nakamura, H. (2006). Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett* 580, 2041–2045. doi: 10.1016/j.febslet.2006.03.003

Peng, Z., Yan, J., Fan, X., Mizianty, M. J., Xue, B., Wang, K., et al. (2015). Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci* 72, 137–151. doi: 10.1007/s00018-014-1661-9

Pentony, M. M., and Jones, D. T. (2010). Modularity of intrinsic disorder in the human proteome. *Proteins* 78, 212–221. doi: 10.1002/prot.22504

Petersen, S. E., and Sporns, O. (2015). Brain Networks and Cognitive Architectures. *Neuron* 88, 207–219. doi: 10.1016/j.neuron.2015.09.027

Privalov, P. L. (1979). Stability of proteins: small globular proteins. *Adv Protein Chem* 33, 167–241. doi: 10.1016/s0065-3233(08)60460-x

Privalov, P. L. (1982). Stability of proteins. Proteins which do not present a single cooperative system. *Adv Protein Chem* 35, 1–104.

Radivojac, P., Vucetic, S., O'connor, T. R., Uversky, V. N., Obradovic, Z., and Dunker, A. K. (2006). Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins* 63, 398–410. doi: 10.1002/prot.20873

Rajagopalan, P. T., and Benkovic, S. J. (2002). Preorganization and protein dynamics in enzyme catalysis. *Chem Rec* 2, 24–36. doi: 10.1002/tcr.10009

Richards, F. M. (1963). Structure of Proteins. *Annu Rev Biochem* 32, 269–300.

Schad, E., Tompa, P., and Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biol* 12, R120.

Shimizu, K., and Toh, H. (2009). Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. *J Mol Biol* 392, 1253–1265. doi: 10.1016/j.jmb.2009.07.088

Singh, G. P., Ganapathi, M., Sandhu, K. S., and Dash, D. (2006). Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes. *Proteins* 62, 309–315. doi: 10.1002/prot.20746

Sporns, O. (2018). Graph theory methods: applications in brain networks. *Dialogues Clin Neurosci* 20, 111–121. doi: 10.31887/dcns.2018.20.2/osporns

Stickle, D. F., Presta, L. G., Dill, K. A., and Rose, G. D. (1992). Hydrogen bonding in globular proteins. *J Mol Biol* 226, 1143–1159. doi: 10.1016/0022-2836(92)91058-w

Strogatz, S. H. (2001). Exploring complex networks. *Nature* 410, 268–276. doi: 10.1038/35065725

Sutcliffe, M. J., and Scrutton, N. S. (2002). A new conceptual framework for enzyme catalysis. Hydrogen tunnelling coupled to enzyme dynamics in flavoprotein and quinoprotein enzymes. *Eur J Biochem* 269, 3096–3102. doi: 10.1046/j.1432-1033.2002.03020.x

Tanford, C. (1968). Protein denaturation. *Adv Protein Chem* 23, 121–282.

Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., and Tawfik, D. S. (2009). Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* 34, 53–59. doi: 10.1016/j.tibs.2008.10.009

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci* 27, 527–533.

Tompa, P., and Kalmar, L. (2010). Power law distribution defines structural disorder as a structural element directly linked with function. *J Mol Biol* 403, 346–350. doi: 10.1016/j.jmb.2010.07.044

Tompa, P., and Rose, G. D. (2011). The Levinthal paradox of the interactome. *Protein Sci* 20, 2074–2079. doi: 10.1002/pro.747

Tompa, P., Dosztanyi, Z., and Simon, I. (2006). Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J Proteome Res* 5, 1996–2000. doi: 10.1021/pr0600881

Tousignant, A., and Pelletier, J. N. (2004). Protein motions promote catalysis. *Chem Biol* 11, 1037–1042. doi: 10.1016/j.chembiol.2004.06.007

Turoverov, K. K., Kuznetsova, I. M., and Uversky, V. N. (2010). The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. *Prog Biophys Mol Biol* 102, 73–84. doi: 10.1016/j.pbiomolbio.2010.01.003

Uversky, V. N. (2010). The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J Biomed Biotechnol* 2010, 568068.

Uversky, V. N. (2013a). A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci* 22, 693–724. doi: 10.1002/pro.2261

Uversky, V. N. (2013b). Intrinsic disorder-based protein interactions and their modulators. *Curr Pharm Des* 19, 4191–4213. doi: 10.2174/1381612811319230005

Uversky, V. N. (2013c). Unusual biophysics of intrinsically disordered proteins. *Biochim Biophys Acta* 1834, 932–951. doi: 10.1016/j.bbapap.2012.12.008

Uversky, V. N. (2015). Functional roles of transiently and intrinsically disordered regions within proteins. *FEBS J* 282, 1182–1189. doi: 10.1111/febs.13202

Uversky, V. N. (2016a). Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins. *J Biol Chem* 291, 6681–6688. doi: 10.1074/jbc.r115.685859

Uversky, V. N. (2016b). p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. *Int J Mol Sci* 17, 1874. doi: 10.3390/ijms17111874

Uversky, V. N. (2016c). Paradoxes and wonders of intrinsic disorder: Complexity of simplicity. *Intrinsically Disord Proteins* 4, e1135015. doi: 10.1080/21690707.2015.1135015

Uversky, V. N. (2019a). Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Frontiers in Physics* 7:10. doi: 10.3389/fphy.2019.00010

Uversky, V. N. (2019b). Protein intrinsic disorder and structure-function continuum. *Prog Mol Biol Transl Sci* 166, 1–17. doi: 10.1016/bs.pmbts.2019.05.003

Uversky, V. N., and Dunker, A. K. (2010). Understanding protein non-folding. *Biochim Biophys Acta* 1804, 1231–1264.

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 18, 343–384. doi: 10.1002/jmr.747

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

van der Groen, O., Tang, M. F., Wenderoth, N., and Mattingley, J. B. (2018). Stochastic resonance enhances the rate of evidence accumulation during combined brain stimulation and perceptual decision-making. *PLoS Comput Biol* 14:e1006301. doi: 10.1371/journal.pcbi.1006301

Vickaryous, M. K., and Hall, B. K. (2006). Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc* 81, 425–455. doi: 10.1017/s1464793106007068

Villa, J., Strajbl, M., Glennon, T. M., Sham, Y. Y., Chu, Z. T., and Warshel, A. (2000). How important are entropic contributions to enzyme catalysis? *Proc Natl Acad Sci U S A* 97, 11899–11904. doi: 10.1073/pnas.97.22.11899

Ward, J. J., Sodhi, J. S., Mcguffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337, 635–645. doi: 10.1016/j.jmb.2004.02.002

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Weaver, W. (1948). Science and complexity. *Am Sci* 36, 536–544.

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293, 321–331. doi: 10.1006/jmbi.1999.3110

Xue, B., Blocquel, D., Habchi, J., Uversky, A. V., Kurgan, L., Uversky, V. N., et al. (2014). Structural disorder in viral proteins. *Chem Rev* 114, 6880–6911.

Xue, B., Dunker, A. K., and Uversky, V. N. (2012). Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* 30, 137–149. doi: 10.1080/07391102.2012.675145

Xue, B., Williams, R. W., Oldfield, C. J., Dunker, A. K., and Uversky, V. N. (2010). Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst. Biol.* 4(Suppl. 1):S1. doi: 10.1186/1752-0509-4-S1-S1

Yang, L. W., and Bahar, I. (2005). Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* 13, 893–904. doi: 10.1016/j.str.2005.03.015

Zhang, L., Li, M., and Liu, Z. (2018). A comprehensive ensemble model for comparing the allosteric effect of ordered and disordered proteins. *PLoS Comput Biol* 14:e1006393. doi: 10.1371/journal.pcbi.1006393

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.

The reviewer GH declared a past co-authorship with one of the authors AG to the handling editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/fgene.2021.656042

Identification of Potential Signatures and Their Functions for Acute Lymphoblastic Leukemia: A Study Based on the Cancer Genome Atlas

Weimin Wang, Chunhui Lyu, Fei Wang, Congcong Wang, Feifei Wu, Xue Li and Silin Gan*

Department of Hematology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Objective: Acute lymphoblastic leukemia (ALL) is a malignant disease most commonly diagnosed in adolescents and young adults. This study aimed to explore potential signatures and their functions for ALL.

Methods: Differentially expressed mRNAs (DEmRNAs) and differentially expressed long non-coding RNAs (DElncRNAs) were identified for ALL from The Cancer Genome Atlas (TCGA) and normal control from Genotype-Tissue Expression (GTEx). DElncRNA–microRNA (miRNA) and miRNA–DEmRNA pairs were predicted using online databases. Then, a competing endogenous RNA (ceRNA) network was constructed. Functional enrichment analysis of DEmRNAs in the ceRNA network was performed. Protein–protein interaction (PPI) network was then constructed. Hub genes were identified. DElncRNAs in the ceRNA network were validated using Real-time qPCR.

Results: A total of 2,903 up- and 3,228 downregulated mRNAs and 469 up- and 286 downregulated lncRNAs were identified for ALL. A ceRNA network was constructed for ALL, consisting of 845 lncRNA-miRNA and 395 miRNA–mRNA pairs. These DEmRNAs in the ceRNA network were mainly enriched in ALL-related biological processes and pathways. Ten hub genes were identified, including SMAD3, SMAD7, SMAD5, ZFYVE9, FKBP1A, FZD6, FZD7, LRP6, WNT1, and SFRP1. According to Real-time qPCR, eight lncRNAs including ATP11A-AS1, ITPK1-AS1, ANO1-AS2, CRNDE, MALAT1, CACNA1C-IT3, PWRN1, and WT1-AS were significantly upregulated in ALL bone marrow samples compared to normal samples.

Conclusion: Our results showed the lncRNA expression profiles and constructed ceRNA network in ALL. Furthermore, eight lncRNAs including ATP11A-AS1, ITPK1-AS1, ANO1-AS2, CRNDE, MALAT1, CACNA1C-IT3, PWRN1, and WT1-AS were identified. These results could provide a novel insight into the study of ALL.

Keywords: acute lymphoblastic leukemia, long non-coding RNAs, functional enrichment analysis, competing endogenous RNAs, hub genes interaction

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Michelle Lacey,
Tulane University, United States
Xianghu Li,
Zhejiang University, China

****Correspondence:***
Weimin Wang
fccwangwm@zzu.edu.cn

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

Received: *20 January 2021*
Accepted: *12 May 2021*
Published: *06 July 2021*

Citation:
Wang W, Lyu C, Wang F, Wang C, Wu F, Li X and Gan S (2021) Identification of Potential Signatures and Their Functions for Acute Lymphoblastic Leukemia: A Study Based on the Cancer Genome Atlas. Front. Genet. 12:656042. doi: 10.3389/fgene.2021.656042

INTRODUCTION

Acute lymphoblastic leukemia (ALL) is a malignant disease most commonly diagnosed in adolescents and young adults, especially in patients younger than 15 years. Despite significant improvements in the management of ALL, the long-term survival rate of ALL patients, especially adult patients, remains low (Jabbour et al., 2018; Richard-Carpentier et al., 2019). Therefore, it is of importance to understand the pathogenesis of ALL and identify novel diagnostic biomarkers and therapeutic targets for ALL.

LncRNA is a type of RNA longer than 200 nucleotides. Dysregulated lncRNA as tumor suppressor genes or oncogenes plays a key role in a variety of biological processes, such as cell proliferation, apoptosis, migration, and invasion. Increasing studies are focusing on the role and mechanism of lncRNA in the occurrence and development of ALL (Trimarchi et al., 2014; Arthur et al., 2017). For instance, lncRNA CASC15 could regulate SOX4 expression in RUNX1-translocated leukemia (Fernando et al., 2017). LncRNA HOTAIR is closely associated with acute leukemia patients' poor prognosis (Zhang et al., 2016). LncRNA HOXA-AS2 induces glucocorticoid resistance by promoting ALL cell proliferation and inhibiting apoptosis (Zhao et al., 2019). Despite the fact that many studies have shown the diagnostic and prognostic values of lncRNAs in ALL, it is still required to further understand their regulatory mechanism. It has been widely accepted that lncRNAs indirectly regulate gene expression through targeted miRNAs (about 20 nucleotides) at the transcriptional or post-transcriptional level. Many miRNAs have been found to play a functional regulatory role in the development of ALL, such as miRNA-126 (Nucera et al., 2016), miRNA-155 (El-Khazragy et al., 2019), and miR-141-3p (Zhou et al., 2019). Yet, the regulatory interactions between lncRNAs and miRNAs in ALL require to be clarified.

The development of transcriptome analysis and RNA sequencing technology is increasing the possibility of identifying lncRNAs that may be involved in the pathogenesis of ALL. Moreover, further studies on the function of abnormally expressed lncRNAs may help understand the pathogenesis of ALL and provide important insights for the treatment of ALL. In this study, we comprehensively analyzed DElncRNAs and DEmRNAs in bone marrow samples of ALL. A ceRNA network was constructed for ALL on the basis of DElncRNA–miRNA and miRNA–DEmRNA pairs. DEmRNAs in the ceRNA network were significantly associated with ALL-related biological processes and pathways. Among DElncRNAs in the ceRNA network, eight lncRNAs including ATP11A-AS1, ITPK1-AS1, ANO1-AS2, CRNDE, MALAT1, CACNA1C-IT3, PWRN1, and WT1-AS were validated by Real-time qPCR, which could become potential diagnostic and therapeutic targets of ALL.

Abbreviations: ALL, acute lymphoblastic leukemia; DEmRNAs, differentially expressed mRNAs; DElncRNAs, differentially expressed long non-coding RNAs; TCGA, The Cancer Genome Atlas; GTEx, Genotype-Tissue Expression; ceRNA, competing endogenous RNA; PPI, protein–protein interaction; FC, fold change; GO, Gene Ontology; BP, biological process; CC, cellular component; MF, molecular function; KEGG, Kyoto Encyclopedia of Genes and Genomes.

MATERIALS AND METHODS

ALL Data Acquisition and Differential Expression Analysis

LncRNA and mRNA RNA-seq data of 494 bone marrows with ALL (hematopoietic and reticuloendothelial systems) were retrieved from TCGA repository[1], which were derived from the IlluminaHiSeq RNA-Seq platform. All the data from three phases together, including 12 cases of phase 1, 468 cases of phase 2, and 14 cases of phase 3 were enrolled in the study. There were 321 (64.98%) males, 172 (34.82%) females, and 1 unknown (0.02%). The age distribution of the ALL group is as follows: 403 cases of 0–14 years old and 91 cases of $\geq$ 14 years old. All data of normal tissue samples were obtained from 407 whole blood in the Genotype-Tissue Expression (GTEx) database[2]. There were 265 (65.11%) males and 142 (34.89%) females in the control group. The age distribution of the control group is as follows: 34 cases of 20–29 years old, 34 cases of 30–39 years old, 72 cases of 40–49 years old, 130 cases of 50–59 years old, 132 cases of 60–69 years old, and 5 cases of 70–79 years old. Complete description of the multiple ethnicity groups, the biospecimen procurement methods, and sample fixation was provided in the GTEx official annotation. Differential expression analyses between ALL samples and normal samples were carried out using the EdgeR package in R (Robinson et al., 2010). The obtained p-values were corrected by false discovery rate (FDR). mRNAs and lncRNAs with adjusted $p < 0.05$ and $|\log 2\text{fold change (FC)}| \geq 2$ were considered as DEmRNAs and DElncRNAs. Volcano plots and heatmaps were generated using the ggplot2 and packages in R, respectively.

ceRNA Network Construction

After identification of DElncRNAs and DEmRNAs, lncRNA–miRNA pairs were predicted by miRcode[3] that provides > 10,000 lncRNAs (Jeggari et al., 2012). Then, miRNAs that targeted DEmRNAs were predicted using TargetScan[4] (Agarwal et al., 2015), miRDB[5] (Wang, 2008), and miRTarBase database[6], which provides an experimentally validated microRNA–target interactions database (Huang et al., 2020). After integration of DElncRNA–miRNA and miRNA–DEmRNA, a ceRNA network was constructed and visualized using the Cytoscape software (version 3.5.1) (Shannon et al., 2003).

Functional Enrichment Analyses of DEmRNAs in the ceRNA Network

Gene Ontology (GO) analysis of DEmRNAs in the ceRNA network was carried out using Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Dennis et al., 2003), including biological process (BP), cellular

[1]https://portal.gdc.cancer.gov/
[2]https://gtexportal.org/home/datasets
[3]http://www.mircode.org/
[4]http://www.targetscan.org/
[5]http://www.mirdb.org/
[6]http://mirtarbase.mbc.nctu.edu.tw/

TABLE 1 | Primer sequence information for Real-time qPCR.

Gene symbol	Primer sequence (5′–3′)
Human GAPDH	5′-CGGAGTCAACGGATTTGGTCGTAT-3′ (forward) 5′-AGCCTTCTCCATGGTGGTGAAGAC-3′ (reverse)
Human ANO1-AS2	5′-CCGGAACAAGAACCTCGCTC-3′ (forward) 5′-GGTCCTCGCCTACCATCCAA-3′ (reverse)
Human PWRN1	5′-ACATTCGAAACCCAGGTGCC-3′ (forward) 5′-GGAAGTGGATGCTGACGCTC-3′ (reverse)
Human MALAT1	5′-GGTTCAGAAGGTCTGAAGCTC-3′ (forward) 5′-CCCAGAAGTGTTTACACTGCT-3′ (reverse)
Human CACNA1C-IT3	5′-GCCAGGACCAAGACACCAAGAC-3′ (forward) 5′-TTGGGCAGGGCTCGGTTCC-3′ (reverse)
Human ITPK1-AS1	5′-AATCCTGTGCGCTGTCATCC-3′ (forward) 5′-GATTGCTCTTGGCTGTGCCT-3′ (reverse)
Human ATP11A-AS2	5′-ACAGTCCCTTCCCTTACGCT-3′ (forward) 5′-TGAACGCTGCACTTGTGGAC-3′ (reverse)
Human CRNDE	5′-GAGGACGTGCTGGGGCT-3′ (forward) 5′-CTGAGTCCATGTCCCGAATC-3′(reverse)
Human WT1-AS	5′-GCCTCTCTGTCCTCTTCTTTGT-3′ (forward) 5′-GCTGTGAGTCCTGGTGCTTAG-3′ (reverse)

component (CC), and molecular function (MF). Moreover, Kyoto Encyclopedia of Genes and Genomes (KEGG) was analyzed using the clusterProfiler in R (Yu et al., 2012). Furthermore, the KEGG results were visualized using the Cytoscape plug-in ClueGO. $p < 0.05$ was set as the cutoff value.

PPI Network

The interactions between proteins were predicted using the Search Tool for the Retrieval of Interacting Genes (STRING) database[7] (minimum required interaction score > 0.4) (Szklarczyk et al., 2019). Furthermore, PPI networks were embodied using the Cytoscape v3.5.0 software. In addition, we used Molecular Complex Detection (MCODE) plugin to identify the hub genes in the PPI network. The criteria were set as follows: MCODE scores > 3 and number of nodes > 4. The top 10 hub genes were identified using the ranking method of degree.

[7]http://string-db.org/

Real-Time qPCR

Bone marrow samples were isolated from 25 ALL patients and 15 healthy participants and red blood cells were removed. Total RNA was extracted from bone marrow samples and then was stored at −80°C. Extracted samples were lysed using 1 ml of Trizol and placed for 5 min on ice. RNA concentration and purity were determined using a NanoDrop UV spectrophotometer. Then, RNA was reverse transcribed into cDNA. Primer sequences of ATP11A-AS1, ITPK1-AS1, ANO1-AS2, CRNDE, MALAT1, CACNA1C-IT3, PWRN1, and WT1-AS were designed and synthesized by Shanghai Shengong Biological Engineering Co., Ltd. (Shanghai, China). The primer sequences are listed in **Table 1**. PCR amplification had the following conditions: 95°C for 3 min; 40 PCR cycle reactions (95°C for 20 s; 60°C for 30 s). GAPDH was used as a control. Relative expression levels of lncRNAs were calculated using $2^{-\Delta\Delta CT}$ method. Differences between the two groups were analyzed using Student's *t*-test. *p*-value < 0.05 was considered statistically significant.

RESULTS

Identification of DElncRNAs and DEmRNAs for ALL

The workflow of this study is shown in **Figure 1**. According to adjusted *p*-value < 0.05 and | log2 fold change (FC)| ≥ 2, 2903 up− and 3228 downregulated mRNAs were identified for ALL, as shown in the volcano plot (**Figure 2A** and

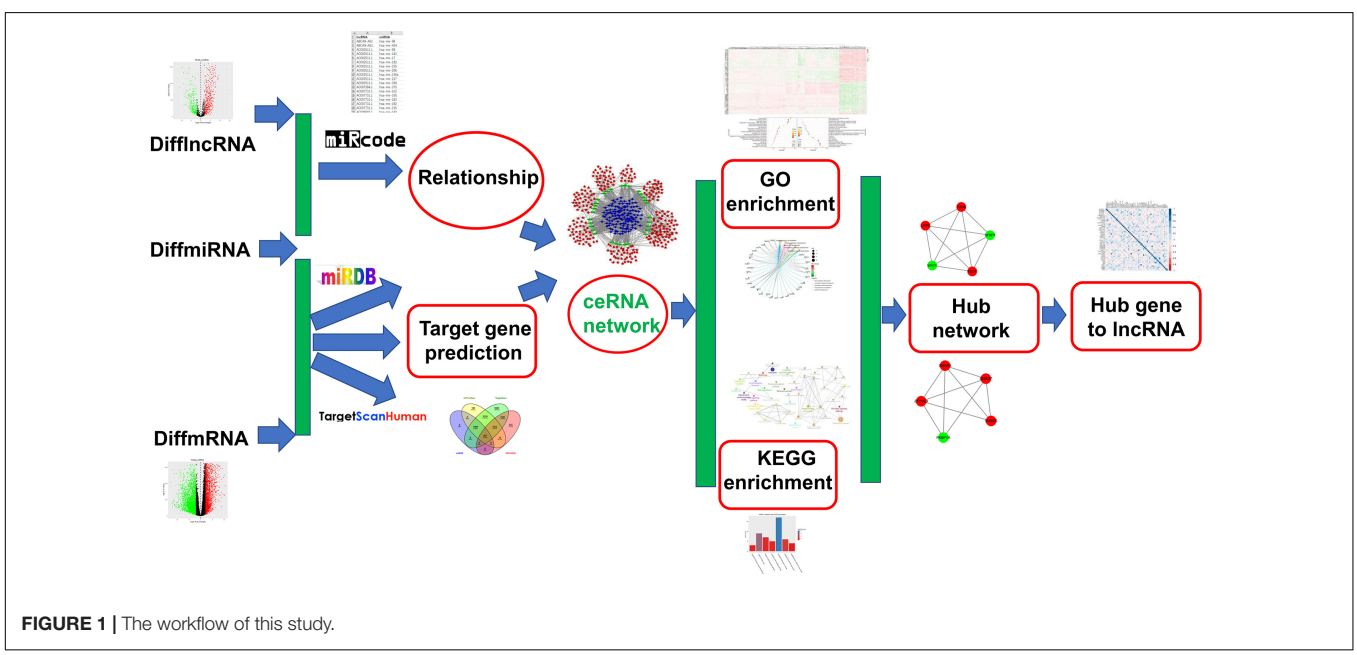


FIGURE 1 | The workflow of this study.

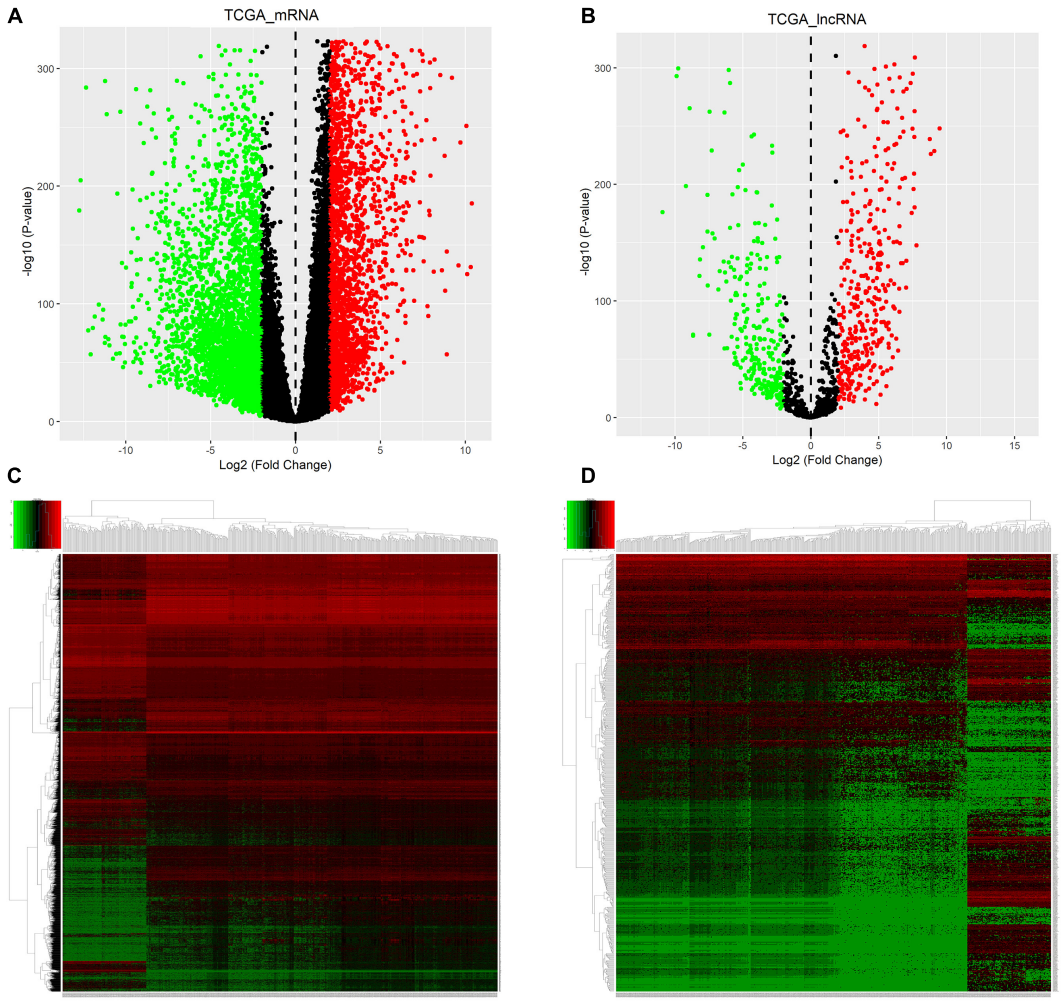


FIGURE 2 | Identification of DEmRNAs and DElncRNAs for ALL. Volcano plot showing DEmRNAs **(A)** and DElncRNAs **(B)** for ALL. As shown in heatmaps, the differences in expression patterns of DEmRNAs **(C)** and DElncRNAs **(D)** between ALL bone marrow samples and normal samples. Red represents upregulation and green represents downregulation. DEmRNAs: differentially expressed mRNAs; DElncRNAs: Differentially expressed lncRNAs; ALL, Acute lymphoblastic leukemia.

Supplementary Material 1). Furthermore, there were 469 up– and 286 downregulated lncRNAs for ALL (**Figure 2B** and **Supplementary Material 2**). Heatmaps depicted the differences in expression patterns of all DEmRNAs (**Figure 2C**) and DElncRNAs (**Figure 2D**) between ALL bone marrow samples and normal samples.

Construction of ceRNA Network for ALL

The miRNAs that targeted DEmRNAs were predicted using TargetScan, miRDB, and miRTarBase databases. After integration of prediction results from the three databases, 297 DEmRNAs were intersected and identified for the construction of ceRNA network (**Figure 3**). Furthermore, DElncRNA–miRNA relationships were predicted using miRcode database. By comprehensively analyzing DElncRNA–miRNA and miRNA–DEmRNA pairs, a ceRNA network was constructed for ALL (**Figure 4**). There were 845 lncRNA–miRNA pairs (**Supplementary Material 3**) and 395 miRNA–mRNA pairs (**Supplementary Material 4**) in the ceRNA network.

Functional Enrichment Analysis of DEmRNAs in the ceRNA Network

As depicted in heatmaps, there were obvious differences in the expression patterns of all DEmRNAs in the ceRNA network between ALL bone marrow samples and normal samples (**Figure 5A**). Bubble diagrams showed the top 40 GO enrichment analysis results enriched by DEmRNAs in the ceRNA network (**Figure 5B**). We found that these mRNAs were mainly enriched in ALL-related biological processes such as transcription, programmed cell death, apoptosis, cell cycle, proliferation, and so on. **Figure 6** depicted the relationships between DEmRNAs and enriched biological processes including morphogenesis of an epithelium, kidney epithelium development, ureteric bud development, mesonephric epithelium development, and mesonephric tubule development. As for KEGG pathway

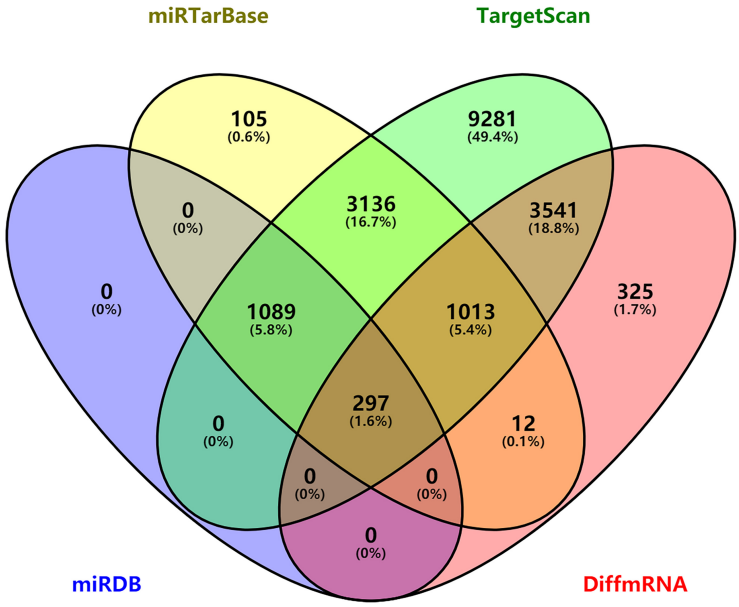


FIGURE 3 | Venn diagram showing 297 differentially expressed mRNAs targeted by miRNAs *via* intersection of prediction results of TargetScan, miRDB, and miRTarBase database.

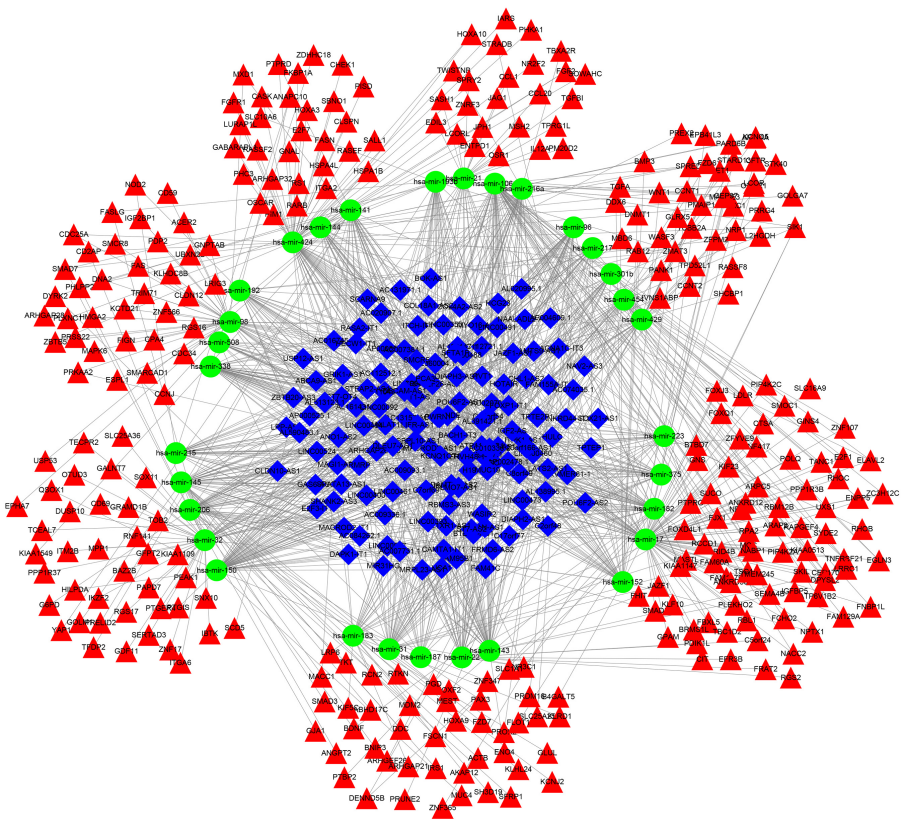


FIGURE 4 | A ceRNA network construction for acute lymphoblastic leukemia. Blue rhombus represents lncRNAs; green circle represents miRNAs and red triangle represents mRNAs.

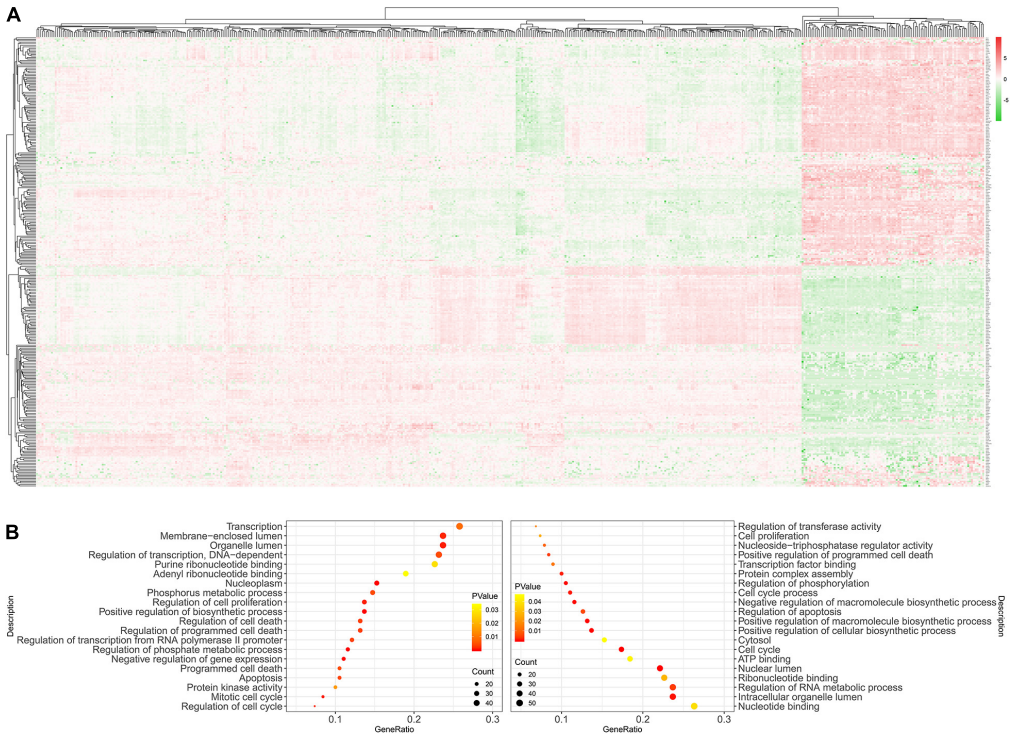


FIGURE 5 | GO enrichment analysis of DEmRNAs in the ceRNA network. **(A)** Heatmaps showing differences in the expression patterns of all DEmRNAs in the ceRNA network between ALL bone marrow samples and normal samples. Red stands for upregulation and green stands for downregulation. **(B)** The top 40 GO enrichment analysis results including biological process, cellular component, and molecular function. DEmRNAs: Differentially expressed mRNAs.

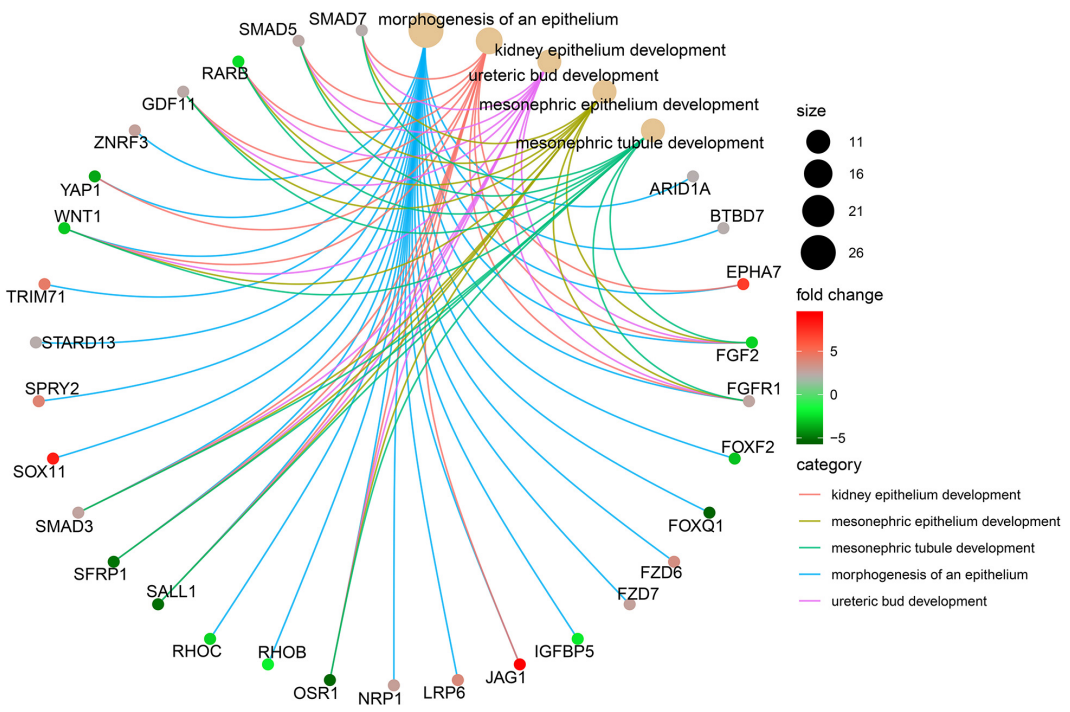


FIGURE 6 | The top five biological processes enriched by DEmRNAs in the ceRNA network. DEmRNAs: Differentially expressed mRNAs.

enrichment analysis results, these DEmRNAs were mainly enriched in pathways in cancer, cell cycle, small cell lung cancer, p53 signaling pathway, Wnt signaling pathway, pentose phosphate pathway, and non-small cell lung cancer (**Figures 7A,B**).

Identification of Hub Genes in the PPI Network

The DEmRNAs in the ceRNA network were imported into STRING database. Then, a PPI network was constructed for ALL (**Figure 8A**). Two PPI subnetworks were then constructed (**Figures 8B,C**). Ten hub genes were identified for ALL, including SMAD3, SMAD7, SMAD5, ZFYVE9, FKBP1A, FZD6, FZD7, LRP6, WNT1, and SFRP1.

Correlation Between Hub Genes and DElncRNAs

Correlation analysis between hub genes and DElncRNAs was performed by corrplot package. The significant correlations between DElncRNAs and hub genes are shown in **Figure 9** and **Supplementary Material 5**. There was strong correlation between WT1-AS and FZD7 (r = 0.751907203; p < 0.0001).

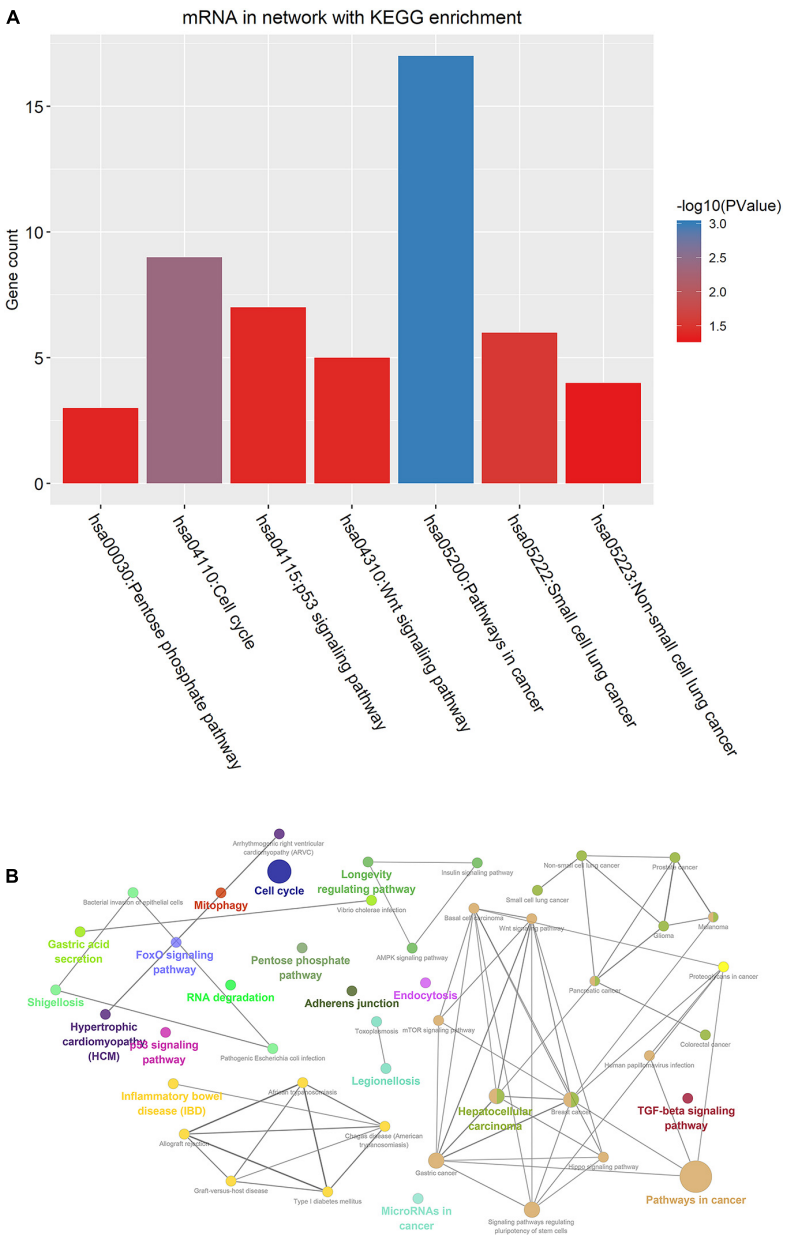


FIGURE 7 | KEGG pathway enrichment analysis of DEmRNAs in the ceRNA network. **(A)** Seven enriched KEGG pathways. **(B)** Visualization of KEGG enrichment analysis results. DEmRNAs: Differentially expressed mRNAs.

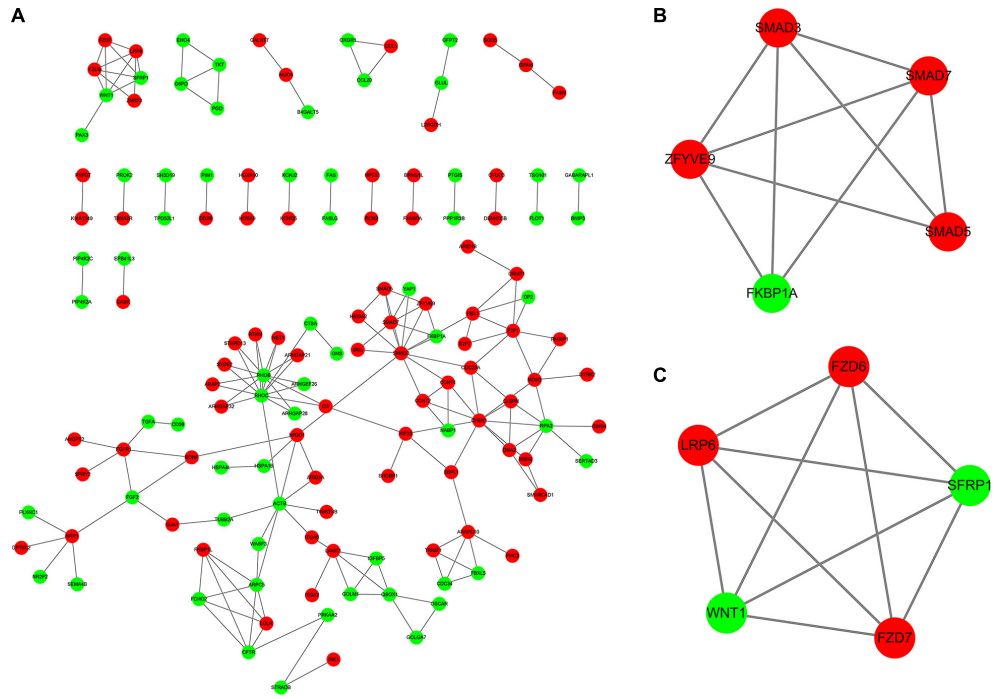


FIGURE 8 | Identification of hub genes in the PPI network. **(A)** Construction of PPI network based on the DEmRNAs in the ceRNA network. **(B,C)** Two PPI subnetworks for ALL. Red represents upregulation and green represents downregulation. DEmRNAs: Differentially expressed mRNAs.

Furthermore, PWRN1 and SMAD3 were significantly correlated ($r = 0.521493415$ and $p = 4.32E{-}08$).

Validation of Eight lncRNAs in Bone Marrow of ALL

Among all DElncRNAs in the ceRNA network, the most significant difference between 10 lncRNAs (AC009093.1, C17orf77, ATP11A-AS1, ITPK1-AS1, ANO1-AS2, ITCH-IT1, CRNDE, MALAT1, CACNA1C-IT3, and PWRN1) in the ceRNA network and WT1-AS (which was closely related to hub gene FZD7) was selected for verification. However, as the primers of AC009093.1, C17orf77, and ITCH-IT1 for RQ-PCR were not ideal, the remaining eight lncRNAs were validated. As **Figure 10** shows, these eight lncRNAs were significantly upregulated in ALL bone marrow samples ($n = 25$) compared to normal samples ($n = 15$) by Real-time qPCR.

DISCUSSION

In this study, we constructed a ceRNA network for ALL based on DElncRNA–miRNA and miRNA–DEmRNA relationships. Among all DElncRNAs in the ceRNA network, eight lncRNAs were validated in ALL bone marrow samples using Real-time qPCR. These lncRNAs might become potential biomarkers for ALL.

To explore potential functions of DEmRNAs in the ceRNA network, we performed functional enrichment analysis. We found that these mRNAs were mainly enriched in ALL-related biological processes such as transcription (Gocho and Yang, 2019), programmed cell death (Hass et al., 2016), apoptosis, cell cycle (Jing et al., 2018), and proliferation (Sun et al., 2019). The DEmRNAs in these biological processes could modulate the development of ALL. Furthermore, these DEmRNAs were significantly associated with pathways in cancer, cell cycle, p53 signaling pathway, Wnt signaling pathway, and pentose phosphate pathway. It has been widely accepted that the p53 signaling pathway is a promising drug target in ALL (Trino et al., 2016). In particular, alterations of the tumor suppressor gene TP53 were frequently found in pediatric ALL (Demir et al., 2020). As for the Wnt signaling pathway, it was significantly correlated with the pathogenesis of ALL (Montano et al., 2018). Recent findings reported that inhibiting Wnt/β catenin could reverse multidrug resistance in children ALL (Fu et al., 2019). Moreover, the pathway is regulated by many factors. For example, miR-181a-5p could promote ALL cell proliferation *via* targeting the Wnt pathway (Lyu et al., 2017). Our results indicated that the DEmRNAs in the ceRNA network could be involved in the pathogenesis of ALL.

We constructed a PPI network for B-ALL on the basis of DEmRNAs in the ceRNA network. Ten hub genes were identified for ALL, including SMAD3, SMAD7, SMAD5, ZFYVE9, FKBP1A, FZD6, FZD7, LRP6, WNT1, and SFRP1. Among them, the loss of the Smad3 protein has been identified as a key feature of acute T-cell lymphoblastic leukemia (Wolfraim et al., 2004). Smad7 is a promising therapeutic target for B-cell ALL (Guo et al., 2018). Furthermore, microRNA-181a might regulate its expression for pediatric ALL (Nabhan et al., 2017). Wnt

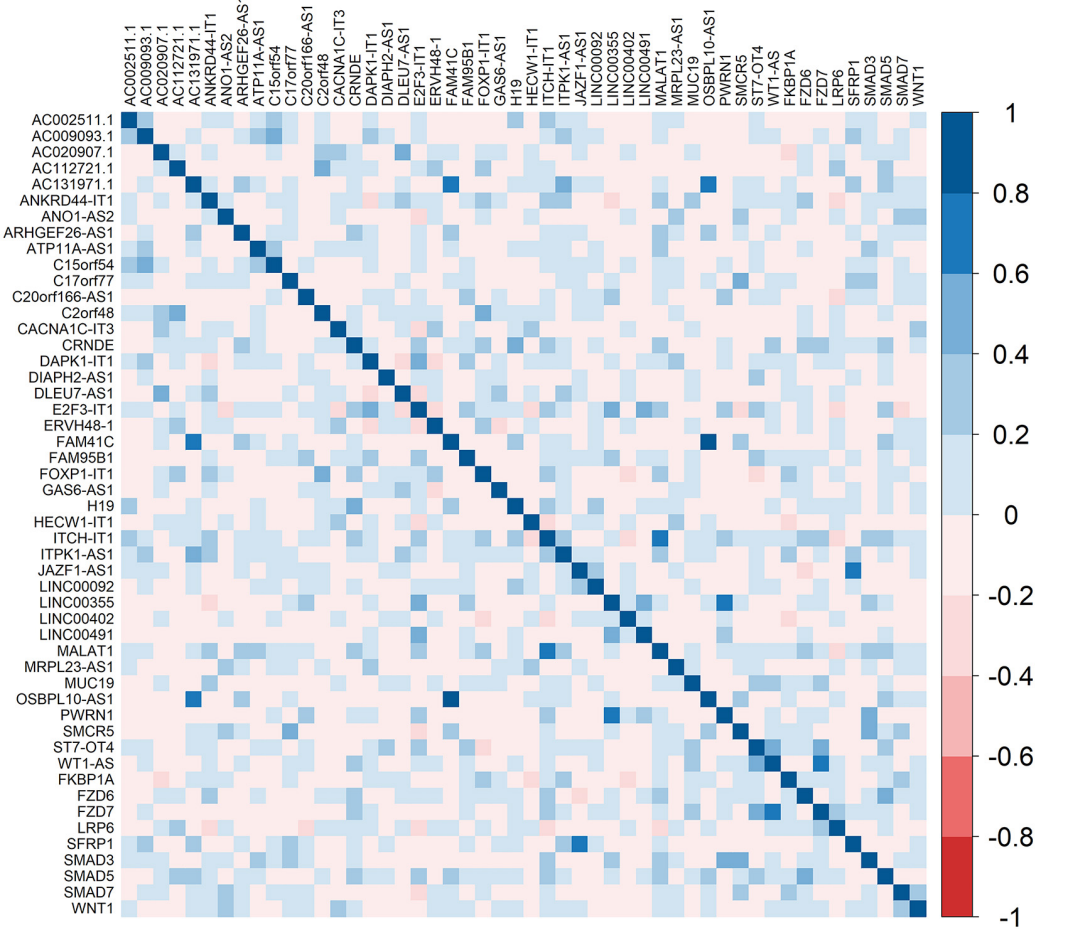


FIGURE 9 | Heatmaps showing the correlation between hub genes and DElncRNAs. DEmRNAs: Differentially expressed mRNAs. The right bar indicates the color legend of Pearson correlation values.

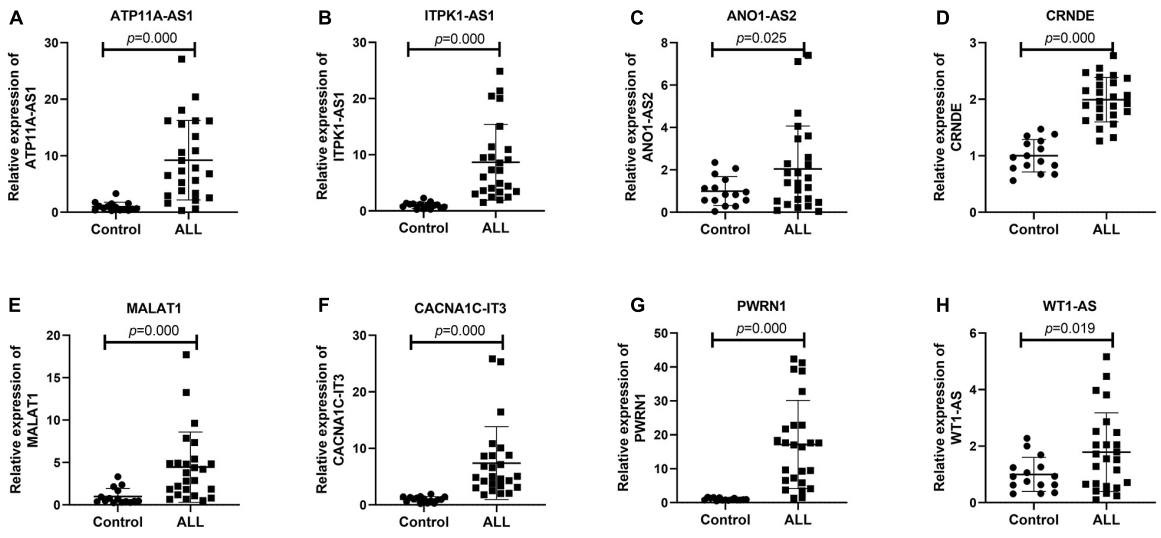


FIGURE 10 | Validation of eight lncRNAs in ALL bone marrow samples using Real-time qPCR. **(A)** ATP11A-AS1; **(B)** ITPK1-AS1; **(C)** ANO1-AS2; **(D)** CRNDE; **(E)** MALAT1; **(F)** CACNA1C-IT3; **(G)** PWRN1; and **(H)** WT1-AS. Control: n = 15; ALL: n = 25.

signaling pathway can enhance hematopoietic cell proliferation (Doubravska et al., 2008). It could mediate growth and prognosis of B-cell progenitor ALL, which could be a potential treatment strategy in ALL (Khan et al., 2007; Mochmann et al., 2011). In the pathway, FZD6, FZD7, LRP6, and WNT1 were marker proteins. LRP6 has been reported to be a candidate tumor suppressor gene in pre-B ALL (Montpetit et al., 2004). Furthermore, low expression of SFRP1 was significantly associated with clinical outcomes of patients with Philadelphia-positive ALL (Martin et al., 2008).

Consistently with differential expression analysis results, eight lncRNAs including ATP11A-AS1, ITPK1-AS1, ANO1-AS2, CRNDE, MALAT1, CACNA1C-IT3, PWRN1, and WT1-AS were significantly upregulated in ALL bone marrow, indicating that these abnormally expressed lncRNAs could be involved in the development of ALL. Among them, CRNDE was upregulated in the bone marrow of B-cell precursor acute lymphoblastic leukemia (BCP-ALL) patients and BCP-ALL cell lines (NALM-6 and RS4;11). Functionally, CRNDE upregulated CREB expression by suppressing miR-345-5p, thus promoting cell proliferation and reducing cell apoptosis in BCP-ALL (Wang W. et al., 2020). A large amount of research has reported that aberrantly expressed MALAT1 was involved in a variety of cancers, such as breast cancer metastasis (Kim et al., 2018), colon cancer (Wu et al., 2018), and non-small cell lung cancer (Li et al., 2018). Abnormally expressed is in significant association with poor prognosis in childhood ALL (Pouyanrad et al., 2019). Furthermore, miR-125b in combination with miR-99a and/or miR-100 could inhibit the expression of MALAT1 in vincristine-resistant children ALL cells (Moqadam et al., 2013). PWRN1 was significantly underexpressed in gastric cancer tissues and cells (Chen et al., 2018). Overexpressed PWRN1 could inhibit the proliferation and metastasis of gastric cancer cells and tumor growth. Furthermore, PWRN1 may regulate miR-425-5p expression by acting as its sponge in gastric cancer cells. ITPK1-AS1 expression could predict gastric cancer patients' survival (Hu et al., 2019). WT1-AS has been characterized as a tumor-suppressive lncRNA in several cancers including cervical squamous cell carcinoma (Zhang et al., 2019), gastric cancer (Du et al., 2016), papillary thyroid carcinoma (Le et al., 2020), non-small cell lung cancer cell (Jiang et al., 2020), and hepatocellular carcinoma (Lv et al., 2015). Besides, WT1-AS can regulate WT1 on oxidative stress injury and apoptosis of neurons in Alzheimer's disease *via* inhibition of the miR-375/SIX4 axis (Wang Q. et al., 2020). However, other lncRNAs have not been reported yet. According to our results, these lncRNAs deserve more research on ALL.

However, there are several limitations in this study. First, since there was no normal control of ALL in TCGA database, data of 407 whole blood in the GTEx database were obtained as control. Given that ALL primarily affects younger individuals, the age distribution of the control group is not ideally matched with the ALL from TCGA database, which may cause confounding. Second, the sample size of this study is small, and larger clinical samples should be used to verify these lncRNAs. In addition, this study lacks functional experiments. In future research, we will further the function and clinical value of these lncRNAs in ALL.

CONCLUSION

In our study, a ceRNA network was constructed for ALL. Among all DElncRNAs in the ceRNA network, eight lncRNAs were validated in ALL bone marrow samples using Real-time qPCR, which might provide a novel insight into the further study of ALL.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Research Ethics Committee of the First Affiliated Hospital of Zhengzhou University (2019-KY-194). The study was conducted in accordance with the Declaration of Helsinki. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WW designed the project, proposed the research concept, and wrote the manuscript. WW, CL, and FWa constructed the bioinformatic analysis, constructed the graphic images and data charts, and performed the statistical processing. CW, FWu, XL, and SG performed the experiments. All authors read and approved the manuscript and agreed to be accountable for all aspects of the research in ensuring that the accuracy and integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (Grant No. 81700138 to WW) and the Medical Science and Technique Foundation of Health Commission of Henan Provincial (Grant No. SBGJ202003041 to WW).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.656042/full#supplementary-material

Supplementary Material 1 | Differentially expressed mRNAs.

Supplementary Material 2 | Differentially expressed long non-coding RNAs.

Supplementary Material 3 | 845 lncRNA–miRNA pairs in the ceRNA network.

Supplementary Material 4 | 395 miRNA–mRNA pairs in the ceRNA network.

Supplementary Material 5 | The specific *p*-values of correlation analysis between hub genes and DElncRNAs.

REFERENCES

Agarwal, V., Bell, G. W., Nam, J., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005. doi: 10.7554/eLife.05005

Arthur, G., Almamun, M., and Taylor, K. (2017). Hypermethylation of antisense long noncoding RNAs in acute lymphoblastic leukemia. *Epigenomics* 9, 635–645. doi: 10.2217/epi-2016-0156

Chen, Z., Ju, H., Yu, S., Zhao, T., Jing, X., Li, P., et al. (2018). Prader-Willi region non-protein coding RNA 1 suppressed gastric cancer growth as a competing endogenous RNA of miR-425-5p. *Clin. Sci.* 132, 1003–1019. doi: 10.1042/CS20171588

Demir, S., Boldrin, E., Sun, Q., Hampp, S., Tausch, E., Eckert, C., et al. (2020). Therapeutic targeting of mutant p53 in pediatric acute lymphoblastic leukemia. *Haematologica* 105, 170–181. doi: 10.3324/haematol.2018.199364

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4:R609. doi: 10.1186/gb-2003-4-9-r60

Doubravska, L., Simova, S., Cermak, L., Valenta, T., Korinek, V., and Andera, L. (2008). Wnt-expressing rat embryonic fibroblasts suppress Apo2L/TRAIL-induced apoptosis of human leukemia cells. *Apoptosis* 13, 573–587. doi: 10.1007/s10495-008-0191-z

Du, T., Zhang, B., Zhang, S., Jiang, X., Zheng, P., Li, J., et al. (2016). Decreased expression of long non-coding RNA WT1-AS promotes cell proliferation and invasion in gastric cancer. *Biochim. Biophys. Acta* 1862, 12–19. doi: 10.1016/j.bbadis.2015.10.001

El-Khazragy, N., Noshi, M. A., Abdel-Malak, C., Zahran, R. F., and Swellam, M. (2019). miRNA-155 and miRNA-181a as prognostic biomarkers for pediatric acute lymphoblastic leukemia. *J. Cell. Biochem.* 120, 6315–6321. doi: 10.1002/jcb.27918

Fernando, T. R., Contreras, J. R., Zampini, M., Rodriguez-Malave, N. I., Alberti, M. O., Anguiano, J., et al. (2017). The lncRNA CASC15 regulates SOX4 expression in RUNX1-rearranged acute leukemia. *Mol. Cancer* 16:126. doi: 10.1186/s12943-017-0692-x

Fu, J., Si, L., Zhuang, Y., Zhang, A., Sun, N., Li, D., et al. (2019). Wnt/betacatenin inhibition reverses multidrug resistance in pediatric acute lymphoblastic leukemia. *Oncol. Rep.* 41, 1387–1394. doi: 10.3892/or.2018.6902

Gocho, Y., and Yang, J. J. (2019). Genetic defects in hematopoietic transcription factors and predisposition to acute lymphoblastic leukemia. *Blood* 134, 793–797. doi: 10.1182/blood.2018852400

Guo, Y., Fang, Q., Ma, D., Yu, K., Cheng, B., Tang, S., et al. (2018). Up-regulation of HO-1 promotes resistance of B-cell acute lymphocytic leukemia cells to HDAC4/5 inhibitor LMK-235 via the Smad7 pathway. *Life Sci.* 207, 386–394. doi: 10.1016/j.lfs.2018.06.004

Hass, C., Belz, K., Schoeneberger, H., and Fulda, S. (2016). Sensitization of acute lymphoblastic leukemia cells for LCL161-induced cell death by targeting redox homeostasis. *Biochem. Pharmacol.* 105, 14–22. doi: 10.1016/j.bcp.2016.01.004

Hu, Z., Yang, D., Tang, Y., Zhang, X., Wei, Z., Fu, H., et al. (2019). Five-long non-coding RNA risk score system for the effective prediction of gastric cancer patient survival. *Oncol. Lett.* 17, 4474–4486. doi: 10.3892/ol.2019.10124

Huang, H., Lin, Y., Li, J., Huang, K., Shrestha, S., Hong, H., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* 48, D148–D154. doi: 10.1093/nar/gkz896

Jabbour, E., Pui, C., and Kantarjian, H. (2018). Progress and innovations in the management of adult acute lymphoblastic leukemia. *Jama Oncol.* 4, 1413–1420. doi: 10.1001/jamaoncol.2018.1915

Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 28, 2062–2063. doi: 10.1093/bioinformatics/bts344

Jiang, X., Wang, J., and Fang, L. (2020). LncRNA WT1-AS over-expression inhibits non-small cell lung cancer cell stemness by down-regulating TGF-beta1. *BMC Pulm. Med.* 20:113. doi: 10.1186/s12890-020-1146-6

Jing, D., Huang, Y., Liu, X., Sia, K. C. S., Zhang, J. C., Tai, X., et al. (2018). Lymphocyte-specific chromatin accessibility pre-determines glucocorticoid resistance in acute lymphoblastic leukemia. *Cancer Cell* 34:906. doi: 10.1016/j.ccell.2018.11.002

Khan, N. I., Bradstock, K. F., and Bendall, L. J. (2007). Activation of Wnt/beta-catenin pathway mediates growth and survival in B-cell progenitor acute lymphoblastic leukaemia. *Br. J. Haematol.* 138, 338–348. doi: 10.1111/j.1365-2141.2007.06667.x

Kim, J., Piao, H., Kim, B., Yao, F., Han, Z., Wang, Y., et al. (2018). Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nat. Genet.* 50:1705. doi: 10.1038/s41588-018-0252-3

Le, F., Luo, P., Ouyang, Q., and Zhong, X. (2020). LncRNA WT1-AS downregulates survivin by upregulating miR-203 in papillary thyroid carcinoma. *Cancer Manag. Res.* 12, 443–449. doi: 10.2147/CMAR.S232294

Li, S., Mei, Z., Hu, H., and Zhang, X. (2018). The lncRNA MALAT1 contributes to non-small cell lung cancer development via modulating miR-124/STAT3 axis. *J. Cell. Physiol.* 233, 6679–6688. doi: 10.1002/jcp.26325

Lv, L., Chen, G., Zhou, J., Li, J., and Gong, J. (2015). WT1-AS promotes cell apoptosis in hepatocellular carcinoma through down-regulating of WT1. *J. Exp. Clin. Cancer Res.* 34:119. doi: 10.1186/s13046-015-0233-7

Lyu, X., Li, J., Yun, X., Huang, R., Deng, X., Wang, Y., et al. (2017). miR-181a-5p, an inducer of Wnt-signaling, facilitates cell proliferation in acute lymphoblastic leukemia. *Oncol. Rep.* 37, 1469–1476. doi: 10.3892/or.2017.5425

Martin, V., Agirre, X., Jimenez-Velasco, A., Jose-Eneriz, E. S., Cordeu, L., Garate, L., et al. (2008). Methylation status of Wnt signaling pathway genes affects the clinical outcome of Philadelphia-positive acute lymphoblastic leukemia. *Cancer Sci.* 99, 1865–1868. doi: 10.1111/j.1349-7006.2008.00884.x

Mochmann, L. H., Bock, J., Ortiz-Tanchez, J., Schlee, C., Bohne, A., Neumann, K., et al. (2011). Genome-wide screen reveals WNT11, a non-canonical WNT gene, as a direct target of ETS transcription factor ERG. *Oncogene* 30, 2044–2056. doi: 10.1038/onc.2010.582

Montano, A., Forero-Castro, M., Marchena-Mendoza, D., Benito, R., and Maria Hernandez-Rivas , J. (2018). New challenges in targeting signaling pathways in acute lymphoblastic leukemia by NGS approaches: an update. *Cancers* 10:1104. doi: 10.3390/cancers10040110

Montpetit, A., Larose, J., Boily, G., Langlois, S., Trudel, N., and Sinnett, D. (2004). Mutational and expression analysis of the chromosome 12p candidate tumor suppressor genes in pre-B acute lymphoblastic leukemia. *Leukemia* 18, 1499–1504. doi: 10.1038/sj.leu.2403441

Moqadam, F. A., Lange-Turenhout, E. A. M., Aries, I. M., Pieters, R., and den Boer, M. L. (2013). MiR-125b, miR-100 and miR-99a co-regulate vincristine resistance in childhood acute lymphoblastic leukemia. *Leukemia Res.* 37, 1315–1321. doi: 10.1016/j.leukres.2013.06.027

Nabhan, M., Louka, M. L., Khairy, E., Tash, F., Ali-Labib, R., and El-Habashy, S. (2017). MicroRNA-181a and its target Smad 7 as potential biomarkers for tracking child acute lymphoblastic leukemia. *Gene* 628, 253–258. doi: 10.1016/j.gene.2017.07.052

Nucera, S., Giustacchini, A., Boccalatte, F., Calabria, A., Fanciullo, C., Plati, T., et al. (2016). miRNA-126 orchestrates an oncogenic program in b cell precursor acute lymphoblastic leukemia. *Cancer Cell* 29, 905–921. doi: 10.1016/j.ccell.2016.05.007

Pouyanrad, S., Rahgozar, S., and Ghodousi, E. S. (2019). Dysregulation of miR-335-3p, targeted by NEAT1 and MALAT1 long non-coding RNAs, is associated with poor prognosis in childhood acute lymphoblastic leukemia. *Gene* 692, 35–43. doi: 10.1016/j.gene.2019.01.003

Richard-Carpentier, G., Kantarjian, H., and Jabbour, E. (2019). Recent advances in adult acute lymphoblastic leukemia. *Curr. Hematol. Malig. Rep.* 14, 106–118. doi: 10.1007/s11899-019-00503-1

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

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

Sun, H., Zhang, Z., Luo, W., Liu, J., Lou, Y., and Xia, S. (2019). NET1 enhances proliferation and chemoresistance in acute lymphoblastic leukemia cells. *Oncol. Res.* 27, 935–944. doi: 10.3727/096504019X15555388198071

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

Trimarchi, T., Bilal, E., Ntziachristos, P., Fabbri, G., Dalla-Favera, R., Tsirigos, A., et al. (2014). Genome-wide mapping and characterization of notch-regulated long noncoding RNAs in acute leukemia. *Cell* 158, 593–606. doi: 10.1016/j.cell.2014.05.049

Trino, S., De Luca, L., Laurenzana, I., Caivano, A., Del Vecchio, L., Martinelli, G., et al. (2016). P53-MDM2 pathway: evidences for a new targeted therapeutic approach in b-acute lymphoblastic leukemia. *Front. Pharmacol.* 7:491. doi: 10.3389/fphar.2016.00491

Wang, Q., Ge, X., Zhang, J., and Chen, L. (2020). Effect of lncRNA WT1-AS regulating WT1 on oxidative stress injury and apoptosis of neurons in Alzheimer's disease via inhibition of the miR-375/SIX4 axis. *Aging (Albany NY)* 12, 23974–23995. doi: 10.18632/aging.104079

Wang, W., Wu, F., Ma, P., Gan, S., Li, X., Chen, L., et al. (2020). LncRNA CRNDE promotes the progression of B-cell precursor acute lymphoblastic leukemia by targeting the miR-345-5p/CREB axis. *Mol. Cells* 43, 718–727. doi: 10.14348/molcells.2020.0065

Wang, X. (2008). miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 14, 1012–1017. doi: 10.1261/rna.965408

Wolfraim, L. A., Fernandez, T. M., Mamura, M., Fuller, W. L., Kumar, R., Cole, D. E., et al. (2004). Loss of Smad3 in acute T-cell lymphoblastic leukemia. *N. Engl. J. Med.* 351, 552–559. doi: 10.1056/NEJMoa031197

Wu, Q., Meng, W., Jie, Y., and Zhao, H. (2018). LncRNA MALAT1 induces colon cancer development by regulating miR-129-5p/HMGB1 axis. *J. Cell. Physiol.* 233, 6750–6757. doi: 10.1002/jcp.26383

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

Zhang, Y., Huang, S., Zhou, H., Chen, C., Tian, L., and Shen, J. (2016). Role of HOTAIR in the diagnosis and prognosis of acute leukemia. *Oncol. Rep.* 36, 3113–3122. doi: 10.3892/or.2016.5147

Zhang, Y., Na, R., and Wang, X. (2019). LncRNA WT1-AS up-regulates p53 to inhibit the proliferation of cervical squamous carcinoma cells. *BMC Cancer* 19:1052. doi: 10.1186/s12885-019-6264-2

Zhao, Q., Zhao, S., Li, J., Zhang, H., Qian, C., Wang, H., et al. (2019). TCF7L2 activated HOXA-AS2 decreased the glucocorticoid sensitivity in acute lymphoblastic leukemia through regulating HOXA3/EGFR/Ras/Raf/MEK/ERK pathway. *Biomed. Pharmacother.* 109, 1640–1649. doi: 10.1016/j.biopha.2018.10.046

Zhou, R., Mo, W., Wang, S., Zhou, W., Chen, X., and Pan, S. (2019). miR-141-3p and TRAF5 network contributes to the progression of t-cell acute lymphoblastic leukemia. *Cell Transplan.* 281(1Suppl), 59S–65S. doi: 10.1177/0963689719887370

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fgene.2021.712150

Construction of Circulating MicroRNAs-Based Non-invasive Prediction Models of Recurrent Implantation Failure by Network Analysis

*Peigen Chen, Tingting Li, Yingchun Guo, Lei Jia, Yanfang Wang and Cong Fang**

Reproductive Medicine Center, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Background: Recurrent implantation failure (RIF) is an obstacle in the process of assisted reproductive technology (ART). At present, there is limited research on its pathogenesis, diagnosis, and treatment methods.

Methods and Results: In this study, a series of analytical tools were used to analyze differences in miRNAs, mRNAs, and lncRNAs in the endometrium of patients in a RIF group and a control group. Then the competing endogenous RNA (ceRNA) network was built to describe the relationship between gene regulation in the endometrium of the RIF group. Based on the results of the logistic regression of co-expression miRNAs between serum and endometrial samples, we built a predictive model based on circulating miRNAs.

Conclusion: The stability and non-invasiveness of the circular miRNA prediction model provided a new method for diagnosis in RIF patients.

Keywords: recurrent implantation failure, competing endogenous RNAs, assisted reproductive technology, GEO, non-invasive prediction model

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Gabriel Tao,
University of Houston, United States
Alper Yilmaz,
Yıldız Technical University, Turkey

***Correspondence:**
Cong Fang
fangcong@mail.sysu.edu.cn

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

Received: *20 May 2021*
Accepted: *18 June 2021*
Published: *23 July 2021*

Citation:
Chen P, Li T, Guo Y, Jia L, Wang Y and Fang C (2021) Construction of Circulating MicroRNAs-Based Non-invasive Prediction Models of Recurrent Implantation Failure by Network Analysis. Front. Genet. 12:712150. doi: 10.3389/fgene.2021.712150

INTRODUCTION

Recurrent implantation failure (RIF) is a thorny issue that couples undergoing *in vitro* fertilization (IVF)/intracytoplasmic sperm injection (ICSI) may face. The generally accepted definition is that women under the age of 40 years have transferred at least four high-quality embryos in at least three fresh or frozen cycles or have transferred a total of 10 high-quality embryos but have not yet achieved clinical pregnancy (Thornhill et al., 2005; Simon and Laufer, 2012; Coughlan et al., 2014). Along with improving *in vitro* fertilization embryo transfer (IVF-ET) technology and increasing clinical pregnancy rate, RIF is still a tough problem in the process of IVF-ET. The normal embryo implantation generally only occurs during the window of implantation (WOI) (Cha et al., 2012), which refers to days 20–24 of the normal menstrual cycle. Abnormalities of the endometrium at this stage are important factors that lead to RIF.

MicroRNAs (miRNA) are a class of non-coding RNA molecules with a length of about 22 nucleotides that are widely found in eukaryotic cells. There have been some studies confirming the role of miRNA in endometrial regulation (Creighton et al., 2010; Kuokkanen et al., 2010;

Revel et al., 2011; Altmae et al., 2013). For example, miR-30b, miR-30d, and miR-494 had been reported to play an important role in the regulation of endometrial function (Altmae et al., 2013). Recent research reported miRNAs associated with RIF, such as miR-34c-5p (Tan et al., 2020) and miR-148A-3P (Zhang et al., 2020).

Moreover, in recent years circulating miRNA has been increasingly used as a non-invasive tool for disease diagnosis and prediction due to its high stability, sensitivity, and specificity (Martinez and Peplow, 2020). In the present study, we aim to use a larger sample size of data in our analysis to explore the regulatory molecular mechanism in the endometrium of RIF patients at the WOI stage. At the same time, we aim to look for peripheral blood miRNAs closely related to RIF and provide a new way for non-invasive early diagnosis of RIF, thereby improving the clinical outcome of patients.

MATERIALS AND METHODS

We used R software (version 3.6.3) (Team, 2018), GraphPad Prism (version 8), and Bioconductor (Gentleman et al., 2004) for all statistical analyses in our study.

Data Acquisition and Preprocessing

Paired serum and endometrial miRNA expression profile data (GSE108966) were obtained from the Gene Expression Omnibus (GEO) database[1]. The paired raw count of endometrial expression profile and corresponding clinical data of a RIF group and a control group were extracted from GSE71331 and GSE71332 and then processed by "Limma" R package (Ritchie et al., 2015) (Agilent-052909 CBC lncRNA mRNA V3, Agilent-046064 Unrestricted Human miRNA V19.0).

Selection of Differentially Expressed Genes

The scanning of differentially expressed (DE) miRNA in the endometrium and the serum was performed by using the "limma" R package (Ritchie et al., 2015) with the following criteria: p-value < 0.05 and $|\log 2\text{-fold change}| > 1$.

Similar to the above process, the differentially expressed genes (DEGs) of GSE71331 and GSE71332 were selected.

Selection and Validation of Co-expression miRNAs Between Serum and Endometrial Samples

The intersection of endometrium DE miRNAs and serum DE miRNAs was taken as intersection miRNAs. To ensure that their expressions were relevant, Pearson correlation analysis is performed by using GraphPad Prism (version 8). Genes with the Pearson correlation coefficient $|r| \geq 0.5$ were considered to be co-expressed miRNAs between the serum and the endometrium.

Weighted Correlation Network Analysis of miRNA of Endometrial Samples

With the "WGCNA" R package (Langfelder and Horvath, 2008), weighted correlation network analysis (WGCNA) was performed on DE miRNAs which were selected based on GSE71332 dataset. The minimum gene dendrogram size of average linkage hierarchical clustering was set as 40. Then the dissimilarity and constructed module dendrograms of these modules were calculated.

To estimate the significance of each module and also measure the relationships between genes and sample traits, the gene significance (GS) of each module was then calculated. The GS and module membership (MM, the correlation between the genes in the module and their expression profiles) of every key gene were calculated with the following thresholds: correlation gene GS > 0.5 and correlation gene MM > 0.8.

Prediction of Target lncRNAs/mRNAs of RIF-Related DE miRNAs

The intersection of the DEGs and the genes of the key modules related to RIF in WGCNA was taken as RIF-related DE miRNAs. Then miRDB[2] (Chen and Wang, 2020), miRTarBase[3] (Hsu et al., 2011), and TargetScan[4] (Agarwal et al., 2015)

Abbreviations: RIF, recurrent implantation failure; ART, assisted reproductive technology; ceRNA, competing endogenous RNA; WOI, window of implantation; WGCNA, weighted correlation network analysis.

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

[2]http://mirdb.org/miRDB/

[3]http://mirtarbase.mbc.nctu.edu.tw/

[4]http://targetscan.org/

TABLE 1 | Clinical characteristics of GSE71331.

	RIF			CON			*p*-value
	Mean	*SD*	n	Mean	*SD*	n	
Age (years)	31.5714	4.8599	7	31	3.5355	5	0.7349
Number of failed cycles	5.1429	3.1320	7	0	0	5	0.0032
Number of transferred embryos	12.4286	6.1062	7	2	0	5	<0.000001
Number of high-quality transferred embryos	7.5714	1.5119	7	1.6	0.5477	5	0.0007
Endometrial thickness on the day of LH surge	0.9429	0.1512	7	1.08	0.1095	5	0.9352
The day of sample (post the day of LH surge)	6.5714	0.7868	7	7.2	0.4472	5	0.7096

RIF, recurrent implantation failure; SD, standard deviation; LH, luteinizing hormone; CON, comment group

were used to predict miRNA-targeting mRNAs. NPInter[5] (Teng et al., 2020) and DIANA-LncBase (Paraskevopoulou et al., 2016) were used to predict miRNA-targeting lncRNAs. The intersection of differential mRNAs/lncRNAs in GSE71331 and the miRNA-targeting mRNAs/lncRNAs were taken as targeting-DE mRNAs/lncRNAs.

Construction of lncRNA-miRNA-mRNA Regulatory Network

LncRNA-miRNA interactomes were then built based on targeting-DE lncRNAs and RIF-related DE miRNAs. Similarly, mRNA-miRNA interactomes were built. Subsequently, lncRNA–miRNA–mRNA regulatory networks were constructed by using cytoscape, version 3.8[6]. Key modules were selected by MCODE using default parameters (Bader and Hogue, 2003).

Functional Enrichment Analysis of Targeted DE mRNAs

Metascape (Zhou et al., 2019)[7] contained many updated functional annotations, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, canonical pathway, Reactome

[5]http://bigdata.ibp.ac.cn/npinter4/

[6]https://cytoscape.org/

[7]http://metascape.org

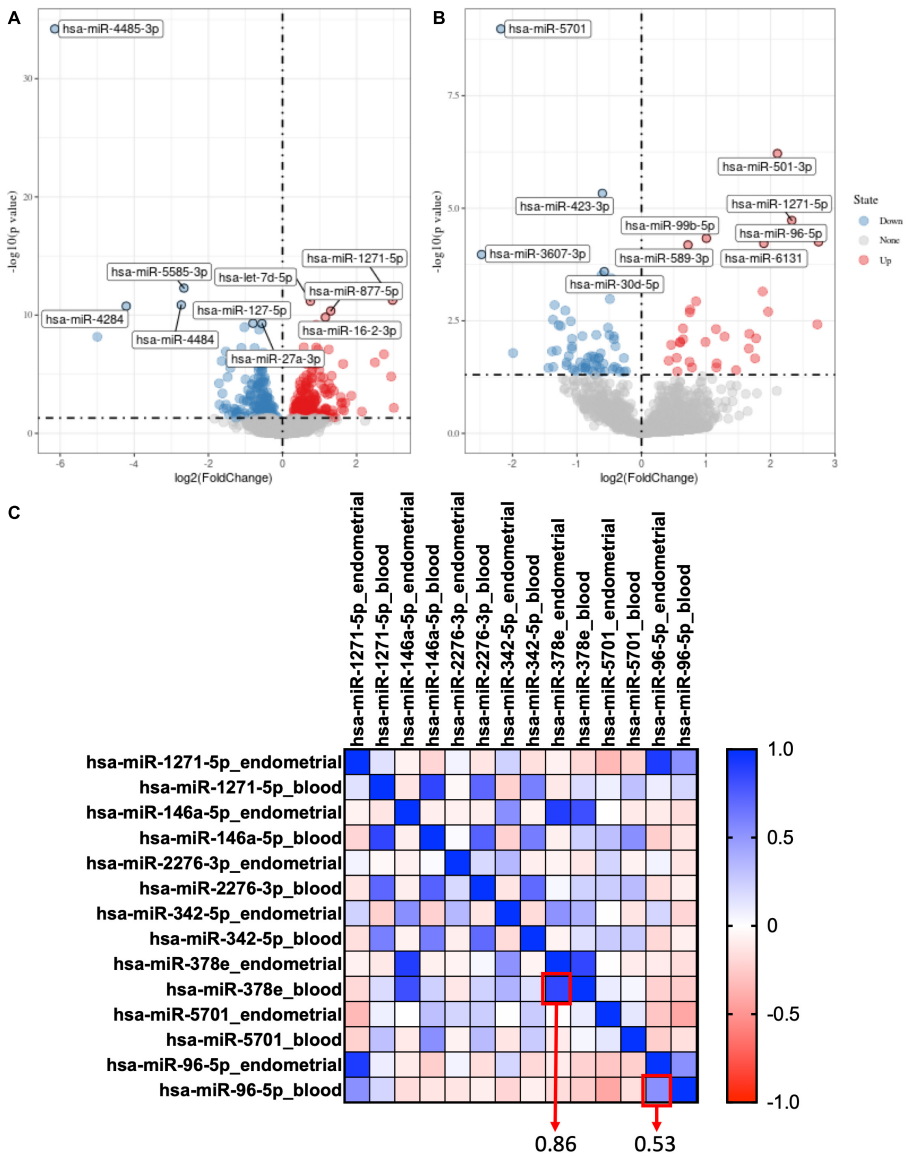


FIGURE 1 | Selection of co-expression miRNAs between serum and endometrial samples. **(A)** The volcano plot of differentially expressed genes in the endometrial sample of GSE108966. **(B)** The volcano plot of differentially expressed genes in the serum sample of GSE108966. **(C)** The heatmap of the co-expression miRNAs between serum and endometrial samples. The numbers inside the boxes stand for correlation coefficient.

pathway, Gene Ontology (GO) biological process, and CORUM (the comprehensive resource of mammalian protein complexes). To understand the biological function of targeted DE mRNAs of GSE71332, Metascape was then used with a p-value of < 0.01 as the cutoff value. Then the terms with a p-value of < 0.01 and a number of genes greater than or equal to 3 were selected as significant terms.

Transcriptional Regulatory Relationship Analysis of Targeted DE mRNAs

TRRUST (transcriptional regulatory relationships unraveled by sentence-based text-mining)[8] is a TF-target regulatory interactions database based on the manual curation of Medline abstracts (Han et al., 2015). We then used TRRUST to screen transcription factors related to targeted DE mRNAs and targeted mRNAs and study their transcription regulation relationships.

Causal Relationship Analysis

DisNor (Lo Surdo et al., 2018)[9] is a web-based tool that can generate and explore protein interaction networks based on disease genes using Mentha protein interaction data and causal interaction information annotated by SIGNOR. DisNor was used to explore the causal relationships among targeted DE mRNAs.

Construction and Validation of Nomogram Based on Circulating miRNAs

Logistic regression analysis was then performed with three selected factors by using "survival" R package (Therneau, 2015) to select the best fit model. Then a nomogram was built to predict the risk of RIF patients by using "rms" R package. At the same time, the consistency index (C-index) was calculated to evaluate the model's ability to distinguish. The consistency of the predicted probability and the actual probability of the model was evaluated by the calibration curves. The predictive performance of the model was evaluated by drawing the receiver operating characteristic (ROC) curve and calculating the area under the curve (AUC) values.

RESULTS

Clinical Characteristics of Samples Used in the Study

All data came from samples taken during the window of implantation. The clinical characteristics of the RIF group and

[8]http://www.grnpedia.org/trrust

[9]https://disnor.uniroma2.it/

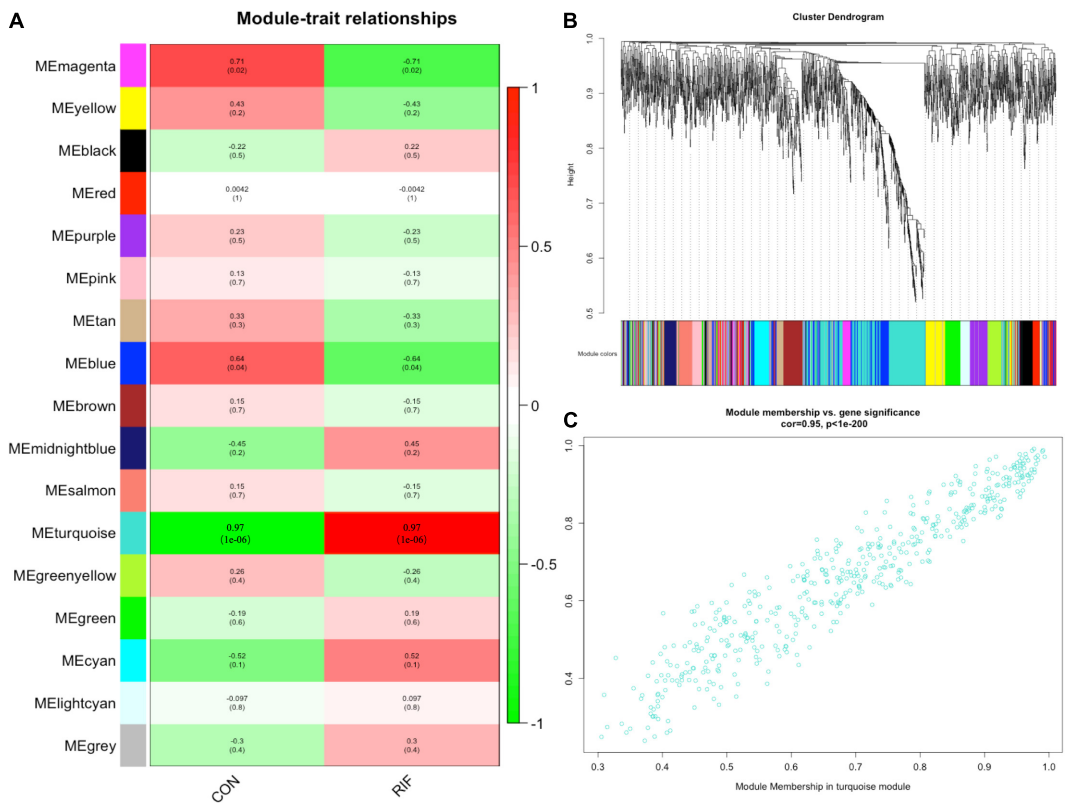


FIGURE 2 | Weighted gene correlation network analysis. **(A)** The relationships between the corresponding modules and clinical phenotypes. **(B)** The modules selected with a minimum module size of 50 for further analysis. **(C)** The module membership in the turquoise module.

the control group in GSE71331 and GSE71332 are listed in **Table 1**. In total, the mRNA and lncRNA expression profiles of seven RIF samples and five control samples were extracted from GSE71331, and the corresponding miRNA expression profiles were extracted from GSE71332.

Selection of Co-expression miRNAs Between Serum and Endometrial Samples and Functional Enrichment Analysis

For GSE108966, 63 downregulation miRNAs and 45 upregulation miRNAs were selected from endometrial samples by Deseq2 with the following criteria: p-value < 0.05 and $|$ log 2-fold change$| > 1$ (**Figure 1A**). Similarly, 28 downregulation miRNAs and 22 upregulation miRNAs were selected from serum samples (**Figure 1B** and **Supplementary Table 1**).

Hsa-miR-378e and hsa-miR-96-5p were selected as co-expression miRNAs between serum and endometrial samples (**Figure 1C**).

Selection of Differentially Expressed Genes

By using "limma" package with p-value < 0.05 and $|$ log 2-fold change$| > 1$, we found that Hsa-miR-378e and hsa-miR-96-5p are also highly expressed in RIF in the profiles of GSE71331 and GSE71332 (**Supplementary Table 2**).

Selection of RIF-Related miRNAs by WGCNA

A gene co-expression network was then constructed based on the samples of GSE71332 by WGCNA to select the most significant gene modules and genes. This procedure can also help to

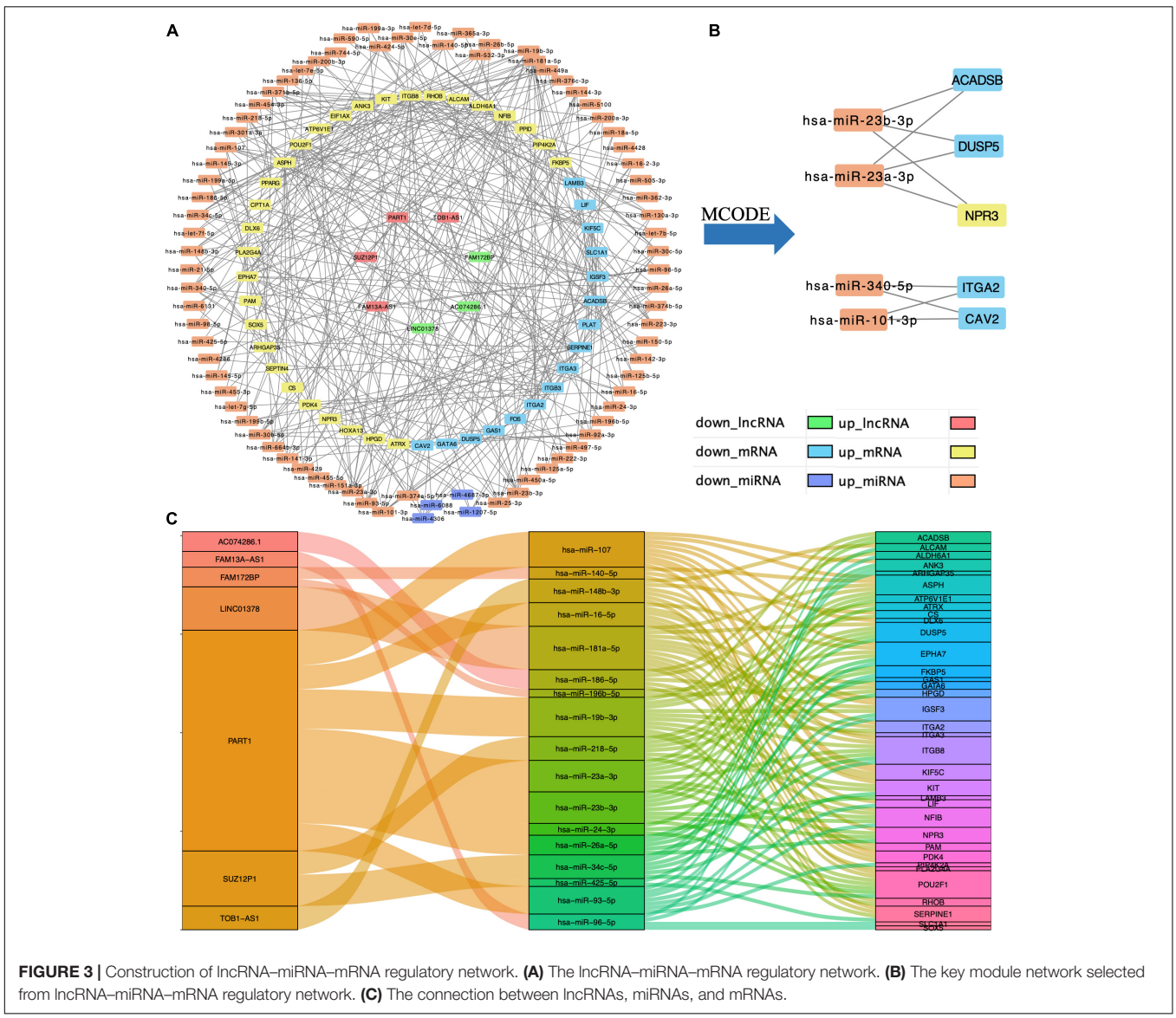


FIGURE 3 | Construction of lncRNA–miRNA–mRNA regulatory network. **(A)** The lncRNA–miRNA–mRNA regulatory network. **(B)** The key module network selected from lncRNA–miRNA–mRNA regulatory network. **(C)** The connection between lncRNAs, miRNAs, and mRNAs.

elucidate the relationship between genes and clinical features. With a soft threshold of β = 7, 16 modules were selected with a minimum module size of 50 for further analysis (**Figure 2B**).

The overall expression gene level was taken as the MS (module significance) to estimate the relationship between the corresponding modules and clinical phenotypes (**Figure 2A**). Based on the results, we found that the turquoise module showed the most significant positive correlation with the RIF (cor = 0.97, $p < 0.0001$) (**Figure 2C**). Therefore, the turquoise module was chosen as the RIF-related module.

Finally, 97 intersection miRNAs between DEG and WGCNA were selected as RIF-related DE miRNAs (**Supplementary Table 3**).

Construction of lncRNA–miRNA–mRNA Regulatory Network

Based on the interaction of the prediction of three databases (miRDB, miTarBase, and TargetScan) and DE mRNAs, 45 mRNAs were selected for network construction. Similarly,

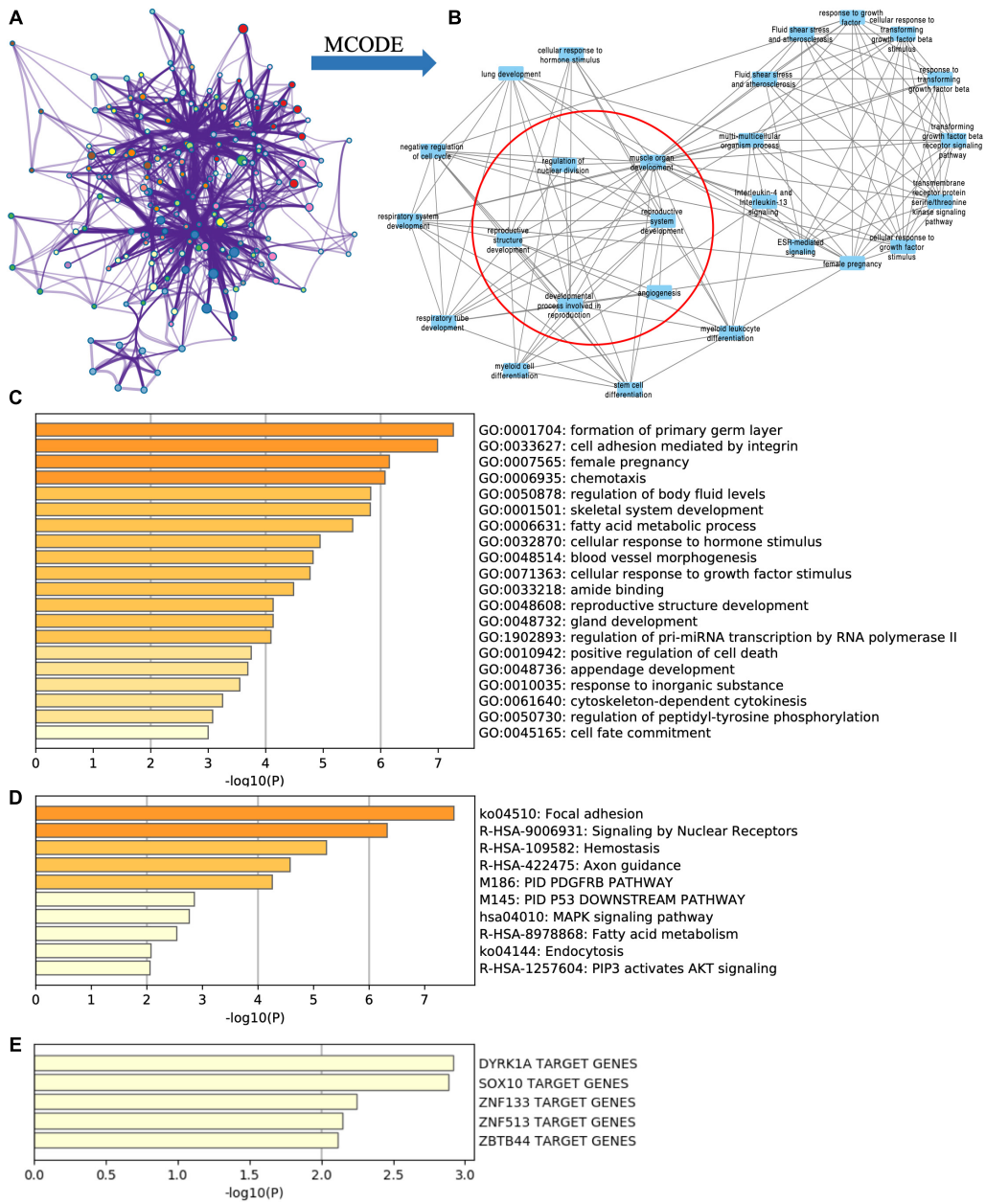


FIGURE 4 | Functional enrichment analysis of targeted DE mRNAs of GSE71331/71332. **(A)** The network of functional terms. **(B)** The key module network selected by MCODE (inside the red circle) and first neighbor node of the key module network (outside the red circle). **(C)** The enriched GO terms of targeted DE mRNAs. **(D)** The enriched pathway of targeted DE mRNAs. **(E)** The transcription factors related to targeted DE mRNAs.

seven lncRNAs were selected. Finally, a lncRNA–miRNA–mRNA regulatory network was constructed based on 80 miRNAs, 45 mRNAs, and 7 lncRNAs by using cytoscape (**Figures 3A,C**). By using the MCODE app of cytoscape, a key module network was selected (**Figure 3B**).

Functional Enrichment Analysis and Causal Relationship Analysis

By using Metascape, we found that the targeted DE mRNAs were enriched mainly in terms of cell adhesion mediated by integrin (GO: 0033627), female pregnancy (GO: 0007565), fatty acid metabolic process (GO: 0006631), and reproductive structure development (GO: 0048608) (**Figure 4C**). The pathways-targeted DE mRNAs were enriched in terms of focal adhesion, signaling by nuclear receptors, MAPK pathway, and PIP3-activated AKT pathway (**Figure 4D**).

Key modules were selected in the functional network (**Figure 4A**) by using the MCODE cytoscape app. We found that one of the key modules was closely related with reproduction function (**Figure 4B** red circle), and their first neighbor nodes are also shown in **Figure 4B** (outside the red circle).

According to TRRUST database, DYRK1A targeted ACADSB, ANK3, ASPH, HOXA13, ITGA3, and NPR3. SOX10 targeted ACADSB, ANK3, LAMB3, and SOX5 (**Figure 4E**).

By using the DisNor tool, we screened the first neighboring genes of the targeted DE mRNAs and built a causal relationship network (**Figure 5**). We found that these genes were closely related to the MAPK signal pathway as we could see in the results of functional enrichment analysis.

Construction and Validation of Nomogram Based on Circulating miRNAs

Based on the results of logistic regression analysis, the best fit models included age, the expression of hsa-miR-96-5p, and the expression of hsa-miR-378e. A nomogram with C-index = 0.865 was established to act as a prediction tool of RIF (**Figure 6A**), which means that our model has a good ability to distinguish the clinical outcome. The calibration curves in **Figure 6B** show a good consistency of the predicted probability and the actual probability of the model with mean absolute error of 0.028, mean squared error of 0.00096, and 0.9 quantile of absolute error of 0.044. **Figure 6C** shows the ROC curve of the model, and the AUC value was 0.865, which means a promising predictive performance.

DISCUSSION

RIF is an obstacle in the process of assisted reproductive technology (ART). Epigenetic regulation of gene expression played an important role in the development of RIF, and one of the most important parts was miRNA. Several factors, including different miRNAs, were selected as key molecules of regulation of endometrial acceptance and implantation (Kang et al., 2015). However, there was still a lack of research on the molecules that affect endometrial tolerance before implantation and the mechanisms of early dialog between the embryo and the uterus (Kang et al., 2015). Due to the limitations of the sample and detection technology, researchers could only study the known

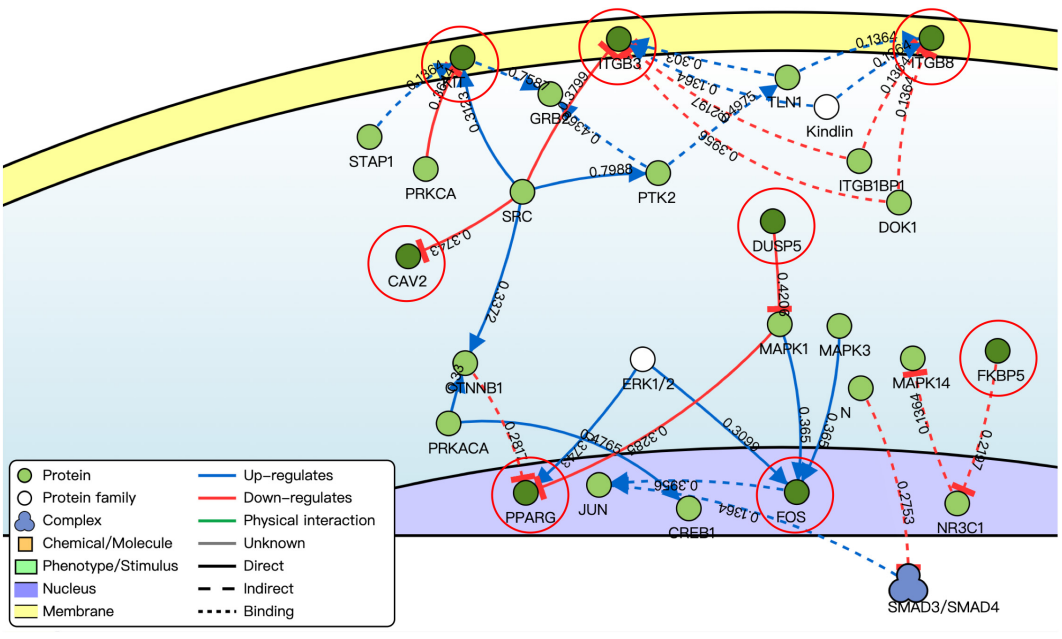


FIGURE 5 | Causal interaction analysis of targeted DE mRNAs of GSE71331/71332 (inside the red circle) and their first neighbor genes (outside the red circle) by using DisNor. The numbers on the line stand for correlation coefficient.

potentially meaningful miRNAs and could not explore them in depth.

In this study, by using WGCNA, we built a lncRNA–miRNA–mRNA regulatory network to analyze the expression and regulation characteristics of miRNAs in the endometrium and the serum. By using the MCODE app in cytoscape, two key modules were selected. We noticed that both hsa-miR-23a and hsa-miR-23b interacted with ACADSB, DUSP5, and NPR3. Fan et al. (2020) reported that the upregulator expression of hsa-miR-23a could suppress hdac2, activate NF-κB, and influence the ability of adhesion, invasion, and proliferation of trophoblasts. Our study showed that hsa-miR-23a played an important role in embryo implantation. At the same time, several studies suggested that hsa-miR-23a and hsa-miR-23b were closely related to the MAPK pathway (Guo et al., 2018; Ma et al., 2019). As many studies reported, the MAPK pathway played an important role in embryo implantation, and it was closely related to the ability of adhesion, invasion, and proliferation of trophoblasts and the procession of endometrium angiogenesis (Baryla et al., 2019; Zhang et al., 2019; Goryszewska et al., 2020). The causal relationship network in **Figure 5** shows that DUSP5 downregulates MAPK1 ($R = 0.42$). According to these results, we could make a hypothesis that lncRNA PART1 may act as a sponge of hsa-miR-23a/b to downregulate DUSP5 to promote RIF.

In this study, the results of functional enrichment analysis of miRNAs target genes also support our conclusions. The targeted mRNAs of hsa-miR-96-5p were mainly enriched in terms of cellular response to organonitrogen compound, negative regulation of cell differentiation, regulation of protein serine/threonine kinase activity, apoptosis pathway, and the MAPK signaling pathway.

Currently, almost all tests for endometrial function in RIF patients are based on endometrial biopsies. Such an inspection operation had a potential impact on the uterine cavity environment. In this study, we developed a non-invasive RIF diagnostic scoring model to assist in the diagnosis and treatment

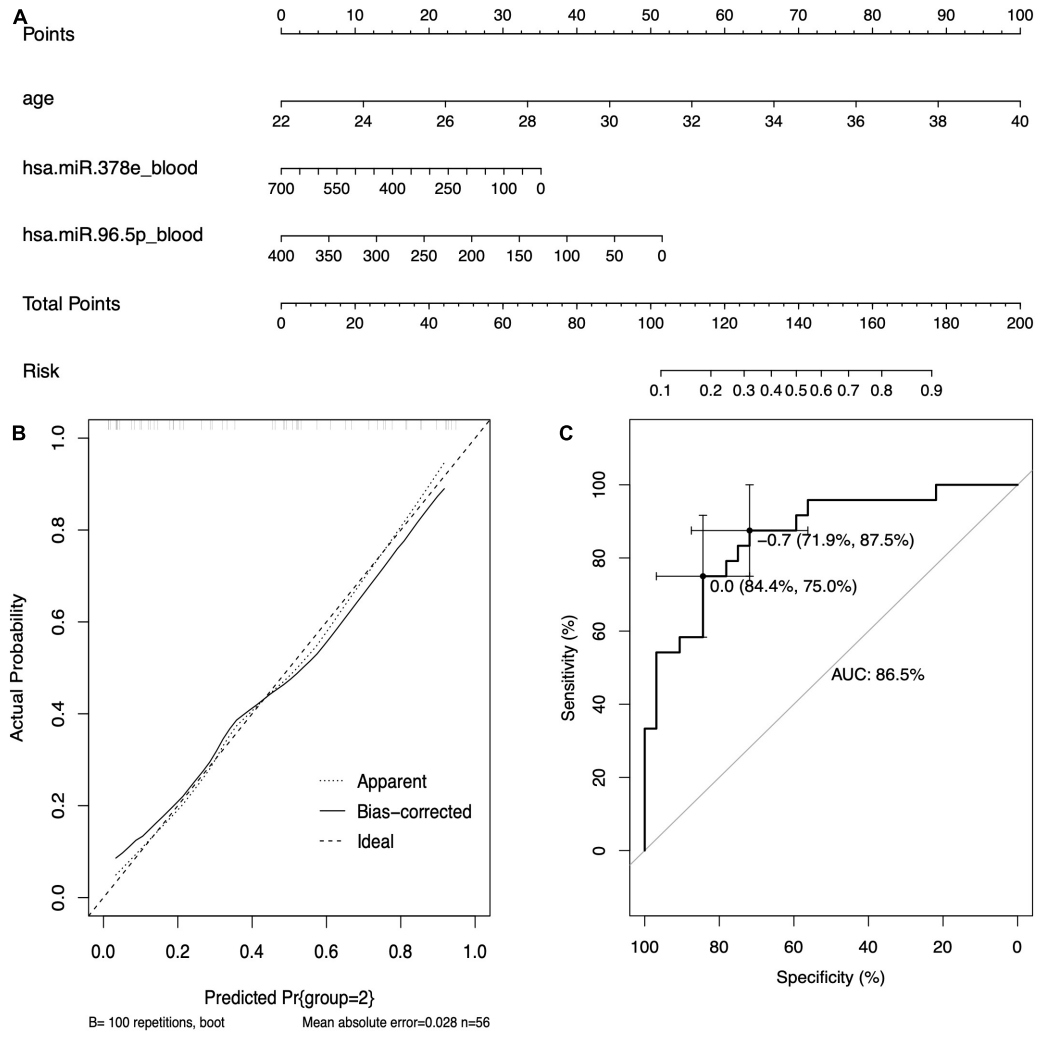


FIGURE 6 | The prediction model based on circulating miRNAs. **(A)** The nomogram of the prediction models based on circulating miRNAs. **(B)** The calibration curves of the prediction model. **(C)** The receiver operating characteristic (ROC) curve of the prediction model.

of RIF patients, and it showed better predictability and accuracy. As far as we know, this was the first RIF predictive scoring model based on circulating miRNA. Clinical trials of models will also be conducted soon. For this study, there were still some shortcomings, such as a lack of adequate laboratory tests to verify the mechanism. We are already starting relevant clinical studies.

CONCLUSION

In this study, we built a circulating miRNA-based prediction and provided a new non-invasive inspection method. We also found that these two miRNAs played an important role in the progress of RIF and found that lncRNA PART1 may act as a sponge of hsa-miR-23a/b to downregulate DUSP5 to promote RIF.

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

CF and PC carried out the study, coordinated the study, participated in the design, and reviewed the manuscript. PC analyzed and interpreted the data. YG and TL drafted the manuscript. LJ and YW collected and analyzed the data. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 81871214), the National Key R&D Program of China (grant no. 2017YFC1001603), and the National Natural Science Foundation of China (grant no. 81070495).

ACKNOWLEDGMENTS

This study has been presented as "pre-print" in "research square" (https://www.researchsquare.com/article/rs-145125/v1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.712150/full#supplementary-material

Supplementary Table 1 | The co-expression miRNAs between serum and endometrial samples.

Supplementary Table 2 | The differentially expressed genes (DEGs) list of GSE71331/71332.

Supplementary Table 3 | The intersection miRNAs between DEG and WGCNA in endometrium of RIF patients.

REFERENCES

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005. doi: 10.7554/eLife.05005

Altmae, S., Martinez-Conejero, J. A., Esteban, F. J., Ruiz-Alonso, M., Stavreus-Evers, A., Horcajadas, J. A., et al. (2013). MicroRNAs miR-30b, miR-30d, and miR-494 regulate human endometrial receptivity. *Reprod. Sci.* 20, 308–317. doi: 10.1177/1933719112453507

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Baryla, M., Kaczynski, P., Goryszewska, E., Riley, S. C., and Waclawik, A. (2019). Prostaglandin F2alpha stimulates adhesion, migration, invasion and proliferation of the human trophoblast cell line HTR-8/SVneo. *Placenta* 77, 19–29. doi: 10.1016/j.placenta.2019.01.020

Cha, J., Sun, X., and Dey, S. K. (2012). Mechanisms of implantation: strategies for successful pregnancy. *Nat. Med.* 18, 1754–1767. doi: 10.1038/nm.3012

Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48, D127–D131. doi: 10.1093/nar/gkz757

Coughlan, C., Ledger, W., Wang, Q., Liu, F., Demirol, A., Gurgan, T., et al. (2014). Recurrent implantation failure: definition and management. *Reprod. Biomed. Online* 28, 14–38. doi: 10.1016/j.rbmo.2013.08.011

Creighton, C. J., Benham, A. L., Zhu, H., Khan, M. F., Reid, J. G., Nagaraja, A. K., et al. (2010). Discovery of novel microRNAs in female reproductive tract using next generation sequencing. *PLoS One* 5:e9637. doi: 10.1371/journal.pone.0009637

Fan, Y., Dong, Z., Zhou, G., Fu, J., Zhan, L., Gao, M., et al. (2020). Elevated miR-23a impairs trophoblast migration and invasiveness through HDAC2 inhibition and NF-kappaB activation. *Life Sci.* 261:118358. doi: 10.1016/j.lfs.2020.118358

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Goryszewska, E., Kaczynski, P., Balboni, G., and Waclawik, A. (2020). Prokineticin 1-prokineticin receptor 1 signaling promotes angiogenesis in the porcine endometrium during pregnancydagger. *Biol. Reprod.* 103, 654–668. doi: 10.1093/biolre/ioaa066

Guo, Y., Min, Z., Jiang, C., Wang, W., Yan, J., Xu, P., et al. (2018). Downregulation of HS6ST2 by miR-23b-3p enhances matrix degradation through p38 MAPK pathway in osteoarthritis. *Cell Death Dis.* 9:699. doi: 10.1038/s41419-018-0729-0

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., et al. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* 5:11432. doi: 10.1038/srep11432

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., et al. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 39, D163–D169. doi: 10.1093/nar/gkq1107

Kang, Y. J., Lees, M., Matthews, L. C., Kimber, S. J., Forbes, K., and Aplin, J. D. (2015). MiR-145 suppresses embryo-epithelial juxtacrine communication at implantation by modulating maternal IGF1R. *J. Cell Sci.* 128, 804–814. doi: 10.1242/jcs.164004

Kuokkanen, S., Chen, B., Ojalvo, L., Benard, L., Santoro, N., and Pollard, J. W. (2010). Genomic profiling of microRNAs and messenger RNAs reveals

hormonal regulation in microRNA expression in human endometrium. *Biol. Reprod.* 82, 791–801. doi: 10.1095/biolreprod.109.081059

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

Lo Surdo, P., Calderone, A., Iannuccelli, M., Licata, L., Peluso, D., Castagnoli, L., et al. (2018). DISNOR: a disease network open resource. *Nucleic Acids Res.* 46, D527–D534. doi: 10.1093/nar/gkx876

Ma, M., Dai, J., Tang, H., Xu, T., Yu, S., Si, L., et al. (2019). MicroRNA-23a-3p Inhibits Mucosal Melanoma Growth and Progression through Targeting Adenylate Cyclase 1 and Attenuating cAMP and MAPK Pathways. *Theranostics* 9, 945–960. doi: 10.7150/thno.30516

Martinez, B., and Peplow, P. V. (2020). MicroRNAs in blood and cerebrospinal fluid as diagnostic biomarkers of multiple sclerosis and to monitor disease progression. *Neural Regen. Res.* 15, 606–619. doi: 10.4103/1673-5374.266905

Paraskevopoulou, M. D., Vlachos, I. S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., et al. (2016). DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.* 44, D231–D238. doi: 10.1093/nar/gkv1270

Revel, A., Achache, H., Stevens, J., Smith, Y., and Reich, R. (2011). MicroRNAs are associated with human embryo implantation defects. *Hum. Reprod.* 26, 2830–2840. doi: 10.1093/humrep/der255

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

Simon, A., and Laufer, N. (2012). Repeated implantation failure: clinical approach. *Fertil. Steril.* 97, 1039–1043. doi: 10.1016/j.fertnstert.2012.03.010

Tan, Q., Shi, S., Liang, J., Zhang, X., Cao, D., and Wang, Z. (2020). MicroRNAs in Small Extracellular Vesicles Indicate Successful Embryo Implantation during Early Pregnancy. *Cells* 9:645. doi: 10.3390/cells9030645

Team, R. C. (2018). *R: A Language and Environment For Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Teng, X., Chen, X., Xue, H., Tang, Y., Zhang, P., Kang, Q., et al. (2020). NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Res.* 48, D160–D165. doi: 10.1093/nar/gkz969

Therneau, T. M. (2015). *A Package for Survival Analysis in S. version 2.38*.

Thornhill, A. R., deDie-Smulders, C. E., Geraedts, J. P., Harper, J. C., Harton, G. L., Lavery, S. A., et al. (2005). ESHRE PGD Consortium 'Best practice guidelines for clinical preimplantation genetic diagnosis (PGD) and preimplantation genetic screening (PGS)'. *Hum. Reprod.* 20, 35–48. doi: 10.1093/humrep/deh579

Zhang, L., Liu, X., Che, S., Cui, J., Ma, X., An, X., et al. (2019). Endometrial Epithelial Cell Apoptosis Is Inhibited by a ciR8073-miR181a-Neurotensis Pathway during Embryo Implantation. *Mol. Ther. Nucleic Acids* 14, 262–273. doi: 10.1016/j.omtn.2018.12.005

Zhang, Q., Ni, T., Dang, Y., Ding, L., Jiang, J., Li, J., et al. (2020). MiR-148a-3p may contribute to flawed decidualization in recurrent implantation failure by modulating HOXC8. *J. Assist. Reprod. Genet.* 37, 2535–2544. doi: 10.1007/s10815-020-01900-9

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

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.

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

frontiers in Genetics

MINI REVIEW
published: 06 October 2021
doi: 10.3389/fgene.2021.742902

Omics and Computational Modeling Approaches for the Effective Treatment of Drug-Resistant Cancer Cells

Hae Deok Jung[1], *Yoo Jin Sung*[1] *and Hyun Uk Kim*[1,2,3]*

[1]Department of Chemical and Biomolecular Engineering (BK21 four), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, [2]KAIST Institute for Artificial Intelligence, KAIST, Daejeon, South Korea, [3]BioProcess Engineering Research Center and BioInformatics Research Center KAIST, Daejeon, South Korea

Chemotherapy is a mainstream cancer treatment, but has a constant challenge of drug resistance, which consequently leads to poor prognosis in cancer treatment. For better understanding and effective treatment of drug-resistant cancer cells, omics approaches have been widely conducted in various forms. A notable use of omics data beyond routine data mining is to use them for computational modeling that allows generating useful predictions, such as drug responses and prognostic biomarkers. In particular, an increasing volume of omics data has facilitated the development of machine learning models. In this mini review, we highlight recent studies on the use of multi-omics data for studying drug-resistant cancer cells. We put a particular focus on studies that use computational models to characterize drug-resistant cancer cells, and to predict biomarkers and/or drug responses. Computational models covered in this mini review include network-based models, machine learning models and genome-scale metabolic models. We also provide perspectives on future research opportunities for combating drug-resistant cancer cells.

Keywords: cancer, drug resistance, omics, computational modeling, network-based model, machine learning, genome-scale metabolic model

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Adil Mardinoglu,
King's College London,
United Kingdom
Satyakam Dash,
The Pennsylvania State University (PSU), United States

***Correspondence:**
Hyun Uk Kim
ehukim@kaist.ac.kr

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

Received: *16 July 2021*
Accepted: *20 September 2021*
Published: *06 October 2021*

Citation:
Jung HD, Sung YJ and Kim HU (2021) Omics and Computational Modeling Approaches for the Effective Treatment of Drug-Resistant Cancer Cells. Front. Genet. 12:742902. doi: 10.3389/fgene.2021.742902

INTRODUCTION

Drug resistance has been a major obstacle for a successful treatment of cancers, as manifested by over 90% mortality of cancer patients that appeared to be associated with drug resistance (Bukowski et al., 2020). Drug resistance is a phenotypic state that arises as a result of a complex interplay between genetic and non-genetic mechanisms (Marine et al., 2020). Such genetic and non-genetic reprogramming consequently leads to drug resistance through various mechanisms (Gatti and Zunino, 2005; Housman et al., 2014; Zheng, 2017; Lim and Ma, 2019; Vasan et al., 2019; Bukowski et al., 2020), including: drug inactivation, for example by an excessive level of glutathione that detoxifies xenobiotics (Jiang et al., 2017; De Luca et al., 2019); alteration of a drug target by mutations or changes in an expression level (Likhite et al., 2006; Costa et al., 2008); drug efflux by transporters (Giddings et al., 2021); enhanced DNA damage repair system (Harte et al., 2014); development of resistance via dysregulated autophagy (Martin et al., 2017; Cai et al., 2019); epithelial-mesenchymal transition (EMT) (Fischer et al., 2015; Zheng et al., 2015); or heterogeneity of a cancer cell population having cancer

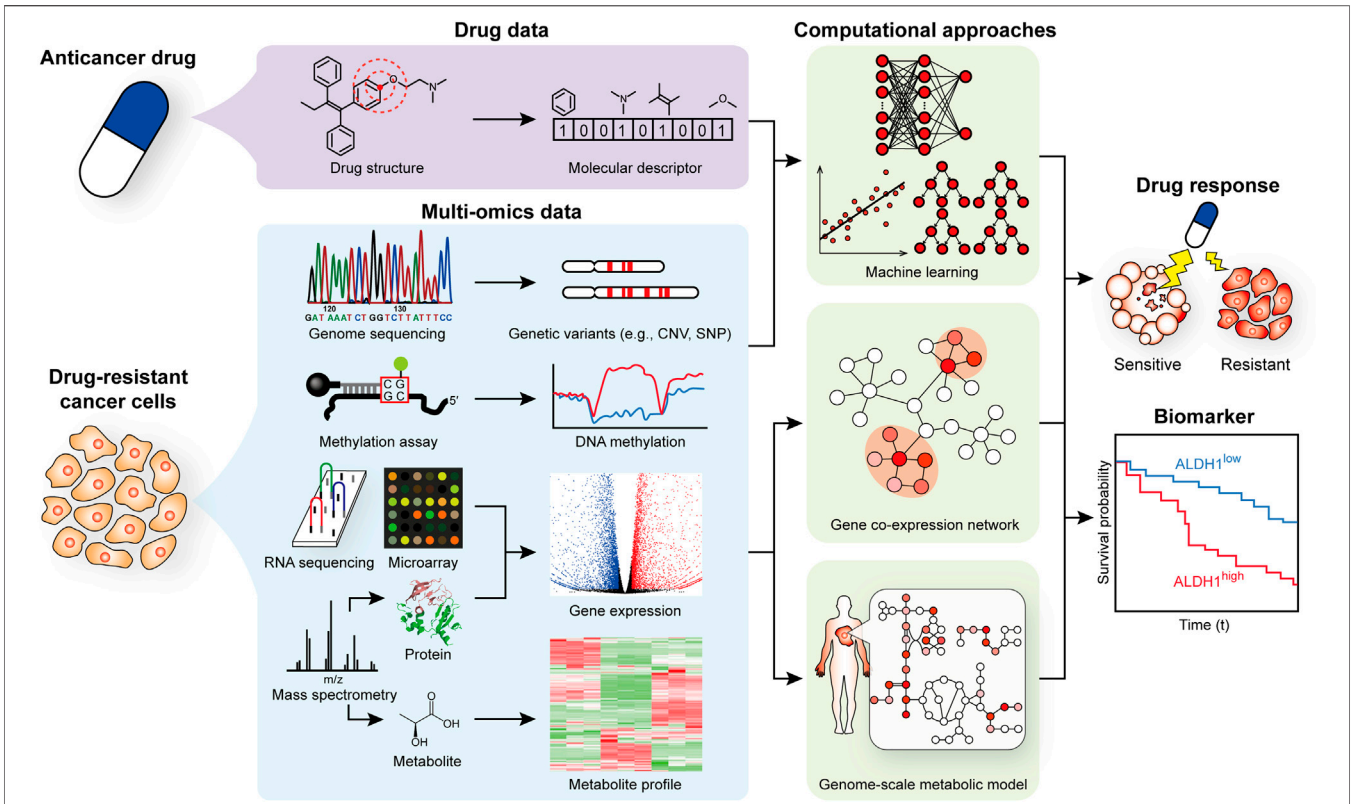


FIGURE 1 | Scheme of omics data generation and computational modeling to better understand and treat drug-resistant cancer cells.

stem cells (Seth et al., 2019; Zhao et al., 2021). A state of drug resistance is indeed a highly complex phenotype that requires multidimensional approaches.

Omics technologies have now become indispensable for characterizing mechanisms of cancer progression, and for identifying effective biomarkers and treatment targets for cancers. For this reason, large-scale projects have been launched to generate omics data of various cancer cells. A recent representative example is the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), which has allowed advanced studies on gene mutations and gene expression profiles across cancers (Consortium, 2020). The resulting various datasets from such large-scale efforts have been found to be useful for studying drug-resistant cancer cells. Relevant representative datasets include the NCI-60 Human Tumor Cell Lines Screen (Shoemaker, 2006), the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2013), TCGA (Cancer Genome Atlas Research et al., 2013), the Cancer Therapeutic Response Portal (CTRP) (Seashore-Ludlow et al., 2015), L1000 profiles from The Library of Integrated Network-Based Cellular Signatures (LINCS) Program (Subramanian et al., 2017), the Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019), and the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Tate et al., 2019). All these datasets have served as a source of novel insights that help characterize and overcome drug-resistant cancer cells. In particular, it is expected that an increasing volume of such large-scale datasets will facilitate development of various computational models that will better systematize our approaches to studying drug-resistant cancer cells.

We here review recent studies that utilized multi-omics and computational modeling approaches to better understand mechanisms associated with the progression of drug resistance, and to identify biomarkers and/or drug responses (**Figure 1** and **Table 1**). Especially, we put more focus on computational modeling that makes predictions for various scenarios for the treatment of drug-resistant cancer cells. We also provide an outlook for further advances on the use of computational models for studying drug-resistant cancer cells.

MULTI-OMICS ANALYSES

Multiple omics data are often generated to examine various biological aspects of drug-resistant cancer cells (**Figure 1**). Target genotypes and phenotypes examined using omics data (**Table 1**) include: cancer-associated mutations (Niehr et al., 2018; Marczyk et al., 2020; Sinkala et al., 2021); changes in the expression level of specific genes (Niehr et al., 2018; Nava et al., 2019; Kagohara et al., 2020; Marczyk et al., 2020; Poojan et al., 2020; Sinkala et al., 2021); changes in chromosome structure (Kagohara et al., 2020; Marczyk et al., 2020; Aissa et al., 2021); epigenetic alterations (e.g., methylation or acetylation states of histone

TABLE 1 | Recent studies on the use of omics data and computational models to better understand and treat drug-resistant cancer cells.

Approaches	Cancer types	Resistance type	Objectives	Drugs	References
Multi-omics analyses					
• ChiP-seq • Single-cell RNA-seq • RNA-seq • Proteome (LC-MS/MS)	• Lung cancer	• Both acquired and intrinsic resistance	• Identification of biomarkers	• Erlotinib, osimertinib, crizotinib, vemurafenib, celastrol, and GSK-1059615	Aissa et al. (2021)
• ATAC-seq • RNA-seq	• Breast cancer	• Intrinsic resistance	• Biological characterization • Identification of therapeutic targets	• Doxorubicin	Kumar et al. (2021)
• Genome sequencing • Methylome (reduced representation bisulfite sequencing) • mRNA microarray and RNA-seq	• 101 Types of cancers from 40,848 patients from cBioPortal	• Not specified	• Biological characterization	• MAPK pathway inhibitors (e.g., selumetinib)	Sinkala et al. (2021)
• RNA-seq • Pooled CRISPR screen (MiSeq)	• Melanoma	• Intrinsic resistance	• Biological characterization	• Vemurafenib	Torre et al. (2021)
• Genome sequencing • Methylome (bisulfite sequencing) • Hi-C • ChiP-seq • RNA-seq	• Breast cancer	• Acquired resistance	• Biological characterization	• Tamoxifen and fulvestrant	Achinger-Kawecka et al. (2020)
• Methylome (EPIC array) • ChiP-seq • RNA-seq • Metabolome (LC-HRMS)	• Breast cancer	• Acquired resistance	• Biological characterization	• Paclitaxel	Deblois et al. (2020)
• ATAC-seq • Single-cell RNA-seq • RNA-seq	• Head and neck squamous carcinoma	• Acquired resistance	• Biological characterization	• Cetuximab	Kagohara et al. (2020)
• Translatome (microarray) • mRNA microarray • Proteome (LC-MS/MS)	• Leukemia	• Not specified	• Biological characterization • Identification of therapeutic targets	• Cytosine arabinoside	Lee et al. (2020)
• Genome sequencing • Methylome (bisulfite sequencing) • ATAC-seq • Single-cell RNA-seq • RNA-seq	• Breast cancer	• Acquired resistance	• Biological characterization • Identification of biomarkers	• Navitoclax	Marczyk et al. (2020)
• ChiP-seq • RNA-seq	• Breast cancer	• Acquired resistance	• Biological characterization	• Doxorubicin and 5-fluorouracil (5-FU)	Mukherjee et al. (2020)
• Single-cell RNA-seq • RNA-seq	• Lung cancer • Gastric cancer	• Acquired resistance	• Biological characterization	• Cisplatin and paclitaxel	Poojan et al. (2020)
• ATAC-seq • ChiP-seq • Single cell RNA-seq • RNA-seq • Click-seq	• Leukemia	• Acquired resistance	• Biological characterization • Identification of therapeutic targets	• Bromodomain and Extra-Terminal motif (BET) inhibitor	Bell et al. (2019)
• RNA-seq • Proteome (nanoLC-MS/MS)	• Lymphoma	• Not specified	• Identification of biomarkers • Identification of therapeutic targets	• Anthracycline-based regimen R-CHOP (i.e., rituximab, cyclophosphamide, doxorubicin, vincristine and prednisone)	Fornecker et al. (2019)

(Continued on following page)

TABLE 1 | (*Continued*) Recent studies on the use of omics data and computational models to better understand and treat drug-resistant cancer cells.

Approaches	Cancer types	Resistance type	Objectives	Drugs	References
• Proteome, phosphoproteome, kinome (LC-MS/MS)	• Breast cancer	• Acquired resistance	• Biological characterization	• 2,5-diaziridinyl-3-hydroxyl-6-methyl-1,4-benzoquinone (RH1)	Kuciauskas et al. (2019)
• ChiP-seq • RNA-seq	• Breast cancer	• Intrinsic resistance	• Biological characterization • Identification of biomarkers	• Trastuzumab	Nava et al. (2019)
• Exome sequencing • Single-cell DNA-seq • Single-cell RNA-seq	• Breast cancer	• Both acquired and intrinsic resistance	• Biological characterization	• Epirubicin, docetaxel, and bevacizumab	Kim et al. (2018)
• Genome sequencing • mRNA microarray • Phosphoproteome (LC-MS/MS)	• Head and neck squamous carcinoma	• Intrinsic resistance • Identification of therapeutic targets	• Biological characterization	• Cisplatin	Niehr et al. (2018)
Network-based modeling					
• GCNA using mRNA microarray data • Cox regression model	• Breast cancer	• Not specified	• Identification of biomarkers	• Trastuzumab and docetaxel	Li et al. (2021a)
• Weighted GCNA using RNA-seq data • Cox regression model	• Breast cancer	• Not specified	• Identification of biomarkers	• Doxorubicin	Li et al. (2021b)
• Gene co-expression network analysis (GCNA) using RNA-seq data • Methylome (BeadChip array) • Genome sequencing	• Breast cancer	• Not specified	• Identification of biomarkers • Biological characterization	• Doxorubicin, cytoxan, and tamoxifen	Cui et al. (2020)
• Weighted GCNA using mRNA microarray data	• Gastric cancer	• Acquired resistance	• Identification of biomarkers	• 5-FU and cisplatin	Qi and Zhang, (2020)
• ceRNA network for correlation between lncRNA and mRNA levels using RNA-seq data	• 19 Types (e.g., Lung cancer, breast cancer, and melanoma)	• Not specified	• Identification of biomarkers • Biological characterization	• 138 Drugs (e.g., vorinostat and bosutinib)	Liu et al. (2019a)
• GCNA using RNA-seq data • Cox regression model	• Glioma	• Acquired resistance	• Identification of biomarkers	• Dibutyryl cyclic adenosine monophosphate	Zhang et al. (2019)
• Weighted GCNA using RNA-seq data	• Breast cancer	• Acquired resistance	• Identification of biomarkers	• Docetaxel	Huang et al. (2018)
Machine learning					
• Deep neural network (DNN) with neighborhood component analysis using CNV, somatic mutation, methylome, mRNA microarray, RNA-seq, and proteome data	• Breast cancer	• Not specified	• Prediction of a drug response	• 100 Drugs (e.g., tamoxifen)	Malik et al. (2021)
• Logistic regression using CNV, somatic mutation, mRNA microarray, drug targets, and drug descriptor data	• 955 Cell lines from GDSC (lung cancer, urogenital, and leukemia) • 491 Cell lines from CCLE	• Not specified	• Prediction of a drug response	• 219 Drugs (e.g., AT-7519) for GDSC cell lines • 24 Drugs (e.g., AZD6244) for CCLE cell lines	Yu et al. (2021)

(Continued on following page)

TABLE 1 | (*Continued*) Recent studies on the use of omics data and computational models to better understand and treat drug-resistant cancer cells.

Approaches	Cancer types	Resistance type	Objectives	Drugs	References
• DNN with multiple elastic nets using mRNA microarray and drug descriptor data	• 983 Cell lines from GDSC • 491 Cell lines from CCLE	• Not specified	• Identification of biomarkers • Prediction of a drug response	• 222 Drugs (e.g., 5-FU) for GDSC cell lines • 12 Drugs for CCLE cell lines	Choi et al. (2020)
• Weighted GCNA, elastic net, and random forest using proteome and phosphoproteome data • Cox regression model	• NCI60 cell line panel • Prediction of a drug response	• Not specified	• Identification of biomarkers • CRC65 cell line panel	• Various drugs (e.g., cytarabine, 5-FU)	Frejno et al. (2020)
• Ridge regression and support vector regression using mRNA microarray and RNA-seq data	• Colorectal cancer • Bladder cancer	• Not specified	• Identification of biomarkers • Prediction of a drug response	• 5-FU for colorectal cancer • Cisplatin for bladder cancer	Kong et al. (2020)
• Ensemble transfer learning (LighGBM, or DNNs with two different architectures) using RNA-seq data and drug descriptor data	• Hundreds of cancer cell lines from CCLE, CTRP, gCSI and GDSC	• Not specified	• Prediction of a drug response	• Hundreds of drugs from CCLE, CTRP, gCSI and GDSC	Zhu et al. (2020)
• Artificial neural network using single-cell metabolome data	• Leukemia	• Intrinsic resistance	• Prediction of a drug response	• Cell adhesion as a indication of drug resistance without addition of a drug	Liu et al. (2019b)
• DNN using mRNA microarray and RNA-seq data	• 1,001 Cell lines from 55 tissues (e.g., leukemia) in GDSC	• Not specified	• Prediction of a drug response	• Bortezomib, PARP inhibitor, cisplatin, and paclitaxel	Sakellaropoulos et al. (2019)
• Random forest using RNA-seq, CNV, and methylome data • Cox regression model	• Bladder cancer • Glioma • Pancreatic cancer • Gastric cancer	• Not specified	• Identification of biomarkers • Prediction of a drug response	• Cisplatin and gemcitabine (for bladder cancer) • Temozolomide (for glioma) • Gemcitabine (for pancreatic cancer) • 5-FU (for gastric cancer)	Xu et al. (2019)
• Elastic net using proteome and kinome data • Cox regression model	• Colorectal cancer	• Not specified	• Identification of biomarkers • Prediction of a drug response	• 577 Drugs (e.g., cetuximab and afatinib)	Frejno et al. (2017)

proteins) (Nava et al., 2019; Kagohara et al., 2020; Marczyk et al., 2020; Poojan et al., 2020; Sinkala et al., 2021); and the presence of heterogeneity of a cell population (Niehr et al., 2018), often increasingly examined at a single-cell resolution (Kagohara et al., 2020; Aissa et al., 2021). In a recent study for cell line heterogeneity, for example, application of single-cell DNA and RNA sequencing (RNA-seq) to 20 triple-negative breast cancer (TNBC) patients revealed that rare pre-existing clones having genotypes associated with chemoresistance were adaptively selected in response to neoadjuvant chemotherapy, which subsequently led to acquired transcriptional reprogramming (Kim et al., 2018). For epigenetic alteration, chromosome conformation capture (Hi-C) along with additional omics analyses were conducted for estrogen receptor positive (ER+) breast cancer, which showed that resistance development to endocrine therapy was accompanied with notable 3-dimensional (3D) epigenome alterations (Achinger-Kawecka et al., 2020). Application of multi-omics analyses has also been extended to examine biological processes in quiescent cancer cells that show drug resistance (Lee et al., 2020; Kumar et al., 2021).

Understanding the biology of drug resistance often helps devise effective treatment strategies for drug-resistant cancer cells. Relevant examples (**Table 1**) include targeting: cancer stem cell phenotypes, in particular stem cell factor receptor c-KIT, for TNBC cells resistant to an anticancer agent RH1 that is currently under clinical trials (Kuciauskas et al., 2019); a range of biological pathways (e.g., metabolism), microenvironment as well as proliferation, migration and invasion of cells, which are all associated with drug resistance for diffuse large B-cell lymphoma patients (Fornecker et al., 2019); zinc finger MYND domain-containing protein 8 (ZMYND8), a putative chromatin reader that appeared to suppress tumorigenic potential and drug resistance induced by doxorubicin (Mukherjee et al., 2020); and EZH2 responsible for histone methylation in taxane-resistant TNBC (Deblois et al., 2020).

As representative examples of overcoming drug resistance on the basis of omics analyses, recent studies additionally conducted

CRISPR-Cas9-based genetic screens to examine cellular plasticity, which was suggested as a therapeutic target for drug-resistant cancer cells (Bell et al., 2019; Torre et al., 2021). Cellular plasticity describes non-genetic transformation of a cellular state into a drug-resistant state by reprogramming gene expression profiles. In a study by Torre et al., CRISPR-Cas9 genetic screens were implemented for melanoma cells to identify genes that affect cell fate decisions by altering cellular plasticity (Torre et al., 2021). In particular, modulating the cellular plasticity was demonstrated for vemurafenib inhibiting B-Raf, encoded by a proto-oncogene, in melanoma. Interestingly, inhibiting DOT1L, associated with the onset of melanoma, before the B-Raf inhibition showed more drug resistance than simultaneous inhibition of DOT1 and B-Raf using pinometostat and vemurafenib, respectively. Subsequent transcriptome analysis of knockout cell lines generated clues for non-genetic mechanisms of drug resistance. Another study by Bell et al. focused on acute myeloid leukemia patients that showed non-genetic drug resistance (Bell et al., 2019). Single-cell RNA-seq, followed by CRISPR-Cas9 screening, led to the identification of genes responsible for transcriptional plasticity that triggered epigenetic resistance. Among the genes identified was Lsd1, the inhibition of which was shown to overcome non-genetic drug resistance. As demonstrated by these two recent studies, implementation of genome engineering in addition to omics analyses provides compelling evidence for targets that can help overcome drug resistance.

COMPUTATIONAL MODELING APPROACHES

While various bioinformatic analyses are available for analyzing omics data, such as enrichment analyses, gene co-expression networks (GCNs) (Cui et al., 2020; Qi and Zhang, 2020) and their variants (e.g., a network of long non-coding RNAs and mRNAs) (Huang et al., 2018; Liu H. et al., 2019) as well as dimensionality reduction (e.g., t-SNE and UMAP), omics data have also been subjected to computational modeling to make predictions for discovering novel mechanisms and devising treatment strategies for drug-resistant cancers (**Figure 1**). Use of survival analysis in combination with GCNs, and development of a gene regulatory network (GRN) model using a set of ordinary differential equations (ODEs), machine learning models, and genome-scale metabolic models (GEMs) are representative computational modeling approaches that have recently been considered for studying drug-resistant cancer cells (**Table 1**).

Network-Based Modeling

GCN has been a popular analysis for understanding gene expression patterns from transcriptome data. GCN is an undirected graph that can be constructed from transcriptome data (e.g., RNA-seq), and connects pairs of genes (nodes in a GCN) with an edge if each pair of genes shows significant co-expression patterns across the transcriptome data. GCN analysis, such as identifying hub genes and/or modules, allows prioritizing candidate genes that may be highly associated with drug resistance of cancer cells. Weighted GCN additionally considers the level of significance in the co-expression relationship between genes in a pair. Often, outcomes from (weighted) GCN analysis are further subjected to other computational analyses, for example survival analysis, to validate the biological and/or clinical significance of the candidate genes. As a recent example, Li et al. focused on *PPP2R2B*, encoding serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B beta isoform, as a potential prognostic biomarker for TNBC on the basis of a series of bioinformatic analyses involving a GCN (Li Z. et al., 2021). Kaplan-Meier survival analysis for this gene revealed that patients with a low expression level of *PPP2R2B* showed shorter survival time than those with a high expression level of *PPP2R2B*. Interestingly, *PPP2R2B* upregulation could attenuate the resistance of TNBC cells to doxorubicin. Likewise, Cox proportional hazards regression model (Cox regression model) was used for genes selected from GCNs to predict prognostic biomarkers for breast cancer, and to suggest genes (e.g., *CCNE2* and *KIF14*) that may help overcome drug resistance (Li Y.-K. et al., 2021).

While GCNs can provide clinically important information when combined with additional predictive models, such as survival analysis above, they have limitations in generating clues on a molecular mechanism associated with development of drug resistance, in particular dynamic interactions between genes. To address this problem, Zhang et al. developed a time-course RNA-seq data-driven computational framework (DryNetMC) to construct GRNs that help elucidate dynamic interactions between genes, and identify key genes associated with mechanisms of drug resistance (Zhang et al., 2019). DryNetMC involves a set of ODEs, a regularized regression method as well as a series of network analyses. Using DryNetMC, GRNs were constructed for dbcAMP-sensitive and dbcAMP-resistant glioma cells based on their time-course RNA-seq data. These differential GRNs were subsequently subjected to a systematic characterization to identify their unique network properties (e.g., node importance) that helped identify key genes (e.g., *KIF2C*, *CCNA2*, *NDC80*, *KIF11*, and *KIF23*) that are predictive of a cancer cell's drug response. Because network-based models, either by using a GCN or other methods (e.g., ODEs), can visualize a biological context (e.g., association between genes), they will continue to be actively used in the analysis of omics data, and likely along with additional predictive models.

Machine Learning

Increasing availability of omics data for drug-resistant cancer cells has also provided unprecedented opportunities for building machine learning models. In general, machine learning models perform classification or regression, depending on a given problem. Recently, prediction of anticancer drug response was attempted by using various types of machine learning methods, such as logistic regression (Frejno et al., 2017; Yu et al., 2021), random forest (Xu et al., 2019) and deep neural network (DNN; e.g., multilayer perceptron) (Malik et al., 2021) on the basis of a range of omics and drug response data (**Table 1**). When developing these machine learning models, transcriptome (RNA-seq or mRNA microarray) was the most frequently adopted dataset, but other types of datasets were also considered, including genome (e.g., gene mutations) (Yu et al., 2021), proteome (Frejno et al., 2020), epigenome (Xu et al.,

2019), mass spectrometry data (Liu R. et al., 2019) and molecular features of a target drug (Zhu et al., 2020).

In a recent study by Kong et al., a machine learning model was developed that can predict a patient's drug response on the basis of the analysis of protein-protein interaction (PPI) network and pharmacogenomic data from 3D organoid culture models (Kong et al., 2020). Specifically, potential biomarkers were first inferred from the PPI network analysis, and their corresponding expression profiles along with drug response data (IC_{50}) were used to train a machine learning model (e.g., ridge regression). The resulting drug responses were validated using survival analysis by focusing on colorectal and bladder cancer patients treated with 5-fluorouracil and cisplatin, respectively. The predicted drug responses also appeared to be consistent with transcriptome profiles from drug-sensitive and drug-resistant isogenic cancer cell lines as well as data on somatic mutations associated with already known biomarkers. In this study, consideration of the network analysis not only helped improve the performance of the developed machine learning model, but also facilitated the interpretation of model prediction outcomes. Likewise, in another study, elastic net and random forest regression were used to predict drug responses from abundance data of proteins and their phosphorylation sites in cancer cell lines (Frejno et al., 2020).

Among machine learning methods, DNNs are increasingly used for various predictions, and they have also been used to predict drug responses. Sakellaropoulos et al. developed a DNN model by using GDSC datasets (i.e., transcriptomic data for 1,001 cancer cell lines and IC_{50} values of 251 drugs) to predict drug responses (Sakellaropoulos et al., 2019). Across several datasets tested, the DNN model showed consistently better performance than elastic net and random forest models. The DNN model was validated by conducting survival analyses for the model-predicted IC_{50} values, which split patients based on their drug responsiveness. Importantly, pathway enrichment analysis using information from the DNN model (i.e., weights that connect the input layer and the first hidden layer) appeared to associate specific biological pathways with mechanisms of action for drugs. In a more recent study, predicting drug response was also attempted by using a DNN model combined with multiple elastic nets (Choi et al., 2020), referred to as Reference Drug-based Neural Network (RefDNN). RefDNN was developed more in the context of drug resistance, which predicts whether a given cell line is resistant to a target drug by processing gene expression profiles and molecular structure of a drug. RefDNN was also shown to help identify biomarker genes associated with drug resistance, and explore a novel anticancer drug via drug repositioning.

Despite its demonstrated performance, machine learning is often challenged with the limited availability of training datasets for many technical fields. This challenge can be addressed to a certain extent by employing transfer learning as recently demonstrated (Zhu et al., 2020). Zhu et al. demonstrated that ensemble transfer learning can improve the prediction of drug responses in the context of drug repositioning (i.e., use of a drug for another cancer that is already known), precision oncology (i.e., use of a drug for a new cancer that has never been treated before) and new drug development (i.e., use of a new drug for already known cancer). In this particular study, LightGBM (Light Gradient Boosting Machine) and two different DNN models were considered for ensemble transfer learning; larger datasets from the CTRP and GDSC were used as source data for initial training of models, and smaller datasets from CCLE and the Genentech Cell Line Screening Initiative (gCSI) served as target data for further refinement and testing of the models. It was shown that ensemble transfer learning-based models almost always outperformed models that were not developed using transfer learning. This study suggests the use of transfer learning for other drug-resistant cancer cells where a training dataset is sufficiently not available.

Genome-Scale Metabolic Modeling

GEM is a computational model that describes gene-protein-reaction (GPR) associations, and can be simulated to predict genome-scale metabolic flux distributions (Gu et al., 2019). GEMs are now available for an increasing number of organisms that are important in biotechnology and biomedicine. Several versions of human GEMs (Ryu et al., 2017; Brunk et al., 2018; Robinson et al., 2020) are currently available, which have been used to examine a target cell's metabolism, and to predict biomarkers and drug targets for various diseases (Cook and Nielsen, 2017; Gu et al., 2019). For a medical application, a generic human GEM, covering all the known GPR associations in human metabolism, is initially integrated with omics data, often transcriptome (e.g., RNA-seq), to build a context-specific GEM, a GEM that is specific to a target cell or tissue (Ryu et al., 2015; Opdam et al., 2017). The resulting context-specific GEM is then simulated for various metabolic studies.

Human GEMs have recently been used to study radiation-resistant tumors (Lewis et al., 2021; Lewis and Kemp, 2021), but not drug-resistant cancer cells, to the best of our knowledge. Lewis et al. newly constructed GEMs for radiation-sensitive and radiation-resistant tumors through multi-omics integration (i.e., transcriptome data, mutational data, kinetic data and thermodynamic data) (Lewis et al., 2021). These context-specific GEMs were used to identify changes in redox cofactor production that give resistance to radiation therapy. In the other study, ensemble machine learning classifiers were developed to predict whether an individual is responsive or resistant to a radiation therapy by considering data of metabolite production rates predicted from context-specific GEMs as well as mutation data, transcriptome data and clinical data from TCGA (Lewis and Kemp, 2021). These two studies obviously suggest that GEM-based approaches can also be considered to identify metabolic signatures of drug-resistant cancer cells, and to predict effective drug targets for these cancer cells.

OUTLOOK

Understanding genotype-phenotype associations in drug-resistant cancer cells is a highly complex problem, and therefore use of multi-omics data has been considered to capture various aspects of these troubling cancer cells. In particular, multi-omics analyses along with additional tools, such as genome engineering (e.g., CRISPR-Cas9), will continue to play an important role in thorough characterization of drug-resistant cancer cells. Also, an increasing volume of omics data will facilitate development of various types of

computational models. As a consequence, prediction outcomes from computational models will allow more systematically designing experiments for drug-resistant cancer cells.

Despite the promises of omics data and computational models, technical challenges exist. First, current coverage of multi-omics data is not sufficient for thoroughly studying a range of drug-resistant cancer cells. In particular, generation of a consistent set of multi-omics data from each single cell is necessary for in-depth study of a target cancer cell and comparison of different types of cancer cells. Also, it will be interesting to examine the effects of using datasets obtained from patients having a specific disease instead of publicly available datasets (e.g., GDSC and CTRP). While currently available machine learning models have been rigorously validated by using public datasets, they might reveal previously unnoticed limitations in a clinical setting because the public datasets are often generated in a highly controlled condition. In particular, additional consideration of non-genetic factors (e.g., age, gender, and lifestyle) may help reveal new insights on drug-resistant cancer cells. Use of patient-specific datasets will allow more widespread use of the state-of-the-art computational models in a clinical setting.

For network-based modeling, including both GCN and GRN, a breakthrough is needed that allows efficiently developing a cell-specific large-scale GRN that can be simulated under various conditions (e.g., gene perturbation). For machine learning, despite its high predictive performance, there is always a challenge of avoiding overfitting and achieving explainability. Explainability in terms of biological processes is particularly important in the field of biomedicine in order to explain prediction outcomes and make medical decisions. In case of human GEMs, because patient-specific omics data (e.g., RNA-seq) are available to a certain extent, human GEMs should be more actively considered to systematically examine metabolism of drug-resistant cancer cells. Availability of multi-omics data will be particularly useful for interpreting human GEMs and their prediction outcomes; because human GEMs only cover a metabolic network, use of multi-omics data can help explain a complex interplay between metabolic and regulatory networks. Prediction outcomes from the simulation of human GEMs will in turn help explain the insights reaped from omics analyses.

Taken together, advances in omics technologies and computational modeling will bring about positive impacts in understanding and treating drug-resistant cancer cells. Feedback from clinicians and biomedical researchers will be additionally useful for the successful development and clinical application of computational models.

AUTHOR CONTRIBUTIONS

HUK, HDJ and YJS wrote the manuscript. HDJ and YJS prepared the figure and table, and critically reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Bio-Synergy Research Project (NRF-2018M3A9C4076475) and the KAIST Cross-Generation Collaborative Lab project of the Ministry of Science and ICT through the National Research Foundation of Korea. This work was also supported by Kwon Oh-Hyun Assistant Professor fund of the KAIST Development Foundation.

REFERENCES

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), 320. doi:10.1038/s41467-019-14098-x

Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., Go, C. C., Rader, A. E., Conrardy, R. D., et al. (2021). Single-cell Transcriptional Changes Associated with Drug Tolerance and Response to Combination Therapies in Cancer. *Nat. Commun.* 12 (1), 1628. doi:10.1038/s41467-021-21884-z

Bell, C. C., Fennell, K. A., Chan, Y.-C., Rambow, F., Yeung, M. M., Vassiliadis, D., et al. (2019). Targeting Enhancer Switching Overcomes Non-genetic Drug Resistance in Acute Myeloid Leukaemia. *Nat. Commun.* 10 (1), 2723. doi:10.1038/s41467-019-10652-9

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., et al. (2018). Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism. *Nat. Biotechnol.* 36 (3), 272–281. doi:10.1038/nbt.4072

Bukowski, K., Kciuk, M., and Kontek, R. (2020). Mechanisms of Multidrug Resistance in Cancer Chemotherapy. *Ijms* 21 (9), 3233. doi:10.3390/ijms21093233

Cai, Q., Wang, S., Jin, L., Weng, M., Zhou, D., Wang, J., et al. (2019). Long Non-coding RNA GBCDRlnc1 Induces Chemoresistance of Gallbladder Cancer Cells by Activating Autophagy. *Mol. Cancer* 18 (1), 82. doi:10.1186/s12943-019-1016-0

Choi, J., Park, S., and Ahn, J. (2020). RefDNN: a Reference Drug Based Neural Network for More Accurate Prediction of Anticancer Drug Resistance. *Sci. Rep.* 10 (1), 1861. doi:10.1038/s41598-020-58821-x

Collisson, E. A., Weinstein, J. N., Weinstein, J. N., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Consortium, I. T. P.-C. A. o. W. G. (2020). Pan-cancer Analysis of Whole Genomes. *Nature* 578 (7793), 82–93. doi:10.1038/s41586-020-1969-6

Cook, D. J., and Nielsen, J. (2017). Genome-scale Metabolic Models Applied to Human Health and Disease. *Wires Syst. Biol. Med.* 9 (6), e1393. doi:10.1002/wsbm.1393

Costa, D. B., Nguyen, K.-S. H., Cho, B. C., Sequist, L. V., Jackman, D. M., Riely, G. J., et al. (2008). Effects of Erlotinib in EGFR Mutated Non-small Cell Lung Cancers with Resistance to Gefitinib. *Clin. Cancer Res.* 14 (21), 7060–7067. doi:10.1158/1078-0432.CCR-08-1455

Cui, H., Kong, H., Peng, F., Wang, C., Zhang, D., Tian, J., et al. (2020). Inferences of Individual Drug Response-Related Long Non-coding RNAs Based on Integrating Multi-Omics Data in Breast Cancer. *Mol. Ther. - Nucleic Acids* 20, 128–139. doi:10.1016/j.omtn.2020.01.038

De Luca, A., Parker, L. J., Ang, W. H., Rodolfo, C., Gabbarini, V., Hancock, N. C., et al. (2019). A Structure-Based Mechanism of Cisplatin Resistance Mediated by Glutathione Transferase P1-1. *Proc. Natl. Acad. Sci. USA* 116 (28), 13943–13951. doi:10.1073/pnas.1903297116

Deblois, G., Tonekaboni, S. A. M., Grillo, G., Martinez, C., Kao, Y. I., Tai, F., et al. (2020). Epigenetic Switch-Induced Viral Mimicry Evasion in Chemotherapy-Resistant Breast Cancer. *Cancer Discov.* 10 (9), 1312–1329. doi:10.1158/2159-8290.CD-19-1493

Fischer, K. R., Durrans, A., Lee, S., Sheng, J., Li, F., Wong, S. T. C., et al. (2015). Epithelial-to-mesenchymal Transition Is Not Required for Lung Metastasis but Contributes to Chemoresistance. *Nature* 527 (7579), 472–476. doi:10.1038/nature15748

Fornecker, L.-M., Muller, L., Bertrand, F., Paul, N., Pichot, A., Herbrecht, R., et al. (2019). Multi-omics Dataset to Decipher the Complexity of Drug Resistance in Diffuse Large B-Cell Lymphoma. *Sci. Rep.* 9 (1), 895. doi:10.1038/s41598-018-37273-4

Frejno, M., Meng, C., Ruprecht, B., Oellerich, T., Scheich, S., Kleigrewe, K., et al. (2020). Proteome Activity Landscapes of Tumor Cell Lines Determine Drug Responses. *Nat. Commun.* 11 (1), 3639. doi:10.1038/s41467-020-17336-9

Frejno, M., Zenezini Chiozzi, R., Wilhelm, M., Koch, H., Zheng, R., Klaeger, S., et al. (2017). Pharmacoproteomic Characterisation of Human colon and Rectal Cancer. *Mol. Syst. Biol.* 13 (11), 951. doi:10.15252/msb.20177701

Gatti, L., and Zunino, F. (2005). Overview of Tumor Cell Chemoresistance Mechanisms. *Methods Mol. Med.* 111, 127–148. doi:10.1385/1-59259-889-7:127

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., Barretina, J., et al. (2019). Next-generation Characterization of the Cancer Cell Line Encyclopedia. *Nature* 569 (7757), 503–508. doi:10.1038/s41586-019-1186-3

Giddings, E. L., Champagne, D. P., Wu, M.-H., Laffin, J. M., Thornton, T. M., Valenca-Pereira, F., et al. (2021). Mitochondrial ATP Fuels ABC Transporter-Mediated Drug Efflux in Cancer Chemoresistance. *Nat. Commun.* 12 (1), 2804. doi:10.1038/s41467-021-23071-6

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., and Lee, S. Y. (2019). Current Status and Applications of Genome-Scale Metabolic Models. *Genome Biol.* 20 (1), 121. doi:10.1186/s13059-019-1730-3

Harte, M. T., Gorski, J. J., Savage, K. I., Purcell, J. W., Barros, E. M., Burn, P. M., et al. (2014). NF-κB Is a Critical Mediator of BRCA1-Induced Chemoresistance. *Oncogene* 33 (6), 713–723. doi:10.1038/onc.2013.10

Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., et al. (2014). Drug Resistance in Cancer: an Overview. *Cancers* 6 (3), 1769–1792. doi:10.3390/cancers6031769

Huang, P., Li, F., Li, L., You, Y., Luo, S., Dong, Z., et al. (2018). lncRNA Profile Study Reveals the mRNAs and lncRNAs Associated with Docetaxel Resistance in Breast Cancer Cells. *Sci. Rep.* 8 (1), 17970. doi:10.1038/s41598-018-36231-4

Jiang, Y., Cheng, J., Yang, C., Hu, Y., Li, J., Han, Y., et al. (2017). An Ultrasensitive Fluorogenic Probe for Revealing the Role of Glutathione in Chemotherapy Resistance. *Chem. Sci.* 8 (12), 8012–8018. doi:10.1039/c7sc03338a

Kagohara, L. T., Zamuner, F., Davis-Marcisak, E. F., Sharma, G., Considine, M., Allen, J., et al. (2020). Integrated Single-Cell and Bulk Gene Expression and ATAC-Seq Reveals Heterogeneity and Early Changes in Pathways Associated with Resistance to Cetuximab in HNSCC-Sensitive Cell Lines. *Br. J. Cancer* 123 (1), 101–113. doi:10.1038/s41416-020-0851-5

Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., et al. (2018). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* 173 (4), 879–893. e813doi:10.1016/j.cell.2018.03.041

Kong, J., Lee, H., Kim, D., Han, S. K., Ha, D., Shin, K., et al. (2020). Network-based Machine Learning in Colorectal and Bladder Organoid Models Predicts Anti-cancer Drug Efficacy in Patients. *Nat. Commun.* 11 (1), 5485. doi:10.1038/s41467-020-19313-8

Kuciauskas, D., Dreize, N., Ger, M., Kaupinis, A., Zemaitis, K., Stankevicius, V., et al. (2019). Proteomic Analysis of Breast Cancer Resistance to the Anticancer Drug RH1 Reveals the Importance of Cancer Stem Cells. *Cancers* 11 (7), 972. doi:10.3390/cancers11070972

Kumar, S., Nandi, A., Singh, S., Regulapati, R., Li, N., Tobias, J. W., et al. (2021). Dll1+ Quiescent Tumor Stem Cells Drive Chemoresistance in Breast Cancer through NF-Kb Survival Pathway. *Nat. Commun.* 12 (1), 432. doi:10.1038/s41467-020-20664-5

Lee, S., Micalizzi, D., Truesdell, S. S., Bukhari, S. I. A., Boukhali, M., Lombardi-Story, J., et al. (2020). A post-transcriptional Program of Chemoresistance by AU-Rich Elements and TTP in Quiescent Leukemic Cells. *Genome Biol.* 21 (1), 33. doi:10.1186/s13059-020-1936-4

Lewis, J. E., Forshaw, T. E., Boothman, D. A., Furdui, C. M., and Kemp, M. L. (2021). Personalized Genome-Scale Metabolic Models Identify Targets of Redox Metabolism in Radiation-Resistant Tumors. *Cel Syst.* 12 (1), 68–81. e11doi:10.1016/j.cels.2020.12.001

Lewis, J. E., and Kemp, M. L. (2021). Integration of Machine Learning and Genome-Scale Metabolic Modeling Identifies Multi-Omics Biomarkers for Radiation Resistance. *Nat. Commun.* 12 (1), 2700. doi:10.1038/s41467-021-22989-1

Li, Y.-K., Hsu, H.-M., Lin, M.-C., Chang, C.-W., Chu, C.-M., Chang, Y.-J., et al. (2021a). Genetic Co-expression Networks Contribute to Creating Predictive Model and Exploring Novel Biomarkers for the Prognosis of Breast Cancer. *Sci. Rep.* 11 (1), 7268. doi:10.1038/s41598-021-84995-z

Li, Z., Li, Y., Wang, X., and Yang, Q. (2021b). PPP2R2B Downregulation Is Associated with Immune Evasion and Predicts Poor Clinical Outcomes in Triple-Negative Breast Cancer. *Cancer Cel Int* 21 (1), 13. doi:10.1186/s12935-020-01707-9

Likhite, V. S., Stossi, F., Kim, K., Katzenellenbogen, B. S., and Katzenellenbogen, J. A. (2006). Kinase-specific Phosphorylation of the Estrogen Receptor Changes Receptor Interactions with Ligand, Deoxyribonucleic Acid, and Coregulators Associated with Alterations in Estrogen and Tamoxifen Activity. *Mol. Endocrinol.* 20 (12), 3120–3132. doi:10.1210/me.2006-0068

Lim, Z.-F., and Ma, P. C. (2019). Emerging Insights of Tumor Heterogeneity and Drug Resistance Mechanisms in Lung Cancer Targeted Therapy. *J. Hematol. Oncol.* 12 (1), 134. doi:10.1186/s13045-019-0818-2

Liu, H., Wang, S., Zhou, S., Meng, Q., Ma, X., Song, X., et al. (2019a). Drug Resistance-Related Competing Interactions of lncRNA and mRNA across 19 Cancer Types. *Mol. Ther. - Nucleic Acids* 16, 442–451. doi:10.1016/j.omtn.2019.03.011

Liu, R., Zhang, G., and Yang, Z. (2019b). Towards Rapid Prediction of Drug-Resistant Cancer Cell Phenotypes: Single Cell Mass Spectrometry Combined with Machine Learning. *Chem. Commun.* 55 (5), 616–619. doi:10.1039/c8cc08296k

Malik, V., Kalakoti, Y., and Sundar, D. (2021). Deep Learning Assisted Multi-Omics Integration for Survival and Drug-Response Prediction in Breast Cancer. *BMC Genomics* 22 (1), 214. doi:10.1186/s12864-021-07524-2

Marczyk, M., Patwardhan, G. A., Zhao, J., Qu, R., Li, X., Wali, V. B., et al. (2020). Multi-Omics Investigation of Innate Navitoclax Resistance in Triple-Negative Breast Cancer Cells. *Cancers* 12 (9), 2551. doi:10.3390/cancers12092551

Marine, J.-C., Dawson, S.-J., and Dawson, M. A. (2020). Non-genetic Mechanisms of Therapeutic Resistance in Cancer. *Nat. Rev. Cancer* 20 (12), 743–756. doi:10.1038/s41568-020-00302-4

Martin, S., Dudek-Peric, A. M., Garg, A. D., Roose, H., Demirsoy, S., Van Eygen, S., et al. (2017). An Autophagy-Driven Pathway of ATP Secretion Supports the Aggressive Phenotype of BRAFV600E Inhibitor-Resistant Metastatic Melanoma Cells. *Autophagy* 13 (9), 1512–1527. doi:10.1080/15548627.2017.1332550

Mukherjee, S., Adhikary, S., Gadad, S. S., Mondal, P., Sen, S., Choudhari, R., et al. (2020). Suppression of Poised Oncogenes by ZMYND8 Promotes Chemo-Sensitization. *Cell Death Dis* 11 (12), 1073. doi:10.1038/s41419-020-03129-x

Nava, M., Dutta, P., Farias-Eisner, R., Vadgama, J. V., and Wu, Y. (2019). Utilization of NGS Technologies to Investigate Transcriptomic and Epigenomic Mechanisms in Trastuzumab Resistance. *Sci. Rep.* 9 (1), 5141. doi:10.1038/s41598-019-41672-6

Niehr, F., Eder, T., Pilz, T., Konschak, R., Treue, D., Klauschen, F., et al. (2018). Multilayered Omics-Based Analysis of a Head and Neck Cancer Model of Cisplatin Resistance Reveals Intratumoral Heterogeneity and Treatment-Induced Clonal Selection. *Clin. Cancer Res.* 24 (1), 158–168. doi:10.1158/1078-0432.CCR-17-2410

Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C., and Lewis, N. E. (2017). A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cel Syst.* 4 (3), 318–329. doi:10.1016/j.cels.2017.01.010

Poojan, S., Bae, S.-H., Min, J.-W., Lee, E. Y., Song, Y., Kim, H. Y., et al. (2020). Cancer Cells Undergoing Epigenetic Transition Show Short-Term Resistance and Are Transformed into Cells with Medium-Term Resistance by Drug Treatment. *Exp. Mol. Med.* 52 (7), 1102–1115. doi:10.1038/s12276-020-0464-3

Qi, W., and Zhang, Q. (2020). Gene's Co-expression Network and Experimental Validation of Molecular Markers Associated with the Drug Resistance of Gastric Cancer. *Biomarkers Med.* 14 (9), 761–773. doi:10.2217/bmm-2019-0504

Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., et al. (2020). An Atlas of Human Metabolism. *Sci. Signal.* 13 (624), eaaz1482. doi:10.1126/scisignal.aaz1482

Ryu, J. Y., Kim, H. U., and Lee, S. Y. (2017). Framework and Resource for More Than 11,000 Gene-Transcript-Protein-Reaction Associations in Human Metabolism. *Proc. Natl. Acad. Sci. USA* 114 (45), E9740–E9749. doi:10.1073/pnas.1713050114

Ryu, J. Y., Kim, H. U., and Lee, S. Y. (2015). Reconstruction of Genome-Scale Human Metabolic Models Using Omics Data. *Integr. Biol. (Camb.)* 7 (8), 859–868. doi:10.1039/c5ib00002e

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., et al. (2019). A Deep Learning Framework for Predicting Response to Therapy in Cancer. *Cel Rep.* 29 (11), 3367–3373. e3364doi:10.1016/j.celrep.2019.11.017

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., et al. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* 5 (11), 1210–1223. doi:10.1158/2159-8290.CD-15-0235

Seth, S., Li, C.-Y., Ho, I.-L., Corti, D., Loponte, S., Sapio, L., et al. (2019). Pre-existing Functional Heterogeneity of Tumorigenic Compartment as the Origin of Chemoresistance in Pancreatic Tumors. *Cel Rep.* 26 (6), 1518–1532. e1519doi:10.1016/j.celrep.2019.01.048

Shoemaker, R. H. (2006). The NCI60 Human Tumour Cell Line Anticancer Drug Screen. *Nat. Rev. Cancer* 6 (10), 813–823. doi:10.1038/nrc1951

Sinkala, M., Nkhoma, P., Mulder, N., and Martin, D. P. (2021). Integrated Molecular Characterisation of the MAPK Pathways in Human Cancers Reveals Pharmacologically Vulnerable Mutations and Gene Dependencies. *Commun. Biol.* 4 (1), 9. doi:10.1038/s42003-020-01552-6

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171 (6), 1437–1452. e1417doi:10.1016/j.cell.2017.10.049

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 47 (D1), D941–D947. doi:10.1093/nar/gky1015

Torre, E. A., Arai, E., Bayatpour, S., Jiang, C. L., Beck, L. E., Emert, B. L., et al. (2021). Genetic Screening for Single-Cell Variability Modulators Driving Therapy Resistance. *Nat. Genet.* 53 (1), 76–85. doi:10.1038/s41588-020-00749-z

Vasan, N., Baselga, J., and Hyman, D. M. (2019). A View on Drug Resistance in Cancer. *Nature* 575 (7782), 299–309. doi:10.1038/s41586-019-1730-1

Xu, Y., Dong, Q., Li, F., Xu, Y., Hu, C., Wang, J., et al. (2019). Identifying Subpathway Signatures for Individualized Anticancer Drug Response by Integrating Multi-Omics Data. *J. Transl. Med.* 17 (1), 255. doi:10.1186/s12967-019-2010-4

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* 41 (Database issue), D955–D961. doi:10.1093/nar/gks1111

Yu, L., Zhou, D., Gao, L., and Zha, Y. (2021). Prediction of Drug Response in Multilayer Networks Based on Fusion of Multiomics Data. *Methods* 192, 85–92. doi:10.1016/j.ymeth.2020.08.006

Zhang, J., Zhu, W., Wang, Q., Gu, J., Huang, L. F., and Sun, X. (2019). Differential Regulatory Network-Based Quantification and Prioritization of Key Genes Underlying Cancer Drug Resistance Based on Time-Course RNA-Seq Data. *Plos Comput. Biol.* 15 (11). e1007435 doi:10.1371/journal.pcbi.1007435

Zhao, Y., Li, Z. X., Zhu, Y. J., Fu, J., Zhao, X. F., Zhang, Y. N., et al. (2021). Single-Cell Transcriptome Analysis Uncovers Intratumoral Heterogeneity and Underlying Mechanisms for Drug Resistance in Hepatobiliary Tumor Organoids. *Adv. Sci.* 8 (11). 2003897 doi:10.1002/advs.202003897

Zheng, H.-C. (2017). The Molecular Mechanisms of Chemoresistance in Cancers. *Oncotarget* 8 (35), 59950–59964. doi:10.18632/oncotarget.19048

Zheng, X., Carstens, J. L., Kim, J., Scheible, M., Kaye, J., Sugimoto, H., et al. (2015). Epithelial-to-mesenchymal Transition Is Dispensable for Metastasis but Induces Chemoresistance in Pancreatic Cancer. *Nature* 527 (7579), 525–530. doi:10.1038/nature16064

Zhu, Y., Brettin, T., Evrard, Y. A., Partin, A., Xia, F., Shukla, M., et al. (2020). Ensemble Transfer Learning for the Prediction of Anti-cancer Drug Response. *Sci. Rep.* 10 (1), 18040. doi:10.1038/s41598-020-74921-0

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/fgene.2021.733783

Parenclitic and Synolytic Networks Revisited

Tatiana Nazarenko[1]*, *Harry J. Whitwell*[2,3,4,5], *Oleg Blyuss*[1,4,6,7] and *Alexey Zaikin*[1,4,5]

[1]Department of Mathematics and Institute for Women's Health, University College London, London, United Kingdom, [2]National Phenome Centre and Imperial Clinical Phenotyping Centre, Department of Metabolism, Digestion and Reproduction, Imperial College London, Hammersmith Campus, London, United Kingdom, [3]Section of Bioanalytical Chemistry, Division of Systems Medicine, Department of Metabolism, Digestion, Imperial College London, South Kensington Campus, London, United Kingdom, [4]Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia, [5]World-Class Research Center "Digital Biodesign and Personalized Healthcare", Sechenov First Moscow State Medical University, Moscow, Russia, [6]School of Physics, Astronomy and Mathematics, University of Hertfordshire, Harfield, United Kingdom, [7]Department of Pediatrics and Pediatric Infectious Diseases, Institute of Child's Health, Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Yuehua Cui,
Michigan State University, United States
Shaoyu Li,
University of North Carolina at Charlotte, United States

***Correspondence:**
Tatiana Nazarenko
t.nazarenko@ucl.ac.uk

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

Received: *30 June 2021*
Accepted: *28 September 2021*
Published: *20 October 2021*

Citation:
Nazarenko T, Whitwell HJ, Blyuss O and Zaikin A (2021) Parenclitic and Synolytic Networks Revisited. Front. Genet. 12:733783. doi: 10.3389/fgene.2021.733783

Parenclitic networks provide a powerful and relatively new way to coerce multidimensional data into a graph form, enabling the application of graph theory to evaluate features. Different algorithms have been published for constructing parenclitic networks, leading to the question—which algorithm should be chosen? Initially, it was suggested to calculate the weight of an edge between two nodes of the network as a deviation from a linear regression, calculated for a dependence of one of these features on the other. This method works well, but not when features do not have a linear relationship. To overcome this, it was suggested to calculate edge weights as the distance from the area of most probable values by using a kernel density estimation. In these two approaches only one class (typically controls or healthy population) is used to construct a model. To take account of a second class, we have introduced synolytic networks, using a boundary between two classes on the feature-feature plane to estimate the weight of the edge between these features. Common to all these approaches is that topological indices can be used to evaluate the structure represented by the graphs. To compare these network approaches alongside more traditional machine-learning algorithms, we performed a substantial analysis using both synthetic data with *a priori* known structure and publicly available datasets used for the benchmarking of ML-algorithms. Such a comparison has shown that the main advantage of parenclitic and synolytic networks is their resistance to over-fitting (occurring when the number of features is greater than the number of subjects) compared to other ML approaches. Secondly, the capability to visualise data in a structured form, even when this structure is not *a priori* available allows for visual inspection and the application of well-established graph theory to their interpretation/application, eliminating the "black-box" nature of other ML approaches.

Keywords: networks, graphs, parenclitic, synolytic, complexity

1 INTRODUCTION

In the era of increasing large and complex (multi-modal) datasets (biological, climatic, medical, etc.), network approaches are becoming very popular. Indeed, representation of complex data in the form of a network, i.e. a graph with nodes and edges, is a powerful tool to visualise data structure, clusters and communities, and all other interdependencies. Graph theory, well established by mathematicians, provides many topological indices to describe possible features of a network. This is especially valuable for complex biological systems, when often some non-specific change can be compensated by changes in other regions of a connected network. By evaluating topological features, the transition between two states such as health or disease be detected. A clear difficulty in this analysis is how to represent the data in the form of a network if links between nodes-features are unknown? Several approaches have been recently suggested and applied to different cases of data analysis.

One approach is correlation graphs, where edge weights are proportional either to the correlation coefficient between the corresponding vectors of features [for a discussion, see Gorban et al. (2021)] or to the correlation between nodes, if each node has some internal structure, e.g. in the case of intra-gene methylation profiles (e.g., see Bartlett and Zaikin, 2016; Bartlett et al., 2014). Recently, a new network approach has gained popularity, first described by Zanin and Boccaletti (2011) and called a *parenclitic* network representation, from the Greek term for "deviation". The main idea of this approach is to establish links between parameters (nodes) without any *a-priori* knowledge of their interactions (Zanin and Boccaletti, 2011) by using residual distances from linear regression models constructed between every pair of parameters as edge-weights. Networks constructed from this linear regression parenclitic approach (LRPA) have been successfully applied to different biological problems. For example, the detection of disease-related genes and metabolites (Zanin and Boccaletti (2011); Zanin et al., 2012; Zanin et al., 2013a; Zanin et al., 2013b; Zanin et al., 2016), brain research (Papo et al., 2014), and to identify signatures of cancer development from human DNA methylation data (Karsakov et al., 2017).

However, for many biological data structures, there is no linear dependence between features, and thus defining a graph in such as way makes interpretation impossible. To overcome this, alternative approaches have been developed. First, it was suggested to use 2-dimensional kernel density estimation (2DKDE) to model the control distribution (KDE Parenclitic approach, KDEPA). This methodology was successfully applied to the problem of diagnosing patients with Ovarian Cancer. (Whitwell et al. 2018).

The advantage of KDEPA over LRPA is that pairs of features do not necessarily have to be correlated, or even grouped into a single cloud. At the same time, KDEPA also has some drawbacks: it is difficult to correctly extend the density distribution beyond what is defined by the underlying data (unlike linear regression which can be extrapolated simply) and, similarly to LRPA, the selection of a threshold (common for all edges) or thresholds (different for each edge) when converting to a binary network for class separation.

As a further development, in Krivonosov et al. (2020) we have introduced a variation of parenclitic networks, that can be called *synolytic* from the Greek word for "ensemble". In some sense, synolytic networks is a graph representation of the simultaneous action of multiple classifier ensembles. We demonstrated previously that any machine learning methods [e.g. support vector machine (SVM)] can be used as the core of the parenclitic approach (a function that describes the separation of controls and cases groups in the plane of two features). We proposed a software implementation with a choice of any kernel and demonstrated its ability to detect the DNA methylation signature of Down Syndrome disease. Moreover, we showed that the characteristics of the constructed networks help to interpret the obtained signatures in relation to aging in individuals from non-Downs Syndrome and Down Syndrome populations. A further development came from not binary networks, but weighted networks, and this method was successfully applied to prediction of survival for severely ill Covid-19 patients (Demichev et al., 2021), and for prediction of prostate cancer progression in patients on active surveillance (Sushentsev et al., 2021). We used SVM as the core, and the probability of belonging to a group of cases as the weights of the edges.

The weighted synolytic network approach (wSA) automatically solves the inherent drawback of KDEPA by normalizing the distance measure in terms of probabilities. Herein, we show that the synolytic approach is comparable, and sometimes better than other machine learning (ML) models. One advantage is in the visualization of results, which allows one to visually identify key features (see examples on **Figure 2A**, producing greater transparency to "black-box" ML algorithms. They offer the opportunity for applying more sophisticated network analysis concepts to study the resulting networks and allow the analysis of networks over time, leading to future ideas of parenclitic-longitudinal data analysis.

In this paper, we compare weighted (w) parenclitic (wLRPA, wKDEPA) and synolytic (wSA) parenclitic models with each other and with other ML methods for solving binary classification problems.

To compare the approaches, we used

1.1 Synthetic Data

Models of N-dimensional spheres of radius 1, where points of the inner sphere of radius 0.5 are denoted by Controls (that is, with class 0), and points with a radius from 0.5 to 1 are denoted by Cases (that is, with 1 class). This model structure was chosen to generate synthetic data to fairly compare different machine learning algorithms within a defined and understandable dataset. To this end, the data is easily visualized in dimensions 2 (a regular circle on a plane) and 3 (a sphere in three-dimensional space), and further they can be easily expanded to any data dimensions in accordance with an understandable principle. In such a model, the radius is a characteristic, implicitly sewn into the full vector of each sample, and which is a measure of the distance of a point from the centre of the sphere, and it is *a*

priori known how far each point is from the spatial multidimensional boundary of the class division.

In addition to "ideal spheres", we also consider "noisy spheres"—spheres with the addition of 50 random variables and "broken spheres"—N-dimensional spheres from which N/2 parameters were replaced with random variables (to study how the models work on data that do not contain the full set of parameters responsible for the difference between the two classes). For each approach separately, we discuss how to choose the best characteristics (providing better quality separation of classes) and then check how these conclusions are reproduced on real data sets.

Studying the characteristics of networks on these data, we initially evaluate how the characteristics of these networks correlate with their radii (that is, how well the network approach reads this implicit characteristic). The higher the correlation, the greater the class separation will be. Confirmation that the characteristics of networks are correlated with radii is an important validation of the correctness of the transformation of raw data to the networks.

We compare parenclitic approaches with other ML models by comparing the results of applying ML models to matrices of raw (initial) data and matrices of strengths (degrees) of vertices derived from the graph-structured determined by the parenclitic model. This showed that models using parenclitic vertex strengths are superior to models on raw data in situations where the sample size is significantly lower than the dimension of the data.

1.2 Real Data

Realizing that the model of synthetic data we have chosen has a very simple structure and the results on it may not be reproduced on real data, we collected a collection of 16 datasets of different dimensions of features (which were randomly selected from the repository of existing datasets), in which the size of the minimum class was >40 samples. We found that wLRPA and wKDEPA approaches did not perform as well in real datasets with a large dimension (in comparison to sample size), whereas the wSA synolytic approach did.

The results obtained in this work create a reliable basis for the application of synolytic approaches to real data. We especially emphasize here the use of such approaches to clinical omics data, where, as a rule, the sample size is typically small in relation to the number of features, which nevertheless can contain a rich source of diagnostic information. We show the advantage of using synolytic approaches to solve classification problems, but we note that the advantage also lies in the fact that this approach allows the visualisation of each patient in the form of a network and opens up additional possibilities for the study of such states using graph theory and network analysis.

2 MATERIALS AND METHODS

2.1 Generation of Different Types of Synthetic Data

Synethetic data was generated using a sphere-model, For all modelling, we considered all possible combinations of *Sphere Dimensions*: (2, 3, 10, 30, 60, 90, 120, 150); number of *Case TRAIN samples*: (15, 65, 115, 165, 215, 265) and number of *Controls TRAIN samples*: (15, 65, 115, 165, 215, 265). Numbers of *Case TEST samples* and *Controls TEST samples* were calculated as 25% of corresponding *TRAIN* numbers.

2.1.1 Ideal Spheres Model

Common model: area bounded by a N-dimension sphere of radius 1 (i.e. each i sample is represented by vector $X_i^1, X_i^2, \ldots, X_i^N$, where $R = \sqrt{X_i^1{}^2 + X_i^2{}^2 + \cdots + X_i^N{}^2} \leq 1$). We define *Controls* as points with radius $0.01 \leq R \leq 0.5$ and *Cases* as points with radius $0.5 \leq R \leq 1$.

2.1.2 Noisy Spheres Model

Each N-dimension "Ideal Sphere" was expanded to 50 "noise" variables (i.e. each sample is represented by vector $X_i^1, X_i^2, \ldots, X_i^N, V_i^1, V_i^2, \ldots, V_i^{50}$, where V^j - vector of random values from uniform distribution in (−1, 1))

2.1.3 Broken Spheres Model

Each N-dimension "Ideal Sphere" was "broken": we only kept half of the variables from there and the other half was changed to random values (i.e. each sample is represented by vector $X_i^1, X_i^2, \ldots, X_i^{\frac{N}{2}}, V_i^1, V_i^2, \ldots, V_i^{\frac{N}{2}}$, where V^j—vector of random values from uniform distribution in (−1, 1))

All generated data are publicly available and described in the **Supplementary Materials**.

2.2 Real Data List

Real datasets were obtained from https://archive.ics.uci.edu and are presented in **Table1**.

The same pre-processing was performed for all datasets:

- All missing values were replaced with the mean of the feature column;
- The features, with standard deviation equal to 0 have been removed;
- If the data was a non-binary classification problem, then the response vector was transformed into a binary one (by highlighting one of the classes): Cortex: "Ts65Dnc"—cases, other controls; Ionosphere: "g"—controls, other cases; QSAR: "RB"—cases, other controls; SONAR: 'R' - cases, other controls; URBAN: "building"—cases, other controls; Vertebral-2c: "AB"—cases, other controls);
- If the data contained a preliminary division into TRAIN and TEST subsets (SPECT, SPECTF and URBAN), then they were collected into a single dataset and TRAIN/TEST labels were disregarded
- For each dataset, we repeatedly (20 times) produced random subsets of 80 samples with equal numbers of Cases and Controls, and then Test and Train labels were assigned equally in each class (that is, each subset consisted of 20 TRAIN Case samples, 20 TRAIN Controls samples, 20 TEST Case samples and 20 TEST Controls samples). Thus, a total of 320 (16 original datasets * 20 subsets) datasets of different feature dimensions, but the same sample size, were obtained.

TABLE 1 | Real datasets description.

N	Dataset	Number of				Area
		Features	Samples	Cases	Controls	
1	Banknote Authentication (2013)	4	1,372	610	762	Computer
2	Blood Transfusion Service Center (2008)	4	748	178	570	Business
3	Vertebral Column (2011)	6	310	210	100	Medicine
4	Breast Cancer Wisconsin (Diagnostic) (1995)	10	699	241	458	Medicine
5	Indian Liver Patient Dataset (ILPD) (2012)	10	583	167	416	Medicine
6	Planning Relax (2012)	12	182	52	130	Computer
7	Climate Model Simulation Crashes (2013)	18	540	494	46	Physical
8	Diabetic Retinopathy Debrecen (2014)	19	1,151	611	540	Medicine
9	SPECTF Heart (2001)	22	267	212	55	Medicine
10	Ionosphere (1989)	33	351	126	225	Physical
11	QSAR Biodegradation (2013)	41	1,055	356	699	Chemical
12	SPECTF Heart (2001)	44	267	212	55	Medicine
13	Connectionist Bench (Sonar, Mines vs. Rocks) (1988)	60	208	97	111	Physical
14	Mice Protein Expression (2015)	77	1,080	510	570	Medicine
15	Urban Land Cover (2014)	147	675	122	553	Physical
16	Arrhythmia Data Set (1998)	260	452	245	207	Medicine

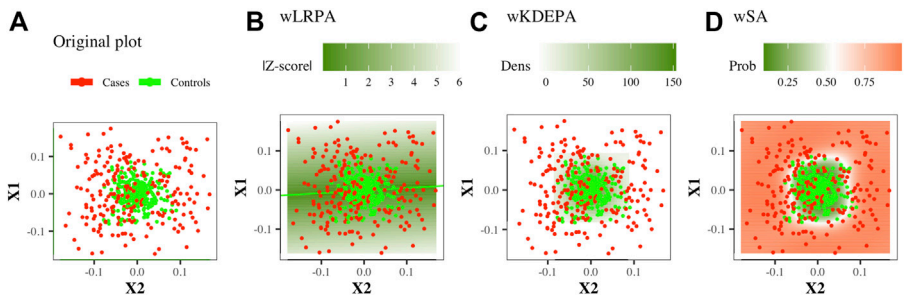


FIGURE 1 | Illustration of the construction of parenclitic models on a pair of features (X^1, X^2) from the *Ideal ball model* with 165/265 Case/Control TRAIN points and 50/80 Case/Control TEST points. Case/controls are red/green points respectively.

All collection of real data and selected subsets of them are publicly available and described in the **Supplementary Materials**.

2.3 Parenclitic Approaches

In this analysis we have used three different Parenclitic networks architecture:

2.3.1 wLRPA

This is a network where only the control group was considered as the basis for determining the normal state on the plane of two features: based on the control group, linear regression were built for every pair of features. The deviation of the control points from it was calculated and the distribution of such deviations was constructed (**Figure 1B**). In previous studies, for each new sample, the edge weight was first determined as the absolute value of z-score, and then binarized (if $|Z\text{–}score| < 3$, then the edge is present in the sample network, otherwise there is no edge). In this work, we will consider weighted networks (that is, the specified binarization will not be carried out, an edge for any sample will always exist and the edge weight will always be equal to $|Z\text{–}score|$).

2.3.2 wKDEPA

This is a network in which again only the control group was considered as the basis for determining the normal state on the plane of two features. For the control group, the 2-dimension kernel density estimation was built for each pair of features (**Figure 1C**), then a function was calculated that converts the density values into an analogue distance, so that the points located in the area of the highest density have the minimum weight. The distance outside the grid was continued (for more details, see Whitwell et al., 2018). For non-weighted networks, for each new sample, the edge weight was first determined as the normalised volume of the density distribution above the point, and then converted to binary form (if the volume is greater than a threshold (which was iteratively selected, so that the characteristics of the resulting networks optimally separate the Case and Control groups), then the edge is present in the sample network, otherwise there is no edge). In this work, we will consider weighted networks (that is, the specified binarization will not be carried out, an edge for any sample will always exist and the edge weight will always be equal to a function of density).

2.3.3 wSA

This is a network in which both groups participate in definition of normal and abnormal states. On the plane of any two features, a radial SVM is used to define the best boundary separating the classes (**Figure 1D**). Automatically, each point in such a model gets a value for the probability of belonging to each class. For each new sample, the edge weight is determined as the probability of belonging to a group of cases.

2.3.4 Networks Characteristics

All characteristics were calculated with using *igraph package* (R).

The values for some characteristics were equal to *NA*, + *Inf*, – *Inf*, in these instances, the values were replaced by 0. Such substitutions could theoretically lead to the loss of differences between classes for some characteristics, although they did not affect the analysis associated with the Strengths of the vertices (these values are always are finite, since the Strength of the vertex is equal to the sum of the weights of the edges included in it).

For each dataset (model spheres and sets of real data), we built model-networks (i.e. built LM, KDE or SVM models in a plane of every pair of features) on the TRAIN folds. Networks were then constructed for each individual (TRAIN and TEST) sample in the dataset and for each of them we calculated:

- Descriptive statistics (*zeros, min, max, mean, standard deviation (sd), coefficient of variation (coefvar) = sd/mean*) of the main network characteristics closeness, betweenness, edge betweenness, page rank, eigen centrality, authority score, strength, edge weights;
- The full vector of strengths (degrees) of the vertices.

We use descriptive statistics of the main network characteristics to demonstrate their correlation with radii on synthetic data (and, as a consequence, the quality of class division into them). We use matrices of full vectors of strengths of vertices for samples in each dataset to compare the results of ML models on them and on the initial raw data.

2.4 ML Models for Comparison with Parenclitic Models

Parenclitic approaches were compared with 3 ML models (*xgbTree*, *nnet*, *glmnet*) from the Caret package in R). We chose these since the principles of their training are based on different static principles and they all produce a selection of features. We trained models using the *train* function within the Caret package, using scaling and centering for data pre-processing, with the selection of hyperparameters set at default and using 5-fold cross-validation.

All networks and their characacteristics from real data and selected subsets of them are publicly available and described in the **Supplementary Materials**.

2.5 Performance Estimation

The performance of ML models was assessed using area under the receiver operator characteristic curve (AUC).

For each dataset (synthetic or real, matrices of raw data or matrices of vertex degrees), we built 3 ML models on TRAIN folds and applied these models to TEST folds. For each result on TEST folds we calculated AUC with "direction" (i.e. *controls* < *cases* or *controls* > *cases*) received on AUC for TRAIN folds. Taking into account the direction along the TRAIN folds results in TEST-fold AUC values <0.5 in some instances.

To calculate the performance of class separation on each characteristic on synthetic data (**Figure 3A**), we use a simple *glm* model on each characteristic separately (obtaining AUC for TEST as described above).

All results (on synthetic or real, on matrices of raw data or matrices of vertex degrees, on separate characteristics) are publicly available and described in the Supplementary Materials.

To highlight the significance of the difference in the results obtained by ML models on raw data and on the degrees of vertices, a two-sided paired Wilcoxon signed-rank test was applyed to AUC values to calculate a *p*-value.

3 RESULTS

3.1 Comparison of Approaches on Model Datasets

3.1.1 Parenclitic

Since the data models were generated in such a way that the characteristic that distinguishes the classes (the radius of the spherical model) is always known, we first investigated how topological characteristics of the networks correlate with these values.

For each point in the sphere-models, the distance from the point to the centre of the *N*-dimensional data (radius) is calculated ("ideal sphere") (see **Section 2.1** for further details). To mimic non-perfect data, "noisy spheres" are also generated, in which 50 random features are added to each sample and "broken spheres" in which *N*/2 parameters are replaced with random variables. The radius for each point in "noisy spheres" and "broken spheres" are not recalculated, and thus the radius value (calculated for each point in the "ideal spheres" data) becomes a less accurate representation of the points position in the data structure.

Data sets were generated varying the number of dimensions, cases and controls, such as for each sphere-model, 288 different datasets are generated (see **Section 2.1**). In each dataset and for each network-characteristic, we calculated the absolute value of Pearson's correlation coefficient between it and the radius of the samples (**Figure 3A**). We specify here following conclusions:

- For wLRPA, the topological characteristic that has the greatest degree of correlation is the maximum weights of the edges. This is despite the fact that in each plane, the control distribution is poorly described by linear regression (**Figure 1B**). When considering the maximum all of edge weights, there is typically always a pair of features for a "case" that is a long way from the regression fit, whereas "controls" are always close to the line. Therefore, considering the most extreme point for every case/

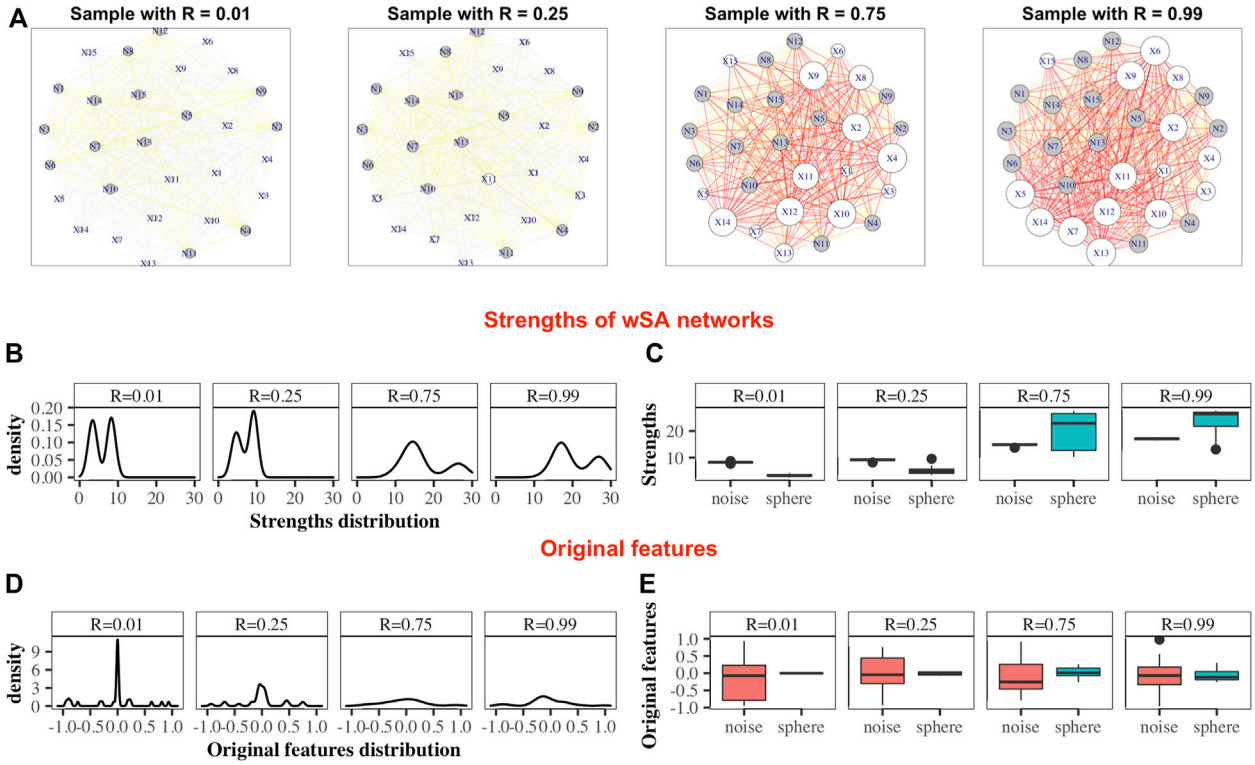


FIGURE 2 | Demonstration of the advantages of the wSA approach: visualization of parameters in the form of networks. Examples wSA networks for samples with R = (0.01, 0.25, 0.75, 0.99) from a *broken sphere model* with 30-dimensions. There were 115/165 Case/Control TRAIN points and 34/50 Case/Control TEST points. **(A)** Examples of networks for two Controls (samples with R = 0.01 and R = 0.25) and two Cases (samples with R = 0.75 and R = 0.99). The sizes of the vertices are equal to their strengths; the colours of the vertices indicate the type of feature in the original data (white is the colour of the parameter of the original sphere, grey is the noise variable); the thickness of the edges corresponds to the edge weights [that is, the probability of a sample belonging to a class of cases on the plane of two features (vertices)]; the colours of the edges additionally emphasize their weight w (grey: if $0 < w \leq 0.25$, yellow: if $0.25 < w \leq 0.5$, orange: if $0.5 < w \leq 0.75$ and red: if $0.75 < w \leq 1$. As can be seen from these examples, the networks of Controls and Cases visually represent different topological objects (the strength of the vertices of the Cases is noticeably greater than the strength of the vertices of the Controls). Moreover, in such networks, the vertices "defining" the difference between the two classes are immediately distinguished (that is, the vertices of the sphere parameters (labeled X) are predominantly larger than the vertices of noise variables (labeled N) for Cases and less for Controls). **(B)** Distributions of nodes strengths of samples networks, demonstrating the similarity of distributions inside the Control group and inside the Case group and the difference in distributions between groups. **(C)** Boxplots of the nodes strengths in the groups of sphere variables and noise variables, demonstrating that the strengths of the most significant vertices differ significantly from the noise ones within the network of each sample (that is, the greater the weight of the vertex in the Cases and the smaller it in the Controls, the greater the role of this parameter for the whole system). Panels **(D)** and **(E)** show the distribution of the features in the dataset. As can be seen, the distributions of Cases and Controls on the original data do not show clear difference, and the boxplots demonstrate that the noise and sphere variables, on average, do not differ for each sample. Thus, we demonstrate that the transition from original data to networks simultaneously solves two problems: it defines the space in which the Cases and Controls are topologically distinguishable and highlights the group of parameters that are most important and significant for the entire system.

control, rather than an average is a good correlator in these synthetic data sets.

- For wKDEPA, a fairly large number of characteristics show a good correlation, in particular, the mean of the edge-weights and the mean of the strengths of the vertices. An interesting finding is that the correlation for the models of "noisy spheres" is very inferior to the other two models (comparing the highest correlation between each topological feature), which most likely indicates overfitting of the wKDEPA on noise variables.
- wSA networks demonstrate the best results out of all three networks. Characteristics such as mean of edge weights, mean of vertex strengths, and mean of closeness show very high correlation across all datasets and any model of Spheres (see examples of such networks on **Figure 2A**) and their strengths distribution on **Figures 2B,C**). From our point of view, this indicates that the construction of wSA is more advantageous and the established rule for the weight of edges (through the probability of belonging to a class of cases) is reasonable as a measure of the distance from the *center of normality*. It is interesting that for some topological characteristics, such as PageRank, the correlation for the ideal sphere is worse. The reason for this may be that the high-quality prediction generated by the SVMs for each feature minimise the variation in edge weights, and this reduces the capacity of some topologies to change (such as PageRank).

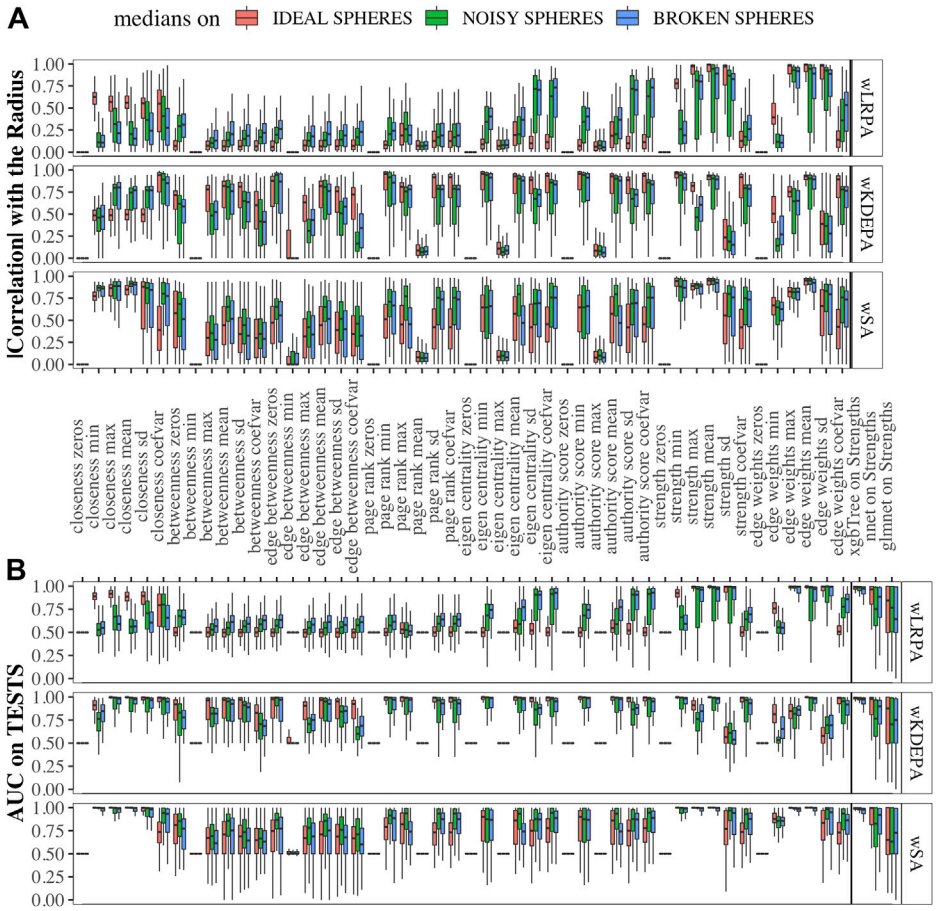


FIGURE 3 | Correlation of network characteristics with sample radii and performance (AUC) of class separation for each network characteristic **(A)** Absolute values of Pearson's correlation coefficient of 48 descriptive statistics of network characteristics with synthetic-samples' radii (only TEST subsets were considered). **(B)** AUCs from glm models calculted from network characteristics for TEST folds, and on ML models using strength characteristics (to the right of the verticle line). Models were calcualted for "ideal spheres" (red), "noisy spheres" (green) and "broken spheres" (blue).

As expected, for those characteristics with a high correlation with the radii, a high quality of class separation was obtained. For each dataset, for each topological characteristic, we trained a glm model on the TRAIN folds and applied it to the TEST folds. With the obtained probability, we calculated the AUCs (area under the receiver operator characteristic curve) and combined them into boxplots for each sphere model (**Figure 3B**). In addition, we added the results of constructing ML models on the strengths of vertices (since each sample within the network can be represented not by the vector of original features, but by the vector of their strengths in the networks (see examples of wSA networks in **Figure 2A** and their strengths distribution in **Figures 2B,C**) which demonstrate an advantage over the distributions of raw features) and found that the results of *xgbTree* for wLRPA and wSA networks are comparable, with the best results on individual characteristics. For wKDEPA, they greatly exceed the results of individual characteristics (and become comparable with the results of wLRPA and wSA). Despite the fact that wKDEPA and wLRPA approaches work a little worse than the wSA, the transformation of the original features to the vertex strengths is an equal substitution, regardless of the choice of the parenclitic approach.

3.1.2 Comparison with Other ML Models on Synthetic Data

We compared the quality of parenclitic approaches (for simplicity, we compared only the results of xgbTree on the vertex strengths, since, as can be see from **Figure 3B**, the nnet and glmnet models worked worse) with 3 ML models on the syntethic datasets (**Figure 4**). The ML model that produced the most accurate classification of this data was xgbTree (**Figure 4A**). The results of all the parenclitic approaches produced exceptionally good classification, however, wKDEPA and wSA had extremely positively skewed distributions, meaning that these more frequently gave better classification than other approaches. When considering the impact of sample size, it can be seen that parenclitic models outperform glmnet, nnet and xgbTree when the sample size is small relatively to the spheres dimension (**Figure 4B**), which most likely indicates that parenclitic approaches are less prone to overfitting. This property itself

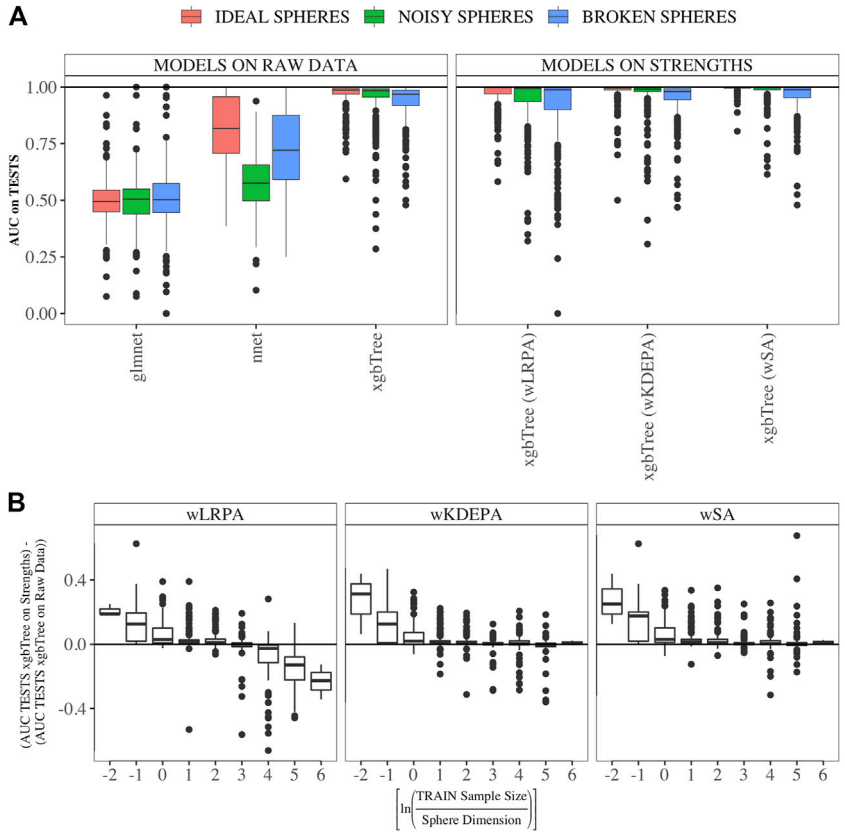


FIGURE 4 | Comparison of the results of parenclitic approaches (the xgbTree model trained on the strengths of vertices) with ML methods (glmnet, nnet, xgbTree) on synthetic data **(A)** Parenclitic analysis demonstrated greater performance (for all three network approaches) than glmnet and nnet on the original data. wSA and wKDEPA approaches demonstrate a slight improvement over xgbTree on the original data. **(B)** The difference of AUC-xgbTree-on-vertex-strengths and AUC-(xgbTree/nnet/glmnet)-on-raw data versus $[ln(\frac{TRAIN\ Sample\ Size}{Dimension})]$, where [ ] denotes the standard rounding function. Parenclitic approaches on average demonstrate superiority to other ML methods in situations where sample size is small raltive to the number of features.

(apart from other advantages of parenclitic approaches) can be valuable in biological and medical problems utilising omics data, where there can be a huge number of features with comparatively few patients. This is exemplified in Demichev et al. (2021), where the use of wSA was more effective than other ML methods, most likely because the sample size was small relative to the large number of features. It was also seen that for wLRPA, the quality of discrimination (compared to other ML techniques) decreased once the number of samples highly exceeded the number of features, whereas there was no such effect for wKDEPA and wSA.

3.2 Comparison of Approaches with Real Data

3.2.1 Parenclitic

We generated models for 16 real data sets and calculated the median AUCs for each of the networks characteristics (**Figure 5A**). For real data, wSA networks performed much better than wLRPA and wKDEPA. Moreover, it is interesting that for wSA networks, the performance of each topological characteristics' AUC mirrored the AUCs from the synthetic data with closeness mean, strengths mean and weights mean performing strongly in all instances. This most likely indicates that the characteristics of the wSA networks have some conservatism (in terms of the quality of separation), regardless of the data type. We would also like to note the repetition of the effect found on the spheres models. For all three parenclitic approaches, the models built with the vertex strengths give the best performance. Most likely, the fact that the medians of AUC for wKDEPA and wLRPA are low for real data, but the quality of models based on the node strengths is high (although lower than for wSA), indicating that, despite the described shortcomings, such networks are correctly distinguishing classes, but the effects on characteristics are not conservative (that is, the quality of the separation for each characteristic depends on the data type).

3.2.2 Comparison with Others ML Models

Comparison of ML models built on the strengths of vertices and on raw data was carried out for each network approach separately.

Inside each real dataset, for each of the 20 subsets we calculated the quality of each model (AUC on TEST fold) on the raw data and its quality on the vertex strengths. For each main

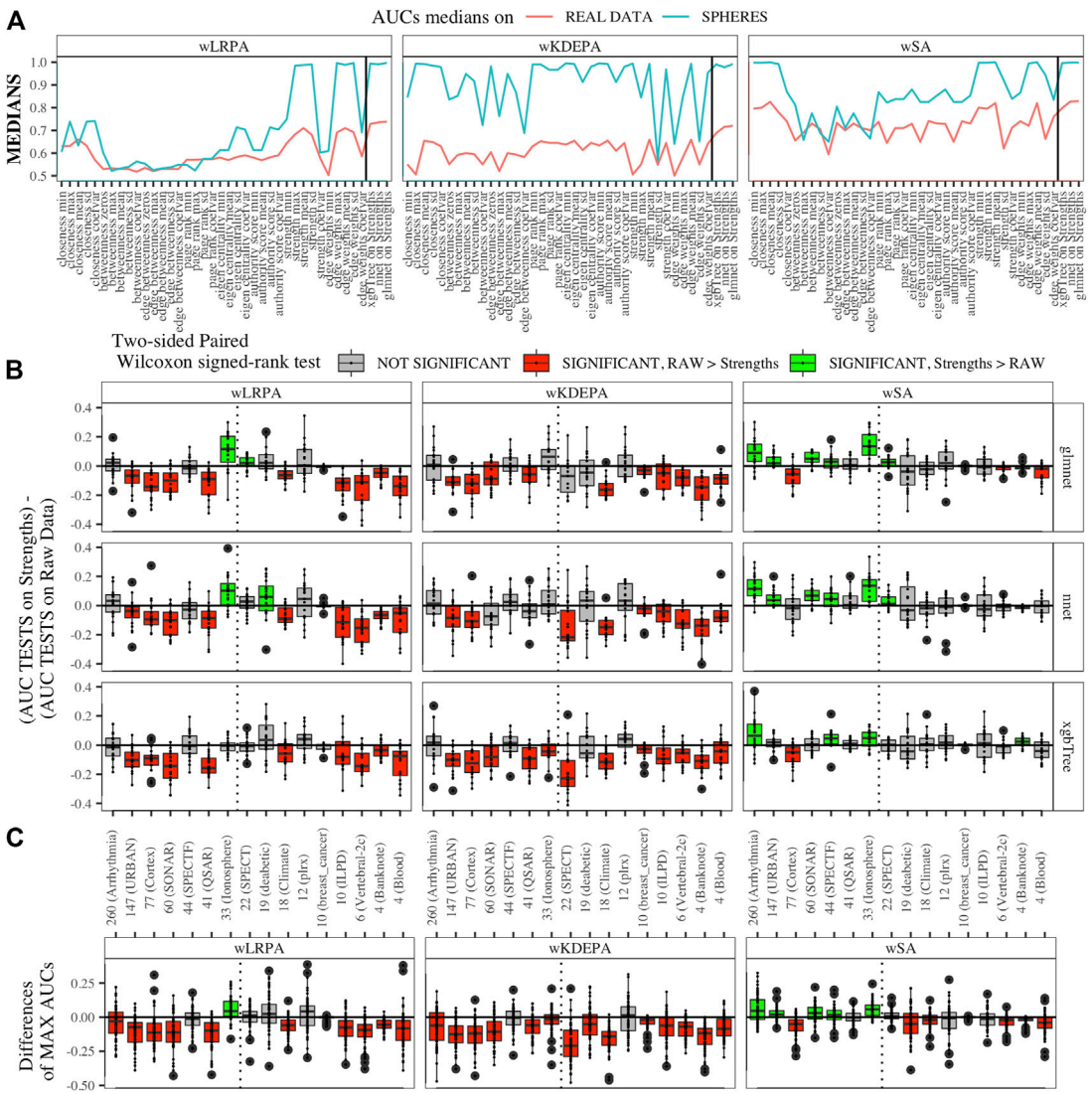


FIGURE 5 | Applying parenclitic approaches to real data **(A)** Medians of AUCs obtained by parenclitic approaches on synthetic data and on real data. On **(B)** and **(C)** we order the datasets by the feature dimension (we display the dimensions on the *x*-axis along with the names of the datasets), we draw a dashed line detaching the datasets to the left that satisfy our expectations to get an advantage of parenclic approaches to them (that is, for which the $[ln(\frac{40}{Dimension})] \leq 0$, is correct, as 40 is the TRAIN fold size for all subsets). **(B)** The difference between the performances of ML models built on vertex strengths and ML models built on raw data. The effect found on the synthetic data was not confirmed for the wLRPA and wKDEPA approaches, but it was mainly obtained for the wSA: for 6/7 datasets the performances are either comparable or give a gain in the wSA. **(C)** The difference between MAX AUC among 3 models on raw data and among 3 models on strengths data. For wSA approach, on 5/7 datasets to the left of the dotted line, the best ML model on vertex strengths shows a significant advantage than the best ML model on raw data; on 1/7 datasets the results were comparable, and on 1/7 datasets the result was worse.

dataset, we get two vectors of length 20 with AUC on TEST for vertices and AUC on TEST for raw data.

First, we calculated the difference between these two vectors for each main dataset and presented them as boxplots in **Figure 5B**. We also computed a two-sided paired Wilcoxon signed-rank test for each pair of such vectors. If p-value ≥ 0.05 (i.e. insignificant), we use gray color for the corresponding box; if p-value < 0.05 (i.e. significant), we additionally check which median results (on raw or on strengths data) more, and use red color if more is raw results, and use green color if more is strengths results.

Datasets are sorted in descending order of number its of features. As it was established on synthetic data (**Figure 4B**) that parenclitic approaches have the greatest advantage on data where the logarithm of the ratio of the sample size in the TRAIN fold to the feature size is no more than 0 (rounded to the nearest real number). Considering that all subsets have sample size of TRAIN folds is 40 samples, among them, datasets with a dimension greater than 33 have this property (indicated by vertical line in **Figure 5B**).

As shown in **Figure 5B**, the wLRPA gave only one advantageous result (green box) for the glmnet and nnet models on datasets with a ≥33 dimensions; moreover, in 4 out

of 7 of these cases this approach turned out to be worse than the model based on non-parenclitic data. A similar situation was seen for the wKDEPA approach (where not a single winning situation was found on any dataset). On the other hand, for the wSA approach, the advantage of using the wSA approach is clearly seen in sample sets with a large proportion of dimensions compared with other observations (the only exception is the Cortex dataset, in all other cases the wSA works better or comparable than the models on non-parenclitic data).

Additionally, for each individual subset, we calculated the best result (maximum AUC) for 3 models on non-parenclitic data and for 3 models on strengths data and examined the difference in such values (**Figure 5C**). For the wSA approach, on 5 out of 7 datasets with dimensions ≥33, the best ML model on vertex strengths (that is, using the parenclitic approach) shows a significant advantage than the best ML model on the raw data; on 1 out of 7 datasets (QSAR) the results were comparable, and on 1 out of 7 datasets (Cortex) the result was worse.

4 DISCUSSION

The results presented in this work show that the quality of the wSA is comparable to or better than other ML models, if we consider them as classifiers, and better than other parenclitic approaches. At the same time, this approach has several advantages. The transformation of the initial data features into individual networks for each sample facilitates the visualization of the relationship between features, identifying the most significant relationships and the most significant features (e.g. **Figure 2A**). This approach allows one to build generalizing networks and get an idea of the whole system of interdependencies between features. New rules can be developed to simplify networks (for example, by removing all edges that are not informative in terms of class separation), or highlight hubs, triangles, and use other advanced network analysis procedures. Moreover, as we were able to show with synthetic data and then confirm on real data, the quality of the wSA as a classifier is higher for those datasets where the sample size is small in comparison with the features size. Using this advantage we have recently applied this approach to the analysis of proteomics data from a large cohort of CoVID-19 patients, in which this is the case (Demichev et al., 2021) (manuscript submitted). In this analysis, we showed that wSA was able to produce accurate classifications, where other ML algorithms were not on the same data.

The disadvantages of parenclitic approaches include high computation times (since the construction of models occurs at each pair of features), and certain data structures for which this method will not work. For example, a simple "chess" three-dimensional cube (please rotate the example here)—where the points of cases and controls are grouped similarly to black and white squares on a chessboard. At the same time, despite the fact that the spatial separation of classes obviously exists, in all of the three projections (any two of the three parameters), parenclitic approaches will be not able to detect a qualitative separation (since the points of cases and controls will be mixed on the two-dimensional plane). Despite the fact that wLMPA and wKDEPA approaches did not work much better for classification problems, they are particularly useful in situations where the case group is not known, as they only use the control group for building models, and therefore highlight groups that deviate greatly from controls. This in contrast to wSA models which require pre-defined case/control groups to construct parenclitic networks.

The development of such approaches for application on longitudinal data is of particular interest. As we have demonstrated with the sphere models, some of the characteristics of the parenclitic networks are highly correlated with radii (which in these models is a measure of the deviation from normality). This may mean that the characteristics of parenclitic networks can themselves be the indicators of the development of the disease and can be traced over time to diagnose the onset of the disease. It has been established that the use of longitudinal models (i.e. models that use all historical data for a subject to predict a future or current state) reduces the time to diagnosis for ovarian cancer (Blyuss et al., 2018; Whitwell et al., 2020). Topologies of parenclitic networks (and combinations of topologies) can naturally be incorporated into longitudinal algorithms. Given the power of these approaches individually, the development of their combined use is now a research priority.

To summarise our approach as an instruction for multi-disciplinary researchers:

- For specialists in the field of medicine and biology, using the wSA approach
  - ∗ As a classifier, in situations where the number of samples is small in comparison with the dimension of analytes [when there are few patients, but there are many measurements of their states, see, for example, (Demichev et al., 2021) (manuscript submitted)];
  - ∗ As a high-quality and simple data visualization, when a visual representation of the state of the system features of an individual in the aggregate is required (we assume that such a representation in the form of networks can give a new understanding of the relationship of features, both among the entire set of subjects, and with an indication of some of their individual properties subjects, as shown in **Figure 2A**);
  - ∗ In situations where it is required to determine the intermediate state of the points during the transition, for example, from a healthy to severely ill state. As we have shown (through the radii on the artificial data, see **Figure 3A**), parenclitic approaches reflect the spatial state on a one-dimensional scale;
  - ∗ When it is required to interpret the transition between two states with respect to some kind of continuous effect (for example, in the work Krivonosov et al. (2020) we showed how the third groups of samples according to the characteristics of networks demonstrate age tendencies between the features selected in binary networks between case and control groups);
- For specialists in the field of machine learning and network approaches, we recommend using the wSA approach, as an interesting method to get new representations of the data. In particular,
  - ∗ One can play with a choice of a model to split classes for each pair of features (currently we used everywhere a radial SVM, but we assume that each edge may have its own model, the main thing is that the edge weight is set as the probability of belonging

to the same class); or with different classes of models, used on the vertices strengths (or other vectors of characteristics) of the networks.

* It is possible to have artificial data for which the synolytic approach in this form is not applicable (for example, a "chess" three-dimensional cube described above). We believe that the Cortex data, on which the wSA approach has not received an advantage, were of similar type. However, it would be interesting to extend this approach to consider not only pairs, but also triplets and quadruples of features with the correct collection of the results into an edge between two features (to continue to obtain a structure on the graph).

Finally, our approach, if combined with artificial neural networks, may contribute to the development of explainable artificial intelligence, because network visualisation assists the understanding in each step of data processing.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

TN and AZ designed the study, TN conducted the graph and data analysis. All authors have written and reviewed the manuscript.

FUNDING

TN, HW, and AZ are supported by a Medical Research Council grant (MR/R02524X/1). HW is supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). OB, HW, and AZ are supported by EPSRC and Cancer Research UK award (EDDCPJT\100022). OB, HW, and AZ acknowledges support by the grant of the Ministry of Education and Science of the Russian Federation Agreement No. 074-02-2018-330(1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.733783/full#supplementary-material

REFERENCES

Arrhythmia Data Set (1998). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Arrhythmia (Accessed April 19, 2021) [Online].

Banknote Authentication (2013). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/banknote+authentication (Accessed April 19, 2021) [Online].

Bartlett, T. E., Olhede, S. C., and Zaikin, A. (2014). A Dna Methylation Network Interaction Measure, and Detection of Network Oncomarkers. *PloS one* 9, e84573. doi:10.1371/journal.pone.0084573

Bartlett, T. E., and Zaikin, A. (2016). Detection of Epigenomic Network Community Oncomarkers. *Ann. Appl. Stat.* 10, 1373–1396. doi:10.1214/16-AOAS939

Blood Transfusion Service Center (2008). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center (Accessed April 19, 2021) [Online].

Blyuss, O., Burnell, M., Ryan, A., Gentry-Maharaj, A., Mariño, I. P., Kalsi, J., et al. (2018). Comparison of Longitudinal Ca125 Algorithms as a First-Line Screen for Ovarian Cancer in the General Population. *Clin. Cancer Res.* 24, 4726–4733. doi:10.1158/1078-0432.CCR-18-0208

Breast Cancer Wisconsin (Diagnostic) (1995). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\%28Diagnostic\%29 (Accessed April 19, 2021) [Online].

Climate Model Simulation Crashes (2013). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes (Accessed April 19, 2021) [Online].

Connectionist Bench (Sonar, Mines vs. Rocks) (1988). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29 (Accessed April 19, 2021) [Online].

Demichev, V., Tober-Lau, P., Nazarenko, T., Aulakh, S. K., Whitwell, H., Lemke, O., et al. (2021). A Proteomic Survival Predictor for COVID-19 Patients in Intensive Care. *medRxiv*. doi:10.1101/2021.06.24.21259374

Diabetic Retinopathy Debrecen (2014). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set (Accessed April 19, 2021) [Online].

Gorban, A. N., Tyukina, T. A., Pokidysheva, L. I., and Smirnova, E. V. (2021). Dynamic and Thermodynamic Models of Adaptation. *Phys. Life Rev.* 37, 17–64. doi:10.1016/j.plrev.2021.03.001

Indian Liver Patient Dataset (ILPD) (2012). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/ILPD+\%28Indian+Liver+Patient+Dataset\%29 (Accessed April 19, 2021) [Online].

Ionosphere (1989). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Ionosphere (Accessed April 19, 2021) [Online].

Karsakov, A., Bartlett, T., Ryblov, A., Meyerov, I., Ivanchenko, M., and Zaikin, A. (2017). Parenclitic Network Analysis of Methylation Data for Cancer Identification. *PloS one* 12, e0169661. doi:10.1371/journal.pone.0169661

Krivonosov, M., Nazarenko, T., Bacalini, M. G., Franceschi, C., Zaikin, A., and Ivanchenko, M. (2020). Dna Methylation Changes with Age as a Complex System: a Parenclitic Network Approach to a Family-Based Cohort of Patients with Down Syndrome. *bioRxiv*. doi:10.1101/2020.03.10.986505

Mice Protein Expression (2015). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression (Accessed April 19, 2021) [Online].

Papo, D., Buldú, J. M., Boccaletti, S., and Bullmore, E. T. (2014). Network Theory in Neuroscience. *Philos. Trans. Biol. Sci.* 369, 1–21. doi:10.1007/978-1-4614-7320-6_713-1

Planning Relax (2012). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Planning+Relax (Accessed April 19, 2021) [Online].

QSAR Biodegradation (2013). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/dataMSsets/QSAR+biodegradation (Accessed April 19, 2021) [Online].

SPECTF Heart (2001). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart (Accessed April 19, 2021) [Online].

Sushentsev, N., Rundo, L., Blyuss, O., Nazarenko, T., Suvorov, A., Gnanapragasam, V. J., et al. (2021). Comparative Performance of Mri-Derived Precise Scores and delta-radiomics Models for the Prediction of Prostate Cancer Progression in

Patients on Active Surveillance. *Eur. Radiol.* (accepted). doi:10.1007/s00330-021-08151-x

Urban Land Cover (2014). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Urban+Land+Cover (Accessed April 19, 2021) [Online].

Vertebral Column (2011). UCI Machine Learning Repository. [Dataset]. Available at: https://archive.ics.uci.edu/ml/datasets/Vertebral+Column (Accessed April 19, 2021) [Online].

Whitwell, H. J., Blyuss, O., Menon, U., Timms, J. F., and Zaikin, A. (2018). Parenclitic Networks for Predicting Ovarian Cancer. *Oncotarget* 9, 22717–22726. doi:10.18632/oncotarget.25216

Whitwell, H. J., Worthington, J., Blyuss, O., Gentry-Maharaj, A., Ryan, A., Gunu, R., et al. (2020). Improved Early Detection of Ovarian Cancer Using Longitudinal Multimarker Models. *Br. J. Cancer* 122, 847–856. doi:10.1038/s41416-019-0718-9

Zanin, M., Menasalvas, E., Boccaletti, S., and Sousa, P. (2013a). Feature Selection in the Reconstruction of Complex Network Representations of Spectral Data. *PloS one* 8, e72045. doi:10.1371/journal.pone.0072045

Zanin, M., and Boccaletti, S. (2011). Complex Networks Analysis of Obstructive Nephropathy Data. *Chaos* 21, 033103. doi:10.1063/1.3608126

Zanin, M., Menasalvas, E., Sousa, P. A. C., and Boccaletti, S. (2012). Preprocessing and Analyzing Genetic Data with Complex Networks: An Application to Obstructive Nephropathy. *Nhm* 7, 473–481. doi:10.3934/nhm.2012.7.473

Zanin, M., Papo, D., Solís, J., Espinosa, J., Frausto-Reyes, C., Anda, P., et al. (2013b). Knowledge Discovery in Spectral Data by Means of Complex Networks. *Metabolites* 3, 155–167. doi:10.3390/metabo3010155

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., et al. (2016). Combining Complex Networks and Data Mining: Why and How. *Phys. Rep.* 635, 1–44. doi:10.1016/j.physrep.2016.04.005

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

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

frontiers in Genetics

REVIEW
published: 21 October 2021
doi: 10.3389/fgene.2021.760501

Network Biology Approaches to Achieve Precision Medicine in Inflammatory Bowel Disease

John P Thomas[1,2,3], *Dezso Modos*[1,2], *Tamas Korcsmaros*[1,2]* *and Johanne Brooks-Warburton*[4,5]

[1]Earlham Institute, Norwich, United Kingdom, [2]Quadram Institute Bioscience, Norwich, United Kingdom, [3]Department of Gastroenterology, Norfolk and Norwich University Hospital, Norwich, United Kingdom, [4]Department of Gastroenterology, Lister Hospital, Stevenage, United Kingdom, [5]Department of Clinical, Pharmaceutical and Biological Sciences, University of Hertfordshire, Hatfield, United Kingdom

Inflammatory bowel disease (IBD) is a chronic immune-mediated condition arising due to complex interactions between multiple genetic and environmental factors. Despite recent advances, the pathogenesis of the condition is not fully understood and patients still experience suboptimal clinical outcomes. Over the past few years, investigators are increasingly capturing multi-omics data from patient cohorts to better characterise the disease. However, reaching clinically translatable endpoints from these complex multi-omics datasets is an arduous task. Network biology, a branch of systems biology that utilises mathematical graph theory to represent, integrate and analyse biological data through networks, will be key to addressing this challenge. In this narrative review, we provide an overview of various types of network biology approaches that have been utilised in IBD including protein-protein interaction networks, metabolic networks, gene regulatory networks and gene co-expression networks. We also include examples of multi-layered networks that have combined various network types to gain deeper insights into IBD pathogenesis. Finally, we discuss the need to incorporate other data sources including metabolomic, histopathological, and high-quality clinical meta-data. Together with more robust network data integration and analysis frameworks, such efforts have the potential to realise the key goal of precision medicine in IBD.

Keywords: inflammatory bowel disease, network biology, protein-protein interaction network, gene coexpression network, multilayered network, precision medicine, gene regulatory network, metabolic network

OPEN ACCESS

Edited by:
Alessio Martino,
National Research Council (CNR), Italy

Reviewed by:
Francesco Bardozzo,
University of Salerno, Italy
Neda Zarayeneh,
Washington State University,
United States

***Correspondence:**
Tamas Korcsmaros
tamas.korcsmaros@earlham.ac.uk

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

Received: *18 August 2021*
Accepted: *08 October 2021*
Published: *21 October 2021*

Citation:
Thomas JP, Modos D, Korcsmaros T and Brooks-Warburton J (2021) Network Biology Approaches to Achieve Precision Medicine in Inflammatory Bowel Disease. Front. Genet. 12:760501. doi: 10.3389/fgene.2021.760501

INTRODUCTION

Inflammatory bowel disease (IBD), comprising Ulcerative Colitis (UC) and Crohn's disease (CD), is a chronic, immune-mediated inflammatory disorder which primarily involves the gastrointestinal tract (Lennard-Jones, 1989; Baumgart and Carding, 2007). It causes significant morbidity and affects almost seven million people worldwide. The prevalence is forecasted to rise steeply in the decades ahead, particularly in newly industrialised countries (GBD 2017 Inflammatory Bowel Disease Collaborators, 2020). IBD arises due to a dysregulated immune response secondary to complex interactions between multiple genetic risk factors, a "dysbiotic" gut microbiota, and environmental factors (Xavier and Podolsky, 2007; Cader and Kaser, 2013). However, the precise mechanistic pathways interlinking these various facets of IBD pathogenesis are still largely unknown (Cader and Kaser, 2013). In addition, despite recent advances in medical management including the use of

biologic and small molecule therapies, a significant proportion of patients who wish to avoid surgery fail to achieve sustained clinical remission (Cosnes et al., 2011). This highlights the need for novel, effective therapeutic strategies in IBD.

Unlike rare, and well-defined monogenic disorders (e.g., cystic fibrosis) which occur due to mutations within a single gene, complex diseases such as IBD arise due to interactions between numerous genetic variants and environmental factors. These interactions occur across several layers that transcend the ecologic, genetic, epigenetic, protein and cellular levels, and work collectively to manifest the disease phenotype. Consequently, IBD demonstrates significant heterogeneity across the population i.e., patients may have varying environmental exposures and express different genetic variants which result in the activation of varying pathogenic pathways. Hence, a one-size-fits-all approach to therapy, as is currently practised, may explain the suboptimal clinical outcomes seen in IBD.

As a result, precision medicine has been identified as a key strategy for improving clinical outcomes in IBD (Denson et al., 2019; Verstockt et al., 2021). Precision medicine aims to harness the biological characteristics of individual patients to tailor the right therapy to the right patient at the right time (Whitcomb, 2019). This would require an understanding of the function of individual biological components and also the holistic effects of their multifactorial interactions to stratify patients (Green et al., 2017; Sudhakar et al., 2021). Whilst still in its infancy, an early example of this approach in IBD is the PROFILE study. In this trial, researchers are utilising a transcriptomic signature of peripheral blood $CD8^+$ T lymphocytes as a biomarker to separate CD patients into two subgroups according to predicted disease course to guide therapeutic strategy i.e. "step up" vs "top down" therapy (Noor et al., 2020). This transcriptomic signature was found to be effective for prognostication through an earlier non-interventional study (Biasci et al., 2019). It is anticipated that multi-omics approaches may be even more robust for directing precision therapies in IBD and other complex disorders (Olivera et al., 2019; Borg-Bartolo et al., 2020). In this effort, over the past decade, researchers across the world have begun profiling the transcriptomics, epigenetics, metabolomics, and proteomics data of large patient cohorts. For IBD, a number of biorepositories have become established such as the IBD BioResource in the United Kingdom (Parkes and IBD BioResource Investigators, 2019), the 1000IBD project in the Netherlands (Spekhorst et al., 2017), and the IBD Multiomics Data project in the USA (Imhann et al., 2019). However, this exponential increase in the availability of molecular data harnessed through "omics" technologies has created one of the biggest challenges we face in biology in the 21st century i.e., what is the best way to make meaningful sense of this data to ultimately improve clinical outcomes in individual patients?

Systems biology and artificial intelligence are two complementary fields that are driving novel computational biology approaches to address this challenge. Systems biology is an interdisciplinary field that allows the systematic study of complex interactions in biological systems using a holistic approach (Ahn et al., 2006; Breitling, 2010). Artificial intelligence, on the other hand, is a domain within computer science which leverages computer systems to perform tasks that normally require human intelligence including problem-solving and decision-making (Meskó and Görög, 2020). Machine learning and deep learning, which are subdomains of artificial intelligence, offer a number of potential solutions to tackle this problem. We have previously reviewed these approaches in depth in the context of IBD (Seyed Tabib et al., 2020). In this narrative review, however, we will focus on the utility of network biology, a subfield of systems biology, to facilitate precision medicine in IBD.

Network biology is one of the fundamental tenets of systems biology, which involves using mathematical graph theory to represent, integrate, and analyse biological processes and data through networks (Pavlopoulos et al., 2011). Depending on the type of data, various biological networks can be produced, such as protein-protein interaction networks, gene regulatory networks, and metabolic networks (Vidal et al., 2011). Using network-based methods as an integration and modelling tool, important molecular interactions can be unravelled. When applied to individual patients, personalised network analysis can lead to the identification of new disease subtypes and therapeutic targets, which facilitates novel drug discovery, biomarker discovery, and drug repurposing as has been seen in cancer (Módos et al., 2017). Hence, network biology can be a valuable tool for analysing multi-omics patient data to achieve the key goal of precision medicine in IBD and other complex disorders (Korcsmaros et al., 2017).

Although in its nascent stages, in this narrative review we will highlight a variety of innovative network biology approaches that are bringing the promise of precision medicine closer to a translational reality in IBD (**Table 1**). First, however, we will briefly discuss some of the fundamental concepts underpinning network biology.

KEY PRINCIPLES OF NETWORK BIOLOGY

A biological network is the representation of a biological system using graphs. It contains biological entities (e.g., cells, proteins or genes) and their interactions with each other (e.g., protein-protein interactions). In network biology, these are called nodes and edges, respectively (Koutrouli et al., 2020). The topology of a network (i.e., the way in which nodes and edges are arranged within a network) can be evaluated to better understand a biological system (**Figure 1**). In biological networks, the topology is usually scale-free i.e., the degree distribution of nodes follows a power law, unlike random networks (Barabási and Oltvai, 2004). This means that some nodes in a biological network may have many interactions called "hubs," whilst other nodes may have fewer connections (Charitou et al., 2016). Furthermore, specific regions of a scale-free network can be more highly interconnected than other parts of the network. These highly connected regions of a network are called modules. Modules often correspond to specific biological functions within the overall system. Specific nodes

TABLE 1 | Characteristics of various network types discussed in this review and their main advantages and disadvantages.

Network type	Node	Edge	Required information to build the network	Pros	Cons
Protein-protein interaction networks	Proteins	Physical interactions	Measurement of the actual protein interactions e.g. using yeast two-hybrid, affinity purification mass spectrometry or small-scale binding experiments	Many different resources, based on physical interactions ensuring larger coverage	Highly incomplete, biases in network generating methods
Metabolic networks	Metabolites	Enzymes, reactions	Measured reactions of the enzymes	Most complete network type, good for systematic modelling	Need to decide what parameter to optimise
Gene regulatory networks	Transcription factors, promoters, enhancers, and target genes	Regulatory interaction	Measurement or modelling of the regulatory interactions e.g. using ChIP-seq, yeast one-hybrid, or through inference from transcriptomics	Various network building approaches to build large coverage and make it research question specific	Highly variable and state-specific, cannot infer feedback loops from transcriptomics only
Gene co-expression networks	Genes	Similarity between the expression of two genes	Gene expression measurement	Needs only transcriptomic data	Correlation does not always equal causation

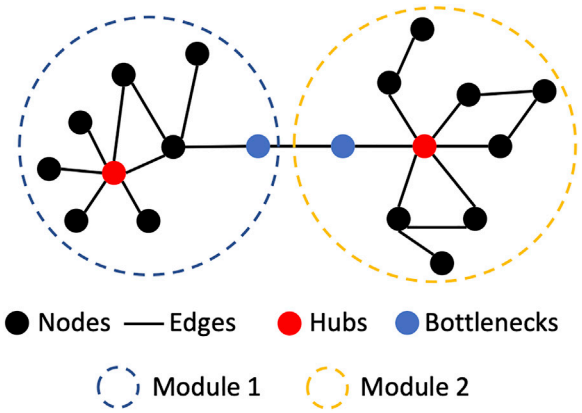


FIGURE 1 | Basic network biological nomenclature and concepts. Hubs are nodes with a high number of interactions (edges). Modules are regions of the network where the nodes interact with members of the region more than with non-members. Bottlenecks are nodes which are connecting two.

that connect distinct modules can also be identified. These are termed "bottleneck nodes" as information needs to traverse through them for one module (or biological subtask) to communicate with another (Csermely et al., 2013).

To further analyse the topology of networks many tools have been developed. One method is to identify network motifs. Network motifs are recurring, significant patterns of interconnections within a network (Milo et al., 2002). Network motifs can provide insights into the type of signalling interactions that occur within different biological networks. For instance, feedforward loops are more common in transcriptional regulatory networks (Hong et al., 2018). Another technique to find the building blocks of a network is to identify graphlets. Graphlets are small, unique (non-isomorphic) subnetworks of a network (Przulj et al., 2004). Using graphlets, the local structure of a network can be better described (Przulj, 2007). Przulj and her colleagues have used graphlets to describe various networks including protein-protein interactions (Przulj et al., 2006) and the world trade network (Sarajlić et al., 2016).

Although it may not be possible to encapsulate all dimensions and features of a complex disease using networks, network analysis can be a valuable approach for better understanding the disease. For instance, disturbance of hubs and bottlenecks in a biological network are likely to have significant consequences on the overall functioning of system. A prime example is the mechanisms driving drug resistance in HER2-amplified breast cancer, in which hub proteins within compensatory circuits and feedback loops were identified (Lee et al., 2012). This led to novel therapeutic strategies for overcoming drug resistance and improving outcomes in these patients (Kirouac et al., 2013). With the successful implementation of network biology in breast cancer and other cancers over the past decade (Yan et al., 2016), researchers are increasingly looking to gain similar translatable insights in complex diseases such as IBD.

USE OF NETWORK BIOLOGY APPROACHES IN IBD

Protein-Protein Interaction Networks

Protein-protein interaction (PPI) networks refer to networks consisting of proteins as nodes and the physical interactions between them as edges (Vidal et al., 2011) (**Table 1**). PPI data can be captured using several different methodologies including experimental approaches such as yeast two-hybrid assays and affinity purification coupled mass spectrometry, as well as computational predictive methods such as text-mining and machine learning approaches (Snider et al., 2015). Several resources containing PPI data are available for use including STRING (Szklarczyk et al., 2019), BioGRID (Chatr-Aryamontri et al., 2015), Bioplex (Huttlin et al., 2021), HAPPI-2 (Chen et al., 2009), HuRI (Luck et al., 2020), and IntAct (Hermjakob et al., 2004) (**Table 2**). PPI networks that are directed can facilitate better modelling of intra- and inter-cellular signalling. To gain

TABLE 2 | Network resources relevant to IBD research.

Name	Description	Website	Latest version (year)
STRING Szklarczyk et al. (2019)	Large PPI database with various sources and confidence scores. It contains text mining data and also other databases. It has both directed and undirected interactions	https://string-db.org/	11.5 (2021)
BioGRID Stark et al. (2006)	Genetic and protein interactions from both high and low throughput experiments	https://thebiogrid.org/	4.4.201 (2021)
BioPlex Huttlin et al. (2021)	Large affinity-purification mass spectrometry based database. It contains undirected PPI data	https://bioplex.hms.harvard.edu/	3.0 (2021)
HAPPI-2 Chen et al. (2017)	Large database collection of PPI data with confidence scores	http://discovery.informatics.uab.edu/HAPPI/	HAPPI 2.0 (2017)
IntAct Kerrien et al. (2012)	Large PPI database collection. Mostly undirected interactions	https://www.ebi.ac.uk/intact/	4.2.18 (2021)
Reactome Jassal et al. (2020)	Large reaction-centric PPI database, concentrating on signalling with well-developed toolsets. It has directed interactions	https://reactome.org/	77 (2021)
WikiPathways Martens et al. (2021)	Community curated database of signalling pathways. It has varying coverage	https://www.wikipathways.org/	September 2021 (2021)
SignaLink Fazekas et al. (2013)	Multi-layered database of signalling pathways with a manually curated core extended by regulatory data, external datasets and predictions	http://signalink.org/	3.0 (2021)
Signor Licata et al. (2020)	Manually curated signalling network	https://signor.uniroma2.it/	2.0 (2020)
CellPhoneDB Efremova et al. (2020)	Network database containing directed intercellular ligand-receptor interactions (i.e. a type of PPI network database)	https://www.cellphonedb.org/	2.1.7 (2021)
Ramilowski et al. Ramilowski et al. (2015)	Directed intercellular ligand-receptor interaction (PPI) network database developed by the FANTOM5 team	https://fantom.gsc.riken.jp/5/suppl/Ramilowski_et_al_2015/	(2015)
DoRothEA Garcia-Alonso et al. (2018)	Transcription factor (TF)-target gene (i.e. GRN) database with varying confidence levels and an easy-to-use application programming interface (API)	https://saezlab.github.io/dorothea	1.5.0 (2021)
TRRUST Han et al. (2015)	Manually curated transcription factor (TF)-target gene (i.e. GRN) database	https://www.grnpedia.org/trrust	2 (2017)
HuRI Luck et al. (2020)	References interactome of human binary protein-protein interactions captured using high throughput yeast two-hybrid assays	http://www.interactome-atlas.org/	April 2020 (2020)
ConsensusPathDB Kamburov et al. (2011)	A meta-database of binary and complex protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions, as well as biochemical pathways, originating from over 30 publicly available resources	http://cpdb.molgen.mpg.de/	Release 35 (2021)
OmniPath Türei et al. (2021)	One-stop solution of intracellular and intercellular interactions. It contains almost all the above mentioned databases and has a programmatically accessible application programming interface (API) both in R and *Python*	https://omnipathdb.org/	2.0 (2021)

information regarding the direction of PPIs, additional experimental data is often required. Several databases have performed a comprehensive manual curation of such experimental data from the literature to provide information on directed PPIs. These include SignaLink (Fazekas et al., 2013), Reactome (Jassal et al., 2020), and the community-driven WikiPathways (Kutmon et al., 2016) (**Table 2**). It is important to note that all network resources have drawbacks depending on the methods that were used to compile the data. Manually curated and text mining-based networks overrepresent certain genes which are hot topics of research - for instance, p53 is often a culprit. On the other hand, unbiased approaches like yeast two-hybrid or affinity purification overrepresent proteins that bind easily to other proteins like heat shock proteins. This can inadvertently implicate heat shock proteins as being associated with all diseases (Csermely et al., 2013). Hence, researchers need to be aware of the scope and bias of the network resources they use for their analysis.

By overlaying additional expression data from RNA sequencing or microarrays, PPI networks can be contextualised to specific pathological states or conditions (**Figure 2**). As a result, proteins are often represented by their transcripts in PPI networks. In this way, PPI networks can be used to detect novel disease-related genes, modules and signalling pathways. However, the application of transcriptomics data to build protein interactions is based on the assumption that a gene transcript accurately represents the amount of protein within the cell. This assumption is only partially true (Kosti et al., 2016).

Network propagation can also be utilised to reveal further disease-associated genes (reviewed by Cowen et al. (2017)). In short, with this approach, a set of known disease-related genes are first mapped to a PPI network and algorithms are used to detect additional proteins (or genes) that are likely to be disease-associated. Such algorithms identify additional proteins (or genes) by finding the interactor partners of the known disease-related genes using a heat propagation algorithm or a random walk approach. These methods assume that proteins (or genes) near a disease-related gene are likely to be associated with the disease as well. This is called guilt by association. Huang et al evaluated various resources that generate PPI networks to see which is the most useful for detecting disease-related genes using network propagation (Huang et al., 2018). They found that the optimal solution came from building a composite network (the parsimonious composite network or PCNet) in which interactions were supported by a minimum of two network resources.

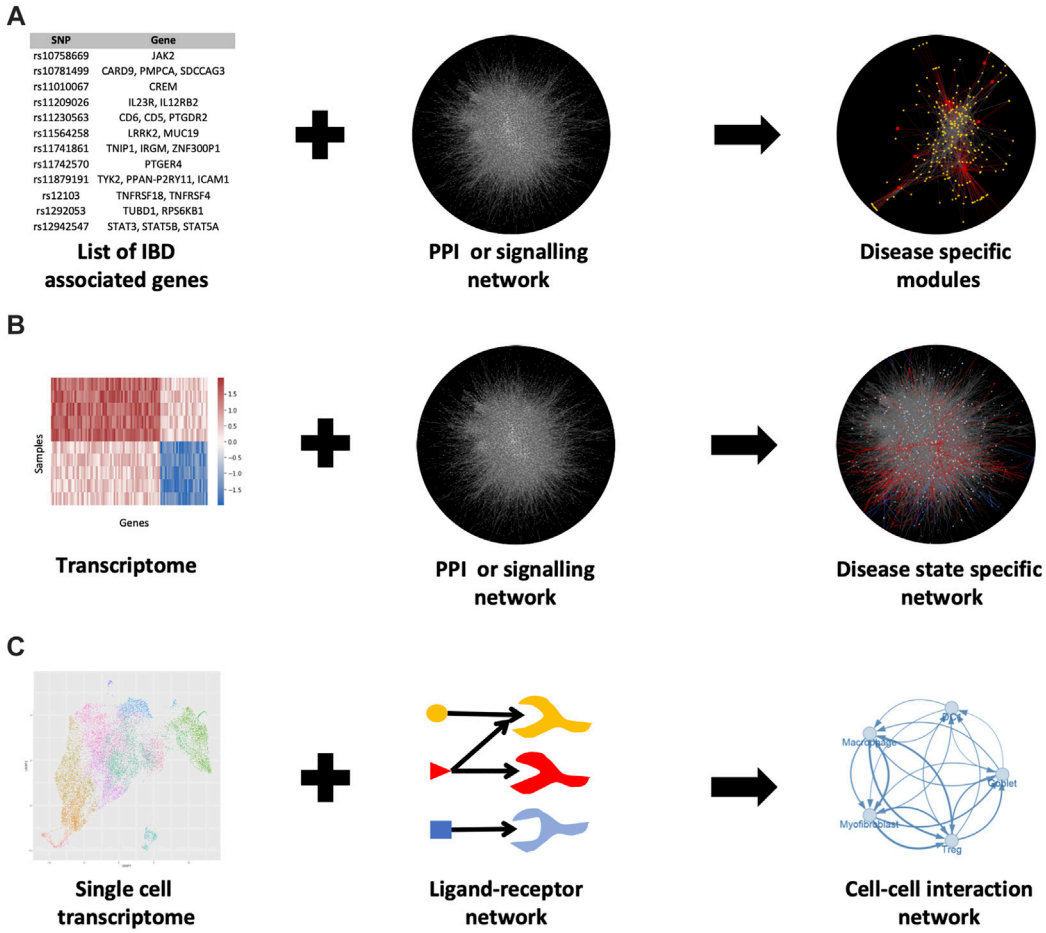


FIGURE 2 | Various methods for generating PPI networks in IBD. **(A)** Known IBD-associated genes can be mapped to a PPI network and the nearby genes in the network can be associated with IBD as well (guilt by association) **(B)** Mapping a transcriptome to the PPI network can elucidate disease-specific modules in the network **(C)** Single-cell RNA-seq data combined with intercellular (ligand-receptor) communication networks can show how various cells are interacting with each other in disease or healthy states. For b) the data from (Olsen et al., 2009) was used. For c) the uniform manifold approximation and projection (UMAP) plot (a nonlinear dimensionality reduction technique for visualising high-dimensional data) was generated using data from Lukassen et al., 2020 (Lukassen et al., 2020).

PPI network-based approaches have been frequently used in IBD research over the past decade such as the study by Eguchi et al. (2018). In this study, the authors determined differentially expressed genes (DEGs) from transcriptomic data of IBD patients and extracted a set of known IBD genes from the DisGeNet database (Piñero et al., 2017) to construct an IBD-relevant PPI network (**Figure 2B**). The authors were able to identify modules within this network by using the DPClusO algorithm (Altaf-Ul-Amin et al., 2012). These IBD gene-enriched modules were used to predict novel IBD-relevant genes and pathways.

In recent years, PPI networks have also been used to generate intercellular communication networks with single-cell RNA sequencing (scRNAseq) data (**Figure 2C**). The method for overlaying PPI networks with scRNAseq data is dependent on the research question being asked i.e., whether the researcher is interested in studying the overall possible ligand-receptor interactions or the condition-specific changes in the strength of interactions between particular cell populations (see review by Armingol et al. (2021)). In either case, databases containing ligand-receptor interactions are required such as CellPhoneDB (Efremova et al., 2020), the FANTOM5 consortium database (Ramilowski et al., 2015) or a one-stop solution OmniPath, which we co-developed recently (Türei et al., 2021) (**Table 2**). OmniPath contains both ligand-receptor interactions as well as downstream intracellular signalling connections (Türei et al., 2021).

An example of such an approach using scRNAseq data in IBD is the study by Smillie et al. (2019). They obtained scRNAseq data from healthy, non-inflamed UC, and inflamed UC colonic biopsies to create PPI networks of intercellular communication (Smillie et al., 2019). The authors first identified ligand-receptor interactions within specific cell types in their scRNAseq datasets by using the FANTOM5 consortium database (Ramilowski et al., 2015). They included only ligand and receptor genes that were significantly differentially expressed between the three conditions and that were also highly-expressed cell subset markers. Using the connections between these filtered ligands and receptors they then constructed cell-cell interaction networks. Statistical analysis of this network revealed significant cell-cell interactions in the

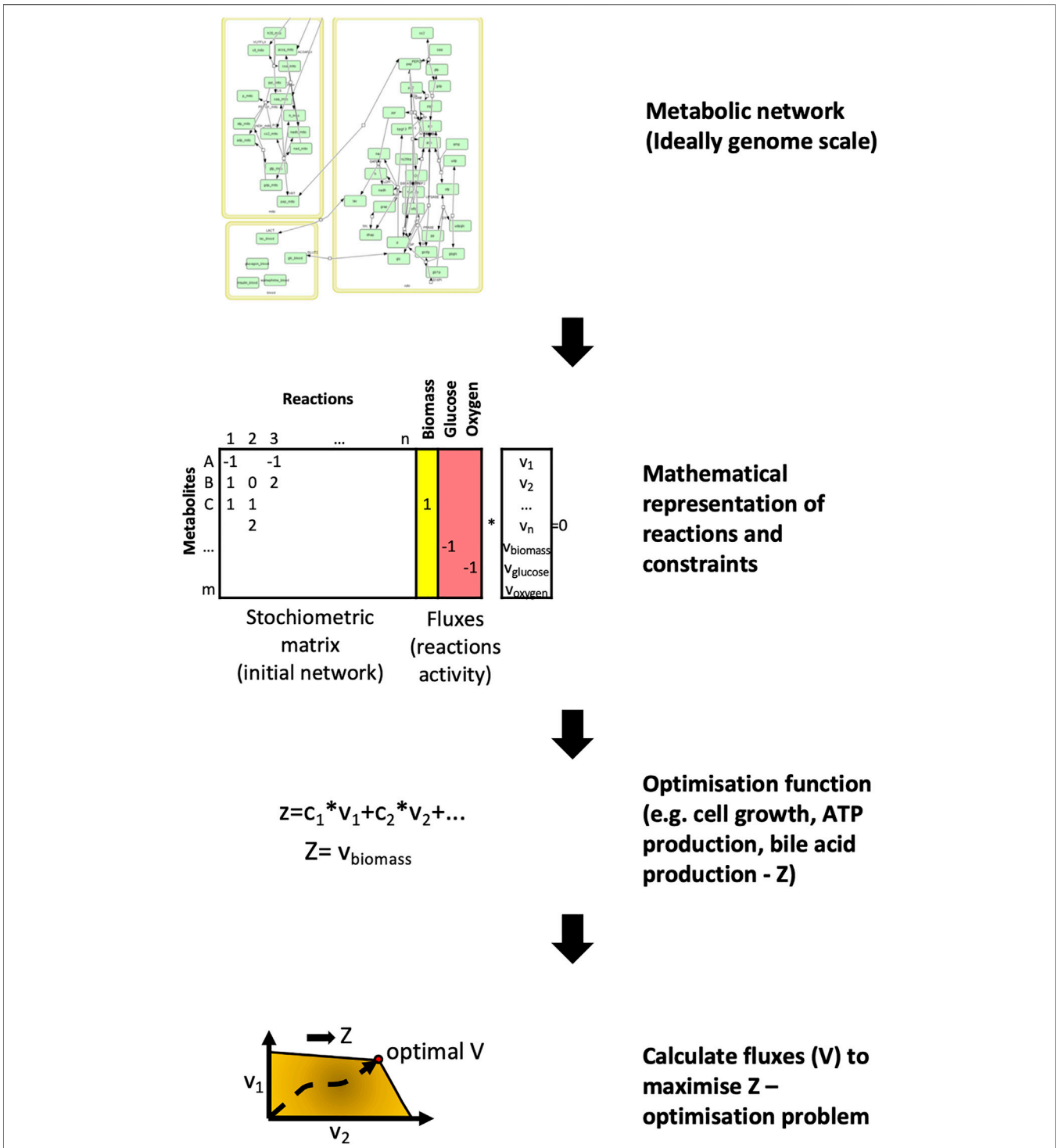


FIGURE 3 | Flux balance analysis - the basics of metabolic network modelling. For metabolic networks the initial step involves collecting the metabolic reactions that form the network. These reactions are represented by a stoichiometric matrix where each reaction is represented by the nodes and metabolites by the edges. The aim of flux balance analysis is to find the optimal vector (flux) that yields the maximum output for a given metabolite or metabolites (Z) through these reactions. For illustration, the glucose metabolism was used from Köenig et al., 2012 (König et al., 2012).

various states. In this way, the authors were able to reveal the rewiring of intercellular connections between healthy and UC states. In the healthy colonic mucosa, intercellular interactions were largely found to be occurring between cell types typically associated with colonic homeostasis such as T regulatory (Treg) cells, dendritic cell type 1 (DC1) cells, as well as $CD8^+$

intraepithelial lymphocytes (IELs) and $CD8^+$ $IL17^+$ T cells. However, in both uninflamed and inflamed states of the UC colonic mucosa, intercellular interactions were shown to be enriched between M-like cells and inflammatory fibroblasts.

To further discern changes in intercellular communication as well as subsequent downstream intracellular signalling in UC patients, we interrogated the scRNAseq data from Smillie et al using OmniPath (Türei et al., 2021). This enabled us to build an integrated network containing both intercellular PPIs and downstream intracellular PPIs in UC patients and healthy controls. This analysis revealed significant rewiring of intercellular communication between myofibroblasts and T regulatory cells (Tregs) in UC patients in comparison to healthy individuals. These changes in intercellular interactions led to major downstream signalling differences in Tregs in UC patients, in particular the TLR4 and TLR3 pathways. These pathways regulate inflammatory cytokine expression and can decrease the abundance of Treg cells (Türei et al., 2021). These findings support the hypothesis that disruption of myofibroblast-mediated regulation of Tregs may play a key role in UC pathogenesis (Pinchuk et al., 2011).

Metabolic Networks

In metabolic networks, nodes represent metabolites whilst edges refer to enzymes that catalyse metabolic reactions between the substrate and product metabolites (Vidal et al., 2011). The most common way of analysing a metabolic network is using flux-balance analysis, which involves calculating the flow of metabolites through the network in steady state (Orth et al., 2010; Anand et al., 2020). (**Figure 3**). The aim of the analysis is to find the best potential flux through the various reactions to maximise the output of a given reaction. These reactions are usually represented by cell mass or energy (ATP production). This results in an optimizable linear equation system giving back metabolic fluxes. The constraints of the model can be modified by gene expression or other experimental results. In recent years the metabolic networks of entire organisms have become available. To model the human host, the Recon2 resource provides a comprehensive global reconstruction of human metabolism (Thiele et al., 2013). For the gut microbiome, the semi-automated AGORA approach makes it possible to reconstruct the metabolism of gut microbial communities from metagenomic data (Magnúsdóttir et al., 2017). These genome-scale metabolic networks make it possible to evaluate the metabolism of the human host and gut bacterial species in the context of IBD, and discover important host-microbiome interactions (Jansma and El Aidy, 2021).

Out of all the network types reviewed here, metabolic networks have the highest completeness in terms of interactions. This makes them ideal for modelling. However, metabolomic studies are far less numerous in comparison to transcriptomics studies as RNA sequencing technologies are now far more high-throughput. In addition, a disadvantage of the standard flux balance analysis is that it needs to be optimised towards a selected metabolic reaction. When investigating IBD, the usual optimisation functions like cell growth are not relevant, so other appropriate targets need to be selected e.g., bile acid production. An alternative solution to avoid this problem is by using the metabolic network as a template and analysing it topologically (Knecht et al., 2016).

In a recent study, Heinken et al used the COBRA (Heirendt et al., 2019) genome-scale metabolic modelling software to evaluate the metabolic potential of the gut microbiome in IBD patients (Heinken et al., 2021). They found that IBD patients with dysbiosis had reduced metabolic diversity with diminished sulphur production, owing to the reduced diversity in microbial strains. In a separate study, Heinken et al also utilised flux balance analysis and genome-scale metabolic modelling to evaluate the differences in bile acid metabolism between IBD patients and healthy controls (Heinken et al., 2019). Here the optimisation function of the flux balance analysis was bile acid biotransformation. They found that one microbial species alone could not generate the whole spectrum of secondary bile acids present in the gut, but microbial pairs could generate most of these bile acids *in silico*. The network modelling also revealed that the dysbiotic microbiome of paediatric IBD patients was depleted of secondary bile acids compared to healthy children, as observed in previous studies (Duboc et al., 2013). The analysis also identified strain-specific bottlenecks that limited primary bile acid (PBA) biotransformation to secondary bile acids (SBA). Disruption of these strains may have important consequences on the inflammatory milieu in IBD, as PBAs and SBAs have been found to exert immune modulatory effects on the gut mucosa through their actions on T regulatory cells and Th17 cells (Hang et al., 2019; Sinha et al., 2020; Song et al., 2020).

An alternative approach of utilising metabolic networks to explore host metabolism in IBD was demonstrated by Knecht et al. They constructed metabolic networks by selecting enzymes which were differentially expressed between healthy controls and paediatric IBD patients from gene expression data (Knecht et al., 2016). They found that metabolic network coherence was high and varied significantly between individuals in the IBD patient cohort in comparison to healthy controls. This could have important implications for drug response in IBD patients, as metabolic networks can play a significant role in determining drug metabolism and response to treatment. Further work is needed to identify whether metabolic networks could act as a novel biomarker for determining drug response in IBD.

Gene Regulatory Networks and Gene Co-expression Networks

A gene regulatory network (GRN) depicts the molecules that govern expression levels of genes as messenger RNA (mRNA) and proteins (Vidal et al., 2011) (**Table 1**). Nodes can represent transcription factor proteins, genes, *cis*- and *trans*- DNA regulatory elements, or microRNA (miRNA). Edges represent physical interactions between these molecular entities and are directed i.e. information is provided regarding whether a molecule inhibits or activates another molecule (Schlitt and Brazma, 2007). GRNs can be mapped using yeast one-hybrid (Y1H), chromatin immunoprecipitation (ChIP) approaches, ChIP-sequencing, and DNA affinity purification (Yeh et al., 2019).

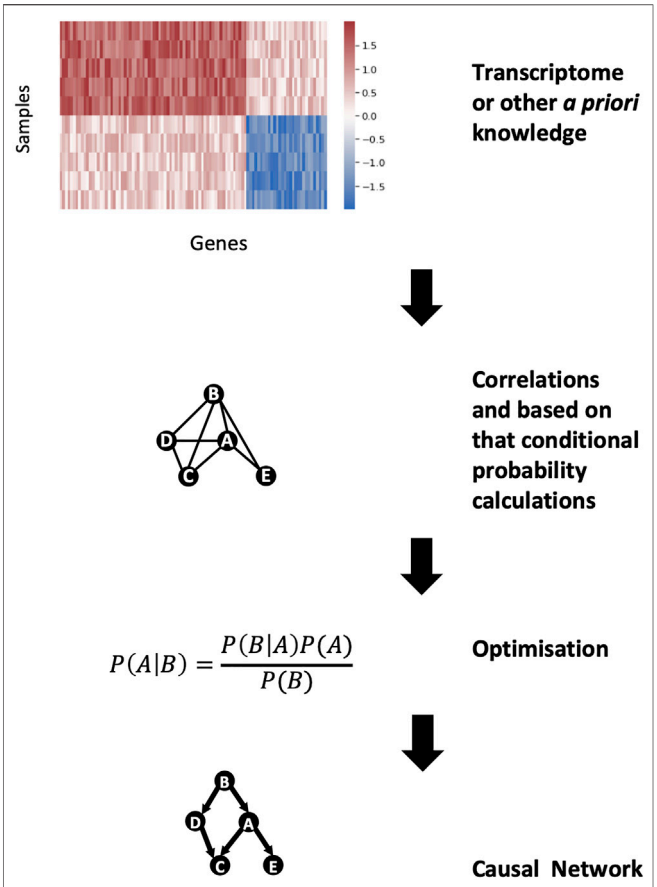


FIGURE 4 | Bayesian network construction for gene regulatory networks. From the high dimension of gene expression data, the correlations between genes can be calculated. These correlations can be modelled as conditional probabilities and, using the Bayes theorem, a casual gene regulatory network can be constructed.

GRNs can be modelled with so-called Bayesian-based network inference approaches to predict the hierarchy and the directionality of the interactions in the network. Bayesian networks are founded on the Bayes theorem, which states that the probability of event A given the occurrence of another event B i.e., P (A|B), is equal to the product of the probability of event B given the occurrence of event A i.e., P(B|A) and the probability of event A i.e., P(A), divided by the probability of event B i.e., P(B) (Bayes, 1763) (**Figure 4**). We can predict the likelihood of event A given the occurrence of event B i.e., P (A|B), if we know how often events A and B occur and how often event B occurs given the prior occurrence of event A. The Bayes theorem can be expanded to be used with transcriptomics data, because the expression of certain genes is dependent on other genes (Friedman et al., 2000). Hence, by applying the Bayes theorem to transcriptomic data it is possible to develop a network to predict which genes are influencing the expression of other genes. As an output, a Bayesian network approach produces a hierarchical graph which reveals the most plausible causal interactions occurring between genes. However, there are two limitations with this approach. Firstly, the Bayesian graph has to be acyclic i.e., it must lack biological feedback loops. Secondly, finding the optimal Bayesian network is a computationally hard optimisation problem as a Bayesian approach results in many equally or similarly good solutions. To tackle the first issue, the research question must be properly defined i.e., research questions involving feedback loops in the biological process cannot be studied using Bayesian network approaches. The second problem can be addressed by reducing the optimisation problem to a limited search space by using predefined biologically meaningful interactions (e.g., interactions from experimentally validated sources).

In gene co-expression networks (GCNs), nodes represent genes and edges connect pairs of genes that are considered co-expressed based on a certain measure (Vidal et al., 2011). Unlike GRNs, edges are undirected and simply indicate a correlation in the expression of two genes, from which causality is inferred. GCNs have become a particularly popular method in recent years as they can be constructed directly from data obtained through high-throughput gene expression experiments such as microarrays or RNA-sequencing (van Dam et al., 2018). The gene co-expression can be measured using a variety of techniques (we encourage the reader to read the comprehensive review by Sonawane et al for a summary of these algorithms (Sonawane et al., 2019)). Of these, the most commonly used algorithm is the Weighted Gene Co-expression Analysis (WGCNA) (Langfelder and Horvath, 2008) (**Figure 5**). In essence, the WGCNA algorithm calculates the correlation between the genes. This correlation is raised on a user-defined power to filter out weak interactions resulting in a scale-free network. The adjacency matrix of this network is used for clustering to find modules which represent co-regulated biological functions. GCNs and GRNs are often used together as they complement each other. The biggest advantage of these networks is that only gene expression data is required and this can be specific for the disease in question. Furthermore, the models can be refined by adding biological constraints such as known regulatory interactions like transcription factor-target gene interactions. However, their largest drawback is the *a priori* assumption that genes which are regulated and expressed together have similar functions. This notion is not always true (Sevilla et al., 2005). GRNs are also based on the assumption that correlation implies causation.

An example of using GCNs and GRNs in IBD is the landmark study by Jostins et al. In this paper, the authors performed a meta-analysis of 15 genome-wide association studies (GWAS) of CD and/or UC, to identify 73 novel and a total of 163 IBD-associated genomic loci (Jostins et al., 2012). The authors undertook network biology analysis of this data to understand how IBD-associated loci may influence pathogenesis. They performed WGCNA of gene expression data obtained from a variety of tissues including stomach, liver, adipose tissue, and blood, and identified 211 co-expression modules. These were then screened against the IBD-associated genomic loci. They identified that IBD-associated loci were particularly enriched in a module consisting of 523 genes from omental adipose tissue obtained from morbidly obese patients (i.e. the "IBD-enriched module").

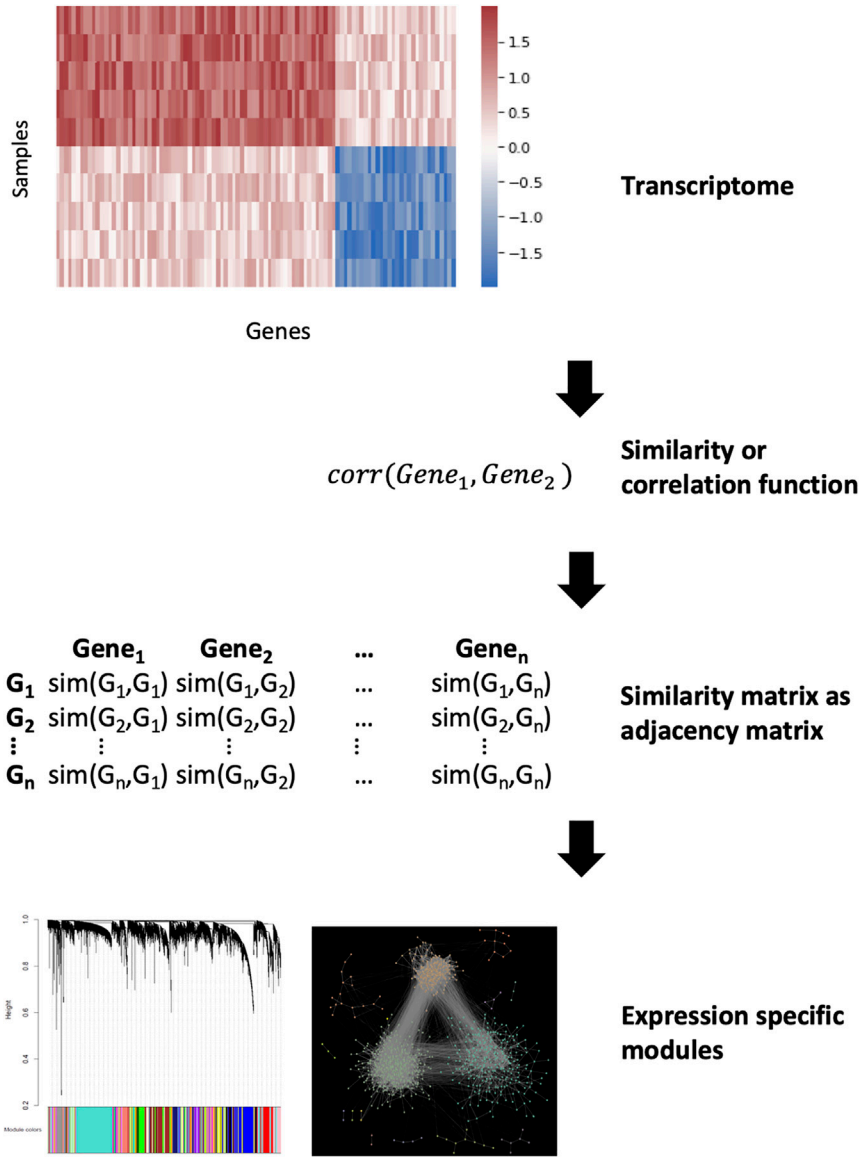


FIGURE 5 | Gene co-expression network analysis. Calculating a similarity between the genes from expression data can be used as an adjacency matrix in a co-expression network. The similarity function depends on the used method but after that the most similar parts of the network can be denoted as modules.

Jostins *et al* also used a Bayesian network inference method to create a GRN for IBD. To do this, they combined both genotype and gene expression data to infer a direction in terms of causality for the effect of single nucleotide polymorphisms (SNPs) on the identified gene expression. The overlap between this network and the genes in the IBD-enriched module revealed a sub-network of genes that were highly expressed in bone marrow-derived macrophages. Thus, by using gene regulatory and gene co-expression networks, the authors were able to annotate IBD-associated GWAS loci to a particular immune cell network and infer causality.

Peters et al employed network biology approaches in three independent cohorts of IBD patients, representing distinct stages of the disease (treatment naïve paediatric patients, patients refractory to biologic therapy, and patients with advanced disease undergoing bowel resection), to identify key driver genes that regulated IBD networks (Peters et al., 2017). The authors integrated data about known IBD-associated SNPs, and expression quantitative trait loci (eQTL) and *cis*-regulatory element (CRE) data from the aforementioned IBD cohorts, to identify candidate causal IBD genes in specific immune cell types. These candidate genes from all immune cell types were then intersected with modules found within GCNs obtained from the three IBD patient cohorts. The authors then identified modules in these networks that were significantly enriched for genes within the macrophage-enriched immune network from Jostins et al. (2012). This enabled them to generate "super-immune" modules by taking the union of these

modules from each cohort. By evaluating common genes of these super-modules, the authors identified a core set of IBD susceptibility genes that were conserved across all three cohorts that were also enriched for the macrophage-enriched immune network and macrophage expression. This was termed the core immune activation module (IAM). By overlaying the core IAM onto Bayesian networks constructed from gene expression data from each cohort, the authors were able to identify an IBD-specific conserved immune component (CIC) in each network. Ultimately, using this approach, the authors identified 133 key driver genes which could regulate the IBD CIC networks, five of which had not been previously associated with IBD including *DOCK2, DOK3, AIF1, GPSM3, NCKAP1L*. The expression of these genes were shown to correlate with disease duration and also were upregulated in inflamed IBD patient intestinal biopsies.

Verstockt et al demonstrated the utility of GCNs to evaluate gene dysregulation at various stages of CD (Verstockt et al., 2019). In this study, transcriptomic and miRNA data were obtained from ileal mucosal biopsies of CD patients at three different stages of their disease i.e., newly diagnosed, recurrent disease following ileal resection, and late-stage disease. The authors conducted a WGCNA on this data which revealed modules that correlated with the three disease stages. The modules positively correlating with the different stages of CD were enriched in genes relating to granulocyte adhesion, diapedesis, and fibrosis. Conversely, genes associated with cholesterol biosynthesis were enriched in the module that negatively correlated with these stages of CD. They also constructed a miRNA-target gene GRN using the Ingenuity Pathway Analysis (IPA) microRNA Target Filter tool. This revealed that dysregulated miRNAs were more abundant in newly diagnosed and late-stage CD in comparison to post-operative recurrent CD. This suggests that surgical resection of the ileum followed by ileo-colonic anastomosis may reset the gene dysregulation occurring in CD.

A recent study by Aschenbrenner et al showed how GCNs could also be used to study cytokine signalling in CD (Aschenbrenner et al., 2021). They utilised transcriptomic data of ileal biopsies from a cohort of treatment naïve paediatric CD patients and non-inflamed controls to investigate the regulation of *IL23*. IL23 is a pro-inflammatory cytokine that has been implicated in IBD pathogenesis. Genetic studies have previously identified IBD-associated SNPs affecting the *IL23R* gene (Duerr et al., 2006). Furthermore, increased production of IL23 by macrophages and dendritic cells have been detected in mouse models of colitis and IBD patients (Maloy and Kullberg, 2008). Aschenbrenner et al conducted a WGCNA to see which modules of the transcriptome from inflamed and non-inflamed tissues correlate with *IL23* expression. This analysis identified 22 gene co-expression modules. Analysis of these modules revealed that *IL23A* expression strongly correlated with the modules enriched in functions for "immune cell differentiation" and "lymphocyte differentiation." These modules were found not to be significantly enriched in CD patients. However, an "inflammatory cytokine" module containing myeloid and stromal marker genes, proinflammatory cytokines (including OSM, IL1B, and IL6) and fibroblast activation protein, was identified that significantly correlated with IL23A expression and were also enriched in CD patients. This work supports the hypothesis that a subgroup of IBD patients may possess a pathogenic myeloid-stromal cell circuit involving OSM as identified in recent landmark studies (West et al., 2017; Smillie et al., 2019).

Multi-Layered Network Approaches

Over the past decade, there has been an increased appetite for the capture of different types of omics data from a single sample as it is believed this could provide greater insights into disease biology. This multi-omics revolution necessitates the combination of various network modelling approaches. Multi-layered networks can be used to integrate the many facets of multi-omics data including the different time scales of biological processes (Hammoud and Kramer, 2020). In recent years, various databases have been developed such as OmniPath (Türei et al., 2021), SignaLink2 (Fazekas et al., 2013), TranscriptomeBrowser (Lepoivre et al., 2012) or ConsensusPathDB (Kamburov et al., 2011), that can be used to generate multi-layered networks to integrate multi-omics data (Santra et al., 2014).

Combining different types of networks together has unravelled important insights into IBD pathogenesis. However, such multi-layered network approaches have largely been performed on a single type of omics data so far i.e., most commonly, gene expression data. This was seen in the earlier landmark study by Jostins et al where GCNs and GRNs were used together as mentioned earlier (Jostins et al., 2012). More recently, Martin et al generated intercellular ligand-receptor networks (a type of PPI network) and GCNs from scRNAseq data obtained from ileal biopsies of patients with ileal CD (Martin et al., 2019). By applying gene co-expression analysis to the scRNAseq data, they first identified a group of cell types which strongly correlated with ileal inflammation in a subset of ileal CD patients and also lack of response to anti-TNF therapy. They termed this group the GIMATS (IgG plasma cells, inflammatory mononuclear phagocytes, activated T cells and stromal cells) module. Next they evaluated intercellular interactions communicating with the GIMATS module by using the scRNAseq data to identify experimentally validated cytokine-cytokine receptor pairs (Ramilowski et al., 2015). This revealed a distinct intercellular network driving the GIMATS module including T cells, mononuclear phagocytes, fibroblasts and endothelial cells.

Cell signalling networks are another important type of multi-layered network consisting of two components: an upstream component which is a directed PPI network containing various intracellular signaling pathways, and a downstream component which is a GRN of transcription factor-target interactions (Csermely et al., 2013). The OmniPath database is particularly useful for generating cell signalling networks as it allows the user to not only access the intracellular PPI network of a cell but also GRNs and even the extracellular ligand-receptor networks from a myriad of databases (Türei et al., 2021). Although examples of cell signalling networks have been limited in IBD thus far, recently we established a novel bioinformatic pipeline termed "iSNP", to create a UC-specific

cell signalling network from patient-derived SNP data (Brooks et al., 2019). In this approach we focused on SNPs located within non-coding regions of the genome, which represent the vast majority of SNPs associated with UC. These non-coding SNPs were annotated to transcription factor binding sites (TFBS) and miRNA-target sites (miRNA-TS) using available databases reporting transcription factor binding profiles and miRNA sequences. Protein-coding genes located within the vicinity of SNP-affected TFBS and those targeted by the SNP-affected miRNA-TS were identified using regulatory interaction data sources. In this way SNP-affected proteins were revealed. Using OmniPath, the first neighbours of these SNP-affected proteins were also pinpointed. Utilising genotyped patient data from an IBD patient cohort in East Anglia in the United Kingdom, we created individual patient-specific cell signalling networks. By applying unsupervised clustering algorithms to these patient-specific cell signalling networks, we revealed that patients clustered into four main groups and identified distinct pathogenic pathways involved in each cluster. Thus, using a novel network biology workflow involving cell signalling networks, we were able to identify distinct regulatory effects of disease-associated non-coding SNPs in subgroups of UC patients.

FUTURE CHALLENGES AND POTENTIAL MITIGATING STRATEGIES TO DEVELOP NETWORK BIOLOGY APPROACHES FOR PRECISION MEDICINE

Despite the recent strides made in unravelling IBD pathogenesis using the aforementioned network biology approaches, there are several challenges that need to be overcome to achieve the goal of precision medicine in IBD (Fiocchi and Iliopoulos, 2021).

First and foremost, research efforts must focus on acquiring patient-specific data from a variety of relevant data sources that could provide a more holistic picture of the disease biology of individual patients. In the past, network biology models used only one or two dimensions of data such as PPI networks, sets of DEGs, or transcriptomic information to reconstruct biological networks (Seyed Tabib et al., 2020). However, recent breakthroughs made in cancer demonstrate that multi-layered networks which incorporate various omics data are likely to yield more powerful and translatable insights for complex diseases (Du and Elemento, 2015). There is a paucity of such approaches in IBD to date, although the aforementioned studies by Jostins et al. (2012), Martin et al. (2019) and Brooks et al. (2019) demonstrate the potential of such methods. In addition, despite the exponential increase in transcriptomics and metatranscriptomics studies in IBD in the past decade, such datasets are often limited by low patient numbers. Recently, a novel meta-analysis framework for transcriptome and metatranscriptome data in IBD has been introduced, called the IBD Transcriptome and Metatranscriptome Meta-Analysis (TaMMA) platform (Massimino et al., 2021). The TaMMA platform collates and integrates transcriptomics (and metatranscriptomics) data from multiple IBD patient cohorts using a standardised pipeline that corrects batch effects and performs differential analysis of the data. This significantly increases the sample size and statistical power for downstream analysis (Modos et al., 2021). This platform, which is available as a user-friendly, open-source web application, can maximise the utility of existing transcriptomics and metatranscriptomics datasets generated from various research centers across the world. Such meta-analysis frameworks could be a powerful way for analysing other omic layers too in the future.

In IBD, it is particularly important to consider the effects of the gut luminal microenvironment which contains bacterial cells up to 10^{13} in number and their repertoire of metabolite products on the host. However, this is an extremely complex ecosystem to model. Adding to this complexity is the dynamic nature of the gut microbiota, which can be affected by the age of the individual and environmental exposures such as diet and drugs. The development of novel genome-scale metabolic models as mentioned earlier as well as strain-specific metabolomics have potential to enhance our understanding of the IBD metabolome and the intestinal microflora (Han et al., 2021; Heinken et al., 2021). In addition to the metabolome, another data source for integration in IBD that should be strongly considered is histopathological data. The importance of integrating histopathological data for precision medicine has been clearly demonstrated in colorectal cancer (CRC) (Thomas et al., 2019). Over the past couple of decades it has been revealed that the type, density, and location of immune cells (i.e., the "immune contexture") within CRC tissues are a better prognostic tool than the traditional Dukes staging for predicting CRC survival and recurrence (Galon et al., 2006; Fridman et al., 2012). Subsequently, transcriptomic data from CRC tissues were integrated with this histological classification, to shed light on the remarkable immunogenomic heterogeneity of CRC (Becht et al., 2016). Similar efforts to amalgamate histopathological and genomic data have thus far been scarce in IBD, but appear to be on the horizon: Friedrich et al have recently revealed distinct pathotypes in IBD that are associated with non-response to several therapies using such an approach (Friedrich et al., 2021). Furthermore, to fully realise the potential of molecular, metabolomic, and histopathological data, it is integral that they are matched with pertinent clinical metadata i.e., information of the patients' treatment(s), age, comorbidities etc (Ahmed, 2020). This has been lacking in many previous studies (Olivera et al., 2019). However, acquiring good quality clinical metadata is challenging due to the use of paper medical records by many hospitals. Also despite the increasing use of electronic medical records (EMRs), hospitals seldom use the same EMR software resulting in interoperability issues and fragmentation of data (Warren et al., 2019). Nevertheless, artificial intelligence, including natural language processing (NLP), may help transform the extraction of clinical metadata from EMRs in the coming decades. Such big data methods can help to better understand and personalise network biology models and can also be used for validation of findings (Olivera et al., 2019; Seyed Tabib et al., 2020) (**Figure 6**).

Another strategy that may yield important insights into IBD disease pathogenesis is to evaluate omics data in IBD in the

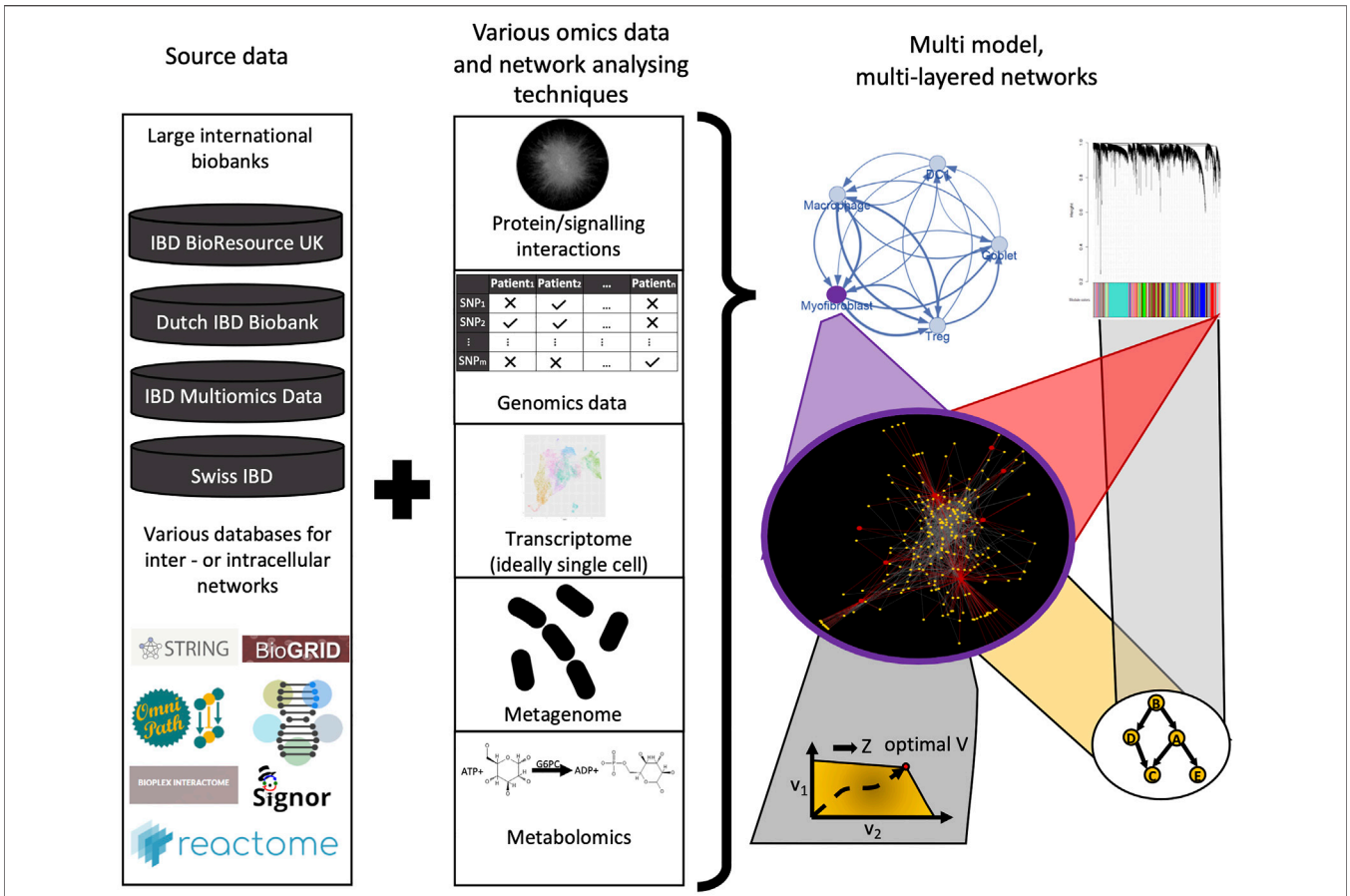


FIGURE 6 | Future perspectives of using network biology and network based modeling in IBD research. From the large amount of omics datasets (genomics, transcriptomics, metabolomics, metagenomics), various interaction networks can be used to develop sophisticated network models, ideally in a multi-layered fashion. Adding granularity with patient metadata from large databases can help to validate these models and will result in better understanding of IBD pathogenesis, novel/personalised therapeutic strategies, and clinical decision-driving signatures.

context of other comorbid disorders. Patients with IBD are more likely to develop other disorders with a significant immune component such as rheumatoid arthritis (RA), psoriasis, asthma and colorectal cancer (García et al., 2020). These disorders share underlying genetic risk factors and environmental exposures which can result in similarities in the immune pathways and cytokines driving inflammatory responses in these conditions (Moni and Liò, 2015). This is reflected in the fact that biologic agents targeting TNFα are effective in IBD as well as inflammatory arthritides such as RA, axial spondyloarthritis and psoriatic arthritis (Schett et al., 2021). However, thus far, there has been limited work which has evaluated multi-omics data between comorbid disease networks involving IBD. Nevertheless, bioinformatics tools have recently been generated that could be readily utilised to generate comorbidity networks from published multi-omics datasets for estimating disease comorbidity risks and patient stratification (Moni and Liò, 2015; Xiao et al., 2018).

One of the major challenges that will need to be addressed in all such approaches is how to integrate the vast amounts of multi-omics data generated from disparate sources to reveal clinically meaningful insights in IBD. Integrating genomics, transcriptomics (ideally single-cell transcriptomics), epigenomics, metabolomics, and metagenomic datasets of patients together with robust clinical meta-data and histopathological data over time will be critical for realising the goal of precision medicine in IBD (**Figure 6**). However, there is often a low degree of agreement between networks generated from different omics datasets, making it difficult to identify salient features that are shared between them. Therefore, more advanced data integration and analysis methods for multi-omics data are necessary.

Recently, a number of novel multi-omic data integration tools have been developed but their use has not yet penetrated the field of IBD. These include early data integration (i.e., combining all datasets into a single dataset first before developing the model) and late data integration (i.e., generating individual models from each dataset first and then finally integrating the models together) methods. Early examples of the former approach which create an aggregative layer within a multiplex network include iCluster (Shen et al., 2009), a joint latent variable model, and a similarity network fusion method by Wang et al. (2014). A weighted

network fusion method has also been developed which incorporates the relative weight or importance of each layer when integrating omics layers (Angione et al., 2016). At present, one of the most common methods of omics integration is an early integration method called non-negative matrix factorisation as implemented in the MOFA package (Argelaguet et al., 2018). In short, in this method a large matrix is first constructed where the columns are the patient samples and the rows are the measurements from the various types of omics data. This large matrix is then deconvoluted into two matrices. The first matrix contains the various omics measurements as rows and factors as columns, with cells referring to the contribution of each omics measurement to a factor. Here, factors represent biological information such as signalling pathways or metabolomic circuits. The second matrix is composed of samples as columns and factors as rows, with cells referring to each factor's value for a sample. Each factor can be traced back to the input measurements whether they are genomics, transcriptomics or metagenomic inputs. This can be used to uncover hidden interactions between various modalities of measurements. Clustering the samples based on the factors helps to reduce the noise that naturally arises when combining disparate data types. This approach was shown to identify major causes of disease heterogeneity in chronic lymphocytic leukaemia (Argelaguet et al., 2018). Late data integration methods have also revealed important insights into disease pathogenesis. An example is the COSMOS tool, in which multiple networks generated from different omics data are integrated using causal reasoning (Dugourd et al., 2021). In this paper, the investigators demonstrated the capability of COSMOS to integrate PPI, GRN and two different metabolic networks from transcriptomics, phosphoproteomics, and metabolomics data in clear cell renal cell carcinoma. Similar non-matrix-based omics methods were used in bacteria such as the MORA approach, which integrates various layers of omics data (transcriptomic, proteomics, metabolomics, genomics) to identify the affected pathways (Bardozzo et al., 2018). This method used mutual synchronisation of binarised omics measurements rather than a matrix deconvolution approach to identify affected pathways.

Recently, Malod-Dognin et al described the application of a novel multi-omics data integration and analysis framework in four different cancer types based on a machine learning technique called non-negative matrix tri-factorisation (NMTF) (Malod-Dognin et al., 2019). For each cancer type, using this approach they were able to integrate three different types of omics tissue-specific molecular interaction networks (i.e., PPI, GCN and gene interaction network) into a single, unified representation of a tissue-specific cell, which they termed "iCell." The NMTF algorithm is an intermediate data integration method i.e., it integrates the information from the various models (networks) and source data (gene expression) giving back valuable information such as clustering of genes or local rewiring of various genes in many networks. It uses an already filtered network for this purpose. The method deconvolutes the adjacency matrices of networks into three smaller matrices per network. Two of the matrices are the same in the various networks and they are transpose of each other that capture sample-specific features, whilst the third matrix displays network-specific features. This was shown to overcome the problems associated with early data integration and late data integration approaches that have been used previously, leading to more accurate predictions. To further analyse these integrated networks, they then utilised graphlets as a more sensitive method for evaluating network topology (Przulj, 2007; Yaveroğlu et al., 2014). The distribution of graphlets can act as a fingerprint for a network, allowing comparisons to be made between networks (Sarajlić et al., 2016). Overall, this innovative integrative and analytical approach was shown to better detect the functional organisation of cancer cells than from a single omics layer and it identified 63 new cancer-related genes.

CONCLUSION

Network biology approaches have provided unique insights into the pathogenesis of IBD which could not have been ascertained through simple evaluation of molecular data. With the recent establishment of several large biorepositories for IBD and the advent of next-generation sequencing, we will soon be able to access high-quality omics patient data with sufficient power to tackle some of the key unanswered questions in the field. It is important that this data is complemented with other relevant data sources, especially reliable clinical metadata. Network biology will be critical for integrating the resulting multifaceted datasets to generate clinically translatable end-points. In recent years, multi-omics integrative methods have been developed and then applied successfully in the field of cancer, but have been limited in IBD and other complex diseases. Further research is required to develop more robust integrative and analytical network biology approaches for various types of omics data. Such efforts will allow us to fully harness the potential of multi-omic patient datasets to provide deeper insights into the pathogenesis of IBD and achieve the goal of precision medicine in this complex disease.

NETWORK BIOLOGY GLOSSARY

Node/vertex: A point in a network. In biological networks it is usually a gene or protein.

Link/edge: The interaction between nodes. In network biology, it can be a physical interaction such as an enzymatic reaction or similarity e.g. correlation between the expression of two genes.

Directed network: The network's edges are directed meaning from node "v" to node "u" is not the same as from node "u" to node "v".

Weighted network: The edges of the network have weight. In network biology, weights often represent the number of interactions between cells or the strength of the interaction between proteins which can depend on the concentration or measured amount of the proteins (in case of proteomic analysis)

or the amount of the genes encoding the protein (in case of transcriptomics analysis).

Signed interaction: It is a type of weight of the network, which informs whether the interaction is positive or negative. A negative sign means the interaction is inhibitory, whereas a positive sign means it is excitatory.

Degree: The number of neighbouring nodes that a particular node connects to in a network.

Hub: A node with high degree.

Path: The set of edges connecting any two nodes.

Shortest path: The path between two nodes which involves the least number of edges.

Betweenness centrality: The number of shortest paths which go through a given node or edge. It is often normalised by the number of all possible shortest paths between all nodes.

Bottleneck: A node with high betweenness centrality but low degree. These are critical nodes in the network because a high amount of information goes through them.

Module/community: A set of nodes in a network which are interacting with each other more strongly than with other nodes outside the module.

Scale free network: A network which has a degree (k) distribution of $P(k) = k^{-\gamma}$. In practice it means that the network has a low number of high degree nodes whilst most of the nodes have a really low degree. Most biological networks closely resemble a scale free distribution.

Adjacency matrix: A matrix which models the network where columns and rows represent nodes and each value is an edge. If the network is undirected, then the adjacency matrix is symmetric, whereas in directed networks the adjacency matrix is asymmetric. If the network is not weighted then the values in the adjacency matrix are 1. However, in a weighted network the values are the weights.

Gene interaction network: A network where the edges represent whether the mutations of the genes together influence a phenotype e.g. synthetic lethality.

Matrix deconvolution: Representing the matrix with multiple smaller order matrices.

Causal reasoning: Finding the best possible path in a network where the signs match with the output of the network.

Graphlet: A local unique (non-isomorphic) structure of a network.

Network motif: An overrepresented local structure of a network (for instance a common graphlet).

AUTHOR CONTRIBUTIONS

JT and DM wrote and reviewed/edited the article before submission. JB-W and TK made substantial contributions to the discussion of content and reviewed/edited the article before submission.

FUNDING

JT is an Academic Clinical Fellow supported by the National Institute of Health Research (NIHR) and has been awarded funding through the Health Education England (HEE) Genomics Education Programme. DM and TK are supported by the Earlham Institute (Norwich, United Kingdom) in partnership with the Quadram Institute (Norwich, United Kingdom) and strategically supported by a United Kingdom Research and Innovation (UKRI) Biotechnological and Biosciences Research Council (BBSRC) Core Strategic Programme Grant for Genomes to Food Security (BB/CSP1720/1) and its constituent work packages, BBS/E/T/000PR9819 and BBS/E/T/000PR9817, as well as a BBSRC ISP grant for Gut Microbes and Health (BB/R012490/1) and its constituent projects, BBS/E/F/000PR10353 and BBS/E/F/000PR10355.

REFERENCES

Ahmed, Z. (2020). Practicing Precision Medicine with Intelligently Integrative Clinical and Multi-Omics Data Analysis. *Hum. Genomics* 14, 35. doi:10.1186/s40246-020-00287-z

Ahn, A. C., Tewari, M., Poon, C.-S., and Phillips, R. S. (2006). The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative? *Plos Med.* 3, e208. doi:10.1371/journal.pmed.0030208

Altaf-Ul-Amin, M., Wada, M., and Kanaya, S. (2012). Partitioning a PPI Network into Overlapping Modules Constrained by High-Density and Periphery Tracking. *ISRN Biomathematics* 2012, 1–11. doi:10.5402/2012/726429

Anand, S., Mukherjee, K., and Padmanabhan, P. (2020). An Insight to Flux-Balance Analysis for Biochemical Networks. *Biotechnol. Genet. Eng. Rev.* 36, 32–55. doi:10.1080/02648725.2020.1847440

Angione, C., Conway, M., and Lió, P. (2016). Multiplex Methods Provide Effective Integration of Multi-Omic Data in Genome-Scale Models. *BMC Bioinformatics* 17, 83. doi:10.1186/s12859-016-0912-1

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-Omics Factor Analysis-A Framework for Unsupervised Integration of Multi-omics Data Sets. *Mol. Syst. Biol.* 14, e8124. doi:10.15252/msb.20178124

Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021). Deciphering Cell-Cell Interactions and Communication from Gene Expression. *Nat. Rev. Genet.* 22, 71–88. doi:10.1038/s41576-020-00292-x

Aschenbrenner, D., Quaranta, M., Banerjee, S., Ilott, N., Jansen, J., Steere, B., et al. (2021). Deconvolution of Monocyte Responses in Inflammatory Bowel Disease Reveals an IL-1 Cytokine Network that Regulates IL-23 in Genetic and Acquired IL-10 Resistance. *Gut* 70, 1023–1036. doi:10.1136/gutjnl-2020-321731

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

Bardozzo, F., Lió, P., and Tagliaferri, R. (2018). A Study on Multi-Omic Oscillations in *Escherichia coli* Metabolic Networks. *BMC Bioinformatics* 19, 194. doi:10.1186/s12859-018-2175-5

Baumgart, D. C., and Carding, S. R. (2007). Inflammatory Bowel Disease: Cause and Immunobiology. *Lancet* 369, 1627–1640. doi:10.1016/S0140-6736(07)60750-8

Bayes, T. (1763). LII. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Phil. Trans. R. Soc.* 53, 370–418. doi:10.1098/rstl.1763.0053

Becht, E., de Reyniès, A., Giraldo, N. A., Pilati, C., Buttard, B., Lacroix, L., et al. (2016). Immune and Stromal Classification of Colorectal Cancer Is Associated

with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clin. Cancer Res.* 22, 4057–4066. doi:10.1158/1078-0432.CCR-15-2879

Biasci, D., Lee, J. C., Noor, N. M., Pombal, D. R., Hou, M., Lewis, N., et al. (2019). A Blood-Based Prognostic Biomarker in IBD. *Gut* 68, 1386–1395. doi:10.1136/gutjnl-2019-318343

Borg-Bartolo, S. P., Boyapati, R. K., Satsangi, J., and Kalla, R. (2020). Precision Medicine in Inflammatory Bowel Disease: Concept, Progress and Challenges. *F1000Res* 9, 54. doi:10.12688/f1000research.20928.1

Breitling, R. (2010). What Is Systems Biology. *Front. Physio.* 1, 9. doi:10.3389/fphys.2010.00009

Brooks, J., Modos, D., Sudhakar, P., Fazekas, D., Zoufir, A., Kapuy, O., et al. (2019). A Systems Genomics Approach to Uncover Patient-specific Pathogenic Pathways and Proteins in a Complex Disease. *BioRxiv*. doi:10.1101/692269

Cader, M. Z., and Kaser, A. (2013). Recent Advances in Inflammatory Bowel Disease: Mucosal Immune Cells in Intestinal Inflammation. *Gut* 62, 1653–1664. doi:10.1136/gutjnl-2012-303955

Charitou, T., Bryan, K., and Lynn, D. J. (2016). Using Biological Networks to Integrate, Visualize and Analyze Genomics Data. *Genet. Sel. Evol.* 48, 27. doi:10.1186/s12711-016-0205-1

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2015). The BioGRID Interaction Database: 2015 Update. *Nucleic Acids Res.* 43, D470–D478. doi:10.1093/nar/gku1204

Chen, J., Mamidipalli, S., and Huan, T. (2009). HAPPI: an Online Database of Comprehensive Human Annotated and Predicted Protein Interactions. *BMC Genomics* 10, S16. doi:10.1186/1471-2164-10-S1-S16

Chen, J. Y., Pandey, R., and Nguyen, T. M. (2017). HAPPI-2: a Comprehensive and High-Quality Map of Human Annotated and Predicted Protein Interactions. *BMC Genomics* 18, 182. doi:10.1186/s12864-017-3512-1

Cosnes, J., Gower–Rousseau, C., Seksik, P., and Cortot, A. (2011). Epidemiology and Natural History of Inflammatory Bowel Diseases. *Gastroenterology* 140, 1785–1794. doi:10.1053/j.gastro.2011.01.055

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network Propagation: a Universal Amplifier of Genetic Associations. *Nat. Rev. Genet.* 18, 551–562. doi:10.1038/nrg.2017.38

Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G., and Nussinov, R. (2013). Structure and Dynamics of Molecular Networks: A Novel Paradigm of Drug Discovery. *Pharmacol. Ther.* 138, 333–408. doi:10.1016/j.pharmthera.2013.01.016

Denson, L. A., Curran, M., McGovern, D. P. B., Koltun, W. A., Duerr, R. H., Kim, S. C., et al. (2019). Challenges in IBD Research: Precision Medicine. *Inflamm. Bowel Dis.* 25, S31–S39. doi:10.1093/ibd/izz078

Du, W., and Elemento, O. (2015). Cancer Systems Biology: Embracing Complexity to Develop Better Anticancer Therapeutic Strategies. *Oncogene* 34, 3215–3225. doi:10.1038/onc.2014.291

Duboc, H., Rajca, S., Rainteau, D., Benarous, D., Maubert, M.-A., Quervain, E., et al. (2013). Connecting Dysbiosis, Bile-Acid Dysmetabolism and Gut Inflammation in Inflammatory Bowel Diseases. *Gut* 62, 531–539. doi:10.1136/gutjnl-2012-302578

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., et al. (2006). A Genome-wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* 314, 1461–1463. doi:10.1126/science.1135245

Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K. B., et al. (2021). Causal Integration of Multi-omics Data with Prior Knowledge to Generate Mechanistic Hypotheses. *Mol. Syst. Biol.* 17, e9730. doi:10.15252/msb.20209730

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). CellPhoneDB: Inferring Cell-Cell Communication from Combined Expression of Multi-Subunit Ligand-Receptor Complexes. *Nat. Protoc.* 15, 1484–1506. doi:10.1038/s41596-020-0292-x

Eguchi, R., Karim, M. B., Hu, P., Sato, T., Ono, N., Kanaya, S., et al. (2018). An Integrative Network-Based Approach to Identify Novel Disease Genes and Pathways: a Case Study in the Context of Inflammatory Bowel Disease. *BMC Bioinformatics* 19, 264. doi:10.1186/s12859-018-2251-x

Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., et al. (2013). SignaLink 2 - a Signaling Pathway Resource with Multi-Layered Regulatory Networks. *BMC Syst. Biol.* 7, 7. doi:10.1186/1752-0509-7-7

Fiocchi, C., and Iliopoulos, D. (2021). IBD Systems Biology Is Here to Stay. *Inflamm. Bowel Dis.* 27, 760–770. doi:10.1093/ibd/izaa343

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The Immune Contexture in Human Tumours: Impact on Clinical Outcome. *Nat. Rev. Cancer* 12, 298–306. doi:10.1038/nrc3245

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.* 7, 601–620. doi:10.1089/106652700750050961

Friedrich, M., Pohin, M., Jackson, M. A., Korsunsky, I., Bullers, S., Rue-Albrecht, K., et al. (2021). IL-1-driven Stromal-Neutrophil Interaction in Deep Ulcers Defines a Pathotype of Therapy Non-responsive Inflammatory Bowel Disease. *BioRxiv*. doi:10.1101/2021.02.05.429804

Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pagès, C., et al. (2006). Type, Density, and Location of Immune Cells within Human Colorectal Tumors Predict Clinical Outcome. *Science* 313, 1960–1964. doi:10.1126/science.1129139

García, M. J., Pascual, M., Del Pozo, C., Díaz-González, A., Castro, B., Rasines, L., et al. (2020). Impact of Immune-Mediated Diseases in Inflammatory Bowel Disease and Implications in Therapeutic Approach. *Sci. Rep.* 10, 10731. doi:10.1038/s41598-020-67710-2

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., et al. (2018). Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res.* 78, 769–780. doi:10.1158/0008-5472.CAN-17-1679

GBD 2017 Inflammatory Bowel Disease Collaborators (2020). The Global, Regional, and National burden of Inflammatory Bowel Disease in 195 Countries and Territories, 1990-2017: a Systematic Analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* 5, 17–30. doi:10.1016/S2468-1253(19)30333-4

Green, S., Şerban, M., Scholl, R., Jones, N., Brigandt, I., and Bechtel, W. (2017). Network Analyses in Systems Biology: New Strategies for Dealing with Biological Complexity. *Synthese* 195, 1751–1777. doi:10.1007/s11229-016-1307-6

Hammoud, Z., and Kramer, F. (2020). Multilayer Networks: Aspects, Implementations, and Application in Biomedicine. *Big Data Anal.* 5, 2. doi:10.1186/s41044-020-00046-0

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., et al. (2015). TRRUST: a Reference Database of Human Transcriptional Regulatory Interactions. *Sci. Rep.* 5, 11432. doi:10.1038/srep11432

Han, S., Van Treuren, W., Fischer, C. R., Merrill, B. D., DeFelice, B. C., Sanchez, J. M., et al. (2021). A Metabolomics Pipeline for the Mechanistic Interrogation of the Gut Microbiome. *Nature* 595, 415–420. doi:10.1038/s41586-021-03707-9

Hang, S., Paik, D., Yao, L., Kim, E., Trinath, J., Lu, J., et al. (2019). Bile Acid Metabolites Control TH17 and Treg Cell Differentiation. *Nature* 576, 143–148. doi:10.1038/s41586-019-1785-z

Heinken, A., Hertel, J., and Thiele, I. (2021). Metabolic Modelling Reveals Broad Changes in Gut Microbial Metabolism in Inflammatory Bowel Disease Patients with Dysbiosis. *NPJ Syst. Biol. Appl.* 7, 19. doi:10.1038/s41540-021-00178-6

Heinken, A., Ravcheev, D. A., Baldini, F., Heirendt, L., Fleming, R. M. T., and Thiele, I. (2019). Systematic Assessment of Secondary Bile Acid Metabolism in Gut Microbes Reveals Distinct Metabolic Capabilities in Inflammatory Bowel Disease. *Microbiome* 7, 75. doi:10.1186/s40168-019-0689-3

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and Analysis of Biochemical Constraint-Based Models Using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702. doi:10.1038/s41596-018-0098-2

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: an Open Source Molecular Interaction Database. *Nucleic Acids Res.* 32, 452D–455D. doi:10.1093/nar/gkh052

Hong, J., Brandt, N., Abdul-Rahman, F., Yang, A., Hughes, T., and Gresham, D. (2018). An Incoherent Feedforward Loop Facilitates Adaptive Tuning of Gene Expression. *eLife* 7, e32323. doi:10.7554/eLife.32323

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., et al. (2018). Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cel Syst.* 6, 484–495. doi:10.1016/j.cels.2018.03.001

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., et al. (2021). Dual Proteome-Scale Networks Reveal Cell-specific Remodeling of the Human Interactome. *Cell* 184, 3022–3040. doi:10.1016/j.cell.2021.04.011

Imhann, F., Van der Velde, K. J., Barbieri, R., Alberts, R., Voskuil, M. D., Vich Vila, A., et al. (2019). The 1000IBD Project: Multi-Omics Data of 1000 Inflammatory Bowel Disease Patients; Data Release 1. *BMC Gastroenterol.* 19, 5. doi:10.1186/s12876-018-0917-5

Jansma, J., and El Aidy, S. (2021). Understanding the Host-Microbe Interactions Using Metabolic Modeling. *Microbiome* 9, 16. doi:10.1186/s40168-020-00955-1

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi:10.1093/nar/gkz1031

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., et al. (2012). Host-microbe Interactions Have Shaped the Genetic Architecture of Inflammatory Bowel Disease. *Nature* 491, 119–124. doi:10.1038/nature11582

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: toward a More Complete Picture of Cell Biology. *Nucleic Acids Res.* 39, D712–D717. doi:10.1093/nar/gkq1156

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct Molecular Interaction Database in 2012. *Nucleic Acids Res.* 40, D841–D846. doi:10.1093/nar/gkr1088

Kirouac, D. C., Du, J. Y., Lahdenranta, J., Overland, R., Yarar, D., Paragas, V., et al. (2013). Computational Modeling of ERBB2 -Amplified Breast Cancer Identifies Combined ErbB2/3 Blockade as Superior to the Combination of MEK and AKT Inhibitors. *Sci. Signal.* 6, ra68. doi:10.1126/scisignal.2004008

Knecht, C., Fretter, C., Rosenstiel, P., Krawczak, M., and Hütt, M.-T. (2016). Distinct Metabolic Network States Manifest in the Gene Expression Profiles of Pediatric Inflammatory Bowel Disease Patients and Controls. *Sci. Rep.* 6, 32584. doi:10.1038/srep32584

König, M., Bulik, S., and Holzhütter, H.-G. (2012). Quantifying the Contribution of the Liver to Glucose Homeostasis: a Detailed Kinetic Model of Human Hepatic Glucose Metabolism. *Plos Comput. Biol.* 8, e1002577. doi:10.1371/journal.pcbi.1002577

Korcsmaros, T., Schneider, M. V., and Superti-Furga, G. (2017). Next Generation of Network Medicine: Interdisciplinary Signaling Approaches. *Integr. Biol. (Camb)* 9, 97–108. doi:10.1039/c6ib00215c

Kosti, I., Jain, N., Aran, D., Butte, A. J., and Sirota, M. (2016). Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci. Rep.* 6, 24799. doi:10.1038/srep24799

Koutrouli, M., Karatzas, E., Paez-Espino, D., and Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* 8, 34. doi:10.3389/fbioe.2020.00034

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., et al. (2016). WikiPathways: Capturing the Full Diversity of Pathway Knowledge. *Nucleic Acids Res.* 44, D488–D494. doi:10.1093/nar/gkv1024

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

Lee, M. J., Ye, A. S., Gardino, A. K., Heijink, A. M., Sorger, P. K., MacBeath, G., et al. (2012). Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks. *Cell* 149, 780–794. doi:10.1016/j.cell.2012.03.031

Lennard-Jones, J. E. (1989). Classification of Inflammatory Bowel Disease. *Scand. J. Gastroenterol.* 24, 2–6. doi:10.3109/00365528909091339

Lepoivre, C., Bergon, A., Lopez, F., Perumal, N. B., Nguyen, C., Imbert, J., et al. (2012). TranscriptomeBrowser 3.0: Introducing a New Compendium of Molecular Interactions and a New Visualization Tool for the Study of Gene Regulatory Networks. *BMC Bioinformatics* 13, 19. doi:10.1186/1471-2105-13-19

Licata, L., Lo Surdo, P., Iannuccelli, M., Palma, A., Micarelli, E., Perfetto, L., et al. (2020). SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 Update. *Nucleic Acids Res.* 48, D504–D510. doi:10.1093/nar/gkz949

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

Lukassen, S., Chua, R. L., Trefzer, T., Kahn, N. C., Schneider, M. A., Muley, T., et al. (2020). SARS-CoV-2 Receptor ACE2 and TMPRSS2 Are Primarily Expressed in Bronchial Transient Secretory Cells. *EMBO J.* 39, e105114. doi:10.15252/embj.2010511410.15252/embj.2020105114

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2017). Generation of Genome-Scale Metabolic Reconstructions for 773 Members of the Human Gut Microbiota. *Nat. Biotechnol.* 35, 81–89. doi:10.1038/nbt.3703

Malod-Dognin, N., Petschnigg, J., Windels, S. F. L., Povh, J., Hemingway, H., Ketteler, R., et al. (2019). Towards a Data-Integrated Cell. *Nat. Commun.* 10, 805. doi:10.1038/s41467-019-08797-8

Maloy, K. J., and Kullberg, M. C. (2008). IL-23 and Th17 Cytokines in Intestinal Homeostasis. *Mucosal Immunol.* 1, 339–349. doi:10.1038/mi.2008.28

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., et al. (2021). WikiPathways: Connecting Communities. *Nucleic Acids Res.* 49, D613–D621. doi:10.1093/nar/gkaa1024

Martin, J. C., Chang, C., Boschetti, G., Ungaro, R., Giri, M., Grout, J. A., et al. (2019). Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* 178, 1493–1508. doi:10.1016/j.cell.2019.08.008

Massimino, L., Lamparelli, L. A., Houshyar, Y., D'Alessio, S., Peyrin-Biroulet, L., Vetrano, S., et al. (2021). The Inflammatory Bowel Disease Transcriptome and Metatranscriptome Meta-Analysis (IBD TaMMA) Framework. *Nat. Comput. Sci.* 1, 511–515. doi:10.1038/s43588-021-00114-y

Meskó, B., and Görög, M. (2020). A Short Guide for Medical Professionals in the Era of Artificial Intelligence. *Npj Digit. Med.* 3, 126. doi:10.1038/s41746-020-00333-z

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 824–827. doi:10.1126/science.298.5594.824

Módos, D., Bulusu, K. C., Fazekas, D., Kubisch, J., Brooks, J., Marczell, I., et al. (2017). Neighbours of Cancer-Related Proteins Have Key Influence on Pathogenesis and Could Increase the Drug Target Space for Anticancer Therapies. *NPJ Syst. Biol. Appl.* 3, 2. doi:10.1038/s41540-017-0003-6

Modos, D., Thomas, J. P., and Korcsmaros, T. (2021). A Handy Meta-Analysis Tool for IBD Research. *Nat. Comput. Sci.* 1, 571–572. doi:10.1038/s43588-021-00124-w

Moni, M. A., and Liò, P. (2015). How to Build Personalized Multi-Omics Comorbidity Profiles. *Front. Cel Dev. Biol.* 3, 28. doi:10.3389/fcell.2015.00028

Noor, N. M., Verstockt, B., Parkes, M., and Lee, J. C. (2020). Personalised Medicine in Crohn's Disease. *Lancet Gastroenterol. Hepatol.* 5, 80–92. doi:10.1016/S2468-1253(19)30340-1

Olivera, P., Danese, S., Jay, N., Natoli, G., and Peyrin-Biroulet, L. (2019). Big Data in IBD: a Look into the Future. *Nat. Rev. Gastroenterol. Hepatol.* 16, 312–321. doi:10.1038/s41575-019-0102-5

Olsen, J., Gerds, T. A., Seidelin, J. B., Csillag, C., Bjerrum, J. T., Troelsen, J. T., et al. (2009). Diagnosis of Ulcerative Colitis before Onset of Inflammation by Multivariate Modeling of Genome-wide Gene Expression Data. *Inflamm. Bowel Dis.* 15, 1032–1038. doi:10.1002/ibd.20879

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What Is Flux Balance Analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Parkes, M.IBD BioResource Investigators (2019). IBD BioResource: an Open-Access Platform of 25 000 Patients to Accelerate Research in Crohn's and Colitis. *Gut* 68, 1537–1540. doi:10.1136/gutjnl-2019-318835

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., et al. (2011). Using Graph Theory to Analyze Biological Networks. *BioData Mining* 4, 10. doi:10.1186/1756-0381-4-10

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

Pinchuk, I. V., Beswick, E. J., Saada, J. I., Boya, G., Schmitt, D., Raju, G. S., et al. (2011). Human Colonic Myofibroblasts Promote Expansion of CD4+ CD25high Foxp3+ Regulatory T Cells. *Gastroenterology* 140, 2019–2030. doi:10.1053/j.gastro.2011.02.059

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. *Nucleic Acids Res.* 45, D833–D839. doi:10.1093/nar/gkw943

Przulj, N. (2007). Biological Network Comparison Using Graphlet Degree Distribution. *Bioinformatics* 23, e177–e183. doi:10.1093/bioinformatics/btl301

Przulj, N., Corneil, D. G., and Jurisica, I. (2006). Efficient Estimation of Graphlet Frequency Distributions in Protein-Protein Interaction Networks. *Bioinformatics* 22, 974–980. doi:10.1093/bioinformatics/btl030

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling Interactome: Scale-free or Geometric? *Bioinformatics* 20, 3508–3515. doi:10.1093/bioinformatics/bth436

Ramilowski, J. A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V. P., et al. (2015). A Draft Network of Ligand-Receptor-Mediated Multicellular Signalling in Human. *Nat. Commun.* 6, 7866. doi:10.1038/ncomms8866

Santra, T., Kolch, W., and Kholodenko, B. N. (2014). Navigating the Multilayered Organization of Eukaryotic Signaling: a New Trend in Data Integration. *Plos Comput. Biol.* 10, e1003385. doi:10.1371/journal.pcbi.1003385

Sarajlić, A., Malod-Dognin, N., Yaveroğlu, Ö. N., and Pržulj, N. (2016). Graphlet-based Characterization of Directed Networks. *Sci. Rep.* 6, 35098. doi:10.1038/srep35098

Schett, G., McInnes, I. B., and Neurath, M. F. (2021). Reframing Immune-Mediated Inflammatory Diseases through Signature Cytokine Hubs. *N. Engl. J. Med.* 385, 628–639. doi:10.1056/NEJMra1909094

Schlitt, T., and Brazma, A. (2007). Current Approaches to Gene Regulatory Network Modelling. *BMC Bioinformatics* 8, S9. doi:10.1186/1471-2105-8-S6-S9

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martínez-Cruz, L. A., et al. (2005). Correlation between Gene Expression and GO Semantic Similarity. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 2, 330–338. doi:10.1109/TCBB.2005.50

Seyed Tabib, N. S., Madgwick, M., Sudhakar, P., Verstockt, B., Korcsmaros, T., and Vermeire, S. (2020). Big Data in IBD: Big Progress for Clinical Practice. *Gut* 69, 1520–1532. doi:10.1136/gutjnl-2019-320065

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis. *Bioinformatics* 25, 2906–2912. doi:10.1093/bioinformatics/btp543

Sinha, S. R., Haileselassie, Y., Nguyen, L. P., Tropini, C., Wang, M., Becker, L. S., et al. (2020). Dysbiosis-Induced Secondary Bile Acid Deficiency Promotes Intestinal Inflammation. *Cell Host Microbe* 27, 659–670. doi:10.1016/j.chom.2020.01.021

Smillie, C. S., Biton, M., Ordovas-Montanes, J., Sullivan, K. M., Burgin, G., Graham, D. B., et al. (2019). Intra and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* 178, 714–730. doi:10.1016/j.cell.2019.06.029

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of Protein Interaction Network Mapping. *Mol. Syst. Biol.* 11, 848. doi:10.15252/msb.20156351

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

Song, X., Sun, X., Oh, S. F., Wu, M., Zhang, Y., Zheng, W., et al. (2020). Microbial Bile Acid Metabolites Modulate Gut RORγ+ Regulatory T Cell Homeostasis. *Nature* 577, 410–415. doi:10.1038/s41586-019-1865-0

Spekhorst, L. M., Imhann, F., Festen, E. A., Bodegraven, A. A. v., Boer, N. K. d., Bouma, G., et al. (2017). Cohort Profile: Design and First Results of the Dutch IBD Biobank: a Prospective, Nationwide Biobank of Patients with Inflammatory Bowel Disease. *BMJ Open* 7, e016695. doi:10.1136/bmjopen-2017-016695

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a General Repository for Interaction Datasets. *Nucleic Acids Res.* 34, D535–D539. doi:10.1093/nar/gkj109

Sudhakar, P., Machiels, K., Verstockt, B., Korcsmaros, T., and Vermeire, S. (2021). Computational Biology and Machine Learning Approaches to Understand Mechanistic Microbiome-Host Interactions. *Front. Microbiol.* 12, 618856. doi:10.3389/fmicb.2021.618856

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

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., et al. (2013). A Community-Driven Global Reconstruction of Human Metabolism. *Nat. Biotechnol.* 31, 419–425. doi:10.1038/nbt.2488

Thomas, J. P., Divekar, D., Brooks, J., and Watson, A. J. M. (2019). Gut Microbes Drive T-Cell Infiltration into Colorectal Cancers and Influence Prognosis. *Gastroenterology* 156, 1926–1928. doi:10.1053/j.gastro.2019.03.035

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., et al. (2021). Integrated Intra- and Intercellular Signaling Knowledge for Multicellular Omics Analysis. *Mol. Syst. Biol.* 17, e9923. doi:10.15252/msb.20209923

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 Bioinform.* 19, bbw139–592. doi:10.1093/bib/bbw139

Verstockt, B., Noor, N. M., Marigorta, U. M., Pavlidis, P., Deepak, P., Ungaro, R. C., et al. (2021). Results of the Seventh Scientific Workshop of ECCO: Precision Medicine in IBD-Disease Outcome and Response to Therapy. *J. Crohns Colitis* 15, 1431–1442. doi:10.1093/ecco-jcc/jjab050

Verstockt, S., De Hertogh, G., Van der Goten, J., Verstockt, B., Vancamelbeke, M., Machiels, K., et al. (2019). Gene and Mirna Regulatory Networks during Different Stages of Crohn's Disease. *J. Crohns Colitis* 13, 916–930. doi:10.1093/ecco-jcc/jjz007

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell* 144, 986–998. doi:10.1016/j.cell.2011.02.016

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

Warren, L. R., Clarke, J., Arora, S., and Darzi, A. (2019). Improving Data Sharing between Acute Hospitals in England: an Overview of Health Record System Distribution and Retrospective Observational Analysis of Inter-hospital Transitions of Care. *BMJ Open* 9, e031637. doi:10.1136/bmjopen-2019-031637

West, N. R., Hegazy, A. N., Hegazy, A. N., Owens, B. M. J., Bullers, S. J., Linggi, B., et al. (2017). Oncostatin M Drives Intestinal Inflammation and Predicts Response to Tumor Necrosis Factor-Neutralizing Therapy in Patients with Inflammatory Bowel Disease. *Nat. Med.* 23, 579–589. doi:10.1038/nm.4307

Whitcomb, D. C. (2019). Primer on Precision Medicine for Complex Chronic Disorders. *Clin. Transl. Gastroenterol.* 10, e00067. doi:10.14309/ctg.0000000000000067

Xavier, R. J., and Podolsky, D. K. (2007). Unravelling the Pathogenesis of Inflammatory Bowel Disease. *Nature* 448, 427–434. doi:10.1038/nature06005

Xiao, H., Bartoszek, K., and Lio', P. (2018). Multi-omic Analysis of Signalling Factors in Inflammatory Comorbidities. *BMC Bioinformatics* 19, 439. doi:10.1186/s12859-018-2413-x

Yan, W., Xue, W., Chen, J., and Hu, G. (2016). Biological Networks for Cancer Candidate Biomarkers Discovery. *Cancer Inform.* 15s3, CIN.S39458. doi:10.4137/CIN.S39458

Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., et al. (2014). Revealing the Hidden Language of Complex Networks. *Sci. Rep.* 4, 4547. doi:10.1038/srep04547

Yeh, C.-S., Wang, Z., Miao, F., Ma, H., Kao, C.-T., Hsu, T.-S., et al. (2019). A Novel Synthetic-Genetic-Array-Based Yeast One-Hybrid System for High Discovery Rate and Short Processing Time. *Genome Res.* 29, 1343–1351. doi:10.1101/gr.245951.118

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

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